frontiers | Frontiers in Artificial Intelligence

TYPE Review
PUBLISHED 24 May 2024
DOI 10.3389/frai.2024.1260952

OPEN ACCESS

EDITED BY
David Tomás,
University of Alicante, Spain

REVIEWED BY
Sagnik Ray Choudhury,
National Board of Medical Examiners, United States
Hiram Calvo,
National Polytechnic Institute (IPN), Mexico

*CORRESPONDENCE
Elize Herrewijnen
e.herrewijnen@uu.nl

RECEIVED 18 July 2023
ACCEPTED 07 May 2024
PUBLISHED 24 May 2024

CITATION
Herrewijnen E, Nguyen D, Bex F and van Deemter K (2024) Human-annotated rationales and explainable text classification: a survey. *Front. Artif. Intell.* 7:1260952. doi: 10.3389/frai.2024.1260952

Human-annotated rationales and explainable text classification: a survey

Elize Herrewijnen[1,2]*, Dong Nguyen[1], Floris Bex[1,3] and Kees van Deemter[1]

[1]Department of Information & Computing Sciences, Utrecht University, Utrecht, Netherlands, [2]National Police Lab AI, Netherlands Police, Driebergen, Netherlands, [3]Tilburg Institute for Law, Technology and Society, Tilburg University, Tilburg, Netherlands

Asking annotators to explain "why" they labeled an instance yields annotator rationales: natural language explanations that provide reasons for classifications. In this work, we survey the collection and use of annotator rationales. Human-annotated rationales can improve data quality and form a valuable resource for improving machine learning models. Moreover, human-annotated rationales can inspire the construction and evaluation of model-annotated rationales, which can play an important role in explainable artificial intelligence.

KEYWORDS

annotator rationales, natural language explanations, explainable artificial intelligence, data collection, machine learning, rationale agreement, text classification, human-annotated rationales

1 Introduction

With an ever-growing number of applications and users, it is important that language-based artificial intelligence (AI) models can be explained in a human-like way. Natural language explanations that are provided by humans, often referred to as "annotator rationales", are a promising resource for building explainable AI (XAI) systems.

Seminal works (Zaidan et al., 2007, 2008; Zaidan and Eisner, 2008) have collected annotator rationales by asking human annotators to highlight parts of a text to justify "why" that text should receive a certain label. The term "annotator rationale" has since been used with different meanings; for example, as human-annotated highlights in a text (Volkova and Yarowsky, 2014; Kutlu et al., 2020), as human-annotated free-text comments (Kartal and Kutlu, 2020), or as highlights generated by a machine learning (ML) model (Yessenalina et al., 2010). In this work, we consider *annotator rationales* to be natural language explanations (i.e., rationales) produced by the annotator (e.g., a human or an ML model).

While annotator rationales have been collected and used within the field of natural language processing (NLP), to our knowledge, no overview to guide those interested in using annotator rationales in NLP exists. Therefore, this article aims to provide insight into the lessons learned when it comes to collecting and using annotator rationales in NLP. We do this by surveying the use of annotator rationales in the field of NLP, specifically for explainable text classification.

1.1 Scope and selection criteria

Rationales have been used in many NLP tasks, e.g., natural language inference (Camburu et al., 2018; Kumar and Talukdar, 2020), next word prediction (Vafa et al., 2021),

question answering (Lamm et al., 2021), and translation quality (Fomicheva et al., 2021). In this survey, we mainly focus on explainable single-input text classification, thereby limiting explanation to identifying or describing relevant parts of a text instance. Multi-input tasks bring additional challenges when generating rationales; for example, it is unclear whether the explanations should refer to all inputs (e.g., the question and the answer in a question-answering task), or a selection of inputs (e.g., only the answer in a question-answering task).

Using Google Scholar, we select relevant literature for our survey by looking into related work that cites (Zaidan et al., 2007) and work that uses the terms "annotator rationales", "rationales" and "natural language explanations". Moreover, we apply the following criteria to select relevant literature:

- The work relates to (explainable) NLP.
- The work collects and/or uses natural language explanations that are provided by humans.
- The work involves a single-input text classification task.

We occasionally include studies outside these criteria, when they give relevant insight into annotator rationales. We survey literature published up to 2023.

1.2 Related surveys

Natural language explanations for AI have been surveyed in related work: there are, for example, surveys on explainable NLP (Danilevsky et al., 2020), general XAI methods that generate natural language explanations (Cambria et al., 2023; Gurrapu et al., 2023), datasets for explainable NLP (Wiegreffe and Marasovic, 2021), and using human rationales for improving ML models (Hartmann and Sonntag, 2022). Complementary to the above surveys, this work focuses specifically on annotator rationales as provided by *humans* and their application in *explainable* text classification.

1.3 Outline

This article is structured as follows: first, we propose a conceptual framework for rationales in Section 2. Then, we discuss how annotator rationales have been collected from human annotators for various tasks, and through various annotation schemes (Section 3). We provide an overview of datasets containing annotator rationales and invite the community to collaboratively update this overview using GitHub.[1] We then proceed to survey how human annotator rationales are used in explainable text classification. We first provide a brief introduction to basic concepts of XAI (Section 4), and then discuss how human annotator rationales are used for generating and evaluating rationales in explainable text classification (Section 5). We conclude our survey with a list of concrete suggestions for the use of rationales in XAI (Section 6).

1 https://utrechtuniversity.github.io/annotator-rationales-survey/

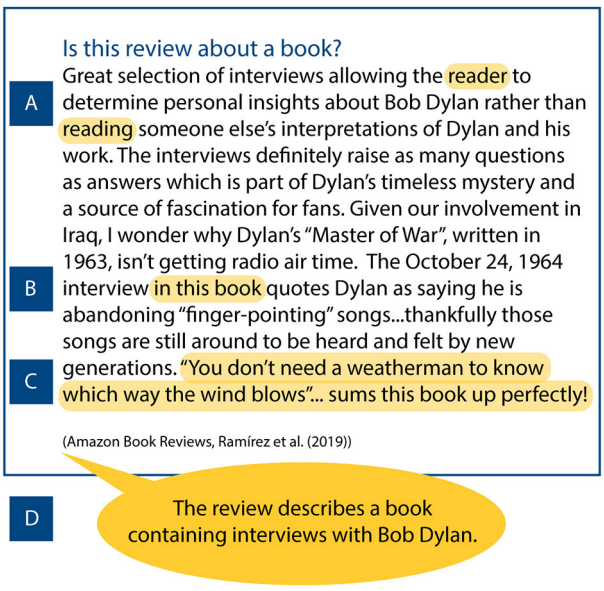

FIGURE 1
A book review from the Amazon Book Reviews dataset (Ramírez et al., 2019). The task is assigning the label `aboutBook` to reviews that are written about books. For this instance, the annotator has labeled the review with the `aboutBook` label. The highlighted parts of the text are examples of extractive word (A), snippet (B), and sentence (C) rationales. The speech bubble (D) shows an abstractive sentence rationale. These rationales explain why the review is annotated with the `aboutBook` label.

2 A framework for rationales

In the following section, we propose a conceptual framework for *rationales*: explanations in natural language format (Ehsan et al., 2019). Figure 1 provides examples of rationales. As humans often explain their decisions through natural language, it is reasonable to assume that this type of explanation is suitable for explaining AI to non-technical users (Miller et al., 2017; Cambria et al., 2023; Mukhtar et al., 2023). Furthermore, rationales can adopt domain-specific jargon, tailoring the explanation to domain expert knowledge (e.g., medical practice Meldo et al., 2020). We describe various kinds of rationales in the following paragraphs.

2.1 Form

2.1.1 Granularity

Rationales appear in various granularities (see Figure 1). The highest granularity is collections of *words* extracted from a text, or keywords describing an instance. *Snippet*-level rationales consist of multiple consecutive words, and *sentence*-level rationales consist of single sentences. Finally, *paragraph*-level rationales are multi-sentence natural language explanations. Note that a rationale can consist of multiple text spans (e.g., snippets), that together make up one rationale explaining a decision. Some granularities may be more suitable than others; for example, Jain et al. (2020) find that humans prefer snippets to words. However, few studies examine the suitability of different rationale granularities.

2.1.2 Extractive and abstractive

We distinguish between two types of rationales; the first is *extractive* rationales, such as words, sentences, and snippets which are parts of the input text. Extractive rationales are also referred to as excerpts (McDonnell et al., 2016) or highlights (Zaidan et al., 2007).

The second type of rationale is an *abstractive* rationale. These are free-text natural language explanations that *refer* to the input text, but are not (exact) parts of the input text. We adopt the term abstractive from automated text summarization (Lin and Ng, 2019; Gurrapu et al., 2023). Abstractive rationales are more difficult to evaluate and may increase annotation time (Kutlu et al., 2020; Wiegreffe et al., 2021), but they allow annotators to intuitively provide explanations using an unrestricted vocabulary.

2.1.3 Categorical and numerical

Most rationales annotated by humans are *categorical*. Human annotators often select specific spans of text that explain their decision, resulting in binary rationales. However, annotations can also occur on a more nuanced level, where spans of text can explain or contradict a decision (Kutlu et al., 2020; Sullivan et al., 2022). Some rationales, in particular extractive rationales, are composed of elements that are assigned *numeric* values. For example, when the word "great" has a value of 0.9 and "okay" has a value of 0.5, this could suggest that "great" is more relevant than "okay". Numerical rationales can be collected or constructed by, for example, combining rationale values across annotators (e.g., 70% agrees "great" is a rationale) (Mathew et al., 2021), asking new annotators to rate rationales based on usefulness (Ramírez et al., 2019), or using machine learning techniques to assign weights to text spans [e.g., the attention mechanism (Bao et al., 2018)].

2.2 Exhaustiveness

Rationales vary in degrees of exhaustiveness. Take for example the annotation instructions by Sen et al. (2020): *"highlight (ALL) the words that are indicative of the chosen sentiment"*; here, the goal is to annotate an *exhaustive* rationale, i.e., *all* text spans (e.g., words, snippets, sentences) that explain a decision (DeYoung et al., 2020; Sen et al., 2021). Alternatively, take the annotation instructions by Abedin et al. (2011): *"annotators are asked to 'do their best to mark enough rationales to provide convincing support for the class of interest', but are not expected to 'go out of their way to mark everything.'"*; here, the rationale is probably not exhaustive, but is nevertheless sufficient to explain the decision.

2.3 Human-annotated rationales and model-annotated rationales

We identify two categories of annotator rationales according to the type of annotator. *Human-annotated* rationales (hARs) are provided by human annotators—e.g., students, domain experts, or crowd workers—and *model-annotated* rationales (mARs) are provided by an ML model.

3 Human-annotated rationales

Next up, we discuss human-annotated rationales (hARs). Specifically, we outline several aims and benefits of collecting hARs (Section 3.1), and the lessons learned from annotation setups collecting hARs (Section 3.2). Table 1 provides an overview of datasets containing hARs.[2]

3.1 Collection aims and benefits

After their introduction by Zaidan et al. (2007), human-annotated rationales have been collected for common tasks like sentiment and topic classification, but also for domain-specific tasks such as legal document classification (see Table 1). Furthermore, hARs have been collected with various aims, which we will discuss in the following section.

3.1.1 Improving ML model performance

First, enriching datasets with hARs can be beneficial to ML model training; using the rationales, the ML model can be guided toward the most useful parts of the input for solving the task. Following Zaidan et al. (2007), many authors use hARs to improve their ML model performance (e.g., Saleem et al., 2012; Tepper et al., 2013; Krening et al., 2016; Chhatwal et al., 2018; Arous et al., 2021; Pruthi et al., 2022). In addition, the required amount of labeled training data can be substantially reduced (e.g., requiring only 10% of the original training data) by using hARs as enriched inputs (Arora and Nyberg, 2009; Sharma et al., 2015; Wang et al., 2022). See Hartmann and Sonntag (2022) for a survey on improving ML model (task) performance with human explanations.

Furthermore, hARs can teach ML models "valid reasons" for a classification, reducing spurious ML model behavior (Mathew et al., 2021; Chen et al., 2022; Joshi et al., 2022) and improving out-of-domain (OOD) performance (Lu et al., 2022).

3.1.2 Task insight

Second, collecting hARs can help gain insight into the annotation task (Yano et al., 2010; Kartal and Kutlu, 2020). For example, Malik et al. (2021) identify annotator groups based on rationales in their legal document classification task: annotators used either "holistic reasoning" or "bare-minimum reasoning". Another example is Kartal and Kutlu (2020), who identify important topics for their tweet classification task using hARs. Furthermore, there is an increasing awareness in NLP that disagreement in labeling is often informative and can point to

2 We only include rationales provided by the decision-maker in this overview, excluding rationales that are provided by other actors and rationales for pre-defined decisions like in Clinciu et al. (2021). Furthermore, we exclude annotator rationales for tasks outside the scope of this work (see Section 1) (e.g., Srivastava et al., 2017; Camburu et al., 2018; Khashabi et al., 2018; Yang et al., 2018; Atkinson et al., 2019; Meldo et al., 2020; Zhang et al., 2020; Yao et al., 2023). See work by Wiegreffe and Marasovic (2021) for an overview of datasets for explainable NLP.

TABLE 1 Overview of datasets with human-annotated rationales in the literature.

Related work	Classification task	Granularity	Form		Value type	Collection aim					Annotator	Name (if available)
						Improving ML	Task insight	Data quality	Gold explanation	Data generation		
Zaidan et al. (2007)	Sentiment	Sn	E		C	✓					O	IMDB
Titov and McDonald (2008)	Sentiment	Se	E		C				✓		O	TripAdvisor*
Yano et al. (2010)	Bias	Sn	E		C		✓				Cw	
Abedin et al. (2011)	Aviation incident causes	Sn	E		C	✓				✓	O	ASRS
McAuley et al. (2012)	Sentiment	S	E		C	✓			✓		De	BeerAdvocate
Saleem et al. (2012)	Medical	Sn	E		C	✓					De	
Xia and Yetisgen-Yildiz (2012)	Medical	S		A	C			✓			De	
Tepper et al. (2013)	Medical	Sn	E		C	✓					De	CPIS/PNA
Marshall et al. (2015)	Bias	Sn	E		C				✓		De	RoB
McDonnell et al. (2016)	Webpage relevance	Se	E	A	C			✓			Cw	
Bao et al. (2018)	Sentiment	Sn	E		C	✓					O	BeerAdvocate*
Carton et al. (2018)	Personal attacks	Sn	E		C				✓		O	
Chhatwal et al. (2018)	Legal	Sn	E	A	C	✓					De	
Kaushik et al. (2019)	Sentiment	Sn	E		C	✓					Cw	IMDB*
Ramírez et al. (2019)	Topic	Sn	E	A	N		✓				Cw	SLR
Ramírez et al. (2019)	Topic	Sn	E	A	C		✓				Cw	Amazon
Wang et al. (2020)	Sentiment	Se		A	C					✓	Cw	SemEval-2014*
Hasanain et al. (2020)	Topic	Se	E	A	C	✓	✓		✓		De	ArTest
Kanchinadam et al. (2020)	Sentiment	Sn	E		C	✓					Cw	IMDB*
Kartal and Kutlu (2020)	Check-worthy claims	Sn		A	C		✓				O	TrClaim-19
Kreiss et al. (2020)	Guilt	Sn	E		C	✓	✓				Cw	SuspectGuilt
Kutlu et al. (2020)	Webpage relevance	Se	E	A	C			✓			Cw	
Sap et al. (2020)	Abusive content	Se		A	C	✓			✓		Cw	SBIC
Sen et al. (2020)	Sentiment	Sn	E		C				✓		Cw	Yelp-HAT
Arous et al. (2021)	Topic	Sn	E		C	✓			✓		Cw	Wiki-Tech
Chalkidis et al. (2021)	Legal	P	E		C				✓		De	ECtHR
Hayati et al. (2021)	Style	W	E		C				✓		Cw	Hummingbird
Jayaram and Allaway (2021)	Stance detection	W	E		C	✓					Cw	VAST*
Mohseni et al. (2021)	Sentiment	Sn	E		C				✓		Cw	IMDB*
Mohseni et al. (2021)	Topic	Sn	E		C				✓		Cw	20News*
Mathew et al. (2021)	Hate speech	Sn	E		N	✓			✓		Cw	HateXplain
Malik et al. (2021)	Legal	Se	E		C				✓		De	ILDC
Sharma et al. (2020)	Empathy expression	Sn	E		C				✓		Cw	EMH
Vidgen et al. (2021)	Abusive content	Sn	E		C				✓		De	CAD
El Zini et al. (2022)	Sentiment	W	E		C				✓		O	RottenTomatoes*
Chiang and Lee (2022)	Sentiment	Sn	E		C				✓		Cw	IMDB*

(Continued)

TABLE 1 (Continued)

Related work	Classification task	Granularity	Form		Value type	Collection aim					Annotator	Name (if available)
						Improving ML	Task insight	Data quality	Gold explanation	Data generation		
Guzman et al. (2022)	Forced labor indicators	Sn	E		C	✓					De	RaFoLa
Jørgensen et al. (2022)	Sentiment	W	E		C				✓		O	SST*
Lu et al. (2022)	Sentiment	Sn	E		C	✓					Cw	IMDB*
Sullivan et al. (2022)	Sentiment	Sn	E		C		✓				Cw	IMDB*
Wang et al. (2022)	Topic	Sn	E		C	✓					O	AIvsCR
Jakobsen et al. (2023)	Sentiment	W	E		C	✓					Cw	DynaSent*
Jakobsen et al. (2023)	Sentiment	W	E		C	✓					Cw	SST*

Granularity is abbreviated as Paragraphs, Sentences, Snippets, and Words. Form is abbreviated as Extractive and Abstractive. Values types are abbreviated as Categorical and Numerical. The annotator type is abbreviated as Crowd worker, Domain expert, and Other. When available, the name of the dataset is provided. The * symbol is used when human-annotated rationales are added to an already existing dataset.

differences in interpretation (Uma et al., 2022). Rationales can provide further insight into reasons for labeling disagreement, like annotator bias or instruction ambiguity (Kartal and Kutlu, 2020), especially when the labeling task is subjective (Sen et al., 2021).

3.1.3 Data quality

Third, requesting annotator rationales from human annotators can improve data quality; forming a rationale requires annotators to consider their annotation more deeply, and collecting hARs thus reduces the number of classification mistakes made by human annotators (Kutlu et al., 2020). Moreover, hARs allow for effective data validation, for example through label aggregation [e.g., discarding labels with abnormal rationales (Sen et al., 2020)] or annotator discussion [e.g., providing rationales as arguments that annotators can respond to (Xia and Yetisgen-Yildiz, 2012; Drapeau et al., 2016; McDonnell et al., 2016; Kutlu et al., 2020)].

3.1.4 Data generation

Fourth, hARs can be a valuable resource for data generation; for example, new data points can be created by removing extractive rationales from input texts (Zaidan et al., 2007), retaining only rationales in an input text (Abedin et al., 2011), or combining rationales from multiple input texts into a new instance (Volkova and Yarowsky, 2014). Furthermore, labeling functions can be constructed from both abstractive and extractive hARs (Li et al., 2015; Hancock et al., 2018). All things considered, hARs can be seen as rich labels: Hancock et al. (2018) even claim that for their experiment, "one explanation can be worth 100 labels", and Sharma and Bilgic (2018) suggest that a document with rationales can be worth as many as 20 documents without rationales.

3.1.5 Gold rationales

Fifth, human-annotated rationales are often collected for use as "gold rationales" (Table 1) to determine the quality of generated ML model explanations. In Section 5.2 we discuss this topic in more detail.

3.2 Insights from human rationale collection

In the following section, we discuss some insights from human rationale collection.

3.2.1 Choice of annotators

As shown in Table 1, hARs are often collected with the aim of ML model improvement and as gold explanations. How beneficial hARs are to the various collection aims may depend on the annotator type; crowdsourcing platforms, like Mechanical Turk (Crowston, 2012), give access to a large group of annotators but allow for little data quality control. Here, requesting hARs from crowd workers may improve data quality by, for example, reducing the chance of annotators "cheating" (e.g., always selecting the second answer) (Kutlu et al., 2020). Apart from that, we expect that domain experts, who possess specific (domain) knowledge, can produce hARs that are highly useful for gaining task insight, generating data, or as gold explanations.

3.2.2 Annotation instructions

Very little work explicitly instructs annotators to provide exhaustive rationales, e.g., *"highlight ALL words that reflect this sentiment"* (Sen et al., 2020) or *"we encouraged annotators to try their best to mark as many rationales as possible"* (Lu et al., 2022).

In most cases, annotator rationales are collected using instructions like *"highlight rationales that are short and coherent, yet sufficient for supporting the label"* (Bao et al., 2018), *"select one or more sentences most useful for your decision"* (Ramírez et al., 2019), *"why do you think so?"* (Hancock et al., 2018), or *"select the k most important words in the argument"* (Jayaram and Allaway, 2021). While it is often not stated explicitly, it is unlikely that collecting rationales using these instructions results in exhaustive rationales. This lack of clarity in instructions may pose issues when using hARs as gold rationales, which we will further discuss in Section 5.2.

3.2.3 Effect on annotation time

One concern when asking annotators to provide hARs in annotation tasks is an increased annotation cost. Multiple studies report that the annotation time at most doubles when requesting extractive hARs for tasks like sentiment or topic classification (Hancock et al., 2018; Arous et al., 2021; Sullivan et al., 2022). This additional annotation time can be reduced when annotators gain experience in annotating rationales (McDonnell et al., 2016; Kutlu et al., 2020). A tentative conclusion is that annotators already subconsciously form rationales when performing the classification task, thus only requiring additional time to write or mark down the rationale (Zaidan et al., 2007; Kutlu et al., 2020). Whether this applies to abstractive hARs and more complex tasks, and how annotator experience is affected (i.e. task difficulty and enjoyment) is an open question. Alternatively to actively annotating rationales, techniques like eye-tracking might allow for passive rationale collection (Eberle et al., 2022); for example, construct a heatmap of relevant text snippets based on the annotator's gaze while performing a task.

3.2.4 Inter-annotator agreement of rationales

Rationales are much more versatile than labels. Chiang and Lee (2022) and Sullivan et al. (2022) show that rationale annotation instructions greatly affect the form and exhaustiveness of resulting human-annotated rationales. When two rationales differ in form or size, it is difficult to calculate their inter-annotator agreement (Dumitrache et al., 2018; Kreiss et al., 2020; Malik et al., 2021; Guzman et al., 2022). Many studies report inter-annotator agreement on rationales using pairwise agreement (Zaidan et al., 2007; McDonnell et al., 2016; Wang et al., 2022) or Intersection-over-Union (also known as the Jaccard index) (Malik et al., 2021; Mathew et al., 2021; Guzman et al., 2022). Furthermore, some possible agreement measures are the Ratcliff-Obershelp metric (McDonnell et al., 2016), Cohen's Kappa (DeYoung et al., 2020), Krippendorff's alpha (Carton et al., 2018), ROUGE (Malik et al., 2021), and worker quality score (WQS) (Jayaram and Allaway, 2021). The above metrics often indicate that inter-annotator agreement for rationales is low, and varies between annotators and tasks (Carton et al., 2018; Malik et al., 2021; Wang et al., 2022). Nevertheless, inter-annotator agreement for rationales usually outperforms random baselines (Kreiss et al., 2020; Mathew et al., 2021). Overall, we expect that rationales collected from multiple annotators can capture useful information about the annotation task, and may be a versatile resource for developing (robust) ML models.

4 Explainable text classification

We now connect rationales to explainable AI (XAI), starting by outlining relevant XAI concepts. XAI revolves around explaining AI models, especially the ones based on machine learning (ML). An explanation can *globally* explain a complete ML model (i.e., elucidating the working of the model as a whole), or *locally* explain a specific input-output instance (e.g., highlighting relevant words in a text). Furthermore, an XAI method can be *model-agnostic*, meaning that it is applicable to any ML model, or *model-dependent*, meaning that it is applicable to a specific (group of) model(s).

4.1 Roles in XAI

In the XAI process, there are various actor roles that can be fulfilled by humans or ML models:

- The *annotator* maps inputs to outputs, also referred to as labeler, classifier, or decision-maker. Examples are human annotators and ML classification models. In this paper, we use the term "mapping" to refer to the annotator's internal decision-making process that maps inputs to outputs.
- The *explainer* explains the output produced by the annotator to the explanation receiver. For example, humans can justify their decision, or a surrogate explanation model like LIME (Ribeiro et al., 2016) can construe an ML model's behavior.
- The *explainee* receives the explainer's explanation. Explanations can be addressed to human users, but also to an ML model that learns a task using the explanation. Human users of AI systems can be broadly divided into three groups: *ML developers* have technical knowledge about the system; *domain experts* have domain-specific knowledge; *lay users* lack both technical and domain knowledge (Ribera and Lapedriza, 2019).
- The *validator* determines the quality of the explanation produced by the explainer. The desired qualities of an explanation depend on the explainee and the goal of the explanation.

In some cases, an actor can play multiple roles; for example, annotators providing a rationale are also explainers, and a human reviewing an explanation is both the explainee and the validator.

4.2 Explanation timing

Zaidan et al. (2007) asked human annotators to provide rationales in addition to labels, thus collecting explanations *at the same time* as labels. Explanations for ML model behavior are often created at a different time than the labels. The timing of an explanation is closely related to the explainer role; *when* the explanation is constructed depends on *whom* the explainer is. Using the illustration in Figure 2, we now discuss different explanation timings.

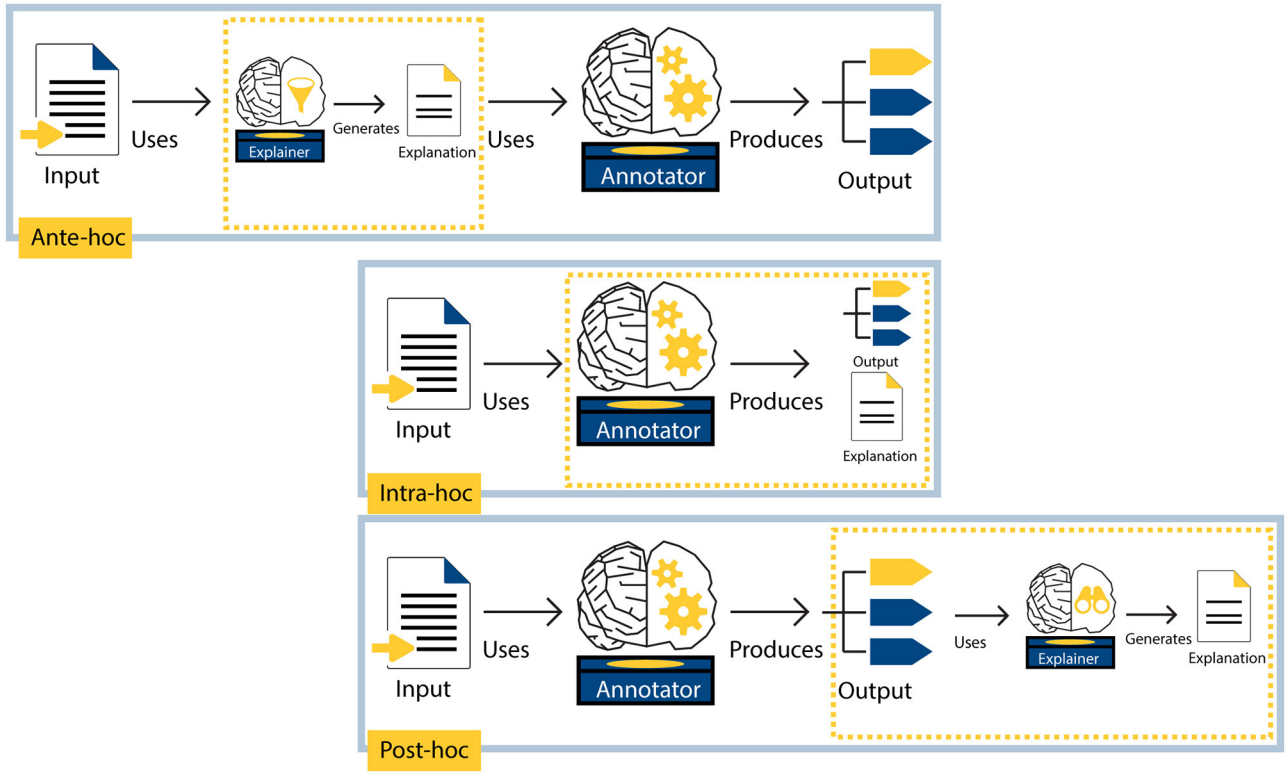

FIGURE 2
Three different timings in which explanations for AI systems can be generated. The **top row** shows an ante-hoc explanation setup, where the explainer model generates an explanation that is used as input by the annotator model to produce an output (e.g., Jain et al., 2020). The **middle row** shows an intra-hoc explanation setup, where the annotator model produces both an output and an explanation (e.g., Wiegreffe and Marasovic, 2021). The **bottom row** shows a *post-hoc* explainer that uses the output produced by the annotator model to generate an explanation (e.g., Ribeiro et al., 2016).

4.2.1 Ante-hoc explanations are created *before* the annotator's mapping

An ante-hoc explainer first generates an explanation, which is then used as input for the annotator model. For example, an ante-hoc explainer model first generates an extractive rationale by identifying rationale sentences in an input text. This extractive rationale then replaces the original input text, which the annotator model (e.g., a classification model) uses to produce an output.

The ante-hoc explanation approach has also been referred to as a pipeline (Wiegreffe et al., 2021), select-then-predict (Chrysostomou and Aletras, 2022), and explain-then-predict (Camburu et al., 2018) setup. The annotator model can be trained separately from the ante-hoc explainer model (Yessenalina et al., 2010; Jain et al., 2020) or the models can be trained jointly (Lei et al., 2016; Bastings et al., 2019). In Figure 2, the top row shows an ante-hoc explainer. Note that ante-hoc explanations do not elucidate the annotator's mapping itself: they only show which input the annotator received to perform the task.

4.2.2 Intra-hoc explanations are created *during* the annotator's mapping

When an explanation is produced *while* the annotator performs the task, such that the roles of explainer and annotator are performed by the same actor, we call this explanation intra-hoc. For example, an ML model annotator may perform a task while "thinking out loud" (Ehsan et al., 2018; Wei et al., 2022), or an ML model may classify a text and provide a free-text explanation at the same time (e.g., Wiegreffe et al., 2021; Wei et al., 2022). We call annotators that are capable of providing rationales while performing a task, thus performing the annotation and explanations tasks simultaneously, *self-rationalizing* annotators.

Alternatively, intra-hoc explanations can be constructed when an ML model is *transparent*. The *transparency* of a model refers to how accessible and interpretable the model's internal mapping from input to output is to humans. Examples of (relatively) transparent ML models include lexicon-based classifiers (Clos et al., 2017) and (small) decision trees.

As described in Section 1, Zaidan et al. (2007) collected human "annotator rationales" by asking human annotators to explain their decisions. Following their terminology, we call explanations that are generated by the annotator itself, *annotator explanations*. Here, the explainer and the annotator are the same actor, so the explainer directly accesses the mapping as performed by the annotator. We therefore consider intra-hoc explanations (e.g., the explanation produced by the annotator in the middle row in Figure 2) to be annotator explanations.

4.2.3 *Post-hoc* explanations are constructed *after* the annotator's mapping

For this type of explanation, the explainer is an external actor that uses the annotator's output to approximate an explanation after the mapping is performed (Ribeiro et al., 2016; Malik et al., 2021). The bottom row in Figure 2 illustrates a *post-hoc* explainer. One advantage of *post-hoc* explainers is that they are usually model-agnostic and applicable to black box models. For example, LIME (Ribeiro et al., 2016), where an ML model learns another ML model's behavior from its outputs, is often used (Carton et al., 2018; Mathew et al., 2021). It is important to keep in mind that *post-hoc* explainers only *approximate* annotator behavior, without actual knowledge about the steps taken inside the annotator model (Jacovi and Goldberg, 2020).

4.3 Faithfulness and plausibility

A key question regarding any explanation is whether the explanation is accurate. In XAI, the term *faithfulness* describes whether an explanation accurately reflects the mapping from input to output as performed by the annotator model (Jacovi and Goldberg, 2020). Determining the faithfulness of explanations is challenging, especially for black box models, where the actual mapping from input to output is unknown (Jacovi and Goldberg, 2020; Yin et al., 2022; Lyu et al., 2024). Furthermore, as complete faithfulness may be unattainable, Jacovi and Goldberg (2020) regard faithfulness more as a "greyscale", rather than as a binary property. The degree to which an explanation is, or can be, (un)faithful will depend on the explanation timing:

- **Ante-hoc:** Ante-hoc explainers do not access the annotator's internal mapping. Instead, the explanation consists of a modified input (e.g. a selection of words) that allows the annotator model to (better) perform the task. Using this modified input, ante-hoc explainers allow the explainee to infer which input features are (ir)relevant for the output (Jain et al., 2020; Chrysostomou and Aletras, 2022). However, the explanations do not explain *how* the annotator generated an output, and may even be barely associated with model outputs (Wiegreffe et al., 2021).
- **Intra-hoc**: A completely faithful explanation can arguably only be constructed when the inner workings of the ML model are known and interpretable, which holds for completely transparent, intra-hoc explainer models (Jacovi and Goldberg, 2020). The faithfulness of intra-hoc explainers like the attention mechanism (Jain and Wallace, 2019; Bibal et al., 2022) and self-rationalizing models (Wiegreffe et al., 2021; Lyu et al., 2023; Turpin et al., 2024) remains unclear and under debate.
- ***Post-hoc***: In contrast to intra-hoc explainers, *post-hoc* explainers do not have access to the annotator's mapping itself. There is therefore no guarantee that *post-hoc* explainers are fully faithful. For example, *post-hoc* explainers that rely on perturbations to approximate the mapping are sensitive to adversarial attacks (Slack et al., 2020). Moreover, *post-hoc* explainers often rely on the Linearity Assumption,[3] not taking into account that removing parts of the input might unintentionally create out-of-distribution inputs (Hase et al., 2021).

Related to faithfulness is *plausibility*, which describes whether humans find the explanation *convincing* (Jacovi and Goldberg, 2020). In the literature, plausibility is used to describe various notions related to human perception of an explanation, for example, the interpretability (Wood-Doughty et al., 2022), persuasiveness (Herman, 2017), sensibility (Zhong et al., 2019), usefulness (Chiang and Lee, 2022), or the degree to which the explanation is similar to human-annotated explanations (Vafa et al., 2021; El Zini et al., 2022; Schlegel et al., 2022).

4.4 Rationales: human-friendly explanations

How an AI system should be explained depends on the explainee and the explanation goal; for example, ML model developers require more technical explanations than domain experts or lay users. For non-technical users, the human-friendliness of an explanation may be much more important than its faithfulness (Carvalho et al., 2019).

Since humans often explain their behavior through natural language (Section 2), rationales can be considered a promising vehicle for conveying explanations about ML models and their behavior in a human-friendly way (Miller et al., 2017). Nonetheless, human-friendliness may be affected by various factors; for example, high-granularity rationales (e.g., words) may require more context to explain in a human-friendly manner, or rationales may be incoherent when the model uses (for humans) illogical heuristics to solve the task. Moreover, humans might prefer using abstractive rationales (e.g., a comment) over extractive rationales (e.g., highlighting words) to explain their decisions. Possible approaches to improving the human-friendliness of rationales may be combining high-granularity rationales like words into sentences (e.g., changing "awesome", "film" to "this is an awesome film") (Meldo et al., 2020; Mukhtar et al., 2023), or adding more contextual information justifying the model's behavior.

5 Human-annotated rationales in explainable text classification

In this section, we discuss the use of hARs in explainable text classification. In Section 5.1, we first briefly discuss rationales generated by ML models, called model-annotated rationales (mARs). Then, we describe various metrics used to determine agreement between mARs and hARs (Section 5.2). Finally, we discuss how hARs can be used to generate mARs (Section 5.3).

3 The assumption that different parts of the input independently influence an annotator's output (Jacovi and Goldberg, 2021).

5.1 Model-annotated rationales

As discussed in Section 2, mARs are natural language explanations provided by an ML model. Note that mARs can be provided by the annotator (i.e., the classification model), but also by another ML model (i.e., the explainer model). Such a separate explainer model can provide explanations before or after the annotator model maps an input to an output (see Section 4.2). Like hARs, model-annotated rationales (mARs) can be categorized according to the framework we introduced in Section 2.

5.1.1 Form

5.1.1.1 Granularity

Model-annotated rationales come in various granularities: words (e.g., Martens and Provost, 2014; Lundberg and Lee, 2017; Hayati et al., 2021), snippets (e.g., Carton et al., 2018; Sharma et al., 2020; Shen et al., 2022), sentences (e.g., Glockner et al., 2020; Malik et al., 2021), and paragraphs (Chalkidis et al., 2021).

5.1.1.2 Extractive and abstractive

Both extractive and abstractive mARs can be generated; identifying features in the input text that the ML model used to make a classification (e.g., Yessenalina et al., 2010; Ribeiro et al., 2016) results in extractive rationales. Abstractive mARs are created when an ML model generates natural language explanations for its predictions (e.g., Costa et al., 2018; Sap et al., 2020).

5.1.1.3 Categorical and numerical

Most hARs are categorical (see Section 2), e.g., annotators have selected text spans that explain their decision. However, mARs are often numerical values assigned to text spans, like attention weights (Bao et al., 2018; Sen et al., 2020) or saliency maps (Mohseni et al., 2021). Nevertheless, mARs can also be categorical. For example, explainers that first select a subset of the input (Jain et al., 2020), or explainers that perform discrete optimization by applying binary masks on the input (Lei et al., 2016; Bastings et al., 2019).

5.1.2 Exhaustiveness

The exhaustiveness of a mAR is highly dependent on the type of explainer; some explainers aim to produce exhaustive mARs by identifying all text spans that explain the annotator's output, for example, attention, gradient-based, or occlusion-based explainers (Bao et al., 2018; Hayati et al., 2021; Malik et al., 2021). Nevertheless, sometimes a selection is made to limit the number of selected text spans. For example, requiring the rationale size to be less than a fixed percentage of the input text (Lei et al., 2016), choosing a target rationale length (Shen et al., 2022), using a threshold to select high-scoring text spans (Chalkidis et al., 2021; Herrewijnen et al., 2021), or selecting a single most informative sentence (Glockner et al., 2020).

5.2 Evaluating model-annotated rationales

Similar to how human-provided labels are often treated as "gold labels", human-annotated rationales are often treated as "gold rationales" (Section 3.1). Model-annotated rationales are often compared against human-annotated rationales to analyse whether models make predictions based on similar reasons as humans. For example, low agreement can indicate that the model is focusing on spurious correlations (Srivastava et al., 2020; Jørgensen et al., 2022). Comparing against hARs can also provide valuable insights into other aspects. For example, Sen et al. (2020) measure the correlation between the distribution of *all* hAR and mAR words, to study whether models focus on similar categories of words (e.g., adjectives). Nevertheless, care should be taken when using hARs to evaluate mARs and using hARs to represent "human" reasoning (Sen et al., 2021). A study by Jakobsen et al. (2023) found systematic disagreements between demographic groups that were asked to annotate rationales, suggesting that uniform "human" reasoning may not exist.

We now discuss how the agreement between hARs and mARs can be determined. We survey approaches used in the literature (see Table 2), and suggest metrics that might be suited to calculate the agreement between different types of rationales.

5.2.1 Agreement between rationales

When comparing mARs against hARs, some aspects from the above sections are more relevant than others. We now discuss in more detail how agreement between rationales with different forms and degrees of exhaustiveness can be calculated. We focus on ways to measure agreement between hARs and mARs at the instance level (i.e., an individual text). Furthermore, we mainly focus on extractive mARs, as they are most often compared against hARs.

5.2.1.1 Form

The form of the rationale plays a large role in choosing a suitable metric to calculate agreement. Rationales with different *granularities* should not be mixed: word-level rationales will probably not agree with sentence-level rationales, as such rationales have a different bandwidth (Guerreiro and Martins, 2021).

As shown in Table 2, *extractive* rationales are often evaluated using evaluation metrics for classification or regression tasks, and *abstractive* rationales are usually evaluated using metrics from the Natural Language Generation (NLG) field. When evaluating extractive rationales, it can be useful to determine whether the *position* of text spans is relevant. For example, when explaining a negative sentiment label for the sentence "*I had great expectations, but this was not a great movie*", the position of *great* matters. In this case, the task can be framed as predicting values for each text span (e.g., token, sentence). In practice, text spans may not match exactly; token-level agreement metrics on human-annotator rationales also often show variability between annotators (see Section 3.2). To allow more flexibility when matching text spans, DeYoung et al. (2020) propose the IOU-F1 metric, which is a more "forgiving" metric to measure overlap between text spans. For example, when two rationales overlap more than 50% (e.g., "a really nice film" and "really nice"), this metric would count this as agreement.

For rationales with *categorical* values (e.g., a text span is part of a rationale or not), agreement is often measured using classification metrics like accuracy, precision, recall, and F1-score (see Table 2). When the values are numerical, agreement has been calculated

using metrics like the mean absolute error (MAE) (Mohseni et al., 2021) and Pearson's R (Hayati et al., 2021).

When hARs are categorical, but mARs are *numerical*, the mARs can be converted to categorical values. However, this may cause information loss: for example, when the words "okay" and "fine" received the scores 0.4 and 0.6, they can be converted to 0 and 1 using a threshold, but this will leave out the relevance of the word "okay". Therefore, we recommend using metrics applicable for comparing numerical to categorical values, like AUC (DeYoung et al., 2020; Sen et al., 2020). Furthermore, when hARs and mARs are both numerical, metrics for measuring the similarity of rankings, like the extrapolated version of the rank-biased overlap (RBO_{EXT}) (Webber et al., 2010; Jørgensen et al., 2022) can be used.

Sometimes the position of text spans is of less importance. For example, when the task is to classify whether a review is about a book (i.e., Figure 1), the phrase "*in this book*" may be a sufficient explanation, no matter where, when, or how often the phrase occurs in the input text. Here, different text similarity metrics could be applied to calculate agreement for both extractive and abstractive rationales. For example, to measure the overlap between words or n-grams in a rationale, metrics like the ROUGE, BLUE, or Meteor can be used (Sap et al., 2020; Malik et al., 2021). Furthermore, to measure the semantic similarity between rationales, metrics like BERTScore, BLEURT, and Word Mover's Distance (WMD) (Sap et al., 2020; Clinciu et al., 2021) are applicable.

5.2.1.2 Exhaustiveness

To correctly interpret agreement metrics, it is important to know the exhaustiveness of a rationale. Suppose human annotators were asked to annotate all sentences that support their decision to classify a movie review as positive or negative. In this case, the goal was to collect exhaustive hARs. When such hARs are then compared to mARs that are less exhaustive (e.g., an explainer that only selects the three most important sentences), it cannot be expected that the mARs contain all sentences included in the hARs (i.e. recall is likely to be low). In this case, precision-oriented metrics [e.g., precision or mean R-Precision (mRP)] may be more suitable. Conversely, when human annotators were not asked to annotate all supporting evidence, but the mARs do include all text spans that support a decision, recall-oriented metrics may be more suitable. When both the hARs and mARs are non-exhaustive, agreement in terms of precision and recall is expected to be lower and more difficult to interpret.

Rationales that use different words may still describe similar concepts. For example, in a review topic classification task (i.e. *is this review about a book?*), a human might have highlighted "a well-written novel" as a rationale, while an ML model explainer identified the snippet 'an engaging book' as a rationale. Then, both rationales are very similar semantically, but use different words. In such cases, evaluation metrics that take into account the semantics of the text (e.g., BERTScore) could be considered to determine agreement between the two rationales.

5.2.2 How should hARs be used in mAR evaluation?

Some work calculates the agreement between multiple human annotators (Carton et al., 2018; Malik et al., 2021), and find that even when human annotators agree on a label, they often do not completely agree when it comes to rationales. If the agreement between hARs is low, we believe it is likely that the agreement between hARs and mARs is also low.

In addition to calculating agreement between hARs and mARs, hARs can be used to put the evaluation scores of mARs into context. For example, mARs are sometimes evaluated by asking users to perform a classification task, replacing the original input with mARs (e.g., Ramírez et al., 2019; Chang et al., 2020; Jain et al., 2020). However, it might be difficult to interpret the resulting scores without a meaningful baseline. Here, hARs can be used as a reference point for comparing the mAR evaluation scores to. For example, some work replaces the original input text with hARs (Jain et al., 2020; Herrewijnen et al., 2021), finding that humans can accurately make predictions based on human rationales. Where some work evaluates mARs according to their readability (Jain et al., 2020) or length (Shen et al., 2022), it can be informative to compare these scores against evaluation scores for hARs (Jain et al., 2020; Wiegreffe et al., 2021).

Finally, hARs should not be viewed as a benchmark for faithful mARs; Carton et al. (2020) apply faithfulness measures to hARs and ML models, and find that human rationales do not fare well under faithfulness evaluation metrics. This can be expected, as task-solving processes may differ between ML models and humans (Sen et al., 2021; Ju et al., 2022).

5.3 Generating model-annotated rationales

Apart from using hARs to *evaluate* mARs, hARs can also serve as examples from which ML models can learn to *generate* mARs. One effect of using hARs to generate mARs, is that the resulting mARs are likely to resemble hARs. A positive aspect of this is that the mARs are more likely to be human-friendly (Section 4.4). A possible downside of this is that the generated mARs might not faithfully represent model behavior, since the mARs are based on human annotator behavior (Section 4.3).

In this section, we will discuss how hARs have been used to train ML models to generate natural language explanations (i.e., rationales). Table 3 provides an overview of work that uses hARs to generate mARs for text classification models.

5.3.1 Ante-hoc

An ante-hoc explainer model *first* generates an explanation, which is then used by an annotator model to perform a task (e.g., classification) (Section 4.2). Using hARs, ante-hoc explainer models can learn to construct rationales. For example, Tepper et al. (2013) train an explainer model on hARs to identify rationale sentences in a text, which are then used as input features for their medical classification model. Furthermore, explainer models can learn to generate abstractive mARs from hARs, which can be used as input for the annotator model (Wiegreffe et al., 2021). The latter is comparable to the data generation aim as described in Section 3.1, but with the focus on explainability.

TABLE 2 Overview of work that uses hARs to evaluate their generated mARs.

Related work	Extractive								Abstractive		
	Categorical					Numerical					
Titov and McDonald (2008)	P	R									
Yessenalina et al. (2010)	P	R	F								
McAuley et al. (2012)		P	R								
Tepper et al. (2013)	P	R	F		IOU-F1						
Marshall et al. (2015)	P	R	F								
Lei et al. (2016)	P										
Bao et al. (2018)				C							
Carton et al. (2018)	P	R	F								
Bastings et al. (2019)	P										
Chang et al. (2020)	P	R	F								
Glockner et al. (2020)	P	R	F								
Paranjape et al. (2020)			F		IOU-F1						
Sap et al. (2020)									R	B	WMD
Sen et al. (2020)						PCC	AUC				
Arous et al. (2021)	P	R									
Chalkidis et al. (2021)			F					mRP			
Guerreiro and Martins (2021)			F								
Hayati et al. (2021)						PCC					
Malik et al. (2021)					IOU-F1				R	B	M
Mathew et al. (2021)			F				AUC				
Mohseni et al. (2021)								MAE			
Sharma et al. (2020)			F		IOU-F1						
Jørgensen et al. (2022)							AUC	RBO_{EXT}			
Shen et al. (2022)	P	R	F								
Bujel et al. (2023)	P		F								

Abbreviations are as follows (from left to right): For extractive categorical rationales: Precision, Recall, F1-score, Cosine similarity, and Intersection-Over-Union. For extractive numerical rationales: Pearson's Correlation Coeficcient, Mean Absolute Error, Area Under the Precision-Recall curve, mean R-Precision, extrapolated Rank-Biased Overlap. For abstractive rationales: Rouge, Blue, Meteor, and Word Mover's Distance.

5.3.2 Intra-hoc

In intra-hoc explanation setups, explanations are constructed *while* the annotator model performs the task (Section 4.2). Annotator models can use hARs as "guidelines" for performing the task. For example, using attention regularization, an attention layer (Bibal et al., 2022) is encouraged to focus on the same text spans as the human rationale examples (Bao et al., 2018; Kanchinadam et al., 2020; Pruthi et al., 2022, inter alia). After training, the attention layer can be inspected to identify rationale tokens. Moreover, a *self-rationalizing* annotator model can learn to simultaneously classify a text and generate a rationale based on pairs of human labels and hARs (Sap et al., 2020; Wiegreffe et al., 2021). One example of a self-rationalizing annotator that generates abstractive rationales, is a large language model that is prompted to produce a chain of thought (CoT). Here, the annotator encapsulates the explanation within the output (Wei et al., 2022).

5.3.3 *Post-hoc*

One understudied research direction is using hARs to generate rationales *after* the annotator model has produced an output (i.e., a prediction). One example from the field of text summarization is work by Li et al. (2020), who construct abstractive summaries from keywords in the input texts. When applied to explainable text classification, such strategies could be applicable to constructing low-granularity mARs (e.g., sentences) from high-granularity mARs (e.g., words) and human examples of low-granularity hARs.

6 Conclusion and discussion

In this survey, we have given an overview of natural language explanations, also called rationales, in explainable text classification. Throughout this survey, we have focussed on

TABLE 3 An overview of work that uses hARs to generate mARs for text classification tasks.

	Dataset	Form	Granularity	Value type	*Post-hoc*	Ante-hoc	Intra-hoc	Method
Tepper et al. (2013)	CPIS/PNA	E	S	C		A		Supervised explainer
Zhang et al. (2016)	IMDB, RoB	E	S	N		A		Supervised explainer
Bao et al. (2018)	BeerAdvocate	E	Sn	N			I	Attention regularization
Strout et al. (2019)	IMDB	E	S	C			I	Supervised explainer
Chang et al. (2020)	IMDB, BeerAdvocate	E	Sn	C		A		Supervised explainer
Glockner et al. (2020)	IMDB	E	Se	C			I	Supervised explainer
Herrewijnen (2020)	IMDB	E	S	C		A		Supervised explainer
Jain et al. (2020)	IMDB	E	Sn	C		A		Supervised explainer
Sap et al. (2020)	SBIC	A	Sn	C			I	NLG model
Arous et al. (2021)	Wiki-tech, Amazon	E	Sn	C	P		I	Attention regularization
Guerreiro and Martins (2021)	IMDB, BeerAdvocate, SST	E	Sn	C			I	Attention regularization
Mathew et al. (2021)	HateExplain	E	Sn	C			I	Attention regularization

Form is abbreviated as Abstractive and Extractive. Granularity is abbreviated as Paragraphs, Sentences, Snippets, and Words. Value type is abbreviated as Categorical and Numerical.

"annotator rationales" as introduced by Zaidan et al. (2007), which are human-annotated highlights explaining "why" a text should receive a particular label. In this section, we provide a concrete list of recommendations for using human-annotated rationales (hARs) in explainable text classification.

6.1 Collect human-annotated rationales by default

While collecting hARs increases required annotation time, it is beneficial to data quality, task insight, and data richness (Section 3.2). Therefore, we call for including rationale collection in labeling tasks by default, whenever possible.

However, there are classification tasks for which it is difficult to collect high-quality hARs. In fact, for some tasks, computational methods are used precisely because the task itself is difficult to carry out by humans. An example is the task of authorship attribution (i.e. deciding who wrote a text), where fine-grained distributional differences in character n-grams or function words have shown to be effective (Grieve, 2007). Future work should explore the collection of human-annotated rationales across a wider variety of tasks, to further our understanding of their applicability, benefits, and limitations.

6.2 Be specific in instructions for collecting human-annotated rationales

The instructions given to a human annotator affect the form (Section 2.1), exhaustiveness (Section 2.2), and inter-annotator agreement (Section 3.2) of the collected rationales. However, the instructions given to human annotators for annotating rationales vary greatly across surveyed work. Knowing which aspects apply to a set of hARs is imperative for using these hARs in explainable text classification (Section 5, Chiang and Lee, 2022). Moreover, we believe that to collect rationales that are consistent in form and exhaustiveness, it is vital to be precise when instructing human annotators to provide rationales. Work that tailors annotation instructions, aiming to model the human decision-making process (e.g., Lamm et al., 2021; Ray Choudhury et al., 2023), might be an inspiring starting point for constructing precise instructions for rationale collection tasks.

6.3 Exploit human-annotated rationales for ML model training

Human-annotated rationales are often collected with the aim of improving ML model training (Section 3.1). Their use during model training has led to improved performance on various classification tasks. Therefore, we believe rationales have great potential for the training of ML models. Furthermore, hARs may prevent models from learning spurious correlations. Whether these benefits hold for other NLP tasks (e.g., Carton et al., 2022), is a question that future research should investigate.

6.4 Be cautious with using hARs as "gold rationales"

In Section 5, we looked into the usage of hARs in explainable AI. As described in Section 5.2.1, hARs can be used as gold rationales that mARs should agree with. However, because the

hARs and mARs are often not comparable (e.g., in terms of their form and in terms of their exhaustiveness), comparing them can be misleading or uninformative (Section 5.2.2). Therefore, we believe that when comparing two rationales, it is imperative to take into account the (differences in) form and exhaustiveness of these rationales.

Furthermore, when using hARs as gold rationales, it needs to be established why, and for what purposes, these hARs can justifiably be considered to be gold rationales. For example, a specific hAR may be considered an exhaustive reason for a decision, but this may not hold for a hAR collected through different annotation instructions.

Finally, agreement between rationales has been calculated using various approaches in the literature, but a unified approach is lacking. Future work should focus on developing clear and uniform metrics for calculating rationale agreement. Aside from calculating the agreement between hARs and mARs, we recommend using hARs as a baseline for various NLG and explainability evaluation metrics (Section 5.2.2). For example, by comparing the readability score of an mAR to the readability score of a hAR, the mAR's scores can be put into context.

6.5 Use hARs as inspiration for generating (human-friendly) mARs

Natural language allows humans to provide explanations that are framed in terms of the knowledge of the explainee (Miller et al., 2017). Accordingly, we believe that rationales are a promising format for explaining ML model behavior in a manner that is human-friendly. Furthermore, because rationales can use domain-specific jargon, we expect that such rationales are an especially suitable explanation format for explaining ML model behavior to domain experts.

In Section 5.3, we briefly discussed how hARs can act as example explanations that explainer models can learn from. One advantage of using hARs as examples for generating mARs, is that the generated mARs are more likely to be human-friendly. Therefore, we believe that hARs form a foundation for explaining the decisions of AI systems to non-technical users working with these systems.

6.6 Final remarks and future work

In this survey, we have highlighted the potential of human-annotated rationales for explainable text classification. Some of our recommendations call for further research. For example, we believe that the scientific community would benefit from the construction of new datasets containing human-annotated rationales. Moreover, we believe that it would be important to investigate how, when, and for what tasks, human-annotated rationales can aid data collection and model training. Finally, our findings suggest that human-annotated rationales are not limited to NLP alone, but that they are a promising tool for other areas of research as well, which has the potential to enrich the entire field of XAI.

Author contributions

EH: Conceptualization, Visualization, Writing—original draft. DN: Conceptualization, Supervision, Writing—review & editing. KD: Supervision, Writing—review & editing. FB: Supervision, Writing—review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. EH was funded by the Netherlands National Police Lab AI.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abedin, M. A. U., Ng, V., and Khan, L. R. (2011). "Learning cause identifiers from annotator rationales," in *Twenty-Second International Joint Conference on Artificial Intelligence* (Washington DC: AAAI Press), 1758–1763.

Arora, S., and Nyberg, E. (2009). "Interactive annotation learning with indirect feature voting," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Student Research Workshop and Doctoral Consortium* (Stroudsburg: Association for Computational Linguistics), 55–60.

Arous, I., Dolamic, L., Yang, J., Bhardwaj, A., Cuccu, G., and Cudré-Mauroux, P. (2021). "MARTA: leveraging human rationales for explainable text classification," in *Proceedings of the AAAI Conference on Artificial Intelligence* (Washington DC: AAAI Press), 5868–5876.

Atkinson, D., Srinivasan, K. B., and Tan, C. (2019). "What gets echoed? Understanding the "pointers" in explanations of persuasive arguments," 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)* (Hong Kong: Association for Computational Linguistics), 2911–2921.

Bao, Y., Chang, S., Yu, M., and Barzilay, R. (2018). "Deriving machine attention from human rationales," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Brussels: Association for Computational Linguistics), 1903–1913.

Bastings, J., Aziz, W., and Titov, I. (2019). "Interpretable neural predictions with differentiable binary variables," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Stroudsburg: Association for Computational Linguistics), 2963–2977.

Bibal, A., Cardon, R., Alfter, D., Wilkens, R., Wang, X., François, T., et al. (2022). "Is attention explanation? An introduction to the debate," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Dublin: Association for Computational Linguistics), 3889–3900.

Bujel, K., Caines, A., Yannakoudakis, H., and Rei, M. (2023). *Finding the Needle in a Haystack: Unsupervised Rationale Extraction from Long Text Classifiers. arXiv [Preprint]. arXiv*:2303.07991.

Cambria, E., Malandri, L., Mercorio, F., Mezzanzanica, M., and Nobani, N. (2023). A survey on XAI and natural language explanations. *Inform. Proc. Manage.* 60:103111. doi: 10.1016/j.ipm.2022.103111

Camburu, O.-M., Rocktäschel, T., Lukasiewicz, T., and Blunsom, P. (2018). "e-SNLI: natural language inference with natural language explanations," in *Advances in Neural Information Processing Systems 31* (NeurIPS 2018) (Montreal, QC: NeurIPS), 9539–9549. doi: 10.1016/j.ipm.2022.103245

Carton, S., Kanoria, S., and Tan, C. (2022). "What to learn, and how: toward effective learning from rationales," in Findings *of the Association for Computational Linguistics: ACL 2022* (Dublin: Association for Computational Linguistics), 1075–1088.

Carton, S., Mei, Q., and Resnick, P. (2018). "Extractive adversarial networks: high-recall explanations for identifying personal attacks in social media posts," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Brussels: Association for Computational Linguistics), 3497–3507.

Carton, S., Rathore, A., and Tan, C. (2020). "Evaluating and characterizing human rationales," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Online: Association for Computational Linguistics), 9294–9307. doi: 10.18653/v1/2020.emnlp-main.747

Carvalho, D. V., Pereira, E. M., and Cardoso, J. S. (2019). Machine learning interpretability: a survey on methods and metrics. *Electronics* 8:832. doi: 10.3390/electronics8080832

Chalkidis, I., Fergadiotis, M., Tsarapatsanis, D., Aletras, N., Androutsopoulos, I., and Malakasiotis, P. (2021). "Paragraph-level rationale extraction through regularization: a case study on european court of human rights cases," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Online: Association for Computational Linguistics), 226–241.

Chang, S., Zhang, Y., Yu, M., and Jaakkola, T. (2020). "Invariant rationalization," in *International Conference on Machine Learning* (Honolulu, HI: International Conference on Machine Learning), 1448–1458.

Chen, H., He, J., Narasimhan, K., and Chen, D. (2022). "Can rationalization improve robustness?," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Seattle: Association for Computational Linguistics), 3792–3805.

Chhatwal, R., Gronvall, P., Huber-Fliflet, N., Keeling, R., Zhang, J., and Zhao, H. (2018). "Explainable text classification in legal document review a case study of explainable predictive coding," in *IEEE International Conference on Big Data (IEEE BigData 2018)* (Seattle: IEEE), 1905–1911.

Chiang, C.-H., and Lee, H. (2022). "Re-examining human annotations for interpretable NLP," in *Explainable Agency in Artificial Intelligence Workshop*, (Arlington, VA: AAAI Press), 25–34.

Chrysostomou, G., and Aletras, N. (2022). "An empirical study on explanations in out-of-domain settings," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Stroudsburg: Association for Computational Linguistics), 6920–6938.

Clinciu, M.-A., Eshghi, A., and Hastie, H. (2021). "A study of automatic metrics for the evaluation of natural language explanations," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (Online: Association for Computational Linguistics), 2376–2387.

Clos, J., Wiratunga, N., and Massie, S. (2017). "Towards explainable text classification by jointly learning lexicon and modifier terms," in *IJCAI-17 Workshop on Explainable AI (XAI)* (Florence: Association for Computational Linguistics), 19.

Costa, F., Ouyang, S., Dolog, P., and Lawlor, A. (2018). "Automatic generation of natural language explanations," in *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion, IUI '18 Companion* (New York: Association for Computing Machinery), 57:1–57:2.

Crowston, K. (2012). "Amazon mechanical turk: a research tool for organizations and information systems scholars," in *Shaping the Future of ICT Research. Methods and Approaches*, eds. A. Bhattacherjee, and B. Fitzgerald (Berlin: Springer Berlin Heidelberg), 210–221.

Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., and Sen, P. (2020). "A survey of the state of explainable AI for natural language processing," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing* (Suzhou: Association for Computational Linguistics), 447–459.

DeYoung, J., Jain, S., Rajani, N. F., Lehman, E., Xiong, C., Socher, R., et al. (2020). "ERASER: a benchmark to evaluate rationalized NLP models," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Stroudsburg: Association for Computational Linguistics), 4443–4458.

Drapeau, R., Chilton, L., Bragg, J., and Weld, D. (2016). "MicroTalk: using argumentation to improve crowdsourcing accuracy," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (Washington DC: AAAI Press), 32–41.

Dumitrache, A., Inel, O., Aroyo, L., Timmermans, B., and Welty, C. (2018). "CrowdTruth 2.0: quality metrics for crowdsourcing with disagreement," in *1st Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing, and Short Paper 1st Workshop on Disentangling the Relation Between Crowdsourcing and Bias Management, SAD+ CrowdBias 2018* (Zürich: CEUR-WS), 11–18.

Eberle, O., Brandl, S., Pilot, J., and Søgaard, A. (2022). Do transformer models show similar attention patterns to task-specific human gaze? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Dublin: PMLR), 4295–4309.

Ehsan, U., Harrison, B., Chan, L., and Riedl, M. O. (2018). "Rationalization: a neural machine translation approach to generating natural language explanations," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES '18* (New York: Association for Computing Machinery), 81–87.

Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., and Riedl, M. O. (2019). "Automated rationale generation: a technique for explainable AI and its effects on human perceptions," in *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray: Association for Computing Machinery), 263–274.

El Zini, J., Mansour, M., Mousi, B., and Awad, M. (2022). "On the evaluation of the plausibility and faithfulness of sentiment analysis explanations," in *Artificial Intelligence Applications and Innovations*, eds. I. Maglogiannis, L. Iliadis, J. Macintyre, and P. Cortez (Cham. Springer International Publishing), 338–349.

Fomicheva, M., Lertvittayakumjorn, P., Zhao, W., Eger, S., and Gao, Y. (2021). "The Eval4NLP shared task on explainable quality estimation: overview and results," in *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems* (Punta Cana: Association for Computational Linguistics), 165–178.

Glockner, M., Habernal, I., and Gurevych, I. (2020). "Why do you think that? Exploring faithful sentence-level rationales without supervision," in *Findings of the Association for Computational Linguistics: EMNLP 2020* (Online: Association for Computational Linguistics), 1080–1095.

Grieve, J. (2007). Quantitative authorship attribution: an evaluation of techniques. *Liter. Lingu. Comp.* 22:251–270. doi: 10.1093/llc/fqm020

Guerreiro, N. M., and Martins, A. F. (2021). "SPECTRA: sparse structured text rationalization," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, (Punta Cana: Association for Computational Linguistics), 6534–6550.

Gurrapu, S., Kulkarni, A., Huang, L., Lourentzou, I., and Batarseh, F. A. (2023). Rationalization for explainable NLP: a survey. *Front. Artif. Intellig.* 6:1225093. doi: 10.3389/frai.2023.1225093

Guzman, E. M., Schlegel, V., and Batista-Navarro, R. T. (2022). "RaFoLa: a rationale-annotated corpus for detecting indicators of forced labour," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (Marseille: European Language Resources Association), 3610–3625.

Hancock, B., Varma, P., Wang, S., Bringmann, M., Liang, P., and Ré, C. (2018). "Training classifiers with natural language explanations," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Melbourne: Association for Computational Linguistics), 1884–1895.

Hartmann, M., and Sonntag, D. (2022). "A survey on improving NLP models with human explanations," in *Proceedings of the First Workshop on Learning with Natural Language Supervision* (Dublin: Association for Computational Linguistics), 40–47.

Hasanain, M., Barkallah, Y., Suwaileh, R., Kutlu, M., and Elsayed, T. (2020). "ArTest: the first test collection for arabic web search with relevance rationales," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '20* (New York: Association for Computing Machinery), 2017–2020.

Hase, P., Xie, H., and Bansal, M. (2021). "The out-of-distribution problem in explainability and search methods for feature importance explanations," in *35th Conference on Neural Information Processing Systems (NeurIPS 2021)* (New Orleans, LA: NeurIPS), 3650–3666. doi: 10.48550/arXiv.2106.00786

Hayati, S. A., Kang, D., and Ungar, L. (2021). "Does BERT learn as humans perceive? Understanding linguistic styles through Lexica," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (Online and Punta Cana: Association for Computational Linguistics), 6323-6331.

Herman, B. (2017). *The Promise and Peril of Human Evaluation for Model Interpretability*. Long Beach, CA: NeurIPS. doi: 10.48550/arXiv.1711.07414

Herrewijnen, E. (2020). *Machine-Annotated Rationales: Faithfully Explaining Machine Learning Models for Text Classification* (Master's thesis), Utrecht University, Utrecht, Netherlands.

Herrewijnen, E., Nguyen, D., Mense, J., and Bex, F. (2021). "Machine-annotated rationales: faithfully explaining text classification," in *Proceedings for the Explainable Agency in AI Workshop at the 35th AAAI Conference on Artificial Intelligence* (Washington DC: AAAI Press), 11–18.

Jacovi, A., and Goldberg, Y. (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* (Online: Association for Computational Linguistics), 4198–4205.

Jacovi, A., and Goldberg, Y. (2021). Aligning faithful interpretations with their social attribution. *Trans. Assoc. Comput. Linguist*. 9, 294–310. doi: 10.1162/tacl_a_00367

Jain, S., and Wallace, B. C. (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, NAACL-HLT 2019*, eds. J. Burstein, C. Doran and T. Solorio (Minneapolis, MN: Association for Computational Linguistics), 3543–3556.

Jain, S., Wiegreffe, S., Pinter, Y., and Wallace, B. C. (2020). "Learning to faithfully rationalize by construction," in *Proceedings of the Association for Computational Linguistics (ACL)* (Online: Association for Computational Linguistics), 4459–4473.

Jakobsen, T., Sasha, Terne Cabello, L., and Søgaard, A. (2023). "Being right for whose right reasons?," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Toronto: Association for Computational Linguistics), 1033–1054.

Jayaram, S., and Allaway, E. (2021). "Human rationales as attribution priors for explainable stance detection," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (Online and Punta Cana: Association for Computational Linguistics), 5540–5554.

Jørgensen, R., Caccavale, F., Igel, C., and Søgaard, A. (2022). "Are multilingual sentiment models equally right for the right reasons?," in *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP* (Abu Dhabi: Association for Computational Linguistics), 131–141.

Joshi, B., Chan, A., Liu, Z., and Ren, X. (2022). "ER-TEST evaluating explanation regularization methods for NLP models," in *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)* (Seattle:Association for Computational Linguistics), 93–109.

Ju, Y., Zhang, Y., Yang, Z., Jiang, Z., Liu, K., and Zhao, J. (2022). "Logic traps in evaluating attribution scores," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Dublin: Association for Computational Linguistics), 5911–5922.

Kanchinadam, T., Westpfahl, K., You, Q., and Fung, G. (2020). "Rationale-based human-in-the-loop via supervised attention," in *Proceedings of the 1st Workshop on Data Science with Human in the Loop (DaSH) at 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (New York, NY: Association for Computational Linguistics).

Kartal, Y. S., and Kutlu, M. (2020). "TrClaim-19: the first collection for turkish check-worthy claim detection with annotator rationales," in *Proceedings of the 24th Conference on Computational Natural Language Learning* (Online: Association for Computational Linguistics), 386–395.

Kaushik, D., Hovy, E., and Lipton, Z. (2019). *Learning the Difference That Makes a Difference with Counterfactually-Augmented Data*. Vienna: International Conference on Learning Representations.

Khashabi, D., Chaturvedi, S., Roth, M., Upadhyay, S., and Roth, D. (2018). "Looking beyond the surface: A challenge set for reading comprehension over multiple sentences," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (New Orleans: Association for Computational Linguistics), 252–262.

Kreiss, E., Wang, Z., and Potts, C. (2020). Modeling Subjective Assessments of Guilt in Newspaper Crime Narratives. In Proceedings of the 24th Conference on Computational Natural Language Learning, pages 56-68. Association for Computational Linguistics. doi: 10.18653/v1/2020.conll-1.5

Krening, S., Harrison, B., Feigh, K. M., Isbell, C. L., Riedl, M., and Thomaz, A. (2016). Learning from explanations using sentiment and advice in RL. *IEEE Trans, Cognitive Dev. Syst*. 9, 44–55. doi: 10.1109/TCDS.2016.2628365

Kumar, S., and Talukdar, P. (2020). "NILE: natural language inference with faithful natural language explanations," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Online: Association for Computational Linguistics), 8730–8742.

Kutlu, M., McDonnell, T., Elsayed, T., and Lease, M. (2020). Annotator rationales for labeling tasks in crowdsourcing. *J. Artif. Intellig. Res*. 69, 143–189. doi: 10.1613/jair.1.12012

Lamm, M., Palomaki, J., Alberti, C., Andor, D., Choi, E., Soares, L. B., et al. (2021). QED: A Framework and dataset for explanations in question answering. *Trans. Assoc. Comp. Linguist*. 9, 790–806. doi: 10.1162/tacl_a_00398

Lei, T., Barzilay, R., and Jaakkola, T. (2016). "Rationalizing neural predictions," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (Austin: Association for Computational Linguistics), 107–117.

Li, H., Zhu, J., Zhang, J., Zong, C., and He, X. (2020). "Keywords-guided abstractive sentence summarization," in *Proceedings of the AAAI Conference on Artificial Intelligence* (Washington DC: AAAI Press), 8196–8203.

Li, Z., Chen, M., Huang, L., and Ng, V. (2015). "Recovering traceability links in requirements documents," in *Proceedings of the Nineteenth Conference on Computational Natural Language Learning* (Beijing: Association for Computational Linguistics), 237–246.

Lin, H., and Ng, V. (2019). "Abstractive summarization: a survey of the state of the art," in *Proceedings of the AAAI Conference on Artificial Intelligence* (Washington DC: AAAI Press), 9815–9822.

Lu, J., Yang, L., Namee, B., and Zhang, Y. (2022). "A rationale-centric framework for human-in-the-loop machine learning," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Dublin: Association for Computational Linguistics), 6986–6996.

Lundberg, S. M., and Lee, S.-I. (2017). "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, eds. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan (Long Beach, CA: NeurIPS), 4765–4774.

Lyu, Q., Apidianaki, M., and Callison-Burch, C. (2024). "Towards faithful model explanation in NLP: a survey," in *Computational Linguistics* (Leiden), 1–70.

Lyu, Q., Havaldar, S., Stein, A., Zhang, L., Rao, D., Wong, E., et al. (2023). "Faithful chain-of-thought reasoning," in *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, eds. J. C. Park, Y. Arase, B. Hu, W. Lu, D. Wijaya, A. Purwarianti, et al. (Nusa Dua: Association for Computational Linguistics), 305–329.

Malik, V., Sanjay, R., Nigam, S. K., Ghosh, K., Guha, S. K., Bhattacharya, A., et al. (2021). "ILDC for CJPE: indian legal documents corpus for court judgment prediction and explanation," 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)* (Online: Association for Computational Linguistics), 4046–4062.

Marshall, I. J., Kuiper, J., and Wallace, B. C. (2015). Automating risk of bias assessment for clinical trials. *IEEE J. Biomed. Health Inform*. 19, 1406–1412. doi: 10.1109/JBHI.2015.2431314

Martens, D., and Provost, F. J. (2014). Explaining data-driven document classifications. *MIS Q*. 38, 73–99. doi: 10.25300/MISQ/2014/38.1.04

Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., and Mukherjee, A. (2021). "HateXplain: a benchmark dataset for explainable hate speech detection," in *Proceedings of the AAAI Conference on Artificial Intelligence* (Washington DC: AAAI Press), 14867–14875.

McAuley, J., Leskovec, J., and Jurafsky, D. (2012). "Learning attitudes and attributes from multi-aspect reviews," in *2012 IEEE 12th International Conference on Data Mining* (IEEE), 1020–1025.

McDonnell, T., Lease, M., Kutlu, M., and Elsayed, T. (2016). "Why is that relevant? Collecting annotator rationales for relevance judgments," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (Washington DC: AAAI Press), 139–148.

Meldo, A., Utkin, L., Kovalev, M., and Kasimov, E. (2020). The natural language explanation algorithms for the lung cancer computer-aided diagnosis system. *Artif. Intell. Med*. 108:101952. doi: 10.1016/j.artmed.2020.101952

Miller, T., Howe, P., and Sonenberg, L. (2017). "Explainable AI: beware of inmates running the asylum," in *Proceedings of the Workshop on Explainable Artificial Intelligence (XAI) (IJCAI 2017)* (Melbourne, VIC: IJCAI).

Mohseni, S., Block, J. E., and Ragan, E. (2021). "Quantitative evaluation of machine learning explanations: a human-grounded benchmark," in *26th International Conference on Intelligent User Interfaces, IUI '21* (New York: Association for Computing Machinery), 22–31.

Mukhtar, A., Hofer, B., Jannach, D., and Wotawa, F. (2023). Explaining software fault predictions to spreadsheet users. *J. Syst. Softw*. 201:111676. doi: 10.1016/j.jss.2023.111676

Paranjape, B., Joshi, M., Thickstun, J., Hajishirzi, H., and Zettlemoyer, L. (2020). "An information bottleneck approach for controlling conciseness in rationale extraction," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Online: Association for Computational Linguistics), 1938–1952.

Pruthi, D., Bansal, R., Dhingra, B., Soares, L. B., Collins, M., Lipton, Z. C., et al. (2022). Evaluating explanations: how much do explanations from the teacher aid students? *Trans. Assoc. Comp. Linguist*. 10, 359–375. doi: 10.1162/tacl_a_00465

Ramírez, J., Baez, M., Casati, F., and Benatallah, B. (2019). "Understanding the impact of text highlighting in crowdsourcing tasks," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (Washington DC: AAAI Press), 144–152.

Ray Choudhury, S., Atanasova, P., and Augenstein, I. (2023). "Explaining interactions between text spans," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, eds. H. Bouamor, J. Pino, and K. Bali (Singapore: Association for Computational Linguistics), 12709–12730.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). ""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* (New York, NY: Association for Computing Machinery), 1135–1144.

Ribera, M., and Lapedriza, A. (2019). "Can we do better explanations? A proposal of user-centered explainable AI," in *IUI Workshops* (New York, NY: Association for Computing Machinery), 38.

Saleem, S., Prasad, R., Vitaladevuni, S., Pacula, M., Crystal, M., Marx, B., et al. (2012). "Automatic detection of psychological distress indicators and severity assessment from online forum posts," in *Proceedings of COLING 2012* (Mumbai: The COLING 2012 Organizing Committee), 2375–2388.

Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., and Choi, Y. (2020). "Social bias frames: reasoning about social and power implications of language," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Online: Association for Computational Linguistics), 5477–5490.

Schlegel, V., Mendez-Guzman, E., and Batista-Navarro, R. (2022). "Towards human-centred explainability benchmarks for text classification," in *Proceedings of the 1st Workshop on Novel Evaluation Approaches for Text Classification Systems (NEATCLasS)* (Limassol: NEATCLasS).

Sen, C., Hartvigsen, T., Yin, B., Kong, X., and Rundensteiner, E. (2020). "Human attention maps for text classification: do humans and neural networks focus on the same words?," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Online: Association for Computational Linguistics), 4596–4608.

Sen, I., Samory, M., Floeck, F., and Wagner, C. (2021). "What human rationales tell us about machine explanations," in *Non Archival Submission at the First Workshop on Bridging Human-Computer Interaction and Natural Language Processing* (Online: Association for Computational Linguistics).

Sharma, A., Miner, A., Atkins, D., and Althoff, T. (2020). "A computational approach to understanding empathy expressed in text-based mental health support," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Online: Association for Computational Linguistics), 5263–5276.

Sharma, M., and Bilgic, M. (2018). Learning with rationales for document classification. *Mach. Learn.* 107, 797–824. doi: 10.1007/s10994-017-5671-3

Sharma, M., Zhuang, D., and Bilgic, M. (2015). "Active learning with rationales for text classification," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Denver: Association for Computational Linguistics), 441–451.

Shen, H., Wu, T., Guo, W., and Huang, T.-H. (2022). "Are shortest rationales the best explanations for human understanding?," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Dublin: Association for Computational Linguistics), 10–19.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020). "Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, AIES '20* (New York, NY: Association for Computing Machinery), 180–186.

Srivastava, M., Hashimoto, T., and Liang, P. (2020). "Robustness to spurious correlations via human annotations," in *International Conference on Machine Learning* (New York City, NY: PMLR), 9109–9119.

Srivastava, S., Labutov, I., and Mitchell, T. (2017). "Joint concept learning and semantic parsing from natural language explanations," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (Copenhagen: Association for Computational Linguistics), 1527–1536,

Strout, J., Zhang, Y., and Mooney, R. (2019). "Do human rationales improve machine explanations?," in *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (Florence: Association for Computational Linguistics), 56–62.

Sullivan, J.r., J., Brackenbury, W., McNutt, A., Bryson, K., Byll, K., et al. (2022). "Explaining why: how instructions and user interfaces impact annotator rationales when labeling text data," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Seattle: Association for Computational Linguistics), 521–531.

Tepper, M., Evans, H. L., Xia, F., and Yetisgen-Yildiz, M. (2013). "Modeling annotator rationales with application to pneumonia classification," in *Proceedings of Expanding the Boundaries of Health Informatics Using AI Workshop of AAAI* (Washington DC: AAAI Press).

Titov, I., and McDonald, R. (2008). "A joint model of text and aspect ratings for sentiment summarization," in *Proceedings of ACL-08: HLT* (Columbus: Association for Computational Linguistics), 308–316.

Turpin, M., Michael, J., Perez, E., and Bowman, S. (2024). "Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting," in *Advances in Neural Information Processing Systems*, 36 (New Orleans, LA: NeurIPS).

Uma, A. N., Fornaciari, T., Hovy, D., Paun, S., Plank, B., and Poesio, M. (2022). Learning from disagreement: a survey. *J. Artif. Int. Res.* 72, 1385–1470. doi: 10.1613/jair.1.12752

Vafa, K., Deng, Y., Blei, D. M., and Rush, A. M. (2021). "Rationales for sequential predictions," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021* (Punta Cana: Association for Computational Linguistics), 10314-10332.

Vidgen, B., Nguyen, D., Margetts, H., Rossini, P., and Tromble, R. (2021). "Introducing CAD: the contextual abuse dataset," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Online: Association for Computational Linguistics), 2289–2303.

Volkova, S., and Yarowsky, D. (2014). "Improving gender prediction of social media users via weighted annotator rationales," in *NeurIPS 2014 Workshop on Personalization* (Vancouver, BC: NeurIPS).

Wang, J., Sharma, M., and Bilgic, M. (2022). "Ranking-constrained learning with rationales for text classification," in *Findings of the Association for Computational Linguistics: ACL 2022* (Dublin: Association for Computational Linguistics), 2034–2046.

Wang, Z., Qin, Y., Zhou, W., Yan, J., Ye, Q., Neves, L., et al. (2020). "Learning from explanations with neural execution tree," in *8th International Conference on Learning Representations* (Hong Kong: OpenReview.net).

Webber, W., Moffat, A., and Zobel, J. (2010). A similarity measure for indefinite rankings. *ACM Trans. Inform. Syst.* (*TOIS)* 28, 1–38. doi: 10.1145/1852102.1852106

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., et al. (2022). "Chain of-thought prompting elicits reasoning in large language models," in *36th Conference on Neural Information Processing Systems (NeurIPS 2022)* (Vancouver, BC: NeurIPS), 35, 24824–24837.

Wiegreffe, S., and Marasovic, A. (2021). "Teach me to explain: a review of datasets for explainable natural language processing," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)* (Online: NeurIPS).

Wiegreffe, S., Marasović, A., and Smith, N. A. (2021). "Measuring association between labels and free-text rationales," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (Punta Cana: Association for Computational Linguistics), 10266–10284.

Wood-Doughty, Z., Cachola, I., and Dredze, M. (2022). "Model distillation for faithful explanations of medical code predictions," in *Proceedings of the 21st Workshop on Biomedical Language Processing* (Dublin: Association for Computational Linguistics), 412–425.

Xia, F., and Yetisgen-Yildiz, M. (2012). "Clinical corpus annotation: challenges and strategies," in *Proceedings of the third workshop on building and evaluating resources for biomedical text mining (BioTxtM2012) in conjunction with the international conference on language resources and evaluation (LREC)* (Baton Rouge: LREC), 21–27.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., et al. (2018). "HotpotQA: a dataset for diverse, explainable multi-hop question answering," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Brussels: Association for Computational Linguistics), 2369–2380.

Yano, T., Resnik, P., and Smith, N. A. (2010). "Shedding (a thousand points of) light on biased language," in *Mturk@HLT-NAACL* (Los Angeles: Association for Computational Linguistics), 152–158.

Yao, B., Jindal, I., Popa, L., Katsis, Y., Ghosh, S., He, L., et al. (2023). "Beyond labels: empowering human annotators with natural language explanations through a novel active-learning architecture," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, eds. H. Bouamor, J. Pino, and K. Bali (Bangkok: Association for Computational Linguistics), 11629–11643.

Yessenalina, A., Choi, Y., and Cardie, C. (2010). "Automatically generating annotator rationales to improve sentiment classification," in *Proceedings of the ACL 2010 Conference Short Papers* (Uppsala: Association for Computational Linguistics), 336–341.

Yin, F., Shi, Z., Hsieh, C.-J., and Chang, K.-W. (2022). "On the sensitivity and stability of model interpretations," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Dublin: Association for Computational Linguistics), 2631–2647.

Zaidan, O., and Eisner, J. (2008). "Modeling annotators: a generative approach to learning from annotator rationales," in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (Honolulu, HI: Association for Computational Linguistics), 31–40.

Zaidan, O., Eisner, J., and Piatko, C. (2007). "Using "Anno tator Rationales" to improve machine learning for text categorization," in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference* (Rochester: Association for Computational Linguistics), 260–267.

Zaidan, O. F., Eisner, J., and Piatko, C. (2008). "Machine learning with annotator rationales to reduce annotation cost," in *Proceedings of the NIPS* 2008 Workshop on Cost Sensitive Learning* (Vancouver, BC: NeurIPS), 260–267.

Zhang, H., Zhao, X., and Song, Y. (2020). "winowhy: a deep diagnosis of essential commonsense knowledge for answering winograd schema challenge," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Online: Association for Computational Linguistics), 5736–5745.

Zhang, Y., Marshall, I., and Wallace, B. C. (2016). "Rationale-augmented convolutional neural networks for text classification," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (Austin: Association for Computational Linguistics), 795–804.

Zhong, R., Shao, S., and McKeown, K. (2019). Fine-grained sentiment analysis with faithful attention. *arXiv [Preprint]. arXiv:*1908.06870.