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

January 2, 2019

UNIVERSITY
OF TWENTE.

Abstract

1 Introduction

State of the world
The big BUT
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

Intro

2.1 Opacity in AI

Intro

2.1.1 Opacity Sources

asd

2.1.2 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]

2.1.3 Issues & Examplary Failures

issues / need for explainability:

- high cost of errors in high-risk domains [2] [8]
- human safety in safety-critical tasks [6] [15]
- automated discrimination [2], biased training data [6] [15]
- censorship [2]
- development: fine-tuning parameters with trial-and-error [4]
- imperceptibly altered input data or altered model: criminals / hackers [6]
- unintuitive systematic errors [6] [56]

- regulations: right to explanation [1] [6] [8]
- source of knowledge that should be accessible if needed [8] [15]
- for HRI (where trust is needed) [15]

need dependent on consequences of classification results [15]:

- **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

[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

2.1.4 Regulations

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]

2.1.5 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

2.1.6 Use Case Scenario

definition of offensive language

2.2 Explanations

Intro

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]

2.2.1 Social Sciences

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 nor 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].

2.2.2 Explanation Types

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

2.3 Explanations in AI

xAI = communication with agents about their reasoning [12]
xAI problems: reverse engineering (model, decision, visualisation/representation) and design of explanations [3]
"Mixed initiative guidance" = human expert working alongside the ML system

[4]
xAI challenge: completeness and interpretability simultaneously (tradeoff), transparency vs. persuasiveness [6]
General trend: the higher the accuracy, the lower the explainability [15]

2.3.1 Why Explanations in AI?

Right for the right reasons: not enough to just be correct, the reasons need to be correct. Example of things going wrong without noticing: [56]

2.3.2 User-Related Goals of Explanations

Scrutability, trust, effectiveness, efficiency, satisfaction, persuasiveness [2]
Explainability improves the user's confidence in the system, user's trust, user's judgement ability on prediction correctness, user satisfaction, user acceptance [10]

2.3.3 Interpretability

Definition Interpretability:

- Accurate proxy for model AND comprehensible for humans [3]
- Dimensions: scope (global vs local), time to understand (target user, use case), prior knowledge (user expertise), dimensionality, accuracy, fidelity (accuracy of explanation / accuracy of model), fairness, privacy, monotonicity, usability [3]
- operations can be understood by a human [10]
- descriptions understandable to humans [6]

Interpretability vs. justification: Why it is a correct decision, not how it came along [10]
Interpretability vs. explainability: Subgroup, showing reasons for behaviour [6]

2.3.4 Barriers to Interpretability

intentional concealment, lack of technical expertise, computational complexity vs. human-scale reasoning abilities [1]
minimum explainability: how features (values) relate to predictions [1]

2.3.5 Explanation Focus

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

2.3.6 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.

2.3.7 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

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]
-

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

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.

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

2.3.8 Evaluation of Explanations

[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

2.4 Trust in AI

Intro

2.4.1 Persuasiveness

persuasiveness of explanation != actual explanation [10]
"High-fidelity explanations, also referred to as faithful, have a strong correspondence between the explanation model and the underlying machine learning model" [51]
"[Persuasive explanations] are less faithful to the underlying model than descriptive explanations in a tradeoff for more freedom on the explanation complexity, structure, and parameters. This freedom permits explanations better tailored to human cognitive function, making them more functionally interpretable." [51]
"Descriptive explanations best satisfy the ethical goal of transparency" [51]

2.4.2 Trust

willingness to put oneself at a risk and believing that the other will be benevolent [TRUST 03]
Placed in agent, not a characteristic inherent to an agent [TRUST 03]
dynamic: evolves as relationship matures [TRUST 03]

attribution of characteristics, e.g. dependability (repeated confirmation in risky situations), reliability (consistency or recurrent behaviour) [TRUST 03]

2.5 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. [1] 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 [1]. 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 [11], intentional reductions of words (e.g. “nite” instead of “night”) [11], abbreviations [3], emojis [2] [10] and smilies [9] [5].
Furthermore, social media content can be unstructured, with word creations that are non in standard dictionaries, like slang words [3] [10], intentional repetitions [11] [4] [7] [8] (e.g. “hhheeeey”), contractions of words [9] [4], and spelling mistakes. Although those new word formations do not appear in the dictionary, they are “intuitive and popular in social media” [6].

[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 [11] [7] [10] [8], retweeting (i.e. answering to) a Tweet [11] [4], and summarizing a Tweet's topic with "#"+topic [11] [10].
Other problems in text mining are the handling of stop words [11] [2] [3], language detection [11], punctuation [2] [4] [7], negation [10], and case folding [2] [3] [8].
Researchers have developed different strategies for preprocessing Tweets. One possible approach is to simply remove URLs, username, hashtags, emoticons, stop words, or punctuation [11] [2] [4] [7] [3] [10]. A reason to eliminate those tokens can be that they assumably do not hold information relevant to the classification goal [4]. Words that only exist for syntactic reasons (this concerns primarily stop words) can be omitted when focussing on sentiment or other semantic characteristics [2]. Mentions of other users are likewise not informative for sentiment analysis and are often removed from the texts [11] [10]. 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 [2]. 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 [11].
Rather than removing tokens, they can also be replaced by a signifier token, e.g. a complete link by "<<<hyperlink>>>" [5]. In Tweets, such signifier tokens are used for mentions of usernames [9] [5] [8], URLs [9] [5] [8], smilies [5] or negations [9]. 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. [5] 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 [2] [5] [3].

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] 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.

[2] 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.

[3] 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.

[4] 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.

[5] Leonard Hövelmann, Stockholmer Allee, 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.

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

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.