AI-Assisted and Explainable Hate Speech Detection for Social Media Moderators – A Design Science Approach

Enrico Bunde
Freie Universität Berlin
enrico.bunde@fu-berlin.de

Abstract

To date, the detection of hate speech is still primarily carried out by humans, yet there is great potential for combining human expertise with automated approaches. However, identified challenges include low levels of agreement between humans and machines due to the algorithms' missing expertise of, e.g., cultural, and social structures. In this work, a design science approach is used to derive design knowledge and develop an artifact, through which humans are integrated in the process of detecting and evaluating hate speech. For this purpose, explainable artificial intelligence (XAI) is utilized: the artifact will provide explanative information, why the deep learning model predicted whether a text contains hate. Results show that the instantiated design knowledge in form of a dashboard is perceived as valuable and that XAI features increase the perception of the artifact's usefulness, ease of use, trustworthiness as well as the intention to use it.

1. Introduction

Today, a large part of human communication takes place in the digital sphere, for instance via social media [1-2], and so does hate speech, which can be harmful for individuals and society as a whole [3]. Ullmann and Tomalin [4], for instance, describe that "[…] offensive posts are only subsequently removed if the complaints are upheld, therefore, they still cause the recipients psychological harm." (p. 1).

Today, automatic hate speech detection is often based on machine learning approaches [1]. However, while, deep learning models achieve a high performance, they also show a low degree of transparency ("black box") due to the complex and self-learning algorithms, which leads to a "trade-off" between performance and explainability [5-7].

In this regard, methods of explainable artificial intelligence (XAI) were developed to make black box approaches explainable, without sacrificing performance [6]. XAI methods allow to generate explanations that can be interpreted by humans without detailed knowledge of the underlying deep learning model [8]. I suggest that XAI features have also versatile potentials in the context of deep learning-based hate speech detection. In this regard, the interaction between the human and hate speech detection system becomes more relevant, which is also the focus in the research field of interactive machine learning [9]. Li et al. [9], for example, showed that human-selected training samples can lead to higher performance faster than randomly selected samples. Moreover, explainability can lead to more trust and comprehensibility for users [5].

While social media users often moderate topical groups they have created, they are usually not supported by any tool to handle the task to identify hate speech in ongoing discussions and to react accordingly [2]. Therefore, the objective of this paper is to answer the following research question: *How does a dashboard of a system for automated hate speech detection needs to be designed to support non-professionals to moderate social media groups?* Non-professional in that context means that they are not employed or paid to moderate a social media group.

In this study, the design science research (DSR) approach by Peffers et al. [10] is followed. Based on two design cycles, a dashboard for social media users is build that integrates an algorithm for hate speech detection based on the Universal Language Model Fine-tuning (ULMFiT), a state-of-the-art deep learning approach [13], which was fine-tuned on a publicly available hate speech dataset with 3,947 samples. The first version of the dashboard interface was derived from insights of the knowledge base and was qualitatively evaluated with 15 participants. In the second design cycle the user feedback was operationalized and the added value of XAI features was tested in an experiment with 200 users. Hence, the contribution of this study lies in the introduction of derived, refined and evaluated design knowledge.

This paper is structured as follows: First the research design is described. This is followed by the problem identification and motivation, succeeded by the section on objectives and design. Afterwards, the design

URI: https://hdl.handle.net/10125/70766
978-0-9981331-4-0

principles are developed and demonstrated (design cycle 1 and 2), followed by the evaluation of the final design and experimental test of the relevance of XAI features. The paper ends with a discussion and conclusion.

2. Research design

2.1. Design science research framework

For this study, a DSR approach was applied and the process steps of Peffers et al. [10] adapted. The objective is to obtain new design knowledge for a class of IT artifacts [11] which also can be utilized by future DSR and information systems (IS) research as input knowledge [12]. The overarching objective is hence to provide an innovative solution to a real-world problem [12] and produce a contribution to the knowledge base [11]. The applied DSR process is illustrated in figure 1. The sequential process followed here, consists of five activities [10]: (i) problem identification and motivation; (ii) definition of objectives and design; (iii) development and demonstration (iv) evaluation; and (v) conclusion and communication.

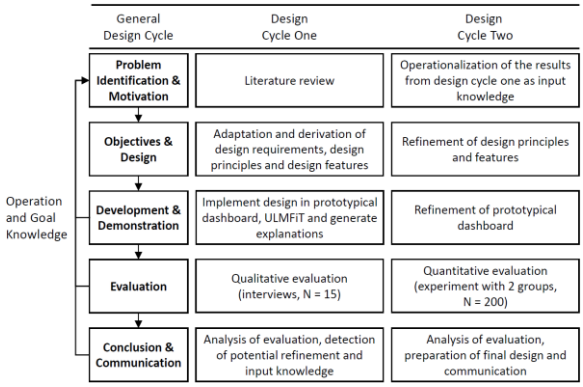

Figure 1. Design science process (based on [10]).

The *first design cycle* was initiated with a literature review to identify the problem, motivate the project, and provide the knowledge base. In this knowledge base, generic design requirements (DRs) were identified and adapted for this DSR project [31]. The objectives are represented by the adapted DRs. To address these requirements, a set of design principles was defined, which provided guidance for the design of the dashboard. These design principles (DPs) were then translated into specific design features (DFs), which were subsequently addressed in a prototypical dashboard. The resulting initial design from the first design cycle was qualitatively evaluated through semi-structured interviews [40]. The participants (N = 15) were provided with the dashboard, for which feedback regarding the design was collected. The first interview participants were recruited in the university environment and through snowball sampling, further interview participants were recommended and identified [40]. The decisive criterion was the experience as a moderator of a social platform, whereby the size or orientation of the platform was not decisive.

The *second design cycle* started by operationalizing the gained knowledge of the evaluation of the first design cycle. Consequently, the refinement of the DPs and DFs was conducted. In the next step, the refined design knowledge was also updated within the prototypical dashboard. This updated dashboard was then evaluated quantitatively. For this purpose, the constructs *perceived usefulness* [33-34], *perceived ease of use* [33-34], *trustworthiness* [35] and *intention to use* [36] were used. Additionally, the role of explainability was investigated by evaluating the dashboard with XAI features and without XAI features. The participants (N = 200) were recruited via MTurk (Amazon Mechanical Turk) and separated into two groups (N_{Group1} = 100; N_{Group2} = 100). To assure that participants did not take part in both groups, unique user IDs were filtered.

2.2. Technical setting

The prototypical dashboard was implemented using Adobe XD, which is a vector-based graphics software for the design of graphical user interfaces for web and mobile apps. The generated examples of hate speech classifications are generated based on ULMFiT [13], a state-of-the-art natural language processing (NLP) model. ULMFiT was chosen as it represents a transfer learning method that can be utilized for various NLP tasks and is able to match the performance of other pretrained models while using less data [13]. This model was implemented with Python and fine-tuned, with the provided AWD-LSTM language model, on a public hate speech dataset that was listed in the work of MacAvaney et al. [2]. Two classes (binary classification) are differentiated in this data set [41] which was part of a Kaggle competition: *hate speech (1,049 samples)* and *no hate speech (2,898 samples)*. Here, only the provided train data was utilized, as the labels are necessary to evaluate the performance. Eventually, the data was split into 80% for fine-tuning and 20% for the test. Metrics were generated with scikit-learn.

3. Problem identification and motivation

3.1. Dangers of hate speech

Davidson et al. [14] define hate speech as "[…] language that is used to express hatred towards a

targeted group or intended to be derogatory, to humiliate, or to insult the members of the group" (p. 1), which represents the working definition for this study. Such content is found to be significantly harmful for individuals and societies as a whole [4]. Hence, there is the need of support through tools that allow for the detection or prevention of hateful content [3]. The consequences of encounters with hateful content is also considered as a threat to national and international security [15]. Major social platforms such as Facebook or Twitter have evaluated hate speech as harmful and therefore implemented general policies to remove such content from the platforms [2]. The German federal government, for example, has introduced the Act to Improve Enforcement of the Law in Social Networks (Network Enforcement Act) in 2017, which sets out fine-enforced compliance rules for the operators of social platforms and includes the topic of hate speech. However, policies or compliance rules are difficult to enforce, if hate speech cannot be detected efficiently.

3.2. Challenges and approaches for automated hate speech detection

Automated hate speech detection is gaining importance as social media content is continuously growing and likewise the spread of hate speech [16]. However, even leading providers of social platforms do not use automated hate speech detection: Udanor and Anyanwu [19] mention how providers such as Twitter or Facebook do not yet apply automated hate speech detection, rather the monitoring is usually done by humans based on posts which have been flagged as potential hate speech by users. Due to the growing amount of content on social platforms, the automated detection of hate speech is a topic that has high potentials for research and practice.

Sahi et al. [20] have investigated the automatic detection of hate towards women on Twitter. There exist further studies, which focus on a specific problem, e.g., detecting abusive language [21] or the risks of racial biases in hate speech detection [22]. To automatically detect hate speech, different approaches are used. Machine learning approaches such as Logistic Regression, Decision Trees, Random Forests or Support Vector Machines are often applied [1; 23]. There are also deep learning approaches, which are increasingly used because of a higher level of performance, such as Convolutional Neural Networks, Recurrent Neural Networks or Long Short-Term Memory [24]. Schmidt and Wiegand [16] have further summarized the features that are being commonly used for hate speech detection, for example, sentiment analysis, word generalization, lexical resources, linguistic features, or multimodal information. However, despite the versatile research in the context of hate speech, there still exist numerous challenges. For instance, Schmidt and Wiegand [16] have identified the need for a benchmark dataset for hate speech detection with a concrete definition of a task. Moreover, Fortuna and Nunes [3] state that classification of hate speech can be more difficult for automatic approaches than for humans [17]. Also, these tasks require expertise in cultural or social structures, why automatic hate speech detection approaches must dynamically adapt to the ever-changing language in social platforms and networks [18]. Hence, it can be said that there are various methods and techniques to approach the problem of automatic hate speech detection and corresponding challenges. The novelty of this work lies in the combination of the deep learning-based hate speech detection with XAI and a dashboard that provides explanations to support moderators of social platforms.

3.3. Deep learning algorithms as black boxes

Deep learning approaches are complex and therefore difficult to comprehend. The criticism of the black box character relates, for example, to the complex task of parameter tuning and the lack of understanding regarding the problem-solving process [25]. In this regard, the research field of XAI develops methods and techniques to improve the transparency and explainability of such approaches [5]. XAI has already been applied in different contexts such as health sector [26] or applications in the context of recommender systems [27]. The reasons and motivations for the use of XAI can be described as, for example, explain to justify, control, improve, discover, verify, or manage as well as to comply to legislation [5, 28].

In the context of XAI and the generated explanations, Miller [29] describes how other research fields such as philosophy, psychology and cognitive science are relevant when it comes to how humans select, understand and present explanations. Thereby, XAI can be described as an interdisciplinary research field. Cheng et al. [30] have investigated explanation interfaces for explaining decision-making algorithms. Thus, XAI and explanation interfaces or dashboards are already the focus of different research streams.

Different approaches exist that aim to explain black box approaches and their outcomes. These XAI methods can be divided into different categories. The existing XAI methods vary in terms of their output and hence usefulness from a developers' or users' perspective. Some methods provide "examples" to explain, others provide "model internals" [5]. Some provide information regarding data features that supported the model's prediction, some provide opposing data features, including "counterfactuals", and again others

provide both [5; 8]. Some methods are model-agnostic, others are model-specific [5-6].

4. Objectives and design

4.1. Adaptation of design requirements

This DSR project aims to contribute theoretically grounded and evaluated design knowledge for dashboards in the application context of decision support for automated hate speech detection. The design knowledge is directed at non-professionals who moderate social platforms and aims to support them in their activity to detect hateful and thus potential harmful content and react accordingly (e.g. delete a post).

To reach this goal, generic design requirements of Meth et al. [31] were adapted. The authors introduced DRs for decision support systems (DSS), which address various human decision makers' goals and are described as important features of any DSS: (i) *increase decision quality by providing advice with high advice quality*; (ii) *reduce human decision maker's cognitive effort by providing decision support*; and (iii) *minimize system restrictiveness by allowing users to control the strategy selection* [31]. In the following the adapted DRs will be described: (DR1) *Increase the automated decision support for non-professional social media moderators (SMMs)*: The dashboard should support the moderator in detecting hateful content. To facilitate trust in the machine learning-based system, the dashboard should offer explanations for the given classifications of hate speech. (DR2) *Minimize cognitive efforts for SMMs required to understand and validate the automated decision support:* The dashboard should provide key reasons for the outcome. (DR3*) Support SMMs with additional information about the author of potential hate speech:* The dashboard should support the moderator with additional contextual information about the user (potential author of hate speech) and his behavior. (DR4) *Retention of the power to make decisions for the SMM:* The dashboard should decrease the system restrictiveness and leave the decision-making power with the moderator by offering appropriate actions that could be taken [31].

4.2. Definition of design principles and features

In the following, the DPs are derived that address specific DRs. Afterwards DFs are developed that address specific DPs and represent features of the artifact. (DP1) *Provide the system with capabilities to explain the present classification.* By implementing XAI techniques to explain automated hate speech detection the users' trust in the system can be improved [5]. Additionally, the improved explainability can lead to a greater support for the work with such systems [6]. Regarding decision support, explainability can also lead to enhanced fairness [32]. To increase the explainability of the system, different techniques and visualizations can be utilized and combined [5-7]. (DP2) *Provide the system with capabilities to provide the key reasons for the outcome and information on the author of potential hate speech.* While DP1 provides a more general and global explanation, DP2 focuses on specific features and information. To provide the key reasons, i.e. most relevant words for hate speech classification, XAI techniques can be utilized such as feature permutation or feature importance [5-8]. Additionally, the system should provide information regarding the user such as the analysis and evaluation of historical posts. Both elements aim to minimize the cognitive efforts for the SMM by providing relevant reasons that can be validated and through additional information on the user, the behavior can be better assessed. (DP3) *Provide the system with the capabilities to support the initiation of appropriate actions.* The final decision-making authority should lie with the SMM. Therefore, appropriate actions should be provided that can be initiated based on the present hate speech case. Additionally, if SMMs detect false classifications, they should be able to correct and re-classify them. This enables an interactive learning process for the underlying system [9] and the explainable system provides support [6-7; 32]. These are the three derived DPs that address specific DRs. Within the next phase, the derivation of the DFs follows.

(DF1) *Utilize XAI techniques to explain the automated hate speech detection*. As many deep learning and state-of-the-art models are opaque and often used as black box, additional techniques and methods are required to explain the outcomes [5-7; 32]. There are various XAI techniques that can be utilized. In this project the provided module of ULMFiT was used to obtain the explanations [39]. (DF 2) *Provide the confidence for the present classification*. The probability of the classification will be represented as the confidence of the AI system which is also an eponymous goal of XAI [32]. This is represented by the probability value of the classification in percentage. (DF3) *Utilize feature importance to obtain the most relevant features for a specific outcome*. Through feature importance techniques, the most relevant features for a specific outcome or classification, e.g. words, can be obtained [5; 8]. Based on these relevant reasons, the moderator can validate the impact of these words. (DF4) *Utilize data analytics and visualization techniques to provide additional information on a user*. By visualizing data, the understanding can be improved [6-7]. To reduce the cognitive effort, visualization

techniques will be utilized to provide the moderator with additional information on the behavior of the user in the past. (DF5) *Provide capacity to initiate actions*. To retain the power to make decisions with the moderator, appropriate actions must be offered. Here, SMMs should be provided by easy to use dashboard elements that initiate actions such as checkboxes, e.g. to delete hateful posts. (DF6) *Allowing the re-classification of cases*. By allowing the SMM to re-classify present cases, e.g. in case of false outcomes, an interactive machine learning loop can be utilized to extend the training data for hate speech detection by human-selected examples which can improve the performance [9]. (DF7) *Provide the possibility to contact affected user*. Additionally, a possibility should be provided for the moderator, to contact the author of potential hate speech. Figure 2 provides an overview of the derived, refined and evaluated design and illustrates their relations. The next section describes the two design cycles and the evolvement of the design knowledge.

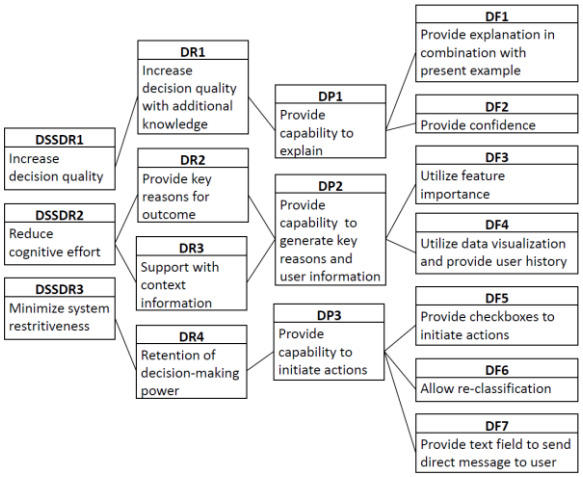

Figure 2. Overview of the derived and refined design knowledge.

5. Development and demonstration

5.1. Design cycle one

Within the *first design cycle*, ULMFiT was implemented and fine-tuned on the hate speech dataset. The fine-tuned ULMFiT model reached following performances for both classes. *Hate speech:* precision: 93,93%; recall: 73,63%; f1-score: 82,55%. *No hate speech:* precision: 96,11%; recall: 95,37%; f1-score: 95,74%. For fine-tuning I followed the recommended steps in the ULMFiT documentation [38]. Additionally, the ULMFiT module for interpretation was utilized to generate the explanations [39]. The explanations and visualizations for the artifact were processed manually and graphically.

The initial prototypical dashboard was implemented and addresses the initial design knowledge (i.e. DFs). The objective was to design the dashboard for layman and support them in the moderation of social platforms (e.g. groups in social media) by identifying hateful content, providing explanations and initiate appropriate actions. Through the capability of re-classifying examples, a dataset for the optimization of the underlying algorithm can be curated and utilized to further improve the performance. This is also one of the concerns within interactive machine learning [9]. The demonstration will showcase how the artifact can solve the identified and described problem [10]. Figure 3 represents the initial design.

The evaluation of the first prototype aims to receive feedback for optimization [10]. This feedback was collected through a qualitative evaluation with 15 participants. After the presentation of the designed dashboard, the participants were asked what they liked about the dashboard and what could be improved. The collected feedback has shown a positive sentiment regarding the design.

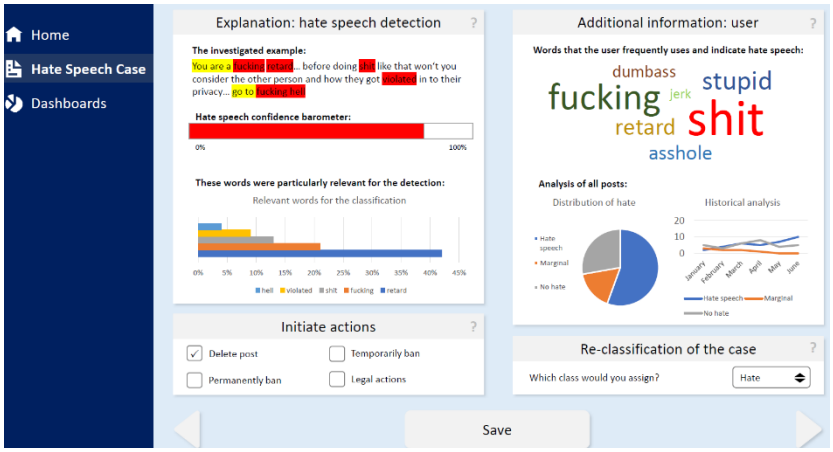

Figure 3. Initial design and prototype evaluated in design cycle one.

Participants stated that the dashboard has a *"clean design and it looks good" (I1)*, it is *"an excellent way of identifying and deleting the hateful contents" (I3)* or that *"the dashboard is intuitive and easy to use" (I7)*. When asked how to improve the presented design some participants stated that the dashboard elements are *"smashed together" (I3)*, it needs more *"interactivity to adjust the presented information and visualization" (I11)* or the information on the user and its presentation are *"too bright, colorful and overwhelming" (I14)*. In addition, the participants stated that the combination of the *"different visualizations are difficult to grasp at first glance" (I8)* and it was perceived as *"static" (I9)*.

In summary, the evaluation of the first design cycle showed that the participants had a positive perception of the dashboard. However, it was also possible to identify additional impulses, which allowed the design knowledge to be refined. This serves as input for the second design cycle.

5.2. Design cycle two

The *second design cycle* started by operationalizing the gained knowledge from the first design cycle and adapting the DPs and DFs to address the feedback and evaluation results. Within this design cycle, the machine learning model of the implementation remained unchanged. However, the design of the dashboard was revised based on the new insights. During the revision of the design the DF1 was refined and the probability was added to represent the confidence. The XAI method used should also explain the classification using an example. Since the dashboard from the evaluation of the first design cycle was described as "too bright, colorful and overwhelming" the implementation of DF3 was adjusted and fewer visualizations are integrated. Additionally, the confidence is no longer depicted as a bar chart, rather it is a confidence score. The DF7 (*providing the SMM with a text field to send direct messages to the author of potential hate speech*) was also introduced based on the evaluation of the first design cycle, since users described this feature as desirable. By integrating more buttons and signaling possibilities of adjustments, the demanded interactivity is addressed, and the individual elements were re-arranged to enhance the clearness and comprehensibility.

The following figure 4 illustrates the refined implementation of the DF1 and DF2, which address DP1: *Provide capability to explain.* DF1 is implemented by providing a clear outcome of the classification, DF2 presenting the analyzed post and the probability represented through a confidence interval. DF3 addresses the DP2*: Provide capability to generate key reasons and user information*. The highlighted words represent the key reasons for the classification and through their coloring their relevance is indicated. Such visualizations can also be generated by utilizing feature importance or feature permutation approaches [5; 8].

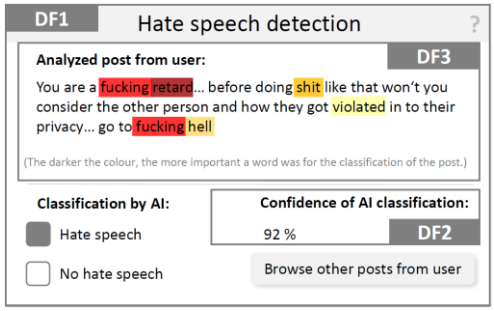

Figure 4. Addressing DP1 and DP2.

The following figure 5, represents the implementation of DF4, which addresses the DP2: *Provide capability to generate key reasons and user information*. Additional information about the user are presented through data visualization. This information supports the SMM to get a better understanding on the communication history as well as actions taken against the user in the past. Consequently, the SMM is supported in getting a comprehensive picture of the user and his behavior.

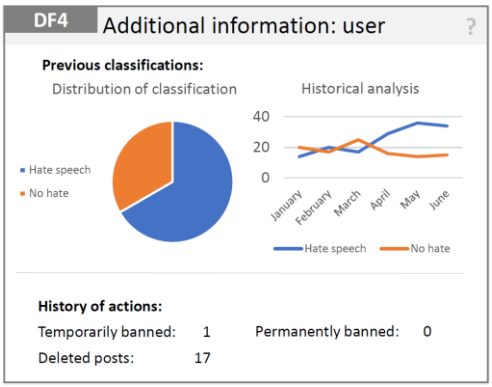

Figure 5. Addressing DP2.

The DP3: *Provide capability to initiate actions*, is addressed by DF5, DF6 and DF7 and is depicted in figure 6. DF5 is implemented by providing checkboxes through which SMMs can easily initiate actions based on the present hate speech case. DF6 is implemented by providing a dashboard element that enables the re-classification of the given text. This could trigger an iterative process in the background to collect data for automated hate speech detection and improve the performance of the deep learning approach [9]. Additionally, the DF7 is provided so that the moderator can send a direct message to the author of potential hate speech.

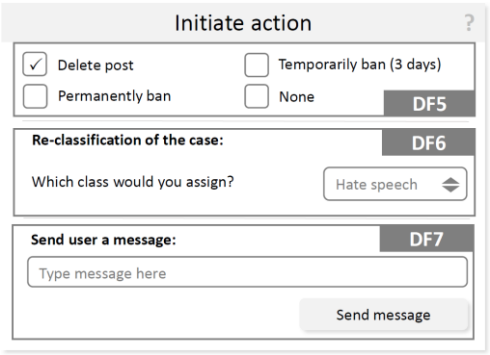

Figure 6. Addressing DP3.

In the following the two versions of the dashboard for the evaluation are presented. As shown in the problem identification and motivation, there is a lack of empirical knowledge on the relevance of XAI features. Hence, in addition to establishing design knowledge, this study aims to provide such empirical knowledge. Therefore, the following section describes the final evaluation with two groups: one group will be presented the final dashboard with XAI features (figure 7), and a second group will be presented the same dashboard but without XAI features (figure 8).

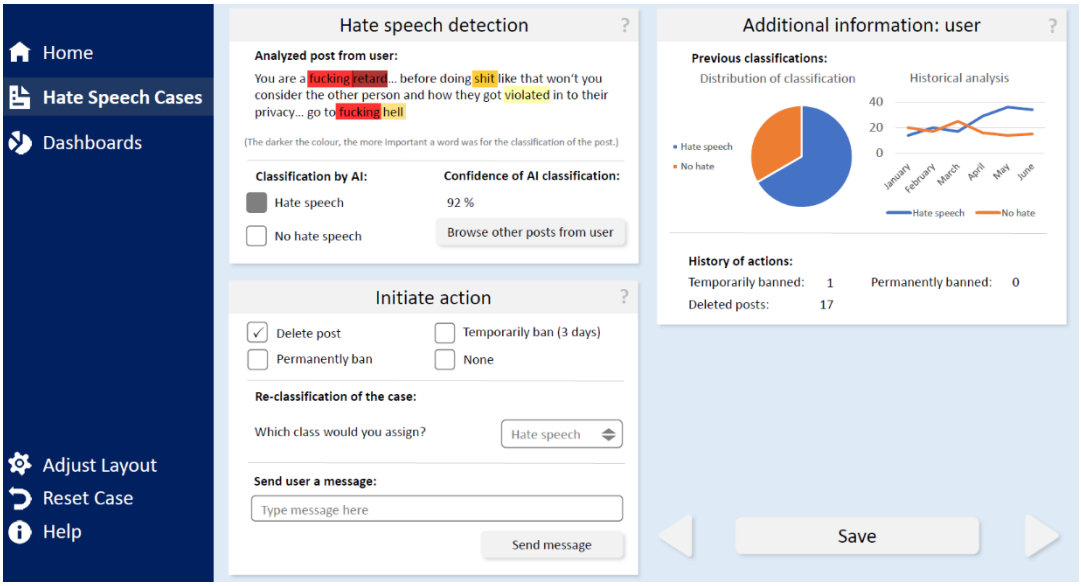

Figure 7. Dashboard with all DFs.

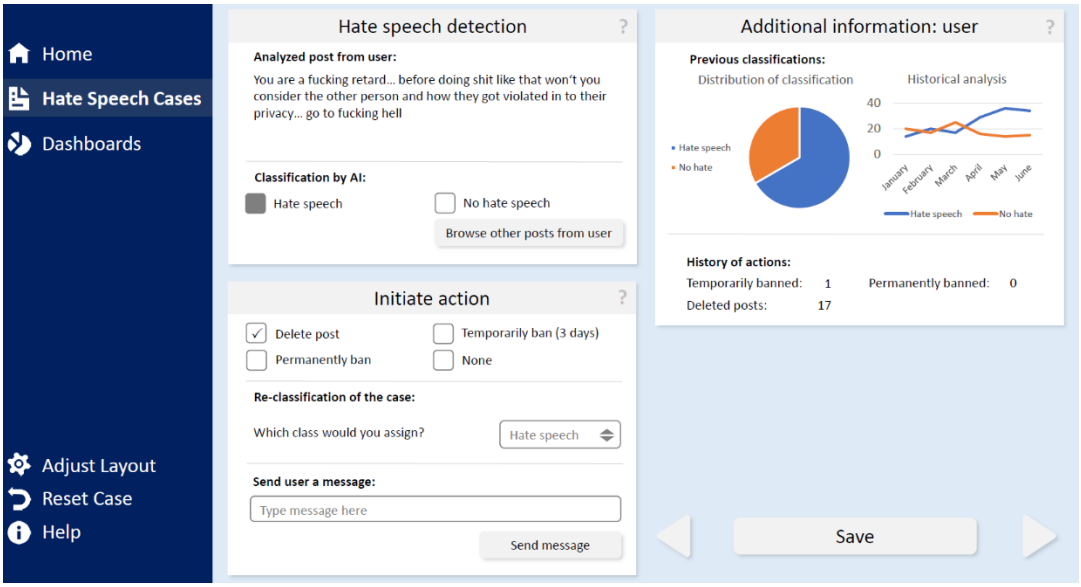

Figure 8. Dashboard with DF4, DF5, DF6, DF7.

6. Evaluation of final design and test for the relevance of explainability

The evaluation of the second design cycle was conducted online. Group 1 (no XAI features) was presented the dashboard as shown in figure 8, group 2 was presented the dashboard as shown in figure 7 (with XAI features). A Likert scale ranging from 1 (I completely disagree) to 5 (I completely agree) was used to evaluate the constructs *perceived usefulness* (e.g., "The AI-based dashboard is useful for detecting hateful content."; [33-34]), *perceived ease of use* (e.g., The AI-based dashboard for hate speech detection is easy to use."; [33-34]), *trustworthiness* (e.g., "The AI-based dashboard can be trusted to carry out hate speech detection faithfully."; [35]), and *intention to use* (e.g., "If available, I intend to use the AI-based dashboard for hate speech detection as a moderator on social platforms in the next six months."; [36]). The participants were recruited through MTurk. 100 users in the first group (no XAI features) and 100 in the second group (with XAI features) participated. Table 1 summarizes descriptive data on the participants.

Table 1. Descriptive data on the participants of the second design cycle (N_{Group1} = 100; N_{Group2} = 100).

Characteristic/ Question	Group 1 (no XAI)	Group 2 (with XAI)
Gender		
Female	32	38
Male	67	62
Other	1	0
Age		
< 20	0	0
20 – 29	41	22
30 – 39	35	37
40 – 49	13	33
50 – 59	8	8
> 59	3	0
Have you ever moderated a social platform or group on social platforms (e.g. social networks)?		
Yes	77	81
No	23	19
Have you ever encountered hateful content on social platforms?		
Yes	86	91
No	14	9

The main objective was to evaluate the final design and assess the impact of explainability features. Therefore, following hypotheses were derived. Figure 9 represents the research model.

H1: Providing a dashboard with explainability features leads to users having increased:

a) perceived usefulness;
b) ease of use;
c) trustworthiness;
d) intention to use,

when compared to a dashboard without explainability features.

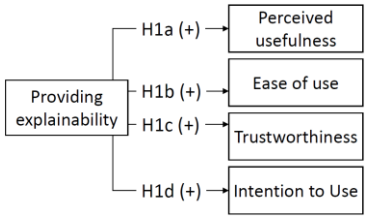

Figure 9. Research model.

The hypotheses were tested by examining differences in the mean values of the two groups. Both groups were compared to each other using Mann-Whitney U tests [42]. Hence, hypothesis H1 was tested by comparing group 1 (no XAI) with group 2 (with XAI). Two-tailed tests were used for the comparison. Table 2 provides an overview of the results. Here, the p-values, Pearson correlation coefficient r (in brackets) [43] as well as an indication if the hypothesis is supported or not supported is provided. Results for H1a indicate that explainability features have a significant and positive effect on the perceived usefulness, ease of use, trustworthiness, and intention to use. Hence, all four hypotheses were supported. The explainability feature has the strongest correlation with intention to use, followed by perceived usefulness, trustworthiness, and ease of use. Table 3 presents the mean and standard deviation for both groups.

Table 2. Results of hypotheses tests (PU = Perceived usefulness, EOU = Ease of use, TRU = Trustworthiness, ITU = Intention to use, Supp. = Supported).

Hypo-thesis	PU	EOU	TRU	ITU
H1a, b, c, d	0.0005 (0.5129)	0.0001 (0.0174)	0.0004 (0.2420)	0.0011 (0.6659)
	Supp.	Supp.	Supp.	Supp.

Table 3. Results of experiment (PU = Perceived usefulness, EOU = Ease of use, TRU = Trustworthiness, ITU = Intention to use, M = Mean, SD = Standard deviation).

Construct	Group 1 (no XAI)		Group 2 (with XAI)	
	M	SD	M	SD
PU	3.71	0.64	4.02	0.51
EOU	3.70	0.62	4.04	0.48
TRU	3.63	0.78	4.01	0.47
ITU	3.68	0.74	3.99	0.54

7. Discussion and conclusion

In this DSR project, design knowledge and an instantiation of according design principles via a deep learning-based dashboard that supports SMMs was introduced. During the development and implementation of the system, many of the challenges and problems described in scientific literature were encountered, e.g. the identification of benchmark dataset for hate speech detection or the relatively small size of the datasets [2]. By integrating XAI techniques such as the ULMFiT interpretation module [39], individual predictions can be explained. Along with additional information the SMM can interpret the model's prediction, help improving the data quality, and generate trust towards the model [25]. Additionally, objectives such as the personalization of explanations could be integrated and examined [44]. The proposed design can be utilized as input knowledge for future DSR or IS research projects. SMMs can validate hate speech detection by the deep learning-based system and make the final decision as to whether it has correctly classified the text or not. By saving these new texts and their corresponding class (e.g. hate speech or no hate speech), the datasets evolve and grow. In doing so, the human beings' knowledge of cultural and social structures can be integrated [17] and the dataset is constantly updated, which also includes the dynamic development of the language [18]. With additional examples from the explainable dashboard, in the long run, there is also the potential that the performance of the AI-system can be increased [9], which leads to a more accurate hate speech detection.

The focus of this study was the design of the dashboard interface, and hence the frontend design. This resulted in the circumstance, that the prototype was not based on interactive machine learning architectures. However, the used examples were generated through a real ULMFiT implementation to demonstrate the technical feasibility. Additionally, such artifacts can be utilized in behavioral science projects to conduct experiments such as the here investigated *perceived usefulness*, *ease of use*, *trustworthiness*, and *intention to use*.

As the content in social media is rapidly growing, practice and science has shown a high demand for automated hate speech detection [2-3]. In the context of such growing online content and hence data volumes, data-driven solutions are being increasingly applied [37]. In this regard, deep learning plays an important role and can lead to major breakthroughs in the field of hate speech detection. At the same time, state-of-the-art AI approaches represent black boxes [6-7; 32]. The here established design principles address this problem as well as other before mentioned challenges regarding the automation of hate speech detection and hence, contribute to solving several challenges identified in this area. Future work could focus, for instance, on the development of an artifact for a longitudinal field study in practice, which could generate valuable insights on how the design is perceived in a real-world setting.

8. References

[1] Burnap, P., and Williams, M. L. 2015. "Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making." *Policy & Internet* (7:2), 223-242.

[2] MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N. and Frieder, O. 2019. "Hate speech detection: Challenges and solutions." *PLoS ONE* (14:8), 1-16.

[3] Fortuna, P., and Nunes, S. 2018. "A Survey on Automatic Detection of Hate Speech in Text." *ACM Computing Surveys* (51:4), 1-30.

[4] Ullmann, S., and Tomalin, M. 2019. "Quarantining online hate speech: technical and ethical perspectives." *Ethics and Information Technology*, 1-12.

[5] Adadi, A., and Berrada, M. 2018. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." *IEEE Access* (6), 52138-52160.

[6] Gunning, D., and Aha, D. 2019. "DARPA´s Explainable Artificial Intelligence (XAI) Program." *AI Magazine* (40:2), 44-58.

[7] Arrieta, A. B., Diaz-Rodriguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. 2020. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." *Information Fusion* (58), pp. 82-115.

[8] Ribeiro, M. T., Singh, S., and Guestrin, C. 2016. "'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*, 1135-1144.

[9] Li, H., Fang, S., Mukhopadhyay, S., Sykin, A. J. and Shen, L. 2018. "Interactive Machine Learning by Visualization: A Small Data Solution." In: *Proceedings of the IEEE International Conference on Big Data*, 3513-3521.

[10] Peffers, K., Tuunanen, T., Rothenberger, M. and Chatterjee, S. 2007. "A Design Science Research Methodology for Information Systems Research." *Journal of Management Information Systems* (24:3), 45-77.

[11] Hevner, A. R., March, S. T., Park, J. and Ram, S. 2004. "Design science in Information Systems research." *MIS Quarterly* (28:1), 75-105.

[12] vom Brocke, J., and Maedche, A. 2019. "The DSR grid: six core dimensions for effectively planning and communicating design science research projects." *Electronic Markets* (29), 379-385.

[13] Howard, J., and Ruder, S. 2018. "Universal Language Model Fin-tuning for Text Classification." In: *Proceedings of the 56th Annual meeting of the Association for Computational Linguistics*, 328-339.

[14] Davidson, T., Warmsley, D., Macy, M. W. and Weber, I. 2017. "Automated Hate Speech Detection and the Problem of Offensive Language." In: *International Conference on Web and Social Media*, 1-4.

[15] Nienierza, A., Reinemann, C., Fawzi, N., Riesmeyer, C., and Neumann, K. 2019. "Too dark to see? Explaining adolescents' contact with online extremism and their ability to recognize it." *Information Communication & Society*, 1-18.

[16] Schmidt, A. and Wiegand, M. 2017. „A Survey on Hate Speech Detection using Natural Language Processing." In: *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 1-10.

[17] Kwok, I. and Wang, Y. 2013. "Locate the hate: Detecting tweets against blacks." In: *Proceedings of the Association for the Advancement of Artificial Intelligence*, 1621-1622.

[18] Raisi, E. and Huang, B. 2016. "Cyberbullying Identification Using Participant-Vocabulary Consistency." *Arxiv*: https://arxiv.org/abs/1606.08084 (28.09.2020).

[19] Udanor, C. and Anyanwu, C. C. 2019. "Combating the challenges of social media hate speech in a polarized society A Twitter ego lexalytics approach." *Data Technologies and Applications* (53:4), 501-527.

[20] Sahi, H., Kilic, Y., and Saglam, R. B. 2018. "Automated Detection of Hate Speech towards Woman on Twitter." In: *3rd International Conference on Computer Science and Engineering*, 533-536.

[21] Uban, A.-S. and Dinu, L. P. 2019. "On Transfer Learning for Detecting Abusive Language Online." *Advances in Computational Intelligence* (11506), 688-700.

[22] Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, N. A. 2019. "The Risk of Racial Bias in Hate Speech Detection." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1668-1678.

[23] Badjatiya, P., Gupta, S., Gupta, M. and Varma, V. 2017. "Deep Learning for Hate Speech Detection in Tweets." In: *Proceedings of the 26th International Conference on World Wide Web Companion*, 759-760.

[24] Pitsilis, G. K., Ramampiaro, H., and Langseth, H. 2018. "Effective hate-speech detection in Twitter data using recurrent neural networks." *Applied Intelligence* (48:12), 4730-4742.

[25] Jiang, L., Liu, S. and Chen, C. 2019. "Recent research advances on interactive machine learning." *Journal of Visualization* (22:2), 401-417.

[26] Lamy, J.-B., Sekar, B., Guezennec, G., Bouaud, J. and Seroussi, B. 2019. "Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach." *Artificial Intelligence in Medicine* (94), 42-53.

[27] Kouki, P., Schaffer, J., Pujara, J., O'Donovan, J. and Getoor, L. 2019. "Personalized explanations for hybrid recommender systems." In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 379-390.

[28] Meske, C., and Bunde, E. 2020. "Transparency and Trust in Human-AI-Interaction: The Role of Model-Agnostic Explanations in Computer Vision-Based Decision Support." In: *International Conference on Human-Computer Interaction, Artificial Intelligence in HCI* (12217), 54-69.

[29] Miller, T. 2018. "Explanation in artificial intelligence: Insights from the social sciences." *Artificial Intelligence* (267), 1-38.

[30] Cheng, H.-F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F. M. and Zhu, H. (2019). „Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders." In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Paper 559, 1-18.

[31] Meth, H., Mueller, B., and Maedche, A. 2015. "Designing a Requirement Mining System," *Journal of the Association for Information Systems* (16:9), 799-837.

[32] Kim, B., Park, J., and Suh, J. 2020. "Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information," *Decision Support Systems* (134), Article 113302.

[33] Davis, F. D. 1989. "Perceived Usefulness, Perceived Ease of use, and User Acceptance of Information Technology." *MIS Quarterly* (13:3), 319-340.

[34] Greven, D., Karahanna, E., and Straub, D. W. 2003. "Trust and TAM in Online Shopping: An Integrated Model." *MIS Quarterly* (27:1), 51-90.

[35] Carter, L., and Belanger, F. 2005. "The utilization of e-government services: citizen trust, innovation and acceptance factors." *Information Systems Journal* (15:1), 5-25.

[36] Venkatesh, V., Morris, M. G., Davis, G. B., and Davis, F. D. 2003. "User Acceptance of Information Technology: Toward a Unified View." *MIS Quarterly* (27:3), 425-478.

[37] Martens, D., and Provost, F. 2014. "Explaining Data-Driven Document Classifications." *MIS Quarterly* (38:1), 73-99.

[38] FastAI *Documentation Text 2020*. [online] https://docs.fast.ai/text.core (28.09.2020).

[39] FastAI *Documentation Interpretation 2020*. [online] https://fastai1.fast.ai/text.interpret.html (28.09.2020).

[40] Patton, M. Q. 2002. *Qualitative research and evaluation methods* (3rd ed.), Thousand Oaks, CA: Sage.

[41] Kaggle 2020. [online] https://www.kaggle.com/c/detecting-insults-in-social-commentary/data?select=train.csv (28.09.2020).

[42] Mann, H. B., and Whitney, D. R: 1947. "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other." *The Annals of Mathematical Statistics* (18:1), 50-60.

[43] Cohen, J. 1992. "A power primer." *Psychological Bulletin* (112:1), 155-159.

[44] Kühl, N., Lobana, J., and Meske, C. 2019. "Do you comply with AI? – Personalized explanations of learning algorithms and their impact on employees' compliance behavior." In: *Proceedings of the 40th International Conference on Information Systems (ICIS)*, 1-6.