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

Supervisor:
Dr. Christin Seifert
Dr. Gwenn Englebienne

December 22, 2018

UNIVERSITY
OF TWENTE.

Abstract

1 Introduction

asd [1] asd [2]

2 Background

Intro

2.1 Opacity in AI

Intro

2.1.1 Opacity Sources

asd

2.1.2 Application Areas

asd

2.1.3 Examplary Failures

asd

2.1.4 Regulations

asd

2.1.5 Use Case Scenario

asd

2.2 Explanations

Intro

2.2.1 Explanation Types

asd

2.2.2 Social Sciences

asd

2.2.3 Psychology

asd

2.3 Explanations in AI

Intro

2.3.1 Explanation Focus

asd

2.3.2 Explanations for Texts

asd

2.4 Trust in AI

Intro

2.4.1 Persuasiveness

asd

2.4.2 Accountability

asd

2.4.3 Trust

asd

2.5 Summary

Summary
Hypotheses

3 Dataset

Intro

3.1 Dataset Selection

asd

3.2 Dataset Construction

asd

3.3 Dataset Preprocessing

asd

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