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 19, 2019

UNIVERSITY
OF TWENTE.

Abstract

Abstract and conclusion section will be written at the end.

Contents

1 Introduction **4**

2 Background **7**
2.1 Interpretability in AI . . . 7
2.2 Need for Explainability in AI . . . 9
2.2.1 Explanation Goals . . . 10
2.2.2 Regulations and Accountability . . . 11
2.2.3 Application Areas . . . 13
2.3 Explanations . . . 14
2.3.1 Human-Human Explanations . . . 15
2.3.2 AI-Human Explanations . . . 17
2.3.3 Explanation Systems . . . 18
2.3.4 Explanation Evaluation . . . 21
2.4 Trust in AI . . . 21
2.4.1 Trust Factors . . . 22
2.4.2 Trust Evaluation . . . 23
2.5 Summary . . . 24

3 Methodology **27**
3.1 Use Case Implications . . . 28
3.1.1 Dataset Selection . . . 29
3.1.2 Twitter Data Preprocessing . . . 30
3.1.3 Offensive Language Detection . . . 31
3.1.4 Explanations . . . 32
3.2 User Study . . . 32
3.2.1 Conditions . . . 33
3.2.2 Measures . . . 33
3.2.3 Procedure . . . 33
3.2.4 Analysis . . . 35
3.2.5 Apparatus . . . 36
3.2.6 Participants . . . 36

4 Materials **37**
4.1 Dataset . . . 37
4.2 Classifier . . . 39
4.3 Explanations . . . 41
4.4 Graphical User Interface . . . 42
4.5 Subset Sampling . . . 44
4.6 Explanation Evaluation . . . 45

5 Results **49**

6 Discussion **54**

7 Conclusion **58**

Acknowledgements

1 Introduction

Deploying machine learning algorithms in applications to support human decision-making is no exception anymore. Automated systems using non-transparent algorithms are no longer restricted solely to computationally heavy applications such as information retrieval or computer graphics [43], but can be found in human-centred areas as well. Medical diagnosis, insurance risk analysis, and self-driving cars are examples of areas with a high potential for utilising machine learning systems [27]. Other areas have already replaced human decision-making with machine learning: Recommender systems for films and music, decision systems for targeted advertisements, and credit rating assessments take decisions without human intervention in the application [24]. Machine learning systems can also be used as a source of knowledge and additional information to support a human in the decision-making process. The collaboration of machine learning systems and humans with the goal to extend the cognitive abilities of humans is called *augmented intelligence* [64].
A human collaborator or an end user - people in contact with the system need to judge how reliable and trustworthy the output is. Understanding what brought about the decisions therefore becomes a challenge for machine learning systems deployed in the real world. *Interpretability* describes how well a machine learning classifier can be understood [39]. One advantage of interpretability is the early detection and avoidance of faulty behaviour. Unexpected algorithmic behaviour can for example originate from biases in the data set, systematic errors in the classifier's design, or intentional alteration for criminal purposes [24]. Especially for high-risk domains such as terrorism detection or mining of health data, detecting anomalies is crucial [54] and ideally happens before deploying and relying on the system. Another reason for avoiding opaque decision systems is fairness, inclusion, and control over personal data [20, 24, 26]. End users should have the possibility to investigate whether they have been judged adequately by an automated system [58]. Likewise, engineers who want to prevent *automated discrimination* profit from transparency [54, 55]. In general, interpretability is not only a beneficial, but a critical characteristic for applications with a potential for serious consequences of faulty decisions [55]. Additionally, explainability fosters *trust* in the system [9, 16, 20, 51, 65] which "make[s] a user (the audience) feel comfortable with a prediction or decision so that they keep using the system" [63].
Explanations for machine learning classifiers have been discussed in literature. Whereas some models are inherently interpretable (e.g. decision trees, Naive Bayes, rule-based systems to some extent of complexity [39]), others are inherently non-interpretable (e.g. neural nets, deep learning algorithms). Additionally, the trend of machine learning algorithms is rather diverting towards more complexity than simplicity [2]. While higher complexity in general relates to higher accuracy [55], it decreases the interpretability of the systems [12]. To overcome opacity of inherently non-interpretable models, they can be explained by *add-on or post-hoc systems*. Add-on systems are machine learning systems that learn to generate human-readable explanations. Other explanation

methods approximate elements of the system on a lower complexity scale, e.g. features with a reduced set of features, or deliver reference cases to put a classification result in perspective of similar or dissimilar cases. As a threshold for minimum explainability, [26] suggests including at least an explanation showing how input features relate to a prediction.

However, with increasing opacity comes the risk of untruthfulness: [42] warns that plausible explanations are not necessarily truthful to the actual mechanisms and structures of the model in question. The more complex a model is, the more it needs to be reduced and simplified to match the attention span and cognitive abilities of humans [40]. Furthermore, system designers could build untruthful yet persuasive explanations on purpose to stimulate trust building. To come to an informed judgement about a system's integrity or fairness, correct (i.e. truthful) understanding of the system is needed. Badly designed explanations likewise lead to false reassurance [11], a problem especially for safety-critical and high-risk domains. [24] describes the challenge of generating explanations that are complete and at the same time truthful as the main challenge of the field of *explainable artificial intelligence (xAI)*.
[41] shows in an experiment of interpersonal communication that humans tend to comply with a request in an automatic way if any explanation for the request is given. The informational content of the explanation does not play a role for the compliance rate - participants were equally likely to comply with a request given an informative (truthful) explanation as they were when given a nonsense explanation without informational content. If the same behaviour of mindlessness can be observed in interaction with decision systems that offer explanations, the risk of inappropriate trust in systems is bigger than previously assumed.

We therefore investigated how different explanations (varying the level of informational content) influence the user's trust into an automatic decision system. Using the scenario of a "social media administrator" with the task to detect offensive language in Tweets, we developed three machine learning classifiers able to process textual input and classify the texts into "offensive" and "not offensive" classes at varying accuracy levels. Furthermore, we implemented and validated the automatic generation of explanations at high fidelity and low fidelity levels. We measured the trust and perceived understanding in a user study with 327 participants in order to compare different classifier-explanation combinations. Our research was driven by the following research questions:

RQ 1: What influence does the accuracy of an automatic decision system have on user's trust?

RQ 2: Do automatically generated explanations influence user trust?
- RQ 2.1: To what extent is user's trust influenced by the presence of explanations?
- RQ 2.2: How does the level of truthfulness of explanations influence user trust?

RQ 3: What role does the truthfulness of an explanation play for the user's

perceived understanding?

Still missing:

- paragraph with key findings
- short reflection of contribution to the xAI research community

This thesis covers theory and related research projects of explainable AI in chapter 2. We give an overview over the regulations supporting transparent machine learning applications and investigate explanations in the context of human-human communication as well as human-machine communication. As trust is central to augmented intelligence, a section is devoted to trust factors and trust evaluation in the field of xAI. Chapter 3 shows the methodology of the research, including a description of a use case scenario and the evaluation setup. The implementation of three machine learning classifiers, the data processing, and the generation and validation of explanations are presented in chapter 4. We then describe the setup and results of the user study in which the influence of accuracy and explanations on user trust and perceived understanding is examined. A detailed discussion of the results is given in the last section of that chapter. Finally, an overall conclusion is drawn in chapter 6.

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 [21, 39]. 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 [11]. 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 [45]. 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. [64] describes the fruitful collaboration between human and machine as *augmented intelligence*.
—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* [39]. *Opacity*, the direct opposite of transparency [42], 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 [11].
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 [39], 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 [9, 64], or it is capable of generating descriptions understandable to humans [24]. 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 [27]. Using an interpretable system as a human means being enabled to make inferences about underlying data [64].
[27] 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 [57]. *Understandability*, *accuracy* of the explanation, and *efficiency* of the explanation together form *interpretability*. *Explainability* is a synonym of *comprehensibility* [68], which is also synonymic to *understandability* [8] and therefore an aspect of interpretability, showing the reasons for the system's behaviour [24]. 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 [9].

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

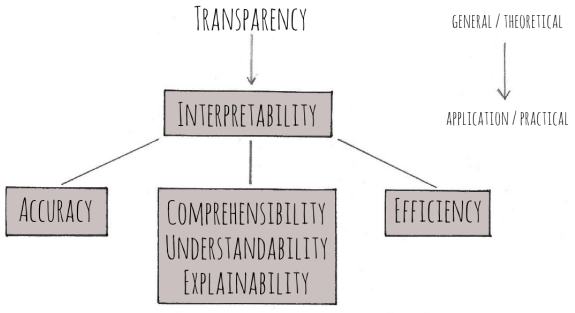

Figure 1: Relation of terms connected to interpretability

about interpretability should also include discussing the interpretability of the human in the loop. [42] 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 [42]. If human interpretability is taken as a point

of reference for the discussion of algorithmic interpretability, [42]'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 [30]. xAI systems face a trade-off challenge as their explanation has to be complete and interpretable at the same time [24]. The attention span and cognitive abilities of humans therefore become an important factor to consider in the design of a xAI system [40]. 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 [24]. The persuasiveness of an explanation is uncoupled from the actual information content of an explanation [9] 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 [31]. 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. [31] 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" [11]. 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 [27]. 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 [12, 55] 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. [62] 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) [51], 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 [64]. [43] 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 [43]. 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 [51]. 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 [65]. 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. [55] 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 [9, 16, 20, 51, 65], 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 [51, 54, 55, 62]. Other positive effects on users are satisfaction and acceptance [9, 16, 65] as well as the ability to predict the system's performance correctly [9].
[43] 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 [51]. Explanations facilitate the identification of unintuitive, systematic errors [24, 54] in the design and redundantise time-consuming trial-and-error procedures for parameter optimisation [43]. Unethical biases in training data leading to automated discrimination [20] can be identified and examined via explanations [24, 54, 55]. Ultimately, the early identification of errors avoids costly errors in high-risk domains [8, 20, 62] and ensures human safety in safety-critical tasks [24, 55].
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 [24]. 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 [26] [24] [8]. Finally, the most general goal of implementing explanations for automatic decision systems is the opening and accessibility of a knowledge source [8] [55]. 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 [26]. The GDPR defines personal data as data relating to an identifiable natural person, i.e. data that can be used to identify a person [49]. Names, location data, or personal identification numbers are all examples of personal data that falls under the GDPR. [26] 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 [20]. 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 [26]. 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. [26] 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" [66] to a positive interpretation [58]. 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 [4]. 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" [4]. 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 [4]. In order to determine autonomy of an algorithm and work towards accountability, [20] 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

[4] 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. [11] 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 [50] and [54], while [27] 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 [12, 50], terrorism detection [54]
- *Processing sensitive data*: banking, e.g. loans [11, 12, 21, 24, 50], medicine & health data [12, 26, 27, 50, 54, 55, 64], insurances [11, 21, 27], navigation [26]
- *Physical safety*: autonomous robotics [27, 55]
- *Knowledge*: education [64], knowledge discovery in research [27]
- *Economy*: manufacturing [64], individual performance monitoring [26], economic situation analysis [26], marketing [11, 21, 24]

But not only systems treating personal data or interacting directly with humans profit from interpretability - [64] 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 [11, 21], as well as in recommender systems [24, 55].
In the past, several machine learning systems have failed due to undetected systematic errors or automated discrimination. [27] 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 [27]. 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 [27], yet [60] 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 [26]; discrimination is therefore "at some level inherent to profiling" [17].
In a study of 600.000 advertisements delivered by Google, [17] 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 [51]. [59] 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 [51]. 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 [46].
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 [36, 46]. Explanations also serve to persuade someone of a belief [46], and can lead to appreciation through understanding [36].
Having understood what brings a certain event about, humans can use their knowledge model to predict the consequences of (similar) events in the future [46]. 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" [51]. Similarly, the knowledge model can serve to prevent unwanted states or events, restore

wanted states, and reproduce observed states or events [36].

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 [36]. 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 [40].
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 [46]. In its purest sense, "complete" means an explanation that uncovers all relevant causes [46], which is rarely the case in the real world.
[36] 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 [46]. 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? [36] 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 [36]. 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 [46]. 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 [36]. 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 [46], which can mean to generalise and to omit information - [36] 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 [46]. 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 [36].
[46] 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.
[40] defines a good explanation as sound, complete, but not overwhelming. While soundness refers to the level of truthfulness, completeness describes the level of disclosure [40]. 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 [36].
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 [36].
An experiment by [41] 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 [41]. [69] 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". [41] 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 [41].

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 [64]:

- 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 [54]. 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 complete model (global explanation). However, global explanation can originate from a set of representative cases [54].
Several suggestions of aspects that can be explained in an automatic decision system context have been made. [9] 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 [9] 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 [8, 24, 46, 51, 55, 64]:

- *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

[55] 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 [9, 39]
- Decision lists [9]
- Naive Bayes [39]
- Rule-Learners [27, 39]
- Compositional generative models [9]

- Linear models [27]

Although those models are not too complex, users who are not familiar with the technical implementation need an understandable representation. [27] 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 [27].
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 [12, 55]. For models that are not inherently interpretable, their explanation can at best be an approximation, but never complete [46]. All elements of the complex model can be approximated [46]. 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*: [55] 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 [42] 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, [22] 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. [12] 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 [51] and [54] 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 [46]. [12] 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 [30].
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. [21] 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. [12] 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) [19], syntactic features such as part-of-speech tags [19], 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 [2, 32], or the rank on a ranked list of word frequencies in the corpus [12]. [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 [12] 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 in a systematic way and the output (classification) observed. [2] 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 [2]. [22] 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 [43], [24], and [64].
Does this make sense? Or should I summarise the methods again here? Or leave out the last sentence?

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. [24] 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.
[8] 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 [55]
- *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; [55] 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 [55].

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 [53]. 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 [44]. Trust is therefore not an objective measure,

but a subjective experience connected to a trustor [7, 47]. 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) [53]. Both dependability and reliability are based on repeated experiences, trust can therefore be described as dynamic: it evolves as the relationships matures [53]. 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 [63].
In the field of computer science, no precise definition of trust in human-machine interaction exists [3]. Most papers agree that trust relates to the assurance that a system performs as expected [47]. 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 [54]. 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 [47].

2.4.1 Trust Factors

For human-human interaction, [44] identifies three factors contributing to the trustworthiness: ability, benevolence, and integrity. Additionally, the trustor's propensity to trust plays a role. [38] uses this model to develop a trust measure for automated systems, incorporating all four aspects. Other work on trust in computer systems mention the following factors contributing to trust [3, 6, 7, 15, 16, 38, 47]:

- *Appeal*: aesthetics, usability
- *Competence*: privacy, security, functionality, correctness
- *Duration*: relationship, affiliation
- *Transparency*: explainability, persuasiveness, perceived understanding,
- *Dependability*
- *Reputation*: warranty, certificates
- *Familiarity*

For trust in automatic classification systems, misclassifications play a special role for trust. If a user expects the system to output correct classifications (i.e. results that align with the user's prediction of the system's behaviour) but the system fails to do so, the "expectation mismatch" leads to a direct decrease in trust [25]. How strong the impact on the trust level is depends on the nature of the mismatch: data-related mismatches weight less strongly than

logic-driven mismatches [25]. [42] argues likewise that trust in machine learning algorithms 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 [42].
Besides transparency, perceived understanding is an important aspect of trust [16]. Explanations in AI aim to create understanding about the system at hand, but since trust is a subjective, it draws on the perceived understanding rather than actual understanding. [16] tested the effects of transparency on user perception, finding a correlation between perceived understanding and trust (as well as with perceived competence and acceptance). Their experiment did not provide evidence for a direct influence of (objective) transparency on trust, however. They hypothesise that actual understanding leads to more knowledge of system boundaries and unfulfilled preferences, which are not apparent in an opaque system.

2.4.2 Trust Evaluation

User trust in a computer system is, just as trust in human-human interaction, a subjective experience. Most trust measures are therefore developed for user studies rather than a list of heuristics to be checked.
For assessing website trustworthiness, [7] developed a technique that uses heuristics and experts to create a trust score per website. Experts examine each graphical element on a website or system and label each feature with a trust factor (reputation, appeal, competence, intention, relationship). Each trust factor has a pre-assigned rank that determines the weight of that factor. Subsequently, the expert judges the "state" of each feature, contributing a coefficient to the weight of the trust factor:

-1 irritant

1 chaotic

2 assuring

3 motivating

0 not present

The overall trust score for the website or system is calculated by multiplying the trust factor rank by the state coefficient and summing the resulting numbers of all features. Although this approach makes it possible to compare different websites with each other, their user study showed that experts find it problematic to assign discrete trust values [7].
[54] measure trust in the LIME add-on explanation system in their user study with open and closed questions:

- 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?

The questions are based on the aspects of competence, perceived understanding, and reliability. It needs to be noted that they report an analysis method specific to the task and dataset used in their experiment (hence not necessarily transferable to other experiments), and that their study participants had previously completed a university course in machine learning. While this questionnaire could be helpful to determine trust of data scientists, it might not be applicable for assessing trust of lay users.
[38] developed and validated a questionnaire to measure trust in automated systems. They translated the trustor and trustee factors from [44] into trust factors for automatic systems:

- Familiarity
- Intention of developers
- Propensity to trust
- Reliability
- Understanding

The questionnaire contains 19 items that are evaluated with a 5-point Likert scale. Other than trust measures for interpersonal trust, this questionnaire accounts for the fact that computer systems are developed by other humans with intentions that influence the trust relation. Additionally, the general attitude towards automated systems (familiarity with automated systems) is included in the questionnaire.
Besides measuring trust with a questionnaire, [65] reports "willingness to accept a computer-generated recommendation" as a proxy for trust.
For measuring perceived understanding, a factor of trust, most researchers use factual statements and ask participants to rate the statements according to their confidence of understanding [5, 10]. Others ask directly for their perception of their subjective understanding [35, 52, 63, 71]. Unlike actual understanding, those answers do not need to be checked and rated by an expert.

2.5 Summary

Machine learning algorithms are nowadays employed in a variety of areas, amongst others to support humans in decision-making and working with large amounts of data. Extending the human cognitive abilities with an automatic decision system in a collaborative setting is called augmented intelligence. Understanding the automatic decision systems is crucial for data scientists and engineers for assuring that the model behaves in the desired way, analysing errors and unwanted behaviour, improving the robustness and performance of the system, and finally ensuring fairness and avoiding automated discrimination. For lay users (usually

without a background in data science), understanding the system is needed to take control over one's personal and sensitive data and assess the fairness of a decision. The General Data Protection Regulation (GDPR) acknowledges the right to be informed about the processing of personal data, ultimately to ensure that the individual can contest discrimination. Understanding the mechanisms of an automated computer system also influences trust, satisfaction, and acceptance in a positive manner. For creating understanding, the system needs to deliver faithful explanations about its behaviour, structure, and data. An explanation, in general, is one or more reason or justification for an action or belief, leading to the creation of mental models.
According to the GDPR, the applicability of explainable AI (xAI) is given wherever personal data is processed with machine learning algorithms. However, applications where automated classification and ranking lead to social consequences for individuals are likewise candidates in need of explainability. The need for explainability is supported by cases in which algorithmic decision making went wrong, e.g. the COMPAS system discriminating ethnic minorities, or the Google Flue Trends system holding systematic modelling errors. Explanations are especially difficult to generate for opaque (not inherently interpretable) systems. Some form of simplification is needed to adapt the explanation to the attention span and cognitive abilities of humans. Overall, three approaches can be distinguished: Add-on systems that model an explanation solely based on the input-output behaviour of a system, referencing similar or dissimilar cases, and approximations that simplify complex facts like high-dimensional feature spaces. While approximations and references lack the ability to provide complete explanations, add-on systems provide no guarantee for fidelity, no matter how plausible their explanations seem to a human. When dealing with texts, feature selection techniques can be used to highlight important words in the text, i.e. words that had a crucial contribution to the classification.

In this research project, we aim to examine the influence of an explanation's fidelity on user trust and perceived understanding. This leads to two subproblems: measuring the fidelity of an explanation as well as measuring user trust and perceived understanding.
Since users can only report on their perception of the explanation, an objective measure to evaluate the fidelity of an explanation is needed. An explanation's fidelity describes how well the explanation represents the classifier. We therefore define a "good" explanation as one that represents the model, at the risk that it is not necessarily meaningful to a human. To generate an explanation for textual input, words that are most informative for explaining the classifier's decision can be highlighted. The fidelity can then be measured by eliminating either the informative or the irrelevant words and observe how the classification result changes.
Unlike fidelity, trust is a subjective experience describing an agent's willingness to put himself or herself at risk while believing that another agent will be benevolent. As trust is experienced by a trustor rather than an inherent characteristic of the trustee, a user study needed to assess user trust. One of the factors of trust

is perceived understanding, a likewise subjective experience only measurable in a user study. Several questionnaires are available to qualitatively and quantitatively assess trust in interpersonal relationships as well as websites and machine learning algorithms. [38] provides a quantitatively analysable questionnaire for trust in automated systems. The questionnaire addresses the trustee as "the system", which is generic enough to be applied to automatic decision systems as well. Furthermore, the questionnaire focusses only on the trustworthiness of a system, but also addresses the influence of familiarity and propensity to trust in general. Measuring trust via a questionnaire requires the participants to reflect on their relationship with the system. It could therefore be useful to use a second method that measures trust based on the (inter)actions. The proxy measure suggested in [65] could be a suitable addition to a questionnaire.

3 Methodology

The following section presents the methodological structure of the research project, driven by the research questions described in section 1.
Since trust is a subjective experience (see section 2.4), it must be evaluated in a user study. For a user study, a use case scenario helps to create a realistic environment and ensure applicability of the results for practitioners. Based on section 2, the following aspects are required for a use case scenario to answer the research questions:

1. *Realistic application scenario* that participants of a user study can relate to and feel into.

2. *Negative consequences for classification errors*: Possible areas are presented in section 2.2.3, but not all of them are feasible to be investigated in this research project. Data with sensitive content are (fortunately) not freely available, and high-risk domains like criminal justice and terrorism detection enclose their data likewise. Datasets used to gain economic advantage over competitors are often kept confidential or being censored. [20].

3. *Augmented intelligence setting* in which a human collaborates with a machine learning classifier, ensuring that the user's "willingness to accept a computer-generated recommendation" [65] is applicable as a proxy measure for trust.

To meet the requirements, we use the scenario of a company's social media channel targetting teenagers and young adults of 15-20 years old. The use case task concerns the identification of offensive texts supported by a machine learning system trained to detect offensiveness. Participants of the study take on the role of a social media moderator and administrator responsible for the content posted within the social media channel.

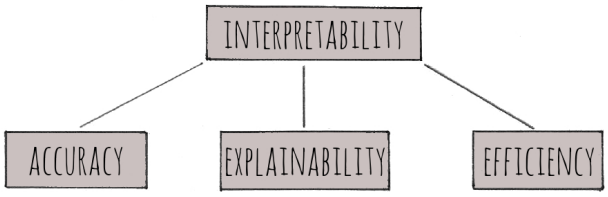

Figure 2: Model of interpretability as defined in [57]

In his model of interpretability, [57] mentions accuracy, explainability, and efficiency as the three main aspects (see figure 2). Interpretability is a positive factor for trust (see section 2.2.1), leading to the assumption that accuracy, explainability, and efficiency have the potential to influence user trust. *Efficiency* relates to the time and mental resources an average user needs to understand

the explanation. While efficiency plays a role for user trust, it is not a particular topic in question for this research project. We aim to keep the efficiency constant and minimal throughout our research and therefore adopt the minimum explanation setup suggested by [26]: showing how the input features relate to the prediction of a classifier.
Accuracy describes the classifier's ability to correctly classify data points (differing from *fidelity*, which describes the explanation's ability to correctly explain the model). To examine the influence of accuracy on user trust, the user's relation with classifiers at varying accuracy levels will be tested. We are especially interested in the extremes as well as a "realistic" accuracy level in between both extremes and therefore build three systems in total: a very good classifier (0.9 accuracy), a medium classifier (0.7 accuracy), and a bad classifier (0.1 accuracy). We explicitly exclude random classification, since we aim to have a condition in which a meaningful (i.e. truthful) explanation is generated - which is impossible for random class assignment.
Finally, *explainability*, as explained in section 2.2, concerns the communication of reasons for an event. For investigating the influence of the explanation's truthfulness, we aim to compare explanations with varying informative content (i.e. varying level of fidelity). The research design is inspired by the copy machine experiment presented in [41], that compared truthful explanations, "placebic" or dishonest explanations, and a condition without any explanation.

3.1 Use Case Implications

The use case scenario takes the user study participants to the topic of offensive language detection for the purpose of youth protection on the internet. [13] define offensive language based on the work of [34] as a composition of three categories:

- *Hateful language*: language that disparages someone on the basis of a protected trait (e.g. ethnicity, religion, nationality, gender, sexuality, disability, age [20])
- *Pornographic language*: language with explicit sexual content for "sexual arousal and erotic satisfaction" [13]
- *Vulgar language*: language with obscenity and profanity referring to "sex or bodily functions" [13]

Although offensive language and hate speech are not accurately separated, [18] argues that hate speech is always connected to the expression of hatred, while offensive language uses words that are hate speech in some contexts, but not in others. An example can be found in slang language. The word "bitch" can be amongst others defamatory, neutral, or an expression of friendship. The "urban dictionary" lists nine different typical applications[1], not all of them being pejorative. In this research project, we adopt the definition of [13] and

[1] https://www.urbandictionary.com/define.php?term=bitch

do not particularly differentiate between hate speech and offensive language to match the use case scenario. Not only hate speech, but also positively meant offensive language should not be broadcasted to a youth audience.
To train machine learning classifiers on offensive language, an annotated dataset is needed. Ideally, the dataset is labelled by human annotators, distinguishes between offensive language and non-offensive language, and contains texts from one or more social media platforms.

3.1.1 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.

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

Table 1: Publicly available datasets for offensive language texts

While the dataset of SwissText has the most fine-grained labelling of its data 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

[1] https://github.com/t-davidson/hate-speech-and-offensive-language
[2] https://www.kaggle.com/c/detecting-insults-in-social-commentary/data
[3] https://datahack.analyticsvidhya.com/contest/practice-problem-twitter-sentiment-analysis/
[4] https://www.swisstext.org/workshops/2018/Hackathon.html

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 choose 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.

3.1.2 Twitter Data 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 [70], intentional reductions of words (e.g. "nite" instead of "night") [70], abbreviations [28], emojis [23] [67] and smilies [61] [32].
Furthermore, social media content can be unstructured, with word creations that are non in standard dictionaries, like slang words [28] [67], intentional repetitions [70] [29] [48] [56] (e.g. "hhheeeey"), contractions of words [61] [29], and spelling mistakes. Although those new word formations do not appear in the dictionary, they are "intuitive and popular in social media" [33].
On Twitter, it is custom to mention other users within a Tweet by adding "@"+username [70] [48] [67] [56], retweeting (i.e. answering to) a Tweet [70] [29], and summarizing a Tweet's topic with "#"+topic [70] [67].
Other problems in text mining are the handling of stop words [70] [23] [28], language detection [70], punctuation [23] [29] [48], negation [67], and case folding [23] [28] [56].
Researchers have developed different strategies for preprocessing Tweets. One possible approach is to simply remove URLs, username, hashtags, emoticons, stop words, or punctuation [70] [23] [29] [48] [28] [67]. A reason to eliminate those tokens can be that they assumably do not hold information relevant to the classification goal [29]. Words that only exist for syntactic reasons (this concerns primarily stop words) can be omitted when focussing on sentiment or other semantic characteristics [23]. Mentions of other users are likewise not informative for sentiment analysis and are often removed from the texts [70] [67]. 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 [23]. 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 [70].
Rather than removing tokens, they can also be replaced by a signifier token, e.g. a complete link by "<<<hyperlink>>>" [32]. In Tweets, such signifier

tokens are used for mentions of usernames [61] [32] [56], URLs [61] [32] [56], smilies [32] or negations [61]. 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. [32] 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 [23] [32] [28].

3.1.3 Offensive Language Detection

Offensive language and hate speech detection systems noted in related literature use a wide range of classifiers.
The data engineers of the dataset we chose tested a variety of classifiers for distinguishing hate speech, offensive language, and non-offensive language [18]. The best results were found for a logistic regression system and a linear support vector machine (SVM), reaching a better performance than Naive Bayes, decision tree, and random forests. They eventually use logistic regression because it offers continuous output and showed promising performance in similar applications. [48] use an ensemble setting with a logistic regression classifier and a random forests system to generate features that are subsequently forwarded to a final logistic regression classifier. Recurrent neural networks (RNN), a type of neural networks, was also used for hate speech detection [19, 56]. [19] compare a RNN and an SVM and found out that the SVM performs better on detecting hate speech, while both classifiers show equal performance for classifying texts without hate speech. Using only a dictionary to filter out offensive words is not a sufficient solution for offensive language detection. Implicit offensiveness describes texts that have an offensive meaning, yet do not contain any words registered in a hate speech database. [37] focus on implicit offensiveness. They examine each word in a text individually, and then determine the class for a text by counting the number of offensive words divided by the total amount of words. For texts encoded in a vector space model (using e.g. fasttext or word2vec), [28] suggest using cosine similarity measured on a scale of 0.0-1.0 with the class boundary at 0.6. Although not focussing on offensive language in particular, but on general sentiment of texts, [12] developed a system architecture called “Learning to Explain” (L2X) that uses a convolutional neural network (CNN) to classify. In a second step, the algorithm determines the k most decisive words in the input text using mutual information analysis.

Our goal are three systems with accuracies of 0.9, 0.7, and 0.1, which hold the potential to create truthful explanations for the decision process. We therefore choose L2X for the very good classifier, and use its inverted version (the same setup but trained on inversely labelled training set) for the bad classifier. We then use a time-tested classifier with average results: logistic regression. The advantage of logistic regression for our purpose is that if offers the coefficients for all input features which helps to generate explanations.

3.1.4 Explanations

To answer the research questions **RQ 2** and **RQ 3**, the level of truthfulness, i.e. fidelity, needs to be varied. Following the setup of [41] in the copy machine experiment, we aim to generate three types of explanations: (1) a good explanation with high fidelity to the underlying model, (2) a bad explanation with placebic, i.e. nonsensical information, and (3) an "explanation" with zero explanatory content.
For generating explanations, we focus on input features and their decisive power regarding the classification result, as suggested in the minimum explanation setup by [26]. Our dataset contains Tweets, hence textual input, encoded with the bag-of-words method. The individual words in a text are therefore the input features. Showing how the words relate to the prediction can be done via highlighting ([2, 13, 22]).
The three explanation types will be generated as follows:

Good explanations All three classifiers (L2X, Inverse L2X, and logistic regression) are able to give information about which of the input features are most decisive for the classification result. The L2X algorithm is especially made to select most decisive features - in this case single words within a text -, while logistic regression offers information about the coefficients, which are equal to the influence of individual features on the classification result. For both algorithms, the decisive words in a text can therefore be determined and subsequently communicated to the user by highlighting.

Bad explanations Similar to the "placebic" explanations used in the copy machine experiment [41], the bad explanations in this research project do not provide useful information. They are, on the surface, visually similar to good explanations, but without being truthful to the underlying model. To generate nonsensical explanations, we randomly draw words from the texts. By using random selection, we assure that no human-readable pattern shows in the bad explanations.

Zero explanations The opposite of good explanations with high informative content is zero information content, thus no explanation at all. It needs to be noted that the explanation is not tied to the classification result (the label). The system can show no explanation but still show the decision.

3.2 User Study

Trust and perceived understanding are subjective experiences and hence must be evaluated with a user study. We therefore aim to answer the research questions in a user study with tasks of the use case scenario explained above. Participants of the study should take the role of a social media administrator. As trust builds up over time and is build on repeated experiences with an agent [53], participants of the user study need to have repeated interaction with the

system. We therefore show them a set of Tweets combined with the classifier's decision and explanation. The study is designed and run on the *SoSci Survey*[5] platform.

3.2.1 Conditions

Explainability and accuracy are the two variables in question for the study. Based on the copy machine experiment [41], three explanation types are generated: (1) no explanation, (2) a placebic explanation, and (3) an informative explanation. We want to test the influence of accuracy on trust with three classifiers: (1) very good with 0.9 accuracy level, (2) medium at 0.7 accuracy, and (3) bad with 0.1 accuracy. Evaluating each classifier-explanation combination leads to 9 conditions that we encode as follows:

Condition
super-good
super-rand
super-no
medium-good
medium-rand
medium-no
bad-good
bad-rand
bad-no

Table 2: List of classifier-explanation combinations evaluated in the user study

3.2.2 Measures

Perceived understanding can be evaluated with a self-assessment questionnaire. For measuring trust in the automatic decision system, we use the trust questionnaire for automtated systems by [57]. However, [41] observed the ignorance of informational content in explanations only in a "mindless", i.e. inattentive state. We therefore use a second measure of trust, the proxy measure of willingness to adopt a classifier's recommendation. The proxy can be measured by exposing the participants to the Tweets once without the system and a second time with the system and observe the changes in classification that were made. A detailed list of all questions used in the survey can be found in appendix X.

3.2.3 Procedure

On both platforms, the participants receive a link to the survey. As soon as the participant opens the survey URL, the survey starts. The survey consists of the following content:

[5] https://www.soscisurvey.de/en/index

1. Introduction & consent form
2. *Scenario 1*: Social media administrator and manual offensive language detection
3. *Tweet block 1*: 15 Tweets for classification, on individual pages (no system)
4. *Scenario 2*: Introduction to automatic decision system supporting the task
5. *Tweet block 2*: Repetition of 15 Tweets for classification, on individual pages (with system)
6. Perceived understanding & trust questionnaire
7. Demographics
8. Outroduction & crowdsourcing completion codes

In general, the study contains three blocks plus an introduction and outroduction section. The first block treats a scenario in which the participant plays the role of a "social media administrator" of a company with a young target group (15-20 years old). The task of the social media administrator is to identify content with offensive language in order to block such comments or Tweets. The next 15 pages of the survey contain one Tweet each, shown on a screenshot of a management tool, and ask the participant to classify the text as offensive or not offensive as shown in figure. The order in which the Tweets are shown is randomised for each participant. There are 10 different sets of Tweets available (without overlap), to avoid effects from the specific wording or topics in the small set of 15 Tweets. At the start of the survey, each participant is randomly assigned to one Tweet set by the system.
The second block introduces the automatic decision system. The participant is again asked to classify 15 "very similar" Tweets, which are, in fact, identical to the ones shown in the first block. This particular formulation aims to liberate the participants from the urge to classify each text with exactly the same label as in the first block. The ordering of the Tweets is random and hence very likely to be different from the ordering of the first block. In total, 9 conditions exist: three systems with each three explanation types (informative, placebic, no explanation) each. Each participant has one condition assigned at the beginning of the survey, such that there is an equal distribution of conditions in finished questionnaires.
Finally, the last block contains three questions regarding perceived understanding, 19 items measuring the user's trust including an attention check, and 5 demographic questions (gender, age, country, ethnicity, English language level).

3.2.4 Analysis

The between-subject setup described in the previous paragraph is tested in a pilot study with 11 participants. The participants are recruited via "Prolific" and receive a compensation of 2.00 GBP (2.28 EUR). They complete the study in "pretest" mode, which shows an additional comment box at the bottom of each survey page.
The main study is set up as a quantitative study without open questions or free text input. Basic frequency analysis is used for the demographic items in order to understand the background of the participants. Three topics are investigated in a statistical manner: perceived understanding (3 items), self-reported trust (19 items), and observed trust via proxy. For the first two, a 5-point Likert scale is employed.
A *Perceived understanding* score is calculated for each participant by taking the mean of the ratings for all three items in the questionnaire. The trust questionnaire used to measure *self-reported trust* contains 14 positive items and 5 inverse items. A single mean score is calculated by taking the average over the positive items and the maximum rating minus the mean of the inverse items. As a second trust measure, *observed trust* is investigated via the proxy of willingness to follow a system's recommendation. The survey contains a block of manual classification without the system, and a second round with the information provided by the automated decision system. In each block, participants classify the same set of Tweets. We can therefore determine how often a participant switches his or her classification out of 15 possible cases, and how often the change is made in agreement with the classifier's prediction but against the truth. Since the three classifiers offer different amount of opportunities to change with the classifier's prediction away from the truth (maximum 14 cases for the bad classifier as opposed to maximum 1 case for the very good classifier), the proxy measure is calculated and normalised as follows for each participant:

$$\frac{changes\_towards\_prediction\_against\_truth}{opportunities\_for\_change\_against\_truth}$$

Cases in which the very good classifier does not make any misclassification (hence no opportunity for the user to change in favour of the classifier and in contradiction to the truth) are excluded, because no valid conclusion can be drawn from these cases.
The goal of the statistical analysis for all three topics (perceived understanding, self-reported trust, observed trust via proxy) is to identify differences between different conditions. To determine whether the samples are normally distributed, we use the Shapiro-Wilk test [6] for normality from the SciPy library). We use the Mann-Whitney U test to compare two samples, since it does not assume normal distribution nor equal sample sizes or variances. For sample sizes above 20 data points, we employ SciPy's approximation[7] of the Mann-Whitney

[6] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.shapiro.html
[7] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html

U test. In case of smaller sample sizes, we use the exact implementation[8] of the Mann-Whitney U test as described in [14].

3.2.5 Apparatus

The user study is set up as an online study, the study can therefore be taken at a self-chosen location on private devices. Participants are asked to completed the survey on a notebook, desktop computer or tablet. For consistency with the use case scenario, screenshots of a fictive social media management platform show the input texts, decisions and explanations. The screenshots have a ratio of 900px (width) to 253px (height). To ensure that improper scaling of the screenshots does not influence the participants' perception, devices with small screens (e.g. smartphones and other mobile devices) are excluded. However, which device participants finally use cannot be verified. No further requirements are made regarding the equipment of the participant's device.

3.2.6 Participants

Participants are recruited via the paid science crowdsourcing platform "Prolific"[9] and "SurveyCircle"[10], an unpaid participant recruitment platform based on mutuality. All participants recruited on the paid platform "Prolific" receive a compensation of 1.40 GBP (1.60 EUR) for an estimated expenditure of time of 12 minutes. Participants from "SurveyCircle" receive a reward of 4.4 Study Points. On both platforms, individuals younger than 18 years are excluded to participate for reasons of consent by a major. The use case scenario includes reading and understanding real-life Tweets with slang words, grammatical and literal errors. The platforms therefore screen for people being fluent in English. The study questionnaire includes an attention check question, asking the participants to answer "completely disagree" in between the trust questionnaire items assessed on a 5-point Likert scale. Data from participants who fail to answer the attention check correctly is excluded from the analysis. Furthermore, only complete responses are analysed, i.e. data from participants who reach the last page of the survey.

[8] https://mail.python.org/pipermail/scipy-dev/2015-March/020475.html
[9] https://prolific.ac
[10] https://www.surveycircle.com

4 Materials

As described in section 3.1.3, we train three classifiers on a dataset of social media texts containing offensive language. Using the trained models, Tweets in a test set can be classified. The models also serve to generate explanations for the classification result. The following section documents the implementation of the classifiers and the generation of explanations. Furthermore, we show the implementation of a graphical user interface (GUI) needed to evaluate the systems in a user study. The user study imposes another constraint: Participants can only be confronted with a limited amount of data points (Tweets). This section therefore also presents the selection of subsets to be shown in the user study.

4.1 Dataset

4.1.1 Dataset Construction

The original dataset was collected by Davidson et al. [18] for their research on defining and differentiating hate speech from offensive language. They constructed a dataset with offensive Tweets and hate speech by conducting a keyword search on Twitter, using keywords registered in the hatebase dictionary[11]. 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[12] (formerly Crowd Flower) who labelled each Tweet as containing hate speech, offensive language, or neither. They reached an inter-annotator agreement of 0.92 [18]. The dataset is publicly available on GitHub[13].
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 interested in offensive and not offensive Tweets. The Tweets labelled as hate speech in this dataset have characteristics that offensive language does not necessarily have, i.e. is always pejorative, while offensive language can also be found in positive contexts (see [18]). 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 3 presents statistical information about the resulting dataset.
The dataset is split into a training set and a test set, with the test set containing 20% of the data points. The training set is used to build the classifier models, while the Tweets of the test set are used to evaluate the classifiers and to generate the data points for the user study.

[11] https://www.hatebase.org
[12] https://www.figure-eight.com
[13] https://github.com/t-davidson/hate-speech-and-offensive-language

	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 3: Statistical characteristics of the constructed dataset

4.1.2 Dataset Preprocessing

To prepare the data to be used in a machine learning application, we adopt the following preprocessing steps (see section 3.1.2) 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)[14].
5. Resolving contractions such as "we're" or "don't" by replacing contractions with their long version[15].
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

[14] https://www.quackit.com/character_sets/emoji/
[15] https://en.wikipedia.org/wiki/Wikipedia:List_of_English_contractions

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.
Applying 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
```

4.2 Classifier

Good System For the system with high accuracy (aiming 0.9 accuracy), we adopt the setup used by [12]. They achieve an accuracy of 0.9 in sentiment analysis on a movie review dataset (IMDB) with a convolutional neural network (CNN) and add their L2X algorithm to generate explanations for the CNN's decisions. The CNN consists of five layers, with an embeddings layer to transform sparse data into dense vectors, a convolution with subsequent pooling (max pooling with factor 2), a fully connected layer with 250 hidden dimensions, and an output layer with two nodes (see figure 3). Similar to [13], we use Categorical Cross-Entropy as loss function, and the Adam algorithm for optimisation. The L2X algorithm and the working example with a CNN on the IMDB dataset is publicly available on GitHub[16]. The CNN is implemented using the *Keras*[17] Python library with a TensorFlow[18] back end. With this setup, we achieve an accuracy of 0.9712 on our test set.

[16] https://github.com/Jianbo-Lab/L2X
[17] https://keras.io
[18] https://www.tensorflow.org

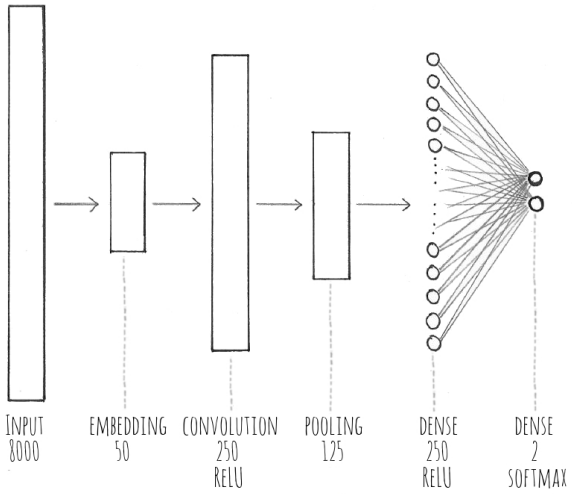

Figure 3: Architecture of the CNN with input vector and 5 layers, depicting layer goal, dimensionality, and activation function where applicable. Architecture adopted from [12].

Medium System In [18], logistic regression is used to identify offensive language and hate speech. They achieve an F1-score of 0.9 on their test set, with an L2 norm optimiser. We adopt the same setup, implementing a logistic regression classifier with the *scikit-learn*[19] Python library. We use the limited-memory BFGS algorithm for optimising and a fixed seed (random_state=0) for shuffling the data. The dataset is encoded in a vector using the bag-of-words (BOW) method. The vocabulary of the training set is used for all further computations and out-of-vocabulary words are not considered. The accuracy on the training set of the logistic regression classifier is 0.9658. However, for the medium classifier, we aim for a system with an accuracy around 0.7*. Therefore, we reduce the resolution of all coefficients c_i with $i \in \{0, I\}$, where I is the number of coefficients (equal to the length of the vocabulary):

$$c_i = \begin{cases} -1 & \text{for } c_i < 0 \\ +1 & \text{for } c_i > 0 \end{cases}$$

The logistic regression classifier determines the probability estimate p as:

$$p = \frac{1}{1 + e^{-(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 ... + \alpha_i x_i)}}$$

[19] https://scikit-learn.org

*We also implemented a dictionary-based approach with characteristic words found by their tf-idf values, and a decision tree, with both methods reaching an accuracy of more than 0.9 - hence not applicable as a medium system.

where α denotes the coefficients and x the feature value. Since we encode the texts with BOW, all x take a value of either 0 (not present in Tweet) or 1 (present in Tweet). The coefficients are binarised and either -1 (decisive for the non-offensive class) or 1 (more influence towards the offensive class). As a result, the exponent of e becomes *positive* if the number of non-offensive words predominates in a text, and *negative* if the Tweet contains more words decisive for the offensive class. With e having a positive exponent, i.e. more words pivoting towards the non-offensive class, p lies in the interval $]0.0, 0.5[$. Contrarily, p is in the interval $]0.5, 1.0[$ for negative exponents (i.e. words decisive for the offensive class predominate the text). This approach is similar to the one used in [37]: After determining the class affiliation of the individual words in a text, they assign the class that has the most words in the text. The accuracy of the binarised logistic regression classifier on the test set is 0.7628.

Bad System The bad system is equal to the good classifier, but trained on a training set with inversed labels. The accuracy on the training set (with original labels) is 0.0288.

4.3 Explanations

We chose the minimum explanation type that shows how input features relate to the classification result. In our use scenario, the input texts are Tweets with a maximum length of 140 characters. The average Tweet in the data set contains between 14 and 15 words (see table 3). We aim to select between $\frac{1}{3}$ and $\frac{1}{4}$ of the texts, which is around 4 words per Tweet. We assume that too few or too much words provide not enough or too much information to derive meaningful patterns.

Good Explanations The *medium classifier*, a logistic regression model, provides coefficients for each word in the training set vocabulary. We binarised the coefficients to reduce the classifier's accuracy to the desired level. For each Tweet, we therefore have a set of words working towards the non-offensive label, and another disjunct set of words adding to the offensive label. To generate an explanation with $k = 4$ words, we draw at random k words from the set of words working towards the predicted label. Theoretically, all words could be member of one set, and a truly truthful explanation would highlight all words. However, we choose to select only k words to reach consistency with the explanations of the good classifier. An alternative would be the usage of the non-binarised coefficients. In that case, the user would have more detailed information than the classifier, which likewise results in a reduction of truthfulness. The medium system causes another issue: The encoding of texts with the BOW method omits the information about position within a sentence. If a word is selected that appears multiple times within a text, all instances contribute equally to the classification result and hence all instances have to be highlighted. The *good classifier* using the *L2X* algorithm analyses the global classification behaviour

to generate local explanations. L2X tries to optimize the selection of features given the classification results - it learns "a distribution over the subset of features given the input vector" [12] using an approximation of mutual information. For consistency with the medium classifier, all instances of a selected words in a Tweet are highlighted.

Bad Explanations As described in section 3.1.4, we generate nonsensical explanations by randomly drawing k words from the Tweets. For consistency with the good explanations, all instances of a selected word are highlighted. The random word selection is done with the random number generator in the *NumPy* Python library.

No Explanations In the case of no information at all, we select zero words per Tweet, hence no word is highlighted at all. Nevertheless, the user still receives the information about the classification result.

4.4 Graphical User Interface

For testing the effect of explanations of an automatic decision tool on users, we aim to create an authentic environment matching the use case scenario (see introduction to section 3). The environment, in our case, is a software tool supporting the work of the "social media administrator". The tool is a means to display the Tweets and the user interface to classify those Tweets. It also serves to show the classification result of the automatic decision system and its explanations.

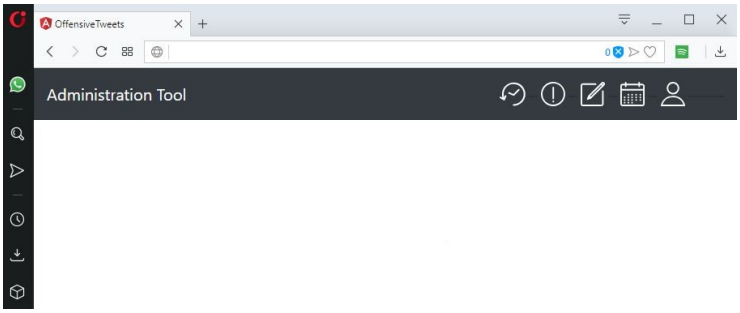

Figure 4: Screenshot of the "Administration Tool" to support the scenario of a social media administrator

The general layout of the software tool should remind the user of a modern web portal incorporating vital functions that are useful for a social media administrator (e.g. showing the social media profiles, showing the feed of the social media channels, reporting incidents, scheduling tasks, etc.). Furthermore, the

user should believe that the system is capable of providing intelligent functionality, such as an automatic decision system. We therefore use the front-end web framework *Bootstrap*[19] to generate a modern and responsive web interface and include a menu bar with items for the various functionalities (see figure 4). The interface design is minimalistic, as to not distract the user from the main task. The classifiers' decisions are visualised both with words and colours for an efficient usage. Texts classified as offensive have a red colouring, while a decision for the non-offensive class has a green colour scheme (see figure 5 and 6). The colours are the extreme colours from the *ColorBrewer*[20] "RdYlGn" scale. The explanations are shown by highlighting words from the texts. We chose to display the Tweets in black font on a white background, while selected words are coloured according to the colour scheme.

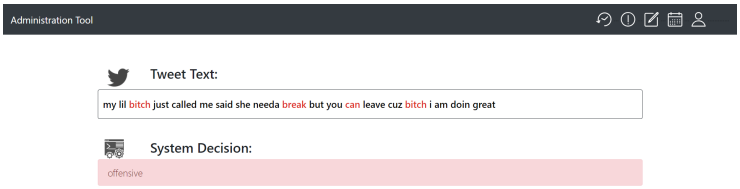

Figure 5: Screenshot of the "Administration Tool" showing an offensive Tweet with explanation for its decision

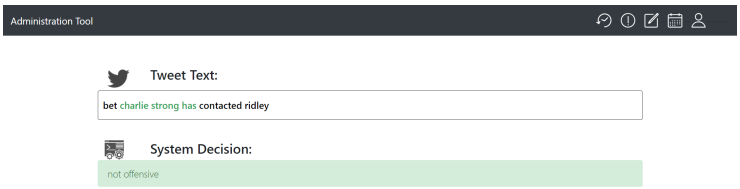

Figure 6: Screenshot of the "Administration Tool" showing a non-offensive Tweet with explanation for its decision

Figure 7: Graphics of manual classification buttons matching the user interface

Since the user study is run online and answers need to be registered within the survey tool, we convert the running system into static images by taking

[19] https://getbootstrap.com
[20] http://colorbrewer2.org

screenshots and incorporating the buttons used for user input directly in the survey tool as graphics (see figure 7). We automated the screenshot generation using *Selenium*[21], a framework for automatic web application testing. The screenshots are taken with a size of 900x253 (in pixels, w x h).

4.5 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 test set, since it has a size of 1665 Tweets. Consequently, a subset of Tweets needs to be drawn from the test set, 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 test set and classifier:

- A class balance of the true labels similar to the test set
- A balance of correctly to incorrectly classified data points similar to the classifier's performance on the complete test set
- No overlap of Tweets in between the 10 subsets and within a single subset
- A feature distribution close to the feature distribution in the complete test set

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 test set.
First, the class balance of the subset is calculated. The difference to the class balance of the whole test set 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 test set. 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 test set 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

[21] https://www.seleniumhq.org

distribution. First, the feature distribution of the complete test set 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.

4.6 Explanation Evaluation

In section 2.3.4, we defined a "good" explanation as one that is truthful to the underlying model. Although the explanations are constructed such that they theoretically represent the underlying model (see section 3.1.4), the possibility exists - especially in the case of the model-agnostic L2X algorithm - that the explanations are meaningful but not truthful.
We therefore set up a series of experiments to examine the quality of the generated explanations. In the experiments, we look at the classifiers' behaviour when being confronted with the Tweets reduced to the words selected as explanation. The first experiment deals with the classifier's accuracy when given the reduced Tweets. In the second experiment, we examine whether the classifiers can reproduce the exact labels when using the reduced Tweets. Finally, we look at the subsets selected for the user study and validate the quality of the explanations only in the subsets.

4.6.1 Evaluation 1: Accuracy on Reduced Texts

The explanations generated for each classifier are meant to show the k features (words) that are decisive for the classifier's decisions. To validate that the selected features are an actual representation of the classifier's reasoning, we investigate the generated explanations by reducing the texts of the test set to only the selected features and subsequently feeding them as new input to the respective classifier. If the selected features are indeed informative for the classifier's behaviour, the accuracy on the reduced test set should be equal to the classifier's accuracy on the original test set. The same test can be run with the placebic, i.e. nonsensical, explanations. They were generated by randomly drawing k words from each text. If the Tweets are reduced to random words, the classifiers should not reproduce the same labels as for the original test set.
Table 4 shows the performance on the test set for all three classifiers and both types of explanations. We vary k from a single word to all words in the texts.

	Good classifier 0.9712		**Medium classifier 0.7628**		**Bad classifier 0.0288**	
k	good	placebic	good	placebic	good	placebic
1	0.9676	0.5826	0.7628	0.6210	0.0330	0.4048
2	0.9646	0.6390	0.7628	0.6012	0.0390	0.3676
3	0.9688	0.6835	0.7628	0.6541	0.0348	0.3219
4	**0.9658**	**0.7345**	**0.7628**	**0.6468**	**0.0348**	**0.2667**
5	0.9682	0.7724	0.7628	0.7165	0.0402	0.2420
6	0.9664	0.8066	0.7628	0.6853	0.0348	0.2090
all	0.9712	0.9718	0.7628	0.7628	0.0288	0.0312

Table 4: Results of evaluating explanations, experiment 1

Table 4 shows that even when reducing the Tweets to a single informative word, the classifiers have a comparable accuracy as on the complete Tweets. When reducing the texts at random, the accuracy is lower (or, in case of the bad classifier, higher). Looking at $k = 4$, the accuracy of the good classifier on the randomly reduced test set is at 0.7345. This is not surprising since the average length of the texts is between 14 and 15 words. Chances are between $\frac{1}{3}$ and $\frac{1}{4}$ that the most informative feature is selected in that case, leading to a relatively high accuracy.
Although this evaluation shows that the explanations at $k = 4$ lead to similar accuracies as the original Tweets, we do not know if only the same number of classification errors was made, or if the errors were made on the same Tweets. That is, we do not know if the classifiers behave the same. We therefore need to evaluate the ability to reproduce the very same classifications (evaluation 2).

4.6.2 Evaluation 2: Ability to Reproduce Classifications

In the second experiment, we repeat experiment 1, but evaluate against the classifiers' original predictions rather than the truth label. The accuracy is therefore not computed using the labels from the data set, but using the classification result from classifying the unreduced test set. In this evaluation, a classification mistake is not a prediction that diverts from the given ground label, but one that diverts from the classifiers' original prediction. We aim to investigate how well the classifiers can reproduce their own classifications when given only the features that are - supposedly - informative for the classifiers' behaviours. If the selected words are very informative for the classifier's prediction, the very same predictions can be reproduced with the reduced dataset and the accuracy should be close to 1.0 . In the case of the nonsensical explanations, the accuracy should be lower than 1.0 .

	Good classifier		**Medium classifier**		**Bad classifier**	
k	good	placebic	good	placebic	good	placebic
1	0.9670	0.5808	1.0000	0.6402	0.9694	0.5898
2	0.9700	0.6366	1.0000	0.6961	0.9694	0.6474
3	0.9718	0.6871	1.0000	0.7417	0.9736	0.6961
4	**0.9748**	**0.7393**	**1.0000**	**0.7664**	**0.9700**	**0.7225**
5	0.9760	0.7730	1.0000	0.8126	0.9718	0.7754
6	0.9766	0.8132	1.0000	0.8252	0.9748	0.8144
all	0.9790	0.9784	1.0000	1.0000	0.9784	0.9796

Table 5: Results of evaluating explanations, experiment 2

As table 5 shows, the supposedly meaningful explanations are enough to reproduce the original prediction in almost all cases and even when reducing the texts to a single word. The nonsensical explanations, on the contrary, have an accuracy around 0.6 for a single selected word. We conclude that the generated explanations are on average a good representation for the relation between input features (words in a Tweet) and label prediction.

4.6.3 Evaluation 3: Subset Evaluation

Although we have shown that the meaningful explanations generated by the systems satisfy our definition of a "good" explanation and those generated at random do not, we looked at the average performance on the complete test set containing 1665 Tweets. The participants of the user study, however, only see a small subset of the data set: They have to judge the systems on the basis of 15 Tweets. The subsets used in the user study are constructed such that they have a feature distribution similar to that of the complete test set, and that the class balance and accuracies of the subset are not biased (see section 4.5). With this last evaluation, we want to investigate whether the assumption about the explanations' quality that we confirmed in evaluation 1 and 2 also hold for the much smaller subsets. We repeat evaluation 2 for each subset, with k at a fixed value of 4, which is the value used in the user study (see section 4.3).

Subset	**Good classifier**		**Medium classifier**		**Bad classifier**	
	good	placebic	good	placebic	good	placebic
0	0.9333	0.5333	1.0000	0.6000	0.9333	0.5333
1	1.0000	0.8000	1.0000	0.8000	1.0000	0.7333
2	1.0000	0.7333	1.0000	0.4667	1.0000	0.7333
3	1.0000	0.8667	1.0000	0.7333	0.9333	0.8667
4	0.9333	0.7333	1.0000	0.5333	0.9333	0.7333
5	0.9333	0.8667	1.0000	0.6667	0.9333	0.8667
6	1.0000	0.7333	1.0000	0.4667	1.0000	0.6667
7	0.9333	0.5333	1.0000	0.8000	1.0000	0.6667
8	1.0000	0.9333	1.0000	0.7333	1.0000	0.8667
9	0.9333	0.6667	1.0000	0.6000	0.9333	0.7333
mean	**0.9666**	**0.7400**	**1.0000**	**0.6400**	**0.9666**	**0.7400**

Table 6: Results of evaluating explanations for subsets, experiment 3

Table 6 shows that the "good" explanations of the subsets comply with our definition of a good explanation. The classifiers almost always output the same classifications for the Tweets in the subsets when being confronted with the reduced version of the texts. Reducing the Tweets to 4 words at random, however, is not enough to reliably show the same behaviour as on the unreduced Tweets. Only subset 8 shows a similar accuracy for the good classifier when comparing the meaningful and the nonsensical explanations. A reason for this result could be the selection method. Selecting 4 words out of 14 at random bears the risk of selecting useful features at some point, and leads to a very nonsensical selection at others. The standard deviation of the performance of placebic information is on average 3 to 4 times higher than the standard deviation of the meaningful explanations, confirming that the selection at random leads to a broad variety of different explanation qualities. Taking out the cases in which meaningful explanations were created "by accident" in the nonsensical condition would, however, influence the randomness of the explanations.
We conclude that in general, the explanations generated to communicate meaningful information about the systems reasoning fulfil our definition of "good" explanations for the subsets and the test set in general. The randomly reduced texts, on the other hand, are not a truthful explanation for the classifiers' behaviours, nor in the complete test set and neither in the subsets.

5 Results

The following section presents the results of the user study. We examined perceived understanding, self-reported trust and an implicit trust measure via the willingness to follow a classifier's recommendation. For each topic, we give the mean score, standard deviation, as well as a comparison of all conditions in a 9x9 matrix.
The matrices show each condition checked for significant difference with every other condition. The colour scale is a visualisation of the differences in means ($\bar{x}_{row} - \bar{x}_{column}$), with negative differences coloured in red and positive differences in green. Per cell, the net difference values are displayed as well. The significance test results are added with asterisks, where one asterisk means significant at $\alpha = 0.05$ significance level, while two asterisks denote significance at $\alpha = 0.01$ significance level.

Demographics In total, 327 participants took part in the main user study with an average age of 29.4 years (SD = 8.8), a gender balance of 56% (males) to 43% (females) and two participants reporting the third gender. 87% of the participants were recruited via "Prolific", while 36 participants enlisted on "SurveyCircle". 57% self-assessed their level of English to be equivalent to a native speaker, 23% as advanced (C1), 14% as upper-intermediate (B2), and 5% as lower than that. All participants claimed to be "fluent" in English. The exclusion criteria (passed attention check and completion of whole questionnaire) invalidated 41 data points, resulting in 286 valid cases.

Perceived Understanding As figure 8 shows, users of the system with a very good classifier and no explanation report the highest perceived understanding. For the very good and the medium classifier, giving no explanations for the decisions leads to a higher perceived understanding than delivering placebic, i.e. random, explanations. In general, users have more confidence in their understanding of the system for the very good and medium classifiers as compared to the bad classifier. One condition, however, does not lead to significantly higher scores than the bad classifier: for the medium classifier with random explanations, users reported the same understanding as for the bad classifier with no explanations. Concerning the bad classifier, giving a truthful explanation for the decision leads to the lowest perceived understanding.

Condition	Mean	SD	Condition	Mean	SD	Condition	Mean	SD
super-good	3.944	0.915	super-rand	3.729	0.860	super-no	4.147	0.701
medium-good	3.818	0.825	medium-rand	2.944	0.963	medium-no	3.700	0.690
bad-good	2.465	1.201	bad-rand	2.500	1.138	bad-no	2.944	1.152

Table 7: Mean scores for perceived understanding measure

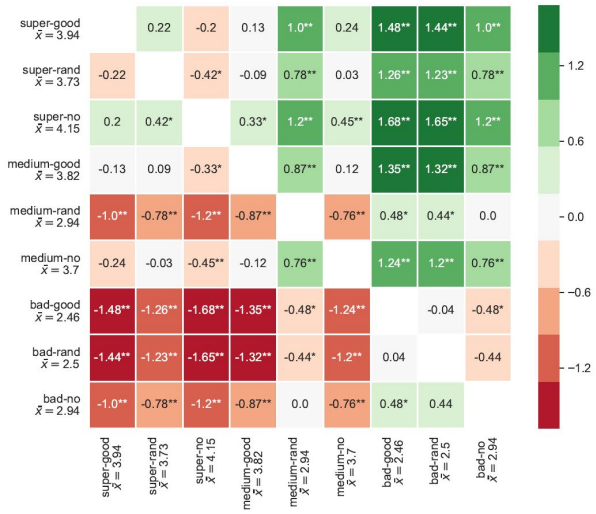

Figure 8: Significance matrices of p-values per condition ordered by classifier

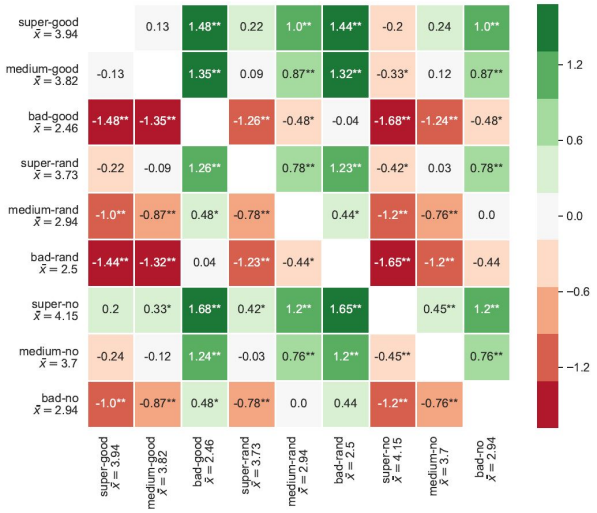

Figure 9: Significance matrices of p-values per condition ordered by explanation

Trust Questionnaire The self-reported trust scores show similar results as the perceived understanding: Besides the medium classifier with random explanations, all systems lead to significantly more trust than the systems employing the bad classifier. The explanations do not play a role regarding user's trust when the bad classifier is used. Looking at the medium classifier, the random explanation leads to a lower trust score than no explanation and a good explanation, with no difference between the latter two. The most trust is evoked by the very good classifier without explanations, significantly more than for any other condition. There is no significant difference between the very good classifier with explanations and the medium classifier with meaningful explanation. For both the bad classifier and the very good classifier, the condition without any explanation again led to the highest scores within the same classifiers. The detailed results are presented in figure 10.

Condition	Mean	SD	Condition	Mean	SD	Condition	Mean	SD
super-good	2.682	0.400	super-rand	2.679	0.482	super-no	2.995	0.512
medium-good	2.633	0.482	medium-rand	2.211	0.509	medium-no	2.630	0.459
bad-good	1.917	0.428	bad-rand	1.951	0.403	bad-no	2.018	0.546

Table 8: Mean scores for self-reported trust measure

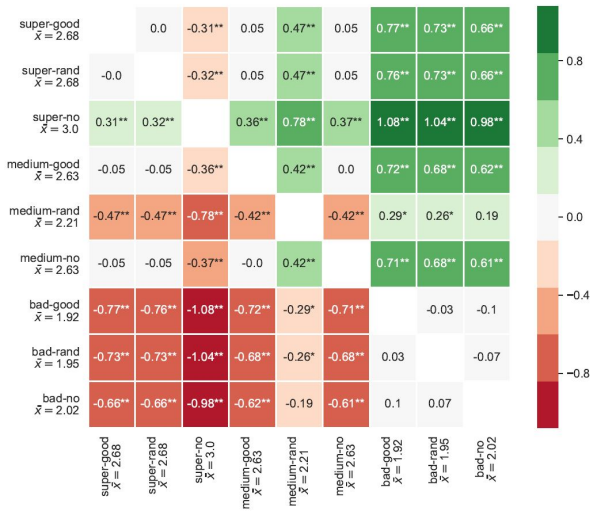

Figure 10: Trust score significance matrices of p-values per condition ordered by classifier

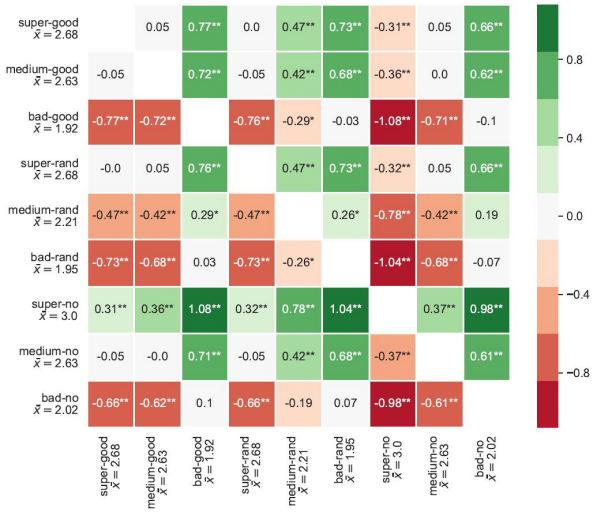

Figure 11: Trust score significance matrices of p-values per condition ordered by explanation

Observed Trust via Proxy The second trust measure uses a proxy to determine the trust a user puts into a system: the willingness to follow a system's recommendation, in this case the decision about offensiveness and non-offensiveness. Figure 12 shows the results of analysing the user's willingness to change a classification to match the system's decision while contradicting the truth. As a comparison, figure 14 deals with changes in classification that were made in favour of both the system and the truth.

The results presented in this section need to be analysed with caution. Consider the case of the very good classifier with meaningful explanations. Out of 30 cases in this condition, 16 did not have the possibility to show a change away from the truth towards the prediction of the classifier, because the classifier in those 16 cases did not make any mistakes. From the remaining 14 cases, 4 showed the behaviour in question, leading to a mean of 0.286. This result is rather high, compared to the other conditions' mean scores. The same issue appears in the data for changing from a faulty classification towards a correct classification in accordance with the bad classifier: Each participant in this condition had at maximum once the possibility to show the behaviour in question. Seen that the number of participants in each group is not large to begin with, which is then reduced by the number of cases where such behaviour is not possible, the remaining sample size is very small for solid statistical analysis.

The highest changing rate in favour of the system but against the true label was detected for users of the very good classifier with a meaningful explanation, but also the highest variance. Users were significantly more likely to adapt the system's faulty decision when confronted with the very good system with random and no explanations than the users of any system with the bad classifier. The

same holds true for users of the medium classifier without explanations.

Condition	Total cases	Cases with opportunities	Avg opportunities	Cases with changes	Avg changes	Cases with changes away	Avg changes away	Normalised
medium-good	33	33	3.48	17	1.12	7	0.30	**0.09**
medium-rand	30	30	3.57	20	1.37	7	0.30	**0.08**
bad-good	38	38	14.45	20	0.95	17	0.71	**0.05**
bad-rand	30	30	14.27	19	1.13	15	0.77	**0.05**
medium - no	30	30	3.40	20	1.00	5	0.23	**0.08**
bad - no	30	30	14.27	19	1.23	16	1.03	**0.07**
super-good	30	14	0.47	20	1.40	4	0.29	**0.29**
super-rand	32	17	0.53	18	1.09	2	0.12	**0.12**
super-no	34	23	0.68	18	1.18	1	0.04	**0.04**

Table 9: Statistics for trust measure via proxy (changes away from truth in favour of system decision)

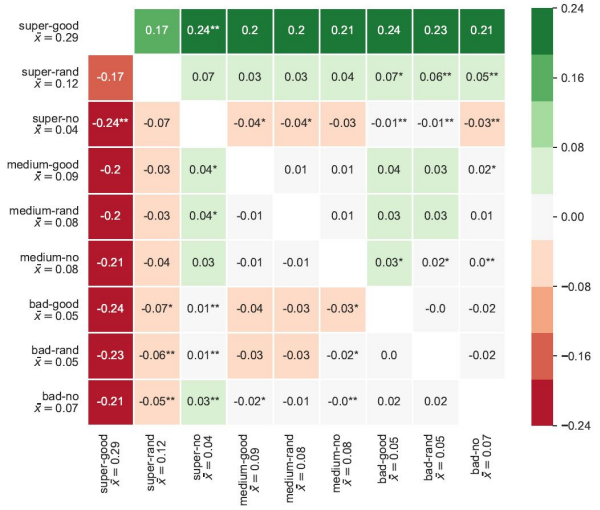

Figure 12: Proxy (away) score significance matrices of p-values per condition ordered by classifier

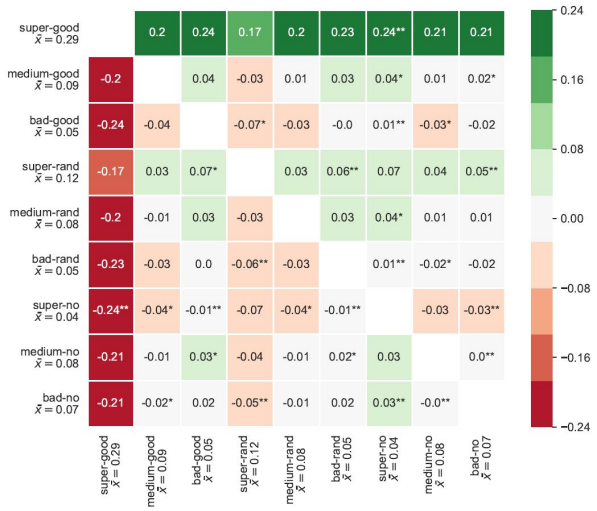

Figure 13: Proxy (away) score significance matrices of p-values per condition ordered by explanation

Looking at the changes made towards the truth in agreement with the classifiers, no significant differences are noted between any condition with the very good and medium classifier. The same holds true for the bad classifier. The very good and medium classifiers, however, evoked significantly more changes towards the

truth than the bad classifier with explanations. The standard deviations of the conditions using the bad classifier are rather high as compared to any other condition.
One condition is exceptional in this analysis: Although the bad classifier without explanation has the highest mean score (i.e. changes towards the truth when the classifier made a correct prediction), the score is not significantly different from the bad classifier with a good and random explanation. The variance of all three systems (bad-no, bad-random, bad-good) are very high as compared to the variances of the other systems. The score deviates, however, from the results of the very good and medium classifier, which have lower mean scores but lower variances. The difference in variance is important to note when comparing the relatively high mean score of the bad classifier without explanation to the conditions with the very good and medium classifiers.

Condition	**Total cases**	**Cases with opportunities**	**Avg opportunities**	**Cases with changes**	**Avg changes**	**Cases with changes towards**	**Avg changes towards**	**Normalised**
medium-good	33	33	11.52	17	1.12	12	0.67	**0.06**
medium-rand	30	30	11.43	20	1.37	12	0.57	**0.05**
bad-good	38	21	0.55	20	0.95	1	0.05	**0.05**
bad-rand	30	22	0.73	19	1.13	1	0.05	**0.05**
medium - no	30	30	11.60	20	1.00	16	0.67	**0.06**
bad - no	30	22	0.73	19	1.23	2	0.09	**0.09**
super-good	30	30	14.53	20	1.40	18	1.07	**0.07**
super-rand	32	32	14.47	18	1.09	14	0.69	**0.05**
super-no	34	34	14.32	18	1.18	15	0.91	**0.06**

Table 10: Statistics for trust measure via proxy (changes towards truth and system decision)

Figure 14: Proxy (towards) score significance matrices of p-values per condition ordered by classifier

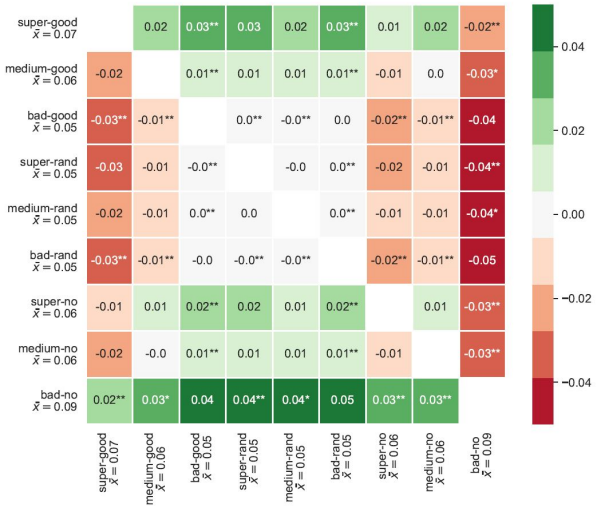

Figure 15: Proxy (towards) score significance matrices of p-values per condition ordered by explanation

6 Discussion

Our results suggest that *both perceived understanding and trust are positively related to the classifier's accuracy in general* (**RQ 1**). The higher the accuracy, the higher the trust in the system. The strongest evidence is found in the reactions to the classifiers without explanations (super-no, medium-no, bad-no). The self-reported trust was the highest for the best classifier and the lowest for the bad classifier, with all scores differing significantly from each other. The same can be observed even when adding the bad explanations: The very good classifier still has significantly higher trust and perceived understanding scores than the other classifiers, with the bad classifier again having the lowest scores. The good explanation, however, influences the trust and perceived understanding differently. Here, the bad classifier still receives significantly worse trust and understanding scores than the other two, yet there is no difference anymore in the scores for the very good and medium classifiers.

An explanation for the similarity of both the trust score and the perceived understanding score for super-good and medium-good could be the persuasiveness of a good explanation. The difference in trust and perceived understanding, that we see between the very good and medium classifier in the condition without explanation, could be compensated by convincing the user of the classifier's trustworthiness through a good explanation.

It seems intuitive to have higher trust in a system that leads to fewer deception, which has also been described in [25] with the "expectation mismatch" (see section 2.4.1). A classifier with high accuracy effectively leads to fewer disappointed expectations, which in turn does not decrease the trust. Furthermore,

the set size of 15 Tweets seems to be enough for users to develop an intuition about the classifier's accuracy. Whether users start with a high trust level and decrease the trust with every mismatch, or have a basic trust level that is increased with expectation matches and decreased with every mismatch remains to be examined in future research. The results also do not deliver information about the proportionality of accuracy and the trust level, which could be a topic of future research as well.

In the copy machine experiment by [41], only the pure presence of an explanation was enough to make people comply with a request resulting in a short waiting time. On the basis of that experiment, we designed three explanations similar to the setup in [41]: No explanation, placebic explanation, and a meaningful explanation for the classifier's behaviour. Similar to the results of the copy machine experiment, we expected to see no difference between trust scores of the meaningful and placebic explanation but a difference between the two explanations and the no explanation settings. The results, however, show a mixed answer.
For the *very good classifier*, the meaningful and placebic explanation indeed led to the same trust score. Other than expected, the no explanation condition showed the best results. The classifier performed at an accuracy of 0.95, which resulted in 44% of the cases in a perfect classification rate within the small subset of 15 Tweets. A possible explanation for the good trust score in the no explanation condition could be the conservation of a perfect image throughout the 15 Tweets. The classifier makes (almost) no mistakes and does not offer any information that could lead to doubts about the classifier's abilities. Both displayed explanation types would then have a disadvantage over the no explanation condition: The good explanations are not necessarily meaningful to a human, as they are based on statistical information rather than semantics or intentions. The placebic explanation is generated at random, which likewise holds potential for doubts and incomprehension. A similar guess was ventured in [16], who suspected that more knowledge about system boundaries and unfulfilled preferences leads to a decrease in trust (see section 2.4.1). The opposite can be observed in the proxy measure for trust, i.e. the changes in labelling that a user made towards the classifier's decision but away from the truth. Here, the no explanation condition led to significantly fewer changes away from the truth as compared to the two conditions giving any type of explanation, which is in turn consistent with the results in the copy machine experiment. It is important to note that the copy machine experiment only worked while people are in a "mindless" state, i.e. an inattentive state of mind. It is possible that users did not in particular pay attention to their trust towards the system during the classification task, but actively reflected on their relationship with the system during the self-report of trust. Being in a mindless state during the proxy measurement while being mindful during the trust questionnaire would explain the conflicting results of both measures.
The results for the *medium classifier* differ from those of the very good classifier. The two conditions with explanations have significantly different trust scores.

The placebic explanation has the lowest score, while the meaningful explanation is ranked at the same trust level as the system without explanation. The classifier delivers faulty classifications in three to four cases out of 15, which presumably raises doubts about the system. The negative effect of placebic explanations could therefore be worsened. The "expectation mismatch" is then twofold, with the wrong classification on the one side and the useless explanation on the other side. With the good explanations, the users receive some information about the underlying reasons for a misclassification. Even if not all information of the explanation is meaningful to a human, it delivers hints to the system's function and malfunction, possibly raising overall trust.
The *bad classifier* did not show evidence of diverting trust scores for any of the three explanation types. The same homogeneity is found in the results of the proxy measurement of trust, for both the changes away from the truth and towards. The trust scores were significantly lower than any other condition, with one exemption. The bad classifier without explanation had comparable trust ratings as the medium classifier with random explanations. Since there is a significant distance between the scores of the medium classifier with random explanations and other conditions of the medium classifier, we conclude that it is due to a property of the medium-random system rather than a phenomenon of the bad classifier. The evidence suggests that users are not fooled by a bad classifier and do not trust it, no matter the explanation given. A plausible assumption could have been that users trust a bad classifier as it is predictable, but do not use it as a basis for their decisions because it is not accurate. This distinction, however, is not found in the results.
Overall, we found evidence that the accuracy of a classifier is more important for trust than the explanation. The explanations did not make a difference for the very good classifier nor the very bad classifier (**RQ 2**). The case of the medium classifier is an interesting one, as we found an influence of the explanations on user trust here. It would be interesting to investigate the relationship of users with the medium classifier in more detail in future research. The findings also show that an evaluation of explanations in xAI should not only be made for extreme cases, but also consider the - supposedly more realistic - cases on the whole spectrum between the extremes.

One of the factors contributing to trust is *perceived understanding*. Our findings show a negative influence of meaningless information on perceived understanding (**RQ 3**). For both the medium classifier and the very good classifier, perceived understanding was the worst when delivering placebic explanations. Although actual understanding is arguably different when comparing a case without any explanation and one with good, meaningful information, the perception of knowledge about both cases is equal here. A mechanism similar to "expectation mismatch" could be in place for perception of knowledge. While building the mental model of the classifiers, no conflicting information have to be consolidated for the good explanation and the no explanation cases. Being confronted with random and therefore meaningless explanations forces the user to unite conflicting information in the mental model. The more conflicts appear,

the lower the confidence in the mental model.
For the bad classifier, perceived understanding ratings are significantly lower as for other classifiers (except for medium-random, which has a low rating as well). However, the system delivering good explanations for the faulty behaviour receives a significantly higher score than no explanation and placebic explanation cases. The positive effect of high accuracy does not hold here because the classifier performs badly on the task. Yet, as the explanations give more information about the inner workings of the classifier, it seems intuitive to evoke more confidence of understanding in this case.

In this research, we used a computational evaluation to validate the fidelity of the automatically generated explanations. Although the "meaningful" explanations demonstrably represent the features that are decisive for the classification, they are not necessarily meaningful to a human observer. We showed in section **??** that the selected words in the texts were enough to reconstruct the behaviour of the classifiers and can therefore serve as a basis for decision. Whether the selection of words is enough for humans to judge is up to discussion. Further research is necessary to determine the actual "meaningfulness" of the generated explanations for humans.
The proxy measurement of trust via changes in classification between the first and second block of Tweets showed ambiguous results with high variance. For future projects dealing with trust in computer systems, it could be useful to measure trust not only via a questionnaire that requires reflection abilities and active processing of the relationship between user and system. Using a trust measure that can be determined without the participant knowing could serve as an additional view on the practical implications of trust.

7 Conclusion

Abstract and conclusion section will be written at the end.

References

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

[2] L. Arras, F. Horn, G. Montavon, K.-R. Müller, and W. Samek. "what is relevant in a text document?": An interpretable machine learning approach. *PloS one*, 12(8):e0181142, 2017.

[3] D. Artz and Y. Gil. A survey of trust in computer science and the semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(2):58–71, 2007.

[4] M. Baldoni, C. Baroglio, K. M. May, R. Micalizio, and S. Tedeschi. Computational accountability. In *CEUR Workshop Proceedings*, volume 1802, pages 56–62. CEUR Workshop Proceedings, 2016.

[5] L. E. Ball and M. D. Leveritt. Development of a validated questionnaire to measure the self-perceived competence of primary health professionals in providing nutrition care to patients with chronic disease. *Family practice*, 32(6):706–710, 2015.

[6] S. Becker, W. Hasselbring, A. Paul, M. Boskovic, H. Koziolek, J. Ploski, A. Dhama, H. Lipskoch, M. Rohr, D. Winteler, et al. Trustworthy software systems: a discussion of basic concepts and terminology. *ACM SIGSOFT Software Engineering Notes*, 31(6):1–18, 2006.

[7] P. Bedi and H. Banati. Assessing user trust to improve web usability. *Journal of computer Science*, 2(3):283–7, 2006.

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

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

[10] E. Broadbent, K. J. Petrie, J. Main, and J. Weinman. The brief illness perception questionnaire. *Journal of psychosomatic research*, 60(6):631–637, 2006.

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

[12] J. Chen, L. Song, M. J. Wainwright, and M. I. Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 883–892. PMLR, 10–15 Jul 2018.

[13] Y. Chen, Y. Zhou, S. Zhu, and H. Xu. Detecting offensive language in social media to protect adolescent online safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 71–80. IEEE, 2012.

[14] Y. K. Cheung and J. H. Klotz. The mann whitney wilcoxon distribution using linked lists. *Statistica Sinica*, pages 805–813, 1997.

[15] C. L. Corritore, R. P. Marble, S. Wiedenbeck, B. Kracher, and A. Chandran. Measuring online trust of websites: Credibility, perceived ease of use, and risk. *AMCIS 2005 Proceedings*, page 370, 2005.

[16] H. Cramer, V. Evers, S. Ramlal, M. Van Someren, L. Rutledge, N. Stash, L. Aroyo, and B. 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.

[17] A. Datta, M. C. Tschantz, and A. Datta. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, 2015(1):92–112, 2015.

[18] T. Davidson, D. Warmsley, M. Macy, and I. 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.

[19] F. Del Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Tesconi. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the 1st Italian Conference on Cybersecurity*, ITASEC17, 2017.

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

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

[22] S. Feng, E. Wallace, A. Grissom II, M. Iyyer, P. Rodriguez, and J. 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.

[23] 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.

[24] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. Explaining explanations: An approach to evaluating interpretability of machine learning. *arXiv preprint arXiv:1806.00069*, 2018.

[25] A. Glass, D. L. McGuinness, and M. Wolverton. Toward establishing trust in adaptive agents. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pages 227–236. ACM, 2008.

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

[27] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 51(5):93, 2018.

[28] P. Gupta, A. Kamra, R. Thakral, M. Aggarwal, S. Bhatti, and V. Jain. A proposed framework to analyze abusive tweets on the social networks. *International Journal of Modern Education and Computer Science*, 10(1):46, 2018.

[29] I. Hemalatha, G. S. 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.

[30] L. A. Hendricks, R. Hu, T. Darrell, and Z. Akata. Generating counterfactual explanations with natural language. In *ICML Workshop on Human Interpretability in Machine Learning*, 2018.

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

[32] L. Hövelmann and C. 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.

[33] X. Hu and H. Liu. Text analytics in social media. In *Mining text data*, pages 385–414. Springer, 2012.

[34] T. Jay and K. Janschewitz. The pragmatics of swearing. *Journal of Politeness Research. Language, Behaviour, Culture*, 4(2):267–288, 2008.

[35] S. Joffe, E. F. Cook, P. D. Cleary, J. W. Clark, and J. C. Weeks. Quality of informed consent: a new measure of understanding among research subjects. *Journal of the National Cancer Institute*, 93(2):139–147, 2001.

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

[37] M. Klenner. Offensive language without offensive words (olwow). *Austrian Academy of Sciences, Vienna September 21, 2018*, 2018.

[38] M. Körber. Theoretical considerations and development of a questionnaire to measure trust in automation. In *Congress of the International Ergonomics Association*, pages 13–30. Springer, 2018.

[39] S. 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.

[40] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W.-K. 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.

[41] E. J. Langer, A. Blank, and B. 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.

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

[43] S. Liu, X. Wang, M. Liu, and J. Zhu. Towards better analysis of machine learning models: A visual analytics perspective. *Visual Informatics*, 1(1):48–56, 2017.

[44] R. C. Mayer, J. H. Davis, and F. D. Schoorman. An integrative model of organizational trust. *Academy of management review*, 20(3):709–734, 1995.

[45] P. Melville, W. Gryc, and R. 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.

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

[47] N. G. Mohammadi, S. Paulus, M. Bishr, A. Metzger, H. Könnecke, S. Hartenstein, T. Weyer, and K. Pohl. Trustworthiness attributes and metrics for engineering trusted internet-based software systems. In *International Conference on Cloud Computing and Services Science*, pages 19–35. Springer, 2013.

[48] J. P. Montani. Tuwienkbs at germeval 2018: German abusive tweet detection. *Austrian Academy of Sciences, Vienna September 21, 2018*, 2018.

[49] C. of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press, 2001.

[50] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Vaughan, and H. Wallach. Manipulating and measuring model interpretability. In *NIPS 2017 Symposium on Interpretable Machine Learning*, 2017.

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

[52] E. Racine, C. Hurley, A. Cheung, C. Sinnott, K. Matvienko-Sikar, C. Baumgartner, N. Rodondi, W. H. Smithson, and P. M. Kearney. Participants' perspectives and preferences on clinical trial result dissemination: The trust thyroid trial experience. *HRB Open Research*, 1, 2018.

[53] J. K. Rempel, J. G. Holmes, and M. P. Zanna. Trust in close relationships. *Journal of personality and social psychology*, 49(1):95, 1985.

[54] M. T. Ribeiro, S. Singh, and C. 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.

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

[56] K. Rother, M. Allee, and A. 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.

[57] S. Rüping. *Learning interpretable models*. Doctoral thesis, Technical University Dortmund, 2006.

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

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

[60] J. Skeem and C. Lowenkamp. Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology*, 54, 11 2016.

[61] J. Smailović, M. Grčar, N. Lavrač, and M. 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.

[62] 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, 2018.

[63] C. L. Van Ess. Perceived knowledge of heart failure and adherence to self-care recommendations. Master thesis, Grand Valley State University, 2001.

[64] E. Ventocilla, T. Helldin, M. Riveiro, J. Bae, V. Boeva, G. Falkmann, and N. Lavesson. Towards a taxonomy for interpretable and interactive machine learning. In *XAI Workshop on Explainable Artificial Intelligence*, pages 151–157, 2018.

[65] E. 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.

[66] S. Wachter, B. Mittelstadt, and L. 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.

[67] H. Watanabe, M. Bouazizi, and T. 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.

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

[69] A. Weller. Challenges for transparency. In *ICML Workshop on Human Interpretability in Machine Learning (WHI 2017)*. ICML, 2017.

[70] G. Xiang, B. Fan, L. Wang, J. Hong, and C. 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.

[71] M. Zamalia and A. L. Porter. Students' perceived understanding and competency in probability concepts in an e-learning environment: An australian experience. *Pertanika Journal of Social Science and Humanities*, 24:73–82, 2016.

Appendix A