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

January 7, 2019

UNIVERSITY
OF TWENTE.

Abstract

Contents

1 Introduction

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

2 Background

—Catchy first sentence.
Machine learning aims to infer generally valid relationships from a finite set of training data and apply those learned relations to new data [9] [20]. 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 [4]. 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 [24]. 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. [33] describes the fruitful collaboration between human and machine as *augmented intelligence*, pointing at the positive aspect of machine learning support.
—Narrowing topic to decision-making and discriminative algorithms and define "decision" as output from ML systems

2.1 Interpretability in AI

Humans cooperating with machines need to understand the principles of the method that is employed - a property referred to as *transparency* [20]. *Opacity*, the direct opposite of transparency [22], 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 [4].
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 [20], 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 [3] [33], or it is capable of generating descriptions understandable to humans [11]. 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 [13]. Using an interpretable system as a human means being enabled to make inferences about underlying data [33].
[13] 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 [30]. *Understandability*, *accuracy* of the explanation, and *efficiency* of the explanation together form *interpretability*. *Explainability* is a synonym of *comprehensibility* [36], which is also synonymic to *understandability* [2] and therefore an aspect of interpretability, showing the reasons for the system's behaviour [11]. 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 [3].

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

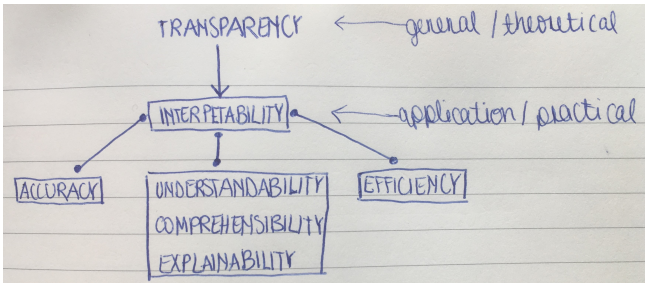

Figure 1: Relation of terms connected to interpretability

ing about interpretability should also include discussing the interpretability of the human in the loop. [22] argues that human behaviour is often mistakenly identified as interpretable, since humans can explain their actions and beliefs. However, the actual operations of the human brain remain opaque, which refutes interpretability [22]. Human interpretability is not the focus of this paper and will therefore not be discussed in the remainder, yet it should serve as a

point of reference for the general discussion of algorithmic interpretability.

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 [16]. xAI systems face a trade-off challenge: Their explanation has to be complete and interpretable at the same time [11]. The attention span an cognitive abilities of humans therefore become an important factor to consider in the design of a xAI systemm [21]. 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 [11]. The persuasiveness of an explanation is uncoupled from the actual information content of an explanation [3] 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 [17]. 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. [17] 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" [4]. 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 [13].

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 [28] [5] 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. [32] 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) [26], 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 [33]. [23] 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 [23]. 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 [26]. 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 [34]. 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. [28] 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 [26] [8] [3] [6] [34], 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 [26] [28] [32] [27]. Other positive effects on users are satisfaction and acceptance [3] [6] [34] as well as the ability to predict the system's performance correctly [3].

Explanations also help engineers and experts to design, debug, and improve an automatic decision system [26]. Explanations facilitate the identification of unintuitive, systematic errors [11] [27] in the design and redundantise time-consuming trial-and-error procedures for parameter optimisation [23]. Unethical biases in training data leading to automated discrimination [8] can be identified and examined via explanations [11] [28] [27]. Ultimately, the early identification of errors avoids costly errors in high-risk domains [8] [2] [32] and ensures human safety in safety-critical tasks [11] [28].
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 [11]. With the European General Data Prodection 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 [12] [11] [2]. Finally, the most general goal of implementing explanations for automatic decision systems is the opening and accessibility of a knowledge source [2] [28]. 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

General Data Prodection Regulation (GDPR): law about processing of personal (related to identifiable person) data, no matter if manually or automatically processed [1]
Sensitive data / protected traits: race, ethnicity, religion, nationality, gender, sexuality, disability, marital status, age [2]
Real-life data contains society's structures and biases, and as classification means separation into groups based on that data, biases are taken into the model [1]

- minimal interpretation: delete sensitive data from dataset
- maximum interpretation: delete sensitive data and correlated variables from dataset

"right to explanation" [1], argument against such interpretation [58] and positive interpretation [37]. Key issue: "data subjects receive meaningful, but properly limited, information" [58] is ambiguous, plus no clear definition of explanation, meaningful, and information. Summary: Precedents are needed to clarify the boundaries.
Problem with explanations: ML algorithms show statistical correlation, not

causality [1]

Accountability information worth disclosing for more accountability [2]:

- 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

2.3 Application Areas

2.3.1 Application Areas

decisions that affect people's lives in critical domains like criminal justice, fair lending, and medicine. [52]
safety-critical industries (self-driving cars, robotic assistants, personalised medicine) [3]

sensitive data processed by algorithms (banks, insurances, health data) [3]

scientific research (making discoveries by understanding data) [3]

individual performance monitoring, health care, economic situation analysis, personal preferences & interests, location & movement [1]

replacing human decision making in advertising, recommendations, finances (loans) [6]

health care, recommender systems, planning, HRI [15]

medicine, finances, criminal justice [19]

education, health care, manufacturing, retail, when machine learning based support systems are used [16]

high-risk domains: medical diagnosis, terrorism detection [24]

"socially consequential mechanisms of classification and ranking" [33]

spam detection, finances (fraud detection, loan), search engines, news trends, marketing, insurance [33] [36]

2.3.2 Exemplary Failures

examples of failures due to missing explanation: [3]:

- St Georges hospital - racist application procedure
- COMPASS crime prediction - racist against blacks (counterargument made in [55]: "group differences in scores may reflect true differences in recidivism risk")
- Amazon prime district selection - defavorising neighborhoods with ethnic minorities
- Automated target identification - decision driven by weather condition
- Animal race
- Mortgage rates of major US banks rate very differently - sign for bad algorithms?

"Discrimination, is at some level, inherent to profiling: the point of profiling is to treat some people differently" [57]:

- Discrimination of women: ads of higher-paid jobs more often shown to men than to women (but no reason given, may be intentional)

[54]

- researcher group is the main reason for variance, not classifier etc., hence human bias in ML

[14]

- Google Flue Trends: systematic modelling error

2.4 Explanations

Explanation = reasons or justification for an action or belief [14]

Function of explanations:

- prediction of consequences of (similar) events in the future [11] [5]
- control of events [5]
- building and refining inner knowledge model [5]

- restauration / prevention of states or events [11]
- comparison of methods [11]
- reproduction of states or events [11]
- assigning guilt [11] [5]
- justification [11] [5]
- persuasion [5]
- pleasure / appreciation [11]

2.4.1 Human-Human Explanations

explanations are not mental model but rather the interpretation of relations [11]
explanations are less general than theories and are application-focussed [11]
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 [5].
Complete explanation == all relevant causes explained [5]

Explanation aspects [11]:

- causal pattern content: common-cause, common-effect, linear chain, homeostatics
- explanatory stance types: mechanical, design, intention stance [5]. Atypical stances can lead to distorted understanding.
- explanatory domain: different fields have different preferences of explanation types
- social-emotional content: can alter acceptance threshold and influence recipient's perception of explained event

What constitutes a **good explanation**? [11] 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 [11]. 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 [5]. 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 [11]. 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 [5], which can mean to generalise and to omit information - [11] 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 [5]. 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 [11].
[5] 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.
[21] defines a good explanation as sound, complete, but not overwhelming. While soundness refers to the level of truthfulness, completeness describes the level of disclosure [21]. 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.

Generally, the more diverse the given evidence, the higher the recipient's **acceptance** of the explanation [11].

Cultural differences exist for the preference of an explanation type, although all explanation types can be understood [11].

Humans build **mental models** of the world, a mental representation of events or elements. Mental models are needed to explain and predict. 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 [21].

mindlessness and explanations [32]

Explanation Types associations between antecedent and consequent, contrast and differences, causal mechanisms [10]
material cause, formal / categorical cause, efficient cause, final cause [5]

2.4.2 AI-Human Explanations

Focus:
[10]:

- feature-level: feature influence, intersection of actual & expected contribution per feature
- sample-level: explanation vector, linguistic explanation for textual data using BOW, subtext as justification for class (trained independently), caption generation
- model-level: rule extraction, prototypes & criticism samples representing model, proxy model (inherently interpretable) with comparable accuracy (NOTE: supposedly meant decision generation, not simple accuracy)

single focus: feature-based explanation best for recommender systems (as compared to similar previous decisions and similar neighbor decisions) [10]
[4]:

- understanding and reassurance (right for the right reasons)
- diagnosis (of errors, inacceptable performance or behaviour)
- refinement (improving robustness and performance)

[6]:

- representation of data & features
- processing of data (operations)
- explanation generation (within model)

[5]:

- computational / operations level
- representational level
- hardware level

[8]:

- learning algorithm behaviour
- model parameters
- model itself
- representation

[15]:

- within algorithm, directly based on model
- feature-based
- secondary, add-on explanation system separate from learning algorithm
- representation

[14]:

- inner workings for transparency
- post-hoc prediction visualisation, e.g. heat maps

[16]:

- dataset / features
- optimizer / learning algorithm
- model
- prediction / result
- evaluator

Explanation selection: it is not possible to show every case, parameter, feature importance to the user, therefore a selection of exemplary cases needs to be made [24]. Global explanation can originate from a set of representative cases [24].

2.4.3 When to explain?

[15] stresses 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.

2.4.4 Explanation Systems

For models that are not inherently interpretable, the explanation can only be an approximation and cannot be complete (definition of non interpretable) [5]. There can be approximations for the computation / operations detecting properties and categorisations, and approximations of the decision behaviour [5].

counterfactual explanation [12] with fact & foil
[4] for overview over solutions for understanding, diagnosis, refinement
[6] for overview of solutions for explaining features, operations, generative explanations

[16] for solutions for dataset, optimizer, model, predictor and evaluator
[14] for set of programs (MYCIN, NEOMYCIN, CENTAUR, EES) that try to model explanations alongside with system
[19] presenting the L2X system
[24] Explanation software: LIME, ELUCIDEBUG

For feature-based models, [19] suggests salience map masks on input features, comparable cases (input and output) as reference (or very dissimilar cases as counterfactuals), and mutual information analysis per feature. For the latter, they use the Kullback-Leibler divergence to calculate the mutual information of two vectors: Learning to explain (L2X).

Inherently interpretable / transparent models:

- decision trees (graphical representation), rules (textual representation), linear models (feature magnitude and sign) [3]
- shallow rule-based models, decision lists, decision trees, feature selection, compositional generative models [10]
- decision trees, Naive Bayes, Rule-Learners [71]

[REF NEEDED] add-on and post-hoc systems might be good as explaining, but this fact in itself does not guarantee a sound, i.e. truthful, explanation, "however plausible they appear" [31].

[15] suggests to develop a new class of learning algorithms that have an inherent "explainability hyperparameter" to achieve high accuracy AND high explainability.

[36] argues that most high-dimensional real-world application data is "concentrated on or near a lower-dimensional manifold" [36], dimension reduction techniques like PCA or other feature selection algorithms can therefore be used to overcome the curse of dimensionality.

explanations for texts: [7] solution to recent development in text mining, where texts are represented in a high-dimensional vector space (e.g. fasttext, word2vec) and classified with neural nets. Compared to BOW/SVM, the W2V/CCN they used yields equally good results, because the CNN is better at identifying characteristic words.
[19] designed a system that uses deep neural networks for classification and mutual information for getting the input feature importance (in their case, single words).

Relevant words: A word is relevant to the text if removing it from the texts and classifying again results in a decrease of the classification score across all texts [56] take the opposite approach by eliminating irrelevant words, which leaves the relevant ones but show that this method does not work for neural classifiers

2.4.5 Explanation Evaluation

[6]:

- application grounded: true context, true task, users
- human-grounded: usability tests, human performance tests
- functionally grounded: no users, proxy

[8] evaluation of model interpretability:

- heuristics: number of rules, number of nodes, minimum description length (model parameters)
- generics: ability to select features, ability to produce class-typical data points, ability to provide information about decision boundaries
- specifics: user testing / perception (BUT: evaluation of visuals and perceived model rather than actual model), e.g. by measuring accuracy of prediction, answer time, answer confidence, understanding of model

[15] rather combination than only a single one:

- algorithm performance score
- user performance score
- user satisfaction score

2.5 Trust in AI

[25] notes that there exists no precise definition of trust in the field of computer science

[TRUST 02] examined the concept of trust in close relationships and define it as the willingness to put oneself at a risk and believing that the other will be benevolent. They grouped aspects of interpersonal trust into a model with three components: faith, dependability, predictability [TRUST 02].
Placed in agent, not a characteristic inherent to an agent [TRUST 02]
Trust is a subjective experience rather than objectively measurable [TRUST 05] [23].
dynamic: evolves as relationship matures [TRUST 02]
attribution of characteristics, e.g. dependability (repeated confirmation in risky situations), reliability (consistency or recurrent behaviour) [TRUST 02]
inappropriate trust can be harmful [17]
Trust as experience, trustworthiness is the characteristic and in case of computer programs consists of factors such as security, privacy, dependability, usability, correctness [TRUST 05] [TRUST 06]. Trust relates to the assurance that a system performs as expected [TRUST 05].

Trust in a system can be misused: e-crime with negative side effects, e.g. data misuse [TRUST 05].

2.5.1 Gaining User Trust

Trust factors: appeal, competence (privacy, security, functionality), transparency, duration (relationship, affiliation), reputation [23]
Concerning algorithms, users can put global trust into the system, which means trusting the model itself. Trust can also be assigned locally, into an individual decision. [24]
Trust dimensions of web systems: target (the entity being evaluated), representation (encoding of trust via social warranty, certificates, etc.), method (security), management (the entity putting trust into the system), computation (evaluation metric), purpose [25]
For classification: expectation mismatch leads to direct decrease in trust [30], strength of decrease depends on the type of mismatch. Data-related mismatch weights less strongly than logic-driven mismatch. [30]
[31] argues that trust in machine learning algorithms also depends on the characteristics of misclassified cases. He points out that an automatic system can be considered trustworthy if it behaves exactly like humans, i.e. it misclassifies the same data points as a human and is correct on those cases that a human would also correctly classify [31].

2.5.2 Trust Evaluation

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

-1 irritant

1 chaotic

2 assuring

3 motivating

0 not present

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

- Do you trust this algorithm to work well in the real world?
- Why do you trust this algorithm to work well in the real world?
- How do you think the algorithm distinguished between the two classes?

- How certain are you of the correctness of your explanation?

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

2.5.3 Perceived Understanding

Perceived understanding important for trust (rather than actual understanding):
"Findings show that the transparent version was perceived as more understandable and perceived understanding correlated with perceived competence, trust and acceptance of the system. Future research is necessary to evaluate the effects of transparency on trust in and acceptance of user-adaptive systems" [59]
Most questionnaires use factual statements to investigate perceived understanding. Participants rate the statements according to their confidence of understanding [UND 03] [UND 07] or directly their subjective understanding [UND 01] [UND 02] [UND 04] [UND 05]

2.6 Use Case Scenario

definition of offensive language [34]
hate speech detection systems

2.7 Summary

Summary

- summary scenario
- systems
- evaluation of explanations and of trust

Hypotheses

3 Dataset

Intro

3.1 Dataset Selection

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

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

Table 1: Publicly available datasets for offensive language texts

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

3.2 Dataset Construction

The original dataset was collected by Davidson et al. [7] for their research on defining and differentiating hate speech from offensive language. They con-

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

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

Table 2: Statistical characteristics of the constructed dataset

3.3 Dataset Preprocessing

Tweets exhibit some special characteristics. First, the maximum length of a single Tweet is 140 characters. Twitter doubled the length in November 2017, yet the dataset was collected before this data and therefore contains only Tweets of 140 characters or shorter. Twitter users found creative ways to make use of the 140 characters given, leading to the usage of short URLs instead of original URLs [37], intentional reductions of words (e.g. “nite” instead of “night”) [37], abbreviations [14], emojis [10] [35] and smilies [31] [18].
Furthermore, social media content can be unstructured, with word creations that are non in standard dictionaries, like slang words [14] [35], intentional repetitions [37] [15] [25] [29] (e.g. “hhheeeey”), contractions of words [31] [15], and spelling mistakes. Although those new word formations do not appear in the dictionary, they are “intuitive and popular in social media” [19].

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

On Twitter, it is custom to mention other users within a Tweet by adding "@"+username [37] [25] [35] [29], retweeting (i.e. answering to) a Tweet [37] [15], and summarizing a Tweet's topic with "#"+topic [37] [35].
Other problems in text mining are the handling of stop words [37] [10] [14], language detection [37], punctuation [10] [15] [25], negation [35], and case folding [10] [14] [29].
Researchers have developed different strategies for preprocessing Tweets. One possible approach is to simply remove URLs, username, hashtags, emoticons, stop words, or punctuation [37] [10] [15] [25] [14] [35]. A reason to eliminate those tokens can be that they assumably do not hold information relevant to the classification goal [15]. Words that only exist for syntactic reasons (this concerns primarily stop words) can be omitted when focussing on sentiment or other semantic characteristics [10]. Mentions of other users are likewise not informative for sentiment analysis and are often removed from the texts [37] [35]. 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 [10]. 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 [37].
Rather than removing tokens, they can also be replaced by a signifier token, e.g. a complete link by "<<<hyperlink>>>" [18]. In Tweets, such signifier tokens are used for mentions of usernames [31] [18] [29], URLs [31] [18] [29], smilies [18] or negations [31]. 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. [18] 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 [10] [18] [14].

The following preprocessing steps are taken in chronological order:

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

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

5. Resolving contractions such as “we’re” or “don’t” by replacing contractions with their long version[9].
6. This dataset uses a few signifiers such as “english translation” to mark a Tweet that has been translated to English, or “rt” to mark a Retweet (i.e. a response to a previous Tweet). Since those information have been added retrospectively, we discard them here and delete the signifiers from the texts.
7. Replacement of all characters that are non-alphabetic and not a hashtag by a whitespace
8. Replacement of more than one subsequent whitespace by a single whitespace
9. Tokenization on whitespaces

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

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

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

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

into a cleaned version:

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

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

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

4 Design

Intro

4.1 Classifier

Intro

Good System L2X

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

Bad System Inverse L2X

4.2 Explanations

Intro

Good System L2X mutual information

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

Bad System Inverse good system

4.3 Graphical User Interface

asdasdasd

4.4 Subset Sampling

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

- A class balance of the true labels similar to the testset,

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

We set the subsample size to 15 Tweets, which is enough to show accuracies to the first decimal place, yet assumably not too much to process for an observer in a user study.
To create a subset, 15 data points are randomly drawn from the testset.
First, the class balance of the subset is calculated. The difference to the class balance of the whole testset needs to be smaller than 0.1.
Additionally, for each classifier in the user study, the prediction accuracy on the subset is compared to the prediction accuracy on the complete testset. If, for all classifiers, the difference is smaller than 0.1, the next check is performed.
To ensure the uniqueness of the subsets, the randomly drawn Tweets are compared with the content of previously found subsets. The subset is only accepted if none of the contained Tweets appear in any previously found subset.
In the last step, the feature distribution of the subset is tested against the features of the complete testset using the *Kullback-Leibler Divergence* (KLD) metric. As the focus is directed towards the explanations (i.e. the highlighted words within a Tweet), only the explanations are used to examine the feature distribution. First, the feature distribution of the complete testset is calculated by constructing a word vector with tuples of words and their respective word counts. The word counts are divided by the total amount of words in the set, such that the sum of regularised counts equals 1. Next, a copy of the word vector is used to count and regularise the word frequencies in the subset. The result are two comparable vectors, yet the vector of the subset is very likely to contain zero counts for words that appear in the complete set but were never selected as explanation in the subset. Since the KLD uses the logarithm, it is undefined for zero counts. We use Laplace smoothing with k=1 to handle zero counts. For each classifier, the KLD is calculated and summed to a total divergence score for the subset.
We generate a quantity of 100 such subsets and order them by their KLD sum. The 10 subsets with the smallest score are chosen as the final set of subsets.

5 Experiment 1: Explanation Evaluation

5.1 Method

Intro

5.2 Results

asdasdasd

6 Experiment 2: Trust Evaluation

Intro

6.1 Method

Intro

Participants

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

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

Procedure

- tasks / survey items
- ordering of tasks

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

One paragraph for statistical analysis.

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

6.2 Results

Intro

asdasdasd

7 Discussion

8 Conclusion

asd

References

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

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

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

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

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

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

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

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

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

[10] 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[32] 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.

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

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

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

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

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