University of Twente

Master Thesis

Trust in automated decision making

How user's trust and perceived understanding is influenced by the quality of automatically generated explanations

Author:
Andrea Papenmeier

Supervisors:
Dr. Christin Seifert
Dr. Gwenn Englebienne

February 1, 2019

UNIVERSITY
OF TWENTE.

Abstract

Contents

1 Introduction

State of the world
The big BUT
— Xerox experiment [32]
Therefore, we did
The key findings are
The contributions of this work are
In HCI, the purpose of empirical contributions is to reveal formerly unknown insights about human behavior in relation to information or technology.

2 Background

—Catchy first sentence.
Machine learning aims to infer generally valid relationships from a finite set of training data and apply those learned relations to new data [13] [26]. While some problems can be solved by manually encoding explicit rules, others require a different approach as explicit decision-making does not deliver highly accurate results [6]. Determining a student's grade in a multiple choice test can be solved by explicitly encoding mathematical rules, yet deciding whether the tonality of a text is positive or negative needs more than a simple rule set to function accurately [32]. The datasets needed to train machine learning models are often large and represented in a high-dimensional feature space, which makes it impossible for a human to carry out the learning task like a machine can. However, machines can be used to extend the cognitive capabilities of humans when working together on those learning tasks. [46] describes the fruitful collaboration between human and machine as *augmented intelligence*, pointing at the positive aspect of machine learning support.
—Narrowing topic to decision-making and discriminative algorithms and define "decision" as output from ML systems

2.1 Interpretability in AI

Humans cooperating with machines need to understand the principles of the method that is employed - a property referred to as *transparency* [26]. *Opacity*, the direct opposite of transparency [29], is a major problem for augmented intelligence. Although opacity can be used voluntarily as a means to self-protection and censorship, it also arises involuntarily due to missing technical expertise and failed human intuition and cognitive abilities [6].
On the application-side of machine learning systems, the question of transparency brings up the notion of *interpretability*. Interpretability refers to how well a "typical classifier generated by a learning algorithm" can be understood [26], as compared to the theoretical principle of the method. That is, an interpretable machine learning system is either inherently interpretable, meaning that its operations and result patterns can be understood by a human [5] [46], or it is capable of generating descriptions understandable to humans [16]. It is also possible to equip a system retrospectively with interpretability by adding a proxy model capable of mirroring the original system's behaviour while being comprehensible for humans [18]. Using an interpretable system as a human means being enabled to make inferences about underlying data [46].
[18] assigns ten desired dimensions to interpretable machine learning systems:

- *Scope*: Global interpretability (understanding the model and operations) and local interpretability (understanding what brought about a single decision)
- *Timing*: Time scope available in the application use case for a target user to understand

- *Prior knowledge*: Level of expertise of target user
- *Dimensionality*: Size of the model and the data
- *Accuracy*: Target accuracy of the system while maintaining interpretability
- *Fidelity*: Accuracy of explanation vs. accuracy of model
- *Fairness*: Robustness against automated discrimination and ethically challenging biases in data
- *Privacy*: Protection of sensible and personal data
- *Monotonicity*: Level of monotonicity in relations of input and output (human intuition is largely monotonic)
- *Usability*: Efficiency, effectiveness, and joy of use

In the context of interpretability for machine learning systems, the terms *understandability*, *comprehensibility*, *explainability*, and *justification* are often mentioned in literature. In this paper, we adopt the definition of [40]. *Understandability*, *accuracy* of the explanation, and *efficiency* of the explanation together form *interpretability. Explainability* is a synonym of *comprehensibility* [50], which is also synonymic to *understandability* [4] and therefore an aspect of interpretability, showing the reasons for the system's behaviour [16]. Figure 1 gives an overview over these terms. Finally, *justification* refers to the evidence for why a decision is correct, which does not necessarily include the underlying reasons and causes [5].

If the human cognition is augmented by a machine learning system, talking

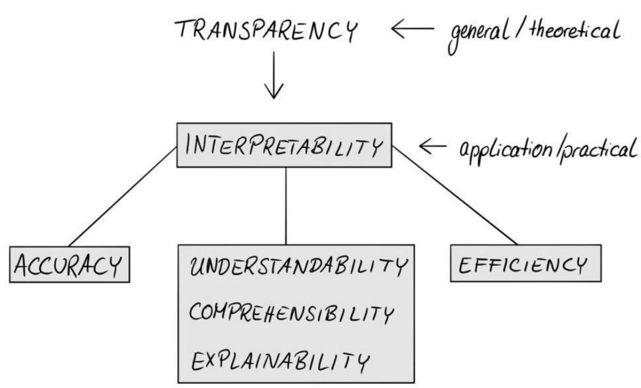

Figure 1: Relation of terms connected to interpretability

about interpretability should also include discussing the interpretability of the human in the loop. [29] argues that human behaviour is often mistakenly identified as interpretable because humans can explain their actions and beliefs. Yet the actual operations of the human brain remain opaque, which contradicts the concept of interpretability [29]. If human interpretability is taken as a point of reference for the discussion of algorithmic interpretability, [29]'s argument should be taken into account. Human interpretability, however, is not the focus of this paper and will therefore not be discussed in more detail here.

2.2 Need for Explainability in AI

A subfield of artificial intelligence research revolves solely around the explainability of intelligent systems: *xAI*, explainable artificial intelligence, for the purpose of enabling communication with agents about their reasoning [21]. xAI systems face a trade-off challenge: Their explanation has to be complete and interpretable at the same time [16]. The attention span and cognitive abilities of humans therefore become an important factor to consider in the design of a xAI systemm [27]. Furthermore, the goal of explaining the system is twofold: create actual knowledge and convince the user that the knowledge is sound and complete. Actual understanding and perceived understanding however do not always go hand in hand: Persuasive systems can convince the user without creating actual transparency [16]. The persuasiveness of an explanation is uncoupled from the actual information content of an explanation [5] and needs to be taken into account in user studies. As users can only report on their perception of the explanation, an objective measure to evaluate the fidelity of an explanation is needed. High-fidelity (also called descriptive) explanations are faithful, in that they represent truthful information about the underlying machine learning model [22]. Persuasive explanations, on the opposite, are less faithful to the underlying model, yet open up possibilities for abstraction, simplification, analogies, and other stylistic devices for communication. [22] notes a dilemma in explanation fidelity: "This freedom permits explanations better tailored to human cognitive function, making them more functionally interpretable", but "descriptive explanations best satisfy the ethical goal of transparency". The xAI practitioner therefore needs to consider a tradeoff between fidelity and interpretability.
Besides low-fidelity persuasiveness, badly designed explanations likewise "provide an understanding that is at best incomplete and at worst false reassurance" [6]. Therefore, not only possible explanations for white box (inherently interpretable) and black box (inherently non-interpretable) systems need to be examined, but also the (visual) design and communication of explanations [18]. In recent years, machine learning algorithms employed in the wild show a trend towards increasing accuracy but also increasing complexity. In general, the higher the accuracy and complexity, the lower the explainability [38] [7] in machine learning. However, users do not necessarily perceive systems with simple explanations as more understandable [1]. The authors of the user study in [1] hypothesise that users detect missing information in simple explanations, which in turn leads to the perception of incomprehensibility. [45] examined user preferences in more detail and concluded that users overall preferred more soundness

and completeness over simplicity, as well as global explanations over local explanations.
Humans involved in the explanation process are not only users, but also domain experts and engineers during the design and training phase. As explanations are user-dependent (not monolithic) [36], the design and evaluation of explanation needs to be conducted in reference to the target users. Including experts in the modelling and training process is not only a way to integrate expert knowledge that is otherwise difficult to model, but can also increase user trust [46]. [30] call the situation where a human expert works alongside the machine learning system to improve it "mixed initiative guidance".

2.2.1 Explanation Goals

Machine learning systems are able to achieve high accuracy on classification tasks, for example in information retrieval, data mining, speech recognition, and computer graphics [30]. Explainability is a means to ensure that machine learning systems are not only right in a high number of cases, but right for the right reasons [36]. High accuracy does not necessarily mean that correct generalisations were learned from the dataset or that no biases were present in the data.
The need for interpretability is dependent on the role of the explanation user and the severity of the consequences of the classification result and possible errors. Since explanations are not monolithic, i.e. have to be adapted to the target user's level of expertise, preferences for explanation types, and cognitive capabilities, the need for interpretability is also dependent on the targeted audience. Furthermore, different users can have different data access rights and have different goals to achieve in their interaction with the system [47]. While an engineer could be interested in technical details, a bank employee assessing loan credibility could be interested in similar cases and relevant characteristics of a single decision case. [38] separates a general need for interpretability into three categories:

- **no need** for interpretability if no consequences arise from faulty decisions
- interpretability is **beneficial** if consequences for individuals arise from faulty decisions
- interpretability is **critical** if serious consequences arise from faulty decisions

The three classes of interpretability needs give an overview about possible consequences, yet are too general to serve as guideline for practitioners. More details about decisive factors are needed.
For users of an automatic decision system, having insights into the system functioning and decision process increases trust [36] [12] [5] [8] [47], even in critical decisions such as medical diagnosis [1]. The level of trust should be in relation to the soundness and completeness of an explanation. Having too much or too

little trust in a system can hinder fruitful interaction between the user and the system [36] [38] [45] [37]. Other positive effects on users are satisfaction and acceptance [5] [8] [47] as well as the ability to predict the system's performance correctly [5].
[30] identifies three goals of explainability in machine learning:

- *Understanding and reassurance*: right for the right reasons
- *Diagnosis*: analysis of errors, unacceptable performance, or behaviour
- *Refinement*: improving robustness and performance

From the point of view of engineers and experts, explanations help to design, debug, and improve an automatic decision system [36]. Explanations facilitate the identification of unintuitive, systematic errors [16] [37] in the design and redundantise time-consuming trial-and-error procedures for parameter optimisation [30]. Unethical biases in training data leading to automated discrimination [12] can be identified and examined via explanations [16] [38] [37]. Ultimately, the early identification of errors avoids costly errors in high-risk domains [12] [4] [45] and ensures human safety in safety-critical tasks [16] [38].
Besides helping users and engineers, explanations also serve general goals of protection, conformity, and knowledge management. Criminals or hackers that aim to disturb the system or take advantage of it can make imperceptible changes to the input data or model at hidden levels. Having a system capable of explaining its behaviour and inner structure helps to identify unwanted alterations [16]. With the European General Data Protection Regulation (GDPR) put into place in 2018, a debate on a *right to explanation* started, which will be discussed in the following section. Although the specific implications of the right to explanation remain unclear, it should still be noted that designing interpretability follows up on that regulation [17] [16] [4]. Finally, the most general goal of implementing explanations for automatic decision systems is the opening and accessibility of a knowledge source [4] [38]. The relations derived by a machine learner (stored in the model) can deliver relevant knowledge about the data at hand.

2.2.2 Regulations and Accountability

The General Data Protection Regulation (GDPR) is a European law dealing with the processing of personal data within the European Economic Area (EEA, includes also all countries of the EU). The law holds for all companies within the EEA, companies with subsidiaries in the EEA, and any company processing personal data of a citizen of the EEA. In this context, "processing" does not only relate to automatic systems but also spans to manual processing of personal data [17]. The GDPR defines personal data as data relating to an identifiable natural person, i.e. data that can be used to identify a person [REF TO LAW TEXT]. Names, location data, or personal identification numbers are all examples of personal data that falls under the GDPR. [17] identifies two consequences of the GDPR: the legal right to non-discrimination, and a right

to explanation.
Algorithmic decisions must not be based on sensitive, personal data (GDPR article 22 paragraph 4) that are nowadays used to identify groups of people with similar characteristics, such as ethnicity, religion, gender, disability, sexuality, and more [12]. Sensitive information can, however, correlate with non-sensitive data. Real-life data almost always reflects a society's structures and biases - explicitly through sensitive information, or implicitly via dependent information. As the task of classification means separating single instances into groups based on the available data, the biases are recovered in the model [17]. A guarantee non-discrimination is therefore difficult to achieve. The GDPR does not specify whether only sensitive data or also correlated variables have to be considered when following the law. [17] identifies both interpretations as possible.
While article 13 of the GDPR specifies a right to obtain information about one's personal information and the processing of that personal information, it assures "meaningful information about the logic involved" in profiling without further defining meaningfulness. Based on the ambiguity of "meaningful", several interpretations exist, ranging from denial of the "right to explanation" [48] to a positive interpretation [41]. In summary, precedents are needed to clarify the boundaries of the law.
Besides legal regulations, ethical considerations also play a role in augmented intelligence. Accountability is the ethical value of acknowledging responsibility for decisions and actions towards another party [3]. It is an inherent factor in human-human interaction; artificial intelligence employed to interact with humans or collaborate with humans in augmented intelligence settings therefore bring about the challenge of "computational accountability" [3]. It is important to note that accountability is not a general issue in the digital world: For something to be held accountable of its own decisions or actions, it needs to act autonomously [REF WENT MISSING; CHECK AGAIN IN NOTES]. In order to determine autonomy of an algorithm and work towards accountability, [12] suggests to disclose the following information for machine learning systems:

- *Human involvement*: who controls the algorithm, who designed it etc., leading to control through social pressure
- *Data statistics*: accuracy, completeness, uncertainty, representativeness, labelling & collection process, preprocessing of data
- *Model*: input, weights, parameters, hidden information
- *Inferencing*: covariance matrix to estimate risk, prevention measures for known errors, confidence score
- *Algorithmic presence*: visibility, filtering, reach of algorithm

[3] argues that causality is a necessary prerequisite for accountability. Machine learning algorithms often learn statistical relations between input features, which at best leads to probabilistic causality, but not certainly to deterministic causality. Whether an automatic decision system itself can be held accountable for its decisions is therefore debatable.

2.2.3 Application Areas

Artificial intelligence and machine learning algorithms are nowadays employed in a variety of areas. As described in 2.2.1, the need for interpretability depends on the potential consequences of the decisions made by an automatic system. [6] summarises the application area as all systems with "socially consequential mechanisms of classification and ranking", pointing in particular to the consequences for humans. A similar view is expressed in [35] and [37], while [18] restricts the application areas in need for interpretability to those that process sensitive, i.e. personal data. In more detail, the following areas in need of interpretable intelligent systems are mentioned in literature:

- *Societal safety*: criminal justice [35] [7], terrorism detection [37]
- *Processing sensitive data*: banking, e.g. loans [35] [16] [7] [6] [13], medicine & health data [35] [18] [17] [38] [7] [46] [37], insurances [18] [6] [13], navigation [17]
- *Physical safety*: autonomous robotics [18] [38]
- *Knowledge*: education [46], knowledge discovery in research [18]
- *Economy*: manufacturing [46], individual performance monitoring [17], economic situation analysis [17], marketing [16] [6] [13]

But not only systems treating personal data or interacting directly with humans profit from interpretability - [46] suggest all machine learning based support systems as suitable candidates for interpretability. Machine learning is already employed in IT-services such as spam detection and search engines [6] [13], as well as in recommender systems [16] [38].
In the past, several machine learning systems have failed due to undetected systematic errors or automated discrimination. [18] lists incidents with machine learning systems, ranging from discrimination in the job application procedure and faulty target identification in automated weapons due to training data biases, to high differences in mortgage decisions by banks.
An interesting case is the American COMPAS system for automated crime prediction. The system predicted a significantly higher relapse rate for black convicts than for whites, which is assumed to result from human bias in the training data [18]. The argument of human bias is often used to object the perceived impartiality of computer systems, and other examples of discrimination of ethnic minorities exist [18], yet [43] counter-argues that differences found in the data set possibly reflect actual differences existing in the real world - which would shift the discussion about auto-discrimination to the field of ethics. Furthermore, the goal of profiling and classification is to separate a data set into groups [17]; discrimination is therefore "at some level inherent to profiling" [9].
In a study of 600.000 advertisements delivered by Google, [9] found a bias against women. Advertisements of higher-paid jobs were more often shown to men than they were to women. Google's targeted advertisements make use of profiling,

i.e. delivering content to users depending on their gender, age, income, location, and other characteristics. In the study, the researchers did not have access to the algorithm and can therefore not determine whether the bias was introduced with the data set, the model, or simply by conforming to the advertisement client's requirement for profiling.

Besides biased training data, systematic modelling errors can account for failures of machine learning systems. Google Flue Trends predicted the amount of humans infected with flue based on the received search queries, leading to large overestimates of actual flue cases [36]. [42] investigated the work of different research groups on the same data set, finding that the main reason for variance in results originates from the composition of the group. Compared to the group composition, the choice of classifier accounted for minor variance. They therefore concluded that the human bias in machine learning systems is the main factor influencing the results.

Deciding whether an automatic decision system meets legal and ethical standards requires knowledge about the system. In the case of Google's targeted advertisements, it is impossible to determine if the algorithm is discriminating women on purpose due to advertiser's requirements, or if the system has internal flaws that lead to unfair treatment. With the GDPR, judging the fairness of an automatic system is not only a concern of the company using machine learning techniques, but also the right of any data subject in the training set and the application.

2.3 Explanations

In the previous sections, we used "explanations" as a generic term. In this section, the concept of an explanation is described in more detail.

In general, an explanation is one or more reasons or justification for an action or belief [36]. Humans need explanations to build up knowledge about events, evaluate events, and ultimately to take control of the course of events.

When being confronted with a new event, artifact, or information in general, humans start building internal models. These mental models are not necessarily truthful nor complete, but represent an individual's interpretation about the event. Explanations are a tool to build and refine the inner knowledge model [33].

Explanations also help to assess events that are happening: We are able to compare methods or events with each other, justify the outcome of an event, and assign responsibility and guilt for past events [33] [25]. Explanations also serve to persuade someone of a belief [33], and can lead to appreciation through understanding [25].

Having understood what brings a certain event about, humans can use their knowledge model to predict the consequences of (similar) events in the future [33]. For an engineer working on a machine learning system, understanding underlying principles and consequences of the system's behaviour is a necessary step in designing a system that is "right for the right reasons" [36]. Similarly, the knowledge model can serve to prevent unwanted states or events, restore wanted states, and reproduce observed states or events [25].

2.3.1 Human-Human Explanations

Humans build mental models of the world, an inner, mental representation of events or elements. It might be noteworthy to point out the difference between the inner knowledge model and an explanation. The mental model is a subjective set of relations resulting from an individual's thought process. An explanation, however, is the interpretation of such relations [25]. Both the mental model and an explanation do not have to be truthful to the real world. We do not need to have complete, holistic mental models in order to use an artifact, but a *functional* model is needed to tell us how to use and make use of it, while a *structural* model stores information about the composition and how it is built [27].

Explanations are a cognitive and social process: The challenge of explaining includes finding a complete but compressed explanation, and transferring the explanation from the explainer to the explainee [33]. In its purest sense, "complete" means an explanation that uncovers all relevant causes [33], which is rarely the case in the real world.

[25] summarises four aspects of explanations:

- *Causal pattern content*: an explanation can reveal information about a common cause with several effects, a common effect brought about by several causes, a linear chain of events influencing each other chronologically, or causes that relate to the inner state of living things (homeostatics), e.g. intent

- *Explanatory stance*: refers to the mechanics, the design, and intention [33]. Atypical explanatory stances can lead to distorted understanding.

- *Explanatory domain*: different fields have different preferences of explanation stances

- *Social-emotional content*: can alter acceptance threshold and influence recipient's perception of explained event

What constitutes a good explanation? [25] describes good explanations as being non-circular, showing coherence, and having a high relevance for the recipient. Circularity are causal chains where an effect is given as cause to itself (with zero or more causal steps in between). Explanations can, but do not have to, explain causal relations [25]. Especially in the case of machine learning algorithms, the learned model shows correlation, not causation. Explanations for statistical models therefore cannot draw on typical causal explanations as found in human-human communication [REF NEEDED]. The probabilistic interpretation of causality comes closest to the patterns learned in statistical models: If an even A caused an event B, then the occurrence of A increases the probability of B occurring. Statistical facts are not satisfactory elements of an explanation, unless explaining the event of observing a fact [33]. Arguably, this holds true for statistical learning. Coherence refers to the systemacity of explanation elements: good explanations do not hold contradicting elements, but elements that influence each other [25]. Finally, relevance is driven by the level of detail given in the explanation. The sender has to adapt the explanation to the recipient's prior knowledge level and cognitive ability to understand the explanation [33], which can mean to generalise and to omit information - [25] calls this adaptation process the "common informational grounding". The act of explaining also includes a broader grounding of shared beliefs and meanings of events and the world [33]. The "compression problem" poses a major challenge in constructing explanations for humans. Humans tend to not comprise all possible causes and aspects of the high-dimensional real world in an explanation, suggesting that there are compression strategies (on the sender's side) and coping strategies (on the recipient's side) in place [25].

[33] notes that besides presenting likely causes, and coherence, a good explanation is simple and general. The latter two characteristics refer to the agreement widely accepted in science that a simple theory (or, in this case, an explanation) is favoured over a more complicated theory if both explain an equal set of events or states.

[27] defines a good explanation as sound, complete, but not overwhelming. While soundness refers to the level of truthfulness, completeness describes the level of disclosure [27]. In order to avoid overwhelming the explainee, the informational grounding process takes place, i.e. a common understanding of related elements and an adaptation of the explanation's detailedness to the explainee's knowledge level. In general, the more diverse the given evidence, the higher the recipient's acceptance of the explanation [25].

The explainees' cultural background is known to influence the preference for an explanation type - explaining foremost the mechanics, the design, or the intention of an event or artifact. Although different explanation types are preferred in different cultures, all explanation types can be understood by all cultures in general [25].

An experiment by [28] shows that humans have behavioural *scripts* in place when confronted with an explanation. The pure presence of an explanation, regardless of the informational content, can make a difference in how people react to requests. In the experiment, people busy with making copies at a copy machine were asked to let another person go ahead. Three conditions were examined: issuing the request of skipping line with a reasonable explanation ("because I am in a rush"), with placebic information (using the structure of an explanation without giving actual explanatory information: "because I need to make copies"), and without any explanation. The compliance rate for cases without any explanation was significantly lower than the compliance in cases where any kind of explanation (placebic or informative) was given, with little difference between the two explanation types [28]. [51] points out the advantage of such explanation - no matter the informative content -: "[t]o make a user (the audience) feel comfortable with a prediction or decision so that they keep using the system". [28] explains this behaviour with behavioural scripts that are triggered when people find themselves in a state of *mindlessness*. In a mindless state, the automatic script “comply if reason is given” is triggered, no matter what the reason is. The mindless state, however, is revoked if the consequences of complying become more severe. In an attentive state, the explanation does make a difference: People were more likely to comply when an informative explanation was given, as compared to the placebic explanation [28].

2.3.2 AI-Human Explanations

Understanding what brought about a machine learning decision can be complex. For explaining the reasons that led to a specific classification, or the classifier in general, different aspects can be highlighted.
A machine learning system generating automatic decisions contains five elements [46]:

- Dataset and subsequent features
- Optimizer or learning algorithm
- Model
- Prediction, or more generally, the result
- Evaluator

All five elements have their share in the automatic decision process and hence hold the potential for explanations. Depending on the recipient of the explanation, purely technical descriptions may not be enough to explain the system's behaviour and mechanisms. While a data scientist or system engineer might need a very complete and sound explanation, a user aiming to judge whether he or she has been treated fairly by the algorithm could be overwhelmed with such an explanation. Furthermore, it is not always possible to show all cases, parameters, and features to a lay user. A selection of information is therefore needed [37]. Explanations becomes more difficult to understand with increasing complexity of the system; Showing the underlying reasons for a single decision (local explanation) can be less complex than showing a holistic explanation of the whole model (global decision). However, global explanation can originate from a set of representative cases [37].
Several suggestions of aspects that can be explained in an automatic decision system context have been made. [5] categorises aspects of a machine learning decisions and respective explanation suggestions into three layers:

- *Feature-level*: feature meaning and influence, actual vs. expected contribution per feature
- *Sample-level*: explanation vector, linguistic explanation for textual data using bag-of-words, subtext as justification for class (trained independently), caption generation (similar to image captions)

- *Model-level*: rule extraction, prototypes & criticism samples representing model, proxy model (inherently interpretable) with comparable accuracy (author's note: supposedly meant comparable decision generation, not simple accuracy)

The categories from [5] make a distinction between the input (feature-level), a local explanation focussing on a single instance (sample-level), and a global view that comprises the whole model and its behaviour (model-level). While those aspects focus rather on the artifacts that play a role in automated decision systems, others divide the explainable elements of AI systems based on the processes and steps [4] [16] [33] [36] [38] [46]:

- *Data & features*: representation of data
- *Operations*: processing of data, computations, learning algorithm
- *Model*: parameters, representation
- *Prediction*: visualisation, e.g. heat maps
- *Secondary / add-on system*: generation of explanation via behaviour, learning algorithm behaviour

[38] stress that different explainability needs call for different timings of the explanation. Showing the explanation **before** a classification or generation task is useful for justifying the next step or explaining the plan. **During** a task, information about the operations and features can help identifying errors for correction and foster trust. Explaining the results of a task **after** the process is useful for reporting and knowledge discovery.

2.3.3 Explanation Systems

Overall, two distinct categories of machine learning systems exist in the context of explainable AI. *Inherently interpretable or transparent* systems do not need an explanation modelled on top, as they can be understood by humans without additional help. *Opaque or shallow* systems are not inherently interpretable by humans and need additional explanation, either by an add-on explanation system, or representations simplifying the actual mechanisms.
Examples of inherently interpretable machine learning models are:

- Decision trees [26] [5]
- Decision lists [5]
- Naive Bayes [26]
- Rule-Learners [26] [18]
- Compositional generative models [5]

- Linear models [18]

Although those models are not too complex, users who are not familiar with the technical implementation need an understandable representation. [18] suggest a graphical representation for decision trees and textual representation of the rules in rule-based systems. For linear models, representing the input feature's magnitude and sign can help users to understand the model [18].
Other than inherently transparent models, opaque models such as random forests, deep learning algorithms or ensemble classifiers are not inherently interpretable for humans. While complexity exceeds the cognitive abilities of humans, an increase in complexity (and therefore opacity) often comes along with a higher accuracy [38] [7]. For models that are not inherently interpretable, their explanation can at best be an approximation, but never complete [33]. All elements of the complex model can be approximated [33]. To achieve explainability of an opaque model, four concepts exist:

- *Add-on or post-hoc systems*: Retrospectively added mechanisms with the goal of generating human-readable explanations.
- *References*: similar or dissimilar cases
- *Approximations*: Simplified elements of the system
- *Inherent hyperparameter*: [38] suggests to develop a new class of learning algorithms that have an inherent "explainability hyperparameter" to achieve high accuracy in addition to high explainability. Although such algorithms do not exist yet, the concept shall be noted here.

Restrospectively added mechanisms with the goal of generating human-readable explanations. Examples of such systems exist, yet [29] points out that understandability of the explanation itself does not guarantee a sound (i.e. truthful) explanation, "however plausible they appear". In an experiment with textual explanations generated for an image classification system, [14] showed that a system with a high accuracy and an added explanatory mechanism generated meaningful descriptions of its decisions. Reducing the texts to their bare minimum, a for a human nonsensical output remained. The neural network used in their experiment, however, continued to provide high accuracy, even on the seemingly nonsensical texts. [7] developed an explanation system based on mutual information analysis. They use the Kullback-Leibler divergence to calculate the mutual information of two vectors and successfully find the influence of words within a text on the prediction. Other systems that try to model explanations alongside with a system are MYCIN, NEOMYCIN, CENTAUR, EES, LIME, and ELUCIDEBUG (see [36] and [37] for a detailed description of those systems).
In human-human explanations, people tend to question underlying principles of events by comparing it to known concepts. "Why A, why not B?" is a common question during this thought process [REF MISSING - one of the social sciences

papers]. [7] suggests showing comparable cases as reference in automatic decision systems. Cases can be compared in terms of their input features, e.g. the words composing a text, and the output, e.g. other cases classified as having the same class. To show the boundaries of a decision, similar cases with a different predicted class can be shown, or very dissimilar cases as in counterfactuals [21]. Approximating elements of an opaque system is another method of achieving interpretability for intransparent systems. Feature reduction techniques lend themselves to reduce the complexity of a system to a human-comprehensible level. [13] argues that most high-dimensional real-world application data is "concentrated on or near a lower-dimensional manifold" anyways; dimension reduction techniques like principle component analysis (PCA) or other feature selection algorithms can therefore be used to overcome the curse of dimensionality. [7] suggests salience map masks on input features to point the attention towards features that are decisive in a sample. In their experiment, they highlight words in texts to point out which ones have the highest impact on the classifier's decision. For textual input, various features are possible: generic text features (e.g. amount of words in text, n-grams) [11], syntactic features such as part-of-speech tags [11], lexicon features (e.g. presence of swear words as listed in a dictionary, polarity as listed in a sentiment lexicon), bag-of-word features which show the presence or absence of a word [2], vector-space models such as word2vec or fasttext [23] [2], or the rank on a ranked list of word frequencies in the corpus [7]. [2] compared two systems with different text representation and characteristic word selection methods. Their support vector machine with a bag-of-words representation yielded equally good results as a convolutional neural network with a vector space representation. With their research, they react on recent developments in text mining, showing a tendency towards the usage of neural nets and vector space models to represent and process textual inputs [2]. Both the work of [2] and the work of [7] described above show that generating explanations is possible at a high soundness level. Selecting relevant words in a text without having access to the complete dataset or inner workings of a classifier is possible as well. In general, the input text is altered by [REF MISSING - one of the xAI papers] remove a supposedly relevant word from all texts and observes how the classification score changes. If there is a significant decrease in accuracy, the removed word is labelled as important to the classification [REF MISSING]. [14] take the opposite approach by eliminating the supposedly irrelevant words from each text in the data set and show that the accuracy does not significantly decrease. Although the latter method did not decrease the classifiers accuracy (in this case a neural net), the remaining words were seemingly nonsensical to human observers.

For a detailed discussion of all available explanation methods, the reader is referred to [30], [16], and [46].

Does this make sense? Or should I summarise the methods again here?

2.3.4 Explanation Evaluation

Depending on the goal of the explanation in artificial intelligence, different demands are made on the explanation. In section 2.2, the concepts of persuasiveness, soundness, and completeness in explanations were introduced. A data scientist might need high soundness and completeness, while a lay user might require less completeness and more persuasiveness. In this paper, we take the stance that persuasiveness resulting from simplicity (and hence less completeness) is a useful tool to adapt the explanation's complexity to the cognitive abilities and level of expertise of a lay user. Persuasiveness should, however, not come along with untruthfulness. We therefore define a "good" explanation as one that is truthfully representing the classifier, no matter the performance of the classifier.
For evaluating how well an explanation lives up to the requirement of being a "good", hence truthful, explanation, several evaluation methods are available. [16] stresses the importance of adapting the evaluation method to the task and goal at hand. Evaluating the explanations' *functionality* can be done without actual users via a *proxy*, e.g. the model and explanation complexity or the explanations' fidelity with respect to the classifier's behaviour. Usability tests or human performance tests assess the effects of the explanations on the user's attitude towards the system. Lastly, for evaluating the system's influence in an application, a user testing in the true context with the true task can be done.
[4] summarises available tests of model interpretability into three categories:

- *Heuristics*: number of rules, number of nodes, minimum description length (model parameters); but also the general algorithm performance [38]
- *Generics*: ability to select features, ability to produce class-typical data points, ability to provide information about decision boundaries
- *Specifics*: user testing and user perception, although this is rather an evaluation of visuals than an evaluation of the actual model, e.g. by measuring accuracy of prediction, answer time, answer confidence, understanding of model; [38] add a user satisfaction score to the list

In most cases, using only one test (e.g. measuring solely the number of rules in a rule-based classifier) is not conclusive. A combination of different measures leads to more solid statements about the quality of explanations [38].

2.4 Trust in AI

One potential positive effect of explainability in AI is increasing user trust (see section 2.2.1). In human-human relationships, trust is understood as the willingness to put oneself at risk while believing that a second party will be benevolent [TRUST02]. Trust is not a characteristic inherent to an agent, but rather placed in an agent (the trustee) by another agent (the trustor). In general, the level of trust results from a trustor's overall trust in others, the propensity to trust, and the trustee's trustworthiness [31]. Trust is therefore not an objective measure,

but a subjective experience connected to a trustor [TRUST 05] [23]. Several characteristics can influence the level of trust: the trustee's dependability (i.e. repeated confirmation of benevolence in risky situations) or reliability (i.e. consistency or recurrent behaviour) [TRUST 02]. Both dependability and reliability are based on repeated experiences, trust can therefore be described as dynamic: it evolves as the relationships matures [TRUST 02]. As it is a subjective experience, there is no guarantee that the trust corresponds to the actual benevolence and trustworthiness of an agent. Inappropriate trust, e.g. trusting a person to live up to promises which he or she has no interest in keeping, can be harmful and have negative consequences [17].
In the field of computer science, no precise definition of trust in human-machine interaction exists [25]. Most papers agree that trust relates to the assurance that a system performs as expected [TRUST 05]. For classification algorithms, trust can be assigned at different scales: global trust means trusting the model itself, while local trust relates to a single decision [24]. Just as in human-human interaction, trust in a computer system can develop inappropriate dimensions. Deliberately creating an inappropriately high trust level can be misused by criminals, e.g. for data tapping [TRUST 05].

2.4.1 Trust Factors

ability, benevolence, integrity [31]
Trust factors: appeal, competence (privacy, security, functionality), transparency, duration (relationship, affiliation), reputation [23]
Trust as experience, trustworthiness is the characteristic and in case of computer programs consists of factors such as security, privacy, dependability, usability, correctness [TRUST 05] [TRUST 06]. Trust dimensions of web systems: target (the entity being evaluated), representation (encoding of trust via social warranty, certificates, etc.), method (security), management (the entity putting trust into the system), computation (evaluation metric), purpose [25]
For classification: expectation mismatch leads to direct decrease in trust [30], strength of decrease depends on the type of mismatch. Data-related mismatch weights less strongly than logic-driven mismatch. [30]
[31] argues that trust in machine learning algorithms also depends on the characteristics of misclassified cases. He points out that an automatic system can be considered trustworthy if it behaves exactly like humans, i.e. it misclassifies the same data points as a human and is correct on those cases that a human would also correctly classify [31]. Perceived understanding important for trust (rather than actual understanding):
"Findings show that the transparent version was perceived as more understandable and perceived understanding correlated with perceived competence, trust and acceptance of the system. Future research is necessary to evaluate the effects of transparency on trust in and acceptance of user-adaptive systems" [59]

2.4.2 Trust Evaluation

[23]: using experts to assign a weighted label to each element on a website or GUI and calculating a score

-1 irritant

1 chaotic

2 assuring

3 motivating

0 not present

But user study showed that experts find it problematic to assign discrete trust values. The advantage of this approach, however, is that it is possible to compare multiple websites [23].
user study with closed and open questions [24]:

- Do you trust this algorithm to work well in the real world?
- Why do you trust this algorithm to work well in the real world?
- How do you think the algorithm distinguished between the two classes?
- How certain are you of the correctness of your explanation?

[TRUST 02] develops a trust scale with 26 items, each belonging to one of the three trust factors (faith, dependability, predictability).
Koerber: same categories but with questions for automatic systems
[TRUST 01] describes online trust (websites) as developing from external factors (website's reputation, navigational architecture, user's prior experience) as well as perceived factors (credibility, ease of use, risk)
"willingness to accept a computer-generated recommendation is considered a proxy measure of trust" [38]

Most questionnaires use factual statements to investigate perceived understanding. Participants rate the statements according to their confidence of understanding [UND 03] [UND 07] or directly their subjective understanding [UND 01] [UND 02] [UND 04] [UND 05]

2.5 Summary

Summary

- summary
- systems
- evaluation of explanations and of trust

Hypotheses

3 Method

Intro

3.1 Use Case Scenario

definition of offensive language [34]
hate speech detection systems

3.2 Evaluation setup

Two evaluations:

- explanation evaluation (generics, ability to select truthfully important words from texts) to asses whether the explanations are “good” in terms of truthful
- user evaluation to assess trust and perceived understanding

4 Implementation

Intro

4.1 Dataset Selection

Few datasets with offensive language texts are publicly available. Table 1 presents an overview of four available datasets, their sizes and class balances. While the dataset of SwissText has the most fine-grained labelling of its data

Corpus	Size	Classes	
Davidson[1]	25,000	hate speech	6%
		offensive	77%
		neither	17%
Imperium[2]	3,947	neutral	73%
		insulting	27%
Analytics Vidhya[3]	31,962	hate speech	7%
		no hate speech	93%
SwissText[4]	159,570	toxic	10%
		severe_toxic	1%
		obscene	5%
		threat	0.3%
		insult	5%
		hate speech	1%
		neither	72.7%

Table 1: Publicly available datasets for offensive language texts

points, details on how the labels were assigned (i.e. number of annotators, inter-annotator agreement score, definition of the classes) are not available. The same holds for the datasets of Analytics Vidhya and Imperium.
In contrast, Davidson's datasets comes with a description of how the data points were collected, how the classes are defined, and uses at least three annotators per text. Furthermore, Davidson's dataset contains the most data points labelled as offensive: roughly 20750 Tweets fall into this category, while the Analytics Vidhya dataset contains 2240 hate speech texts, SwissText 1600, and Imperium 1000.
Throughout the literature, different definitions of hate speech and offensive language are given. For using a dataset in a user study with the scenario of a social media administrator, the definition of the label has to be clear. We therefore chose to work with the dataset of Davidson et al., as it offers the most detailed description of its labels and how the labels were obtained.

4.2 Dataset Construction

The original dataset was collected by Davidson et al. [10] for their research on defining and differentiating hate speech from offensive language. They con-

structed a dataset with offensive Tweets and hate speech by conducting a keyword search on Twitter, using keywords registered in the hatebase dictionary[5]. The timelines of Twitter users identified with the keyword search were scraped, resulting in a dataset of over 8 million Tweets. They selected 25 000 Tweets at random and had at least 3 annotators from Figure Eight[6] (formerly Crowd Flower) who labelled each Tweet as containing hate speech, offensive language, or neither. They reached an inter-annotator agreement of 0.92 [10]. The dataset is publicly available on GitHub[7].
The biggest class in the dataset are the offensive language Tweets (77%), while non-offensive Tweets represent 17%, and hate speech 6% of the dataset.
For our research, we are only interested in offensive and not offensive Tweets. We therefore excluded Tweets labelled as hate speech for the further construction of our dataset. We produced a balanced dataset by selecting only Tweets with the maximum inter-annotator agreement from each of the two remaining classes, and randomly drew Tweets from the bigger class (offensive Tweets) until the size of the subset was equal to the size of the smaller class (non-offensive Tweets). Table 2 presents statistical information about the resulting dataset.

	Not Offensive Class	**Offensive Class**
Size (absolute)	4,162	4,162
Size (relative)	50.00%	50.00%
Total words	58,288	61,504
Unique words	6,437	9,855
Average words per Tweet	14.00	14.78

Table 2: Statistical characteristics of the constructed dataset

4.3 Dataset Preprocessing

Tweets exhibit some special characteristics. First, the maximum length of a single Tweet is 140 characters. Twitter doubled the length in November 2017, yet the dataset was collected before this data and therefore contains only Tweets of 140 characters or shorter. Twitter users found creative ways to make use of the 140 characters given, leading to the usage of short URLs instead of original URLs [52], intentional reductions of words (e.g. “nite” instead of “night”) [52], abbreviations [19], emojis [15] [49] and smilies [44] [23].
Furthermore, social media content can be unstructured, with word creations that are non in standard dictionaries, like slang words [19] [49], intentional repetitions [52] [20] [34] [39] (e.g. “hhheeeey”), contractions of words [44] [20], and spelling mistakes. Although those new word formations do not appear in the dictionary, they are “intuitive and popular in social media” [24].

[5] https://www.hatebase.org
[6] https://www.figure-eight.com
[7] https://github.com/t-davidson/hate-speech-and-offensive-language

On Twitter, it is custom to mention other users within a Tweet by adding "@"+username [52] [34] [49] [39], retweeting (i.e. answering to) a Tweet [52] [20], and summarizing a Tweet's topic with "#"+topic [52] [49].
Other problems in text mining are the handling of stop words [52] [15] [19], language detection [52], punctuation [15] [20] [34], negation [49], and case folding [15] [19] [39].
Researchers have developed different strategies for preprocessing Tweets. One possible approach is to simply remove URLs, username, hashtags, emoticons, stop words, or punctuation [52] [15] [20] [34] [19] [49]. A reason to eliminate those tokens can be that they assumably do not hold information relevant to the classification goal [20]. Words that only exist for syntactic reasons (this concerns primarily stop words) can be omitted when focussing on sentiment or other semantic characteristics [15]. Mentions of other users are likewise not informative for sentiment analysis and are often removed from the texts [52] [49]. Depending on the dataset size, normalising the texts strongly by removing punctuation and emojis, as well as lowercasing the texts, can decrease the vocabulary size [15]. Especially on Twitter with its restricted text size, users tend to use shortened URLs. Short URLs have a concise, but often cryptic form, and redirect to the website with the original, long URL. While website links can encode some information on a topic, this information is lost when using a shortened URL. Removing the shortened URLs without replacement can be a step in preprocessing Tweets [52].
Rather than removing tokens, they can also be replaced by a signifier token, e.g. a complete link by "<<<hyperlink>>>" [23]. In Tweets, such signifier tokens are used for mentions of usernames [44] [23] [39], URLs [44] [23] [39], smilies [23] or negations [44]. Using signifier tokens eliminates some information, i.e. which user was mentioned or which website was linked, but retains the information that a mention or link exists. Tokens can also be grouped by using signifier tokens, i.e. tokens with similar content are summarised with a single token. [23] uses this technique to group smilies with similar sentiment and Twitter usernames related to the same company.
Case folding is often addressed by converting Tweets to lower case [15] [23] [19].

The following preprocessing steps are taken in chronological order:

1. Conversion of all texts to lower cases
2. Replacement of URLs by a dummy URL ("URL")
3. Replacement of referenced user names and handles by a dummy handle ("USERNAME")
4. This dataset encodes emojis in unicode decimal codes, e.g. "😀" for a grinning face. In order to keep the information contained in emojis, each emoji is replaced by its textual description (upper cased and without whitespaces to ensure unity for tokenizing)[8].

[8] https://www.quackit.com/character_sets/emoji/

5. Resolving contractions such as "we're" or "don't" by replacing contractions with their long version[9].

6. This dataset uses a few signifiers such as "english translation" to mark a Tweet that has been translated to English, or "rt" to mark a Retweet (i.e. a response to a previous Tweet). Since those information have been added retrospectively, we discard them here and delete the signifiers from the texts.

7. Replacement of all characters that are non-alphabetic and not a hashtag by a whitespace

8. Replacement of more than one subsequent whitespace by a single whitespace

9. Tokenization on whitespaces

After training the classifiers, the URL and username tokens are replaced by a more readable version ("http://website.com/website" and "@username", respectively) to make it easier for participants of the user study to envision themselves in the scenario of a social media administrator reading real-world Tweets. Replacing the tokens by their original URLs and usernames would give the participants more information than the classifiers had; we therefore chose to use a dummy URL and username.

Following the preprocessing steps, the following Tweet is processed from its original form:

```
"@WBUR: A smuggler explains how he helped fighters along the

"Jihadi Highway": http://t.co/UX4anxeAwd"
```

---

into a cleaned version:

```
@username a smuggler explains how he helped fighters along the

jihadi highway http://website.com/website
```

4.4 Classifier

Intro

Good System L2X

[9] https://en.wikipedia.org/wiki/Wikipedia:List_of_English_contractions

Medium System Logistic Regression with binary (1 / -1) coefficients

Bad System Inverse L2X

4.5 Explanations

Intro

Good System L2X mutual information

Medium System randomly choosing k words from the words with positive (offensive) or negative (not offensive) class

Bad System Inverse good system

4.6 Graphical User Interface

asdasdasd

4.7 Subset Sampling

For evaluating the different system-explanation conditions, users have to experience the system. However, it is not feasible to present them with the complete testeset, since it has a size of 1665 Tweets. Consequently, a subset of Tweets needs to be drawn from the testset, with a size that a human observer can understand and process within the time frame of a user study.
We furthermore aim to find 10 suitable subsets and assign participants randomly to one of the subsets, in order to reduce possible side effects from biases specific to single Tweets.
There are several requirements for the subsamples, originating from the conflict of reducing the sample for a human observer, yet still yielding a good representation of the testset and classifier:

- A class balance of the true labels similar to the testset,
- a balance of correctly to incorrectly classified data points similar to the classifier's performance on the complete testset,
- no overlap of Tweets within the set of 10 subsets,
- a feature distribution as close to the feature distribution in the complete testset.

We set the subsample size to 15 Tweets, which is enough to show accuracies to the first decimal place, yet assumably not too much to process for an observer in a user study.

To create a subset, 15 data points are randomly drawn from the testset.
First, the class balance of the subset is calculated. The difference to the class balance of the whole testset needs to be smaller than 0.1.
Additionally, for each classifier in the user study, the prediction accuracy on the subset is compared to the prediction accuracy on the complete testset. If, for all classifiers, the difference is smaller than 0.1, the next check is performed.
To ensure the uniqueness of the subsets, the randomly drawn Tweets are compared with the content of previously found subsets. The subset is only accepted if none of the contained Tweets appear in any previously found subset.
In the last step, the feature distribution of the subset is tested against the features of the complete testset using the *Kullback-Leibler Divergence* (KLD) metric. As the focus is directed towards the explanations (i.e. the highlighted words within a Tweet), only the explanations are used to examine the feature distribution. First, the feature distribution of the complete testset is calculated by constructing a word vector with tuples of words and their respective word counts. The word counts are divided by the total amount of words in the set, such that the sum of regularised counts equals 1. Next, a copy of the word vector is used to count and regularise the word frequencies in the subset. The result are two comparable vectors, yet the vector of the subset is very likely to contain zero counts for words that appear in the complete set but were never selected as explanation in the subset. Since the KLD uses the logarithm, it is undefined for zero counts. We use Laplace smoothing with k=1 to handle zero counts. For each classifier, the KLD is calculated and summed to a total divergence score for the subset.
We generate a quantity of 100 such subsets and order them by their KLD sum. The 10 subsets with the smallest score are chosen as the final set of subsets.

4.8 Explanation Evaluation

5 User Study: Trust Evaluation

Intro

5.1 Method

Intro

Participants

- amount, mean age, SD age
- recruitment method
- exclusion criteria
- compensation for participation

Apparatus A paragraph about the experiment setup (physically), system requirements and technology used. For example the pixel dimensions of screenshots.

Procedure

- tasks / survey items
- ordering of tasks

Design & Analysis One paragraph for experiment design (statistically).

One paragraph for statistical analysis.

- data points per participant and in total
- statistical test
- corrections / disqualifications

5.2 Results

Intro

asdasdasd

6 Discussion

7 Conclusion

asd

References

[1] Hiva Allahyari and Niklas Lavesson. User-oriented assessment of classification model understandability. In *11th scandinavian conference on Artificial intelligence*. IOS Press, 2011.

[2] Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. "what is relevant in a text document?": An interpretable machine learning approach. *PloS one*, 12(8):e0181142, 2017.

[3] Matteo Baldoni, Cristina Baroglio, Katherine M May, Roberto Micalizio, and Stefano Tedeschi. Computational accountability. In *CEUR Workshop Proceedings*, volume 1802, pages 56–62. CEUR Workshop Proceedings, 2016.

[4] Adrien Bibal and Benoît Frénay. Interpretability of machine learning models and representations: an introduction. In *Proceedings on ESANN*, pages 77–82, 2016.

[5] Or Biran and Courtenay Cotton. Explanation and justification in machine learning: A survey. In *IJCAI-17 Workshop on Explainable AI (XAI)*, page 8, 2017.

[6] Jenna Burrell. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1):2053951715622512, 2016.

[7] Jianbo Chen, Le Song, Martin J Wainwright, and Michael I Jordan. Learning to explain: An information-theoretic perspective on model interpretation. 80:883–892, 10–15 Jul 2018.

[8] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten Van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, 18(5):455, 2008.

[9] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, 2015(1):92–112, 2015.

[10] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515, 2017.

[11] Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. Hate me, hate me not: Hate speech detection on facebook. 2017.

[12] Nicholas Diakopoulos. Accountability in algorithmic decision making. *Communications of the ACM*, 59(2):56–62, 2016.

[13] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.

[14] Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, 2018.

[15] T. Ghorai. An information retrieval system for fire 2016 microblog track. In *Workshop Proceedings working notes of Forum for Information Retrieval Evaluation (FIRE)*, volume 1737 of *CEUR '16*, pages 81–83. CEUR-WS.org, 2016.

[16] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An approach to evaluating interpretability of machine learning. *arXiv preprint arXiv:1806.00069*, 2018.

[17] Bryce Goodman and Seth Flaxman. Eu regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38, 06 2016.

[18] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 51(5):93, 2018.

[19] Priya Gupta, Aditi Kamra, Richa Thakral, Mayank Aggarwal, Sohail Bhatti, and Vishal Jain. A proposed framework to analyze abusive tweets on the social networks. *International Journal of Modern Education and Computer Science*, 10(1):46, 2018.

[20] I Hemalatha, GP Saradhi Varma, and A Govardhan. Preprocessing the informal text for efficient sentiment analysis. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, 1(2):58–61, 2012.

[21] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Generating counterfactual explanations with natural language. In *ICML Workshop on Human Interpretability in Machine Learning*, 2018.

[22] B Herman. The promise and peril of human evaluation for model interpretability. In *NIPS 2017 Symposium on Interpretable Machine Learning*, 2017.

[23] Leonard Hövelmann and Christoph M Friedrich. Fasttext and gradient boosted trees at germeval-2017 on relevance classification and document-level polarity. *Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*, page 30, 2017.

[24] Xia Hu and Huan Liu. Text analytics in social media. In *Mining text data*, pages 385–414. Springer, 2012.

[25] Frank C Keil. Explanation and understanding. *Annu. Rev. Psychol.*, 57:227–254, 2006.

[26] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160:3–24, 2007.

[27] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. Too much, too little, or just right? ways explanations impact end users' mental models. In *Visual Languages and Human-Centric Computing (VL/HCC), 2013 IEEE Symposium on*, pages 3–10. IEEE, 2013.

[28] Ellen J Langer, Arthur Blank, and Benzion Chanowitz. The mindlessness of ostensibly thoughtful action: The role of "placebic" information in interpersonal interaction. *Journal of personality and social psychology*, 36(6):635, 1978.

[29] Zachary Lipton. The mythos of model interpretability. In *ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*. ICML, 2016.

[30] Shixia Liu, Xiting Wang, Mengchen Liu, and Jun Zhu. Towards better analysis of machine learning models: A visual analytics perspective. *Visual Informatics*, 1(1):48–56, 2017.

[31] Roger C Mayer, James H Davis, and F David Schoorman. An integrative model of organizational trust. *Academy of management review*, 20(3):709–734, 1995.

[32] Prem Melville, Wojciech Gryc, and Richard D Lawrence. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1275–1284. ACM, 2009.

[33] Tim Miller. Explanation in artificial intelligence: insights from the social sciences. *arXiv preprint arXiv:1706.07269*, 2017.

[34] Joaqun Padilla Montani. Tuwienkbs at germeval 2018: German abusive tweet detection. *Austrian Academy of Sciences, Vienna September 21, 2018*, 2018.

[35] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *NIPS 2017 Symposium on Interpretable Machine Learning*, 2017.

[36] Alun Preece. Asking 'why' in ai: Explainability of intelligent systems–perspectives and challenges. *Intelligent Systems in Accounting, Finance and Management*, 25(2):63–72, 2018.

[37] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM, 2016.

[38] Ariella Richardson and Avi Rosenfeld. A survey of interpretability and explainability in human-agent systems. In *XAI Workshop on Explainable Artificial Intelligence*, pages 137–143, 2018.

[39] Kristian Rother, Marker Allee, and Achim Rettberg. Ulmfit at germeval-2018: A deep neural language model for the classification of hate speech in german tweets. *Austrian Academy of Sciences, Vienna September 21, 2018*, 2018.

[40] S Rüping. Learning interpretable models, 2006.

[41] Andrew D Selbst and Julia Powles. Meaningful information and the right to explanation. *International Data Privacy Law*, 7(4):233–242, 2017.

[42] Martin Shepperd, David Bowes, and Tracy Hall. Researcher bias: The use of machine learning in software defect prediction. *IEEE Transactions on Software Engineering*, 40(6):603–616, 2014.

[43] Jennifer Skeem and Christopher Lowenkamp. Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology*, 54, 11 2016.

[44] Jasmina Smailović, Miha Grčar, Nada Lavrač, and Martin Žnidaršič. Predictive sentiment analysis of tweets: A stock market application. In *Human-computer interaction and knowledge discovery in complex, unstructured, Big Data*, pages 77–88. Springer, 2013.

[45] J van der Waa, J van Diggelen, K van den Bosch, and M Neerincx. Contrastive explanations for reinforcement learning in terms of expected consequences. *XAI 2018*, page 165.

[46] Elio Ventocilla, Tove Helldin, Maria Riveiro, Juhee Bae, Veselka Boeva, Göran Falkmann, and Niklas Lavesson. Towards a taxonomy for interpretable and interactive machine learning. In *XAI Workshop on Explainable Artificial Intelligence*, pages 151–157, 2018.

[47] Eric S Vorm. Assessing demand for transparency in intelligent systems using machine learning. In *2018 Innovations in Intelligent Systems and Applications (INISTA)*, pages 1–7. IEEE, 2018.

[48] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2):76–99, 2017.

[49] Hajime Watanabe, Mondher Bouazizi, and Tomoaki Ohtsuki. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access*, 6:13825–13835, 2018.

[50] Claus Weihs and UM Sondhauss. Combining mental fit and data fit for classification rule selection. In *Exploratory Data Analysis in Empirical Research*, pages 188–203. Springer, 2003.

[51] Adrian Weller. Challenges for transparency. *arXiv preprint arXiv:1708.01870*, 2017.

[52] Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1980–1984. ACM, 2012.