The right kind of explanation: Validity in automated hate speech detection

Laugwitz, Laura

Erstveröffentlichung / Primary Publication
Sammelwerksbeitrag / collection article

Empfohlene Zitierung / Suggested Citation:
Laugwitz, L. (2023). The right kind of explanation: Validity in automated hate speech detection. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 383-402). Berlin https://doi.org/10.48541/dcr.v12.23

Digital
Communication
Research.de

Recommended citation: Laugwitz, L. (2023). The right kind of explanation: Validity in automated hate speech detection. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 383–402). Digital Communication Research.
https://doi.org/10.48541/dcr.v12.23

Abstract: To quickly identify hate speech online, communication research offers a useful tool in the form of automatic content analysis. However, the combined methods of standardized manual content analysis and supervised text classification demand different quality criteria. This chapter shows that a more substantial examination of validity is necessary since models often learn on spurious correlations or biases, and researchers run the risk of drawing wrong inferences. To investigate the overlap of theoretical concepts with technological operationalization, explainability methods are evaluated to explain what a model has learned. These methods proved to be of limited use in testing the validity of a model when the generated explanations aim at sense-making rather than faithfulness to the model. The chapter ends with recommendations for further interdisciplinary development of automatic content analysis.

DOI 10.48541/dcr.v12.23

Laura Laugwitz

The Right Kind of Explanation

Validity in automated hate speech detection

1 Automated content analysis: Mind the epistemological gap

As one of the core methods of communication research, content analysis has, for decades, provided the tool to describe, analyze, and compare content conveyed by the media (Krippendorff, 2019; Lacy et al., 2015). In conjuncture with growing amounts of digital communication, accessible tutorials, and evolving computing capacities (van Atteveldt & Peng, 2018), content analysis of text in particular is increasingly supported by computational methods (boyd & Crawford, 2012; Trilling & Jonkman, 2018) to analyze larger amounts of data faster and in a more standardized fashion. Hate speech detection is one of the fields in which automated content analysis stands to reason, since it is not simply the subsequent analysis of hateful communication that is of interest in research but also its quick identification (e.g., Davidson et al., 2017), moderation (e.g., Paasch-Colberg et al., 2020), and prevention (e.g., Schmitt et al., 2018). As other contributions in this collection show, the identification of hate speech is a challenge due to varying definitions (see Part 2), context (see Litvinenko), latent features (see Baider; Becker & Troschke), linguistic limits (see Baden), and bias (see Kim & Stoll). Computational methods add another level of challenges, as researchers not only

need to learn how to choose and implement these methods with regard to hate speech detection but also to understand and evaluate their results. While it is reasonable to collaborate with machine learning experts or computer scientists to implement computational methods well, there are fundamental differences in how computer science and social sciences approach the production of knowledge and, thus, how each evaluates the models they build.

Validity and reliability in computational methods are key issues for communication research (Scharkow, 2013; van Atteveldt & Peng, 2018), whereas machine learning experts are focused mainly on a sub-category of reliability, namely reproducibility, of their work (Henderson et al., 2018; Lipton & Steinhardt, 2019; Stodden, 2010). Since there is no clear equivalent for testing validity in text classification, communication researchers risk drawing the wrong inferences from automatically labeled data if they do not develop methods to ensure that such labels are based on "what words mean in the context of their use" (Krippendorff, 2019, S. 218).

These different focal points on criteria for research quality are rooted in different epistemologies: Communication research is largely conducted through the lens of critical rationalism in which a hypothesis's acceptability is tested *a priori* through logic and comparison with other theories as well as *a posteriori* through empirical tests (Chauviré, 2005). Machine learning, by contrast, mainly operates within a technocratic paradigm, dismissing the idea that *a priori* knowledge about the behavior of a program is possible and instead relying on gaining *a posteriori* knowledge through testing (Eden, 2007). It follows that the former understands quality as the production of reliable, valid, and intersubjectively comprehensible knowledge (Brosius et al., 2009), and the latter understands quality as the development of a satisfactory, reusable application (Stodden, 2010). Thus, these different approaches result not only in different quality criteria for research but also lead to various points of friction in interdisciplinary work.

In this chapter, I will examine the consequences of such epistemological differences, then focus on different quality criteria and how this may be alleviated when using supervised text classification for hate speech detection. In gathering various established and novel methods to establish validity in supervised text classification, I will show that current explainability approaches in development by computer scientists provide a useful starting point for deepening the understanding of a model's decision process. However, only a few of these approaches satisfy social science's imperative to examine a model's validity. This leaves

a void that ought to be filled via diligent collaboration between communication and computer scientists.

2 Validity of automated content analysis for hate speech detection

Automated content analysis can merge communication studies' manual content analysis and machine learning's supervised text classification; a model is trained to reproduce the labeling of concepts developed for and coded within a manually created corpus (Boumans & Trilling, 2016; Scharkow, 2013). In manual content analysis, human coders are given instructions in the shape of a codebook and are trained thoroughly for their classification task. The main strategies to ensure content validity—suggesting that the theoretical constructs are exhaustive and adequate (Krippendorff, 2019) – involve creating these codebooks based on a comprehensive literature review, training coders, improving the codebooks through their feedback, and checking for the use of catch-all or open categories after classification. These strategies, as well as intercoder reliability scores, reasonably create qualitative and quantifiable confidence that coders have integrated the knowledge derived from theory and research into their mental models. Additionally, the results may be compared with results from similar studies using the same or other methods (Krippendorff, 2019). Ensuring validity in content analysis specifically for hate speech is a challenge due to different reasons. First, hate speech is a complex construct, and separating it from adjacent concepts, such as incivility, toxicity, or offensive speech in theory, is still in progress; doing so in practice has only been attempted by a few researchers (e.g., Stoll et al., 2020). Second, some dimensions of hate speech show a manifestation in specific words (Davidson et al., 2017), whereas others are more latent (Nielsen, 2002) or contradictory (van Aken et al., 2018), such as generalizations or irony. Even in thorough manual content analysis, intercoder reliability can vary immensely (Poletto et al., 2020; Ross et al., 2017), which poses a challenge in reliably measuring the difficult concept of hate speech. Third, and most important to this paper, translating any attributes or dimensions of hate speech into technologically traceable features is a challenge that is rarely recognized. In a literature review, Fortuna and Nunes (2018) showed different strategies currently in use for hate speech detection. One strand leverages existing generic methods from natural language processing,

such as topic classification, sentiment analysis, named entity recognition, and deep learning, to identify hate speech. A second, much smaller strand identifies specific linguistic features, such as othering language, objectivity-subjectivity, declarations of superiority of an in-group, and particular stereotypes. The latter strand approximates operationalizations from manual content analysis closer than the former; however, the former is much more common.

In supervised text classification, no fixed rules or instructions are given to the machine learning model. Rather, it derives classification rules inductively from previously coded material (Scharkow, 2013). While multiple strategies to measure reliability for automated content analysis and to increase reproducibility have recently been suggested (Krippendorff, 2019; Mitchell et al., 2019; Pineau, 2020; Scharkow, 2013), testing the validity of supervised text classification has yet to be expanded. Most simply, some researchers rely on manual coding as the gold standard, and rule that validity is established if the automated results are similar enough to the manual classification (Lee et al., 2020). The assumption here is that quality can be assured through the creation of a valid and reliable manually coded dataset, since an algorithm will then simply reproduce these classifications. However, computer science itself is currently raising doubts about whether models actually learn content-related features at all, or instead are trained on spurious correlations and artifacts in the dataset (Lapuschkin et al., 2019), raising the issue of validity without explicitly naming it. Other scholars, then, suggest applying their model to a second dataset to test the validity of inferences (Pilny et al., 2019). Beyond these, some communication scholars have also attempted to examine content validity. To show whether a model has learned to identify concepts previously derived from theory, the weights for individual features can be examined (Stoll et al., 2020).

Examining examples from sentiment analysis, which aims to automatically identify mood in text and is occasionally used as a proxy to identify hate speech, Liu and Avci (2019) claimed that models may assign a negative mood to text as soon as it contains identity terms, such as "Jew" or "Black." Similarly, a hate speech classifier may learn to classify any sentence containing the term "Islam" as toxic (Waseem & Hovy, 2016). While this type of error may be acceptable for an application designed to help companies in the private sector identify potentially problematic discussions, such a lack of validity is fatal in communication research, as inferences based on invalid classifications will result in flawed inferences. More

recently, it has been shown that the reasoning of well-performing models for hate speech classification tasks does not necessarily align with the reasoning that coders for manual classification have provided (Mathew et al., 2020), adding to the limited trust in the validity of these models. Relying on a manually coded dataset with high validity is not sufficient to examine whether the theoretical constructs informed the model's classification decision. Applying the same model to another dataset may hint at its capacity to generalize. However, trying to understand how a model made a decision, for example, via feature weights, appears to be the most fruitful and necessary strategy to date in examining the content validity of supervised text classification. In the following chapter, I present strategies currently explored in machine learning that intend to show the reasoning behind a model's classification decisions.

3 Explaining supervised text classification

A recurring theme in machine learning is the question of why a model made a specific decision (local explanation) or how it works in general (global explanation), which is currently mostly found under the umbrella term "explainability." Efforts to increase the explainability of automated results (Samek & Müller, 2019) have gained more relevance in machine learning since the wider use of automated decision-making in business, banking, and the public sector (Mittelstadt et al., 2016). Following the introduction of the general data protection regulation in 2018, users even have a right to be provided with an explanation for an automated decision (§ 14 2 g GDPR). Two main strategies are currently pursued in explainability research (Vilone & Longo, 2020): Model-agnostic methods are meant to provide generic solutions, creating explanations that do not require access to the model itself. In using solely the input and output of a model, they can be considered reverse engineering. Model-specific methods examine particular aspects of a model, such as revealing word relations in different layers of a neural network. They require access to the model and are dependent on its functionality. Beyond explainability strategies, interpretable methods use *ab initio* algorithms that can be understood by humans (Rudin, 2019), such as linear classifiers, Bayesian classifiers, or support vector machines. Based on the systemic literature reviews by Guidotti et al. (2019) and Vilone and Longo (2020), five model-agnostic methods,

five methods specific to neural networks, and three interpretable models can be identified. These solutions aim to explain models for supervised text classification. They will be summarized to show the general breadth of solutions and give communication scholars an idea of the state of research in machine learning.

3.1 Model-agnostic methods

Developing an additional simplified model based on the input and output of the original model is the general strategy for model-agnostic methods. More specifically, the partition aware local model (PALM) consists of two kinds of models: a meta-model partitions the training dataset, and then individual sub-models approximate local patterns of the partitions (Krishnan & Wu, 2017). The meta-model is a decision tree that can be used to compare single misclassifications with the relevant training data (if available), and thus offers human users an intuition for the relation between training input and classification.

With a similar focus on proximity, local interpretable model-agnostic explanations (LIME) trains a linear classifier for a single classification by approximating further cases from the immediate neighborhood of the example (Ribeiro et al., 2016). Thus, complex models are broken down into single, locally interpretable models. Anchors is an extension of LIME that leverages if-then-rules that anchor an explanation locally to a point at which a change of values to other features of that instance does not lead to a different classification (Ribeiro et al., 2018). Similar instances almost always share the same classification, thus providing examples of how features are relevant to a classification.

Simple rules are also used in a model explanation system (MES), which assumes that explanations are simple logical statements and uses a Monte Carlo algorithm to find the best explanation for a single classification via a scoring system (Turner, 2016). Although it is intended to work for text as well, Turner (2016) has only provided examples of computer vision and credit scoring (tabular data). Whether meaningful automatically generated explanations for text can be achieved with MES is an open question.

The last model-agnostic approach is based on game theory, using the idea that each feature represents a player, and each classification represents the profit. Shapley values indicate how this profit must be fairly distributed between features.

For this purpose, every possible feature combination and its effect on classification are compared. Thus, if all classifications are considered, a statement can actually be made about the global relevance of each feature as well as the local relevance of the feature in a single classification. Practically, these values are impossible to compute due to the large set of features and their possible combinations; however, simplified versions of this approach are used as intuition in various explainability methods (e.g., Chen et al., 2019).

3.2 Model-specific methods

Methods that inspect or retrace the partial mechanics of a model are called model-specific. Each approach depends on the model itself; thus, there is no general solution that can be transferred to a different type of model. However, all relevant solutions from the literature have been developed for some versions of a neural network. For example, a rationale generator is trained in parallel with a neural network, learning to select a subset of the input sequence as an explanation for classification (Lei et al., 2016). The rationale then contains a reduced set of meaningful words, which should result in the same classification as the original input sequence if given to the classifier.

This strategy to identify salient input features is also applied in DeepLIFT (Deep Learning Important FeaTures). Here, the principle of layer-wise relevance propagation traces a single classification of a neural network backwards through said network. DeepLIFT then analyzes the difference in the activation values of single neurons for that input-output pair compared to a reference input-output pair (Shrikumar et al., 2017) and indicates which features (e.g., individual words) were most in favor of a classification or its opposite. This approach has the potential to provide counterfactual explanations if the reference input-output pair is intentionally chosen. However, for the application to text, it seems customary to simply choose zero values (Lertvittayakumjorn & Toni, 2019; Sundararajan et al., 2017).

Integrated gradients explain a neural network by analyzing its sensitivity to differences in input (Sundararajan et al., 2017). They create a sequence of gradients leading from the baseline to the input and compute their average, thus measuring the correlation between the uncertainty in the output of a classifier and its input.

With the recent and extremely rapid success of transformer models, the visualization of attention and hidden states has gained popularity for explanatory purposes (see van Aken et al., 2020). Transformer models are a special form of deep neural networks that first learn basic language structures before being trained in specific tasks (see Minaee et al., 2020). Van Aken et al. (2020) proposed considering the feature embeddings of individual layers to visualize the learning process of a transformer model for individual classifications. For this purpose, the vectors with which the individual inputs are technically represented are reduced in dimension after each layer with principal component analysis and mapped on a two-dimensional surface. Across the layers, the proximity of different words becomes apparent (see van Aken et al., 2020), exposing the inner structures of the neural network. It is then left to the researcher to decide on the layer in which a structure is clear enough to be used as an explanation for a classification.

Another approach that employs human interaction is concept-based explanations. In testing with concept activation vectors, people are asked to choose examples and counterexamples for certain concepts (e.g., stripes in pictures) after training a model (see Kim et al., 2018). An additional linear classifier will then be trained to discriminate between activations for each set of examples, generating global explanations for the influence of concepts on classes. However, these concepts depend on what the researcher chooses in terms of content, and it is unclear whether they cover all concepts relevant to the model. Further development of this method adds a step of unsupervised learning, automatically extracting concepts that are sufficiently predictive for classification (see Yeh et al., 2019).

Communication research might benefit from experimenting with a combination of the analysis of different layers in neural networks, where different linguistic concepts are also recognized in different layers (van Aken et al., 2019), and the concept-based analysis of Yeh et al, (2019). Different linguistic layers could be responsible for different concepts. To detect hate speech, for example, it would be possible that manifest insults could be identified at early levels, while latent concepts such as dehumanization would only be identified in later layers. Empirical testing of this assumption could be extremely valuable.

3.3 Interpretable models

Instead of using complex neural networks or proprietary systems for text classification, researchers can also make use of models that are interpretable by default. *Ante hoc* methods are intended to keep models explainable from the beginning and are therefore also called white box or transparent models (Rosenfeld & Richardson, 2019). They include decision trees and decision rules or k-nearest neighbors, as well as discrete choice probabilities, such as logistic regression or Naive Bayes (Molnar, 2020). However, these models are typically heavily domain-specific and, if the data are not well structured and clear, they can require an enormous amount of computational effort (Rudin, 2019). In a recent and commendable study on incivility and impoliteness in text, Stoll et al. (2020) used Naive Bayes for global explanations that give weight to individual words. The weight of a feature is calculated by its probability of appearing in a given classification. The global explanation thus outputs a list of features that are relevant for a class. Risch et al. (2020) also used this strategy in comparison with explainable models and found that the Naive Bayes model showed the lowest performance. Unlike Stoll et al. (2020), who used various preprocessing methods, however, Risch et al. (2020) did not show whether further steps were taken to improve the data, which would be necessary for interpretable models according to Rudin (2019).

Instead of these weighted features, interpretable models can also be used to create prototypes. Bien and Tibshirani (2011) developed the prototype selection approach in which, instead of focusing on reducing the number of features to a manageable amount, the data itself is bundled by selecting a prototype from the neighborhood of each instance that has the same label. The authors aimed to have as few prototypes as possible and ensure that no instance had a prototype with a different label. Prototype selection requires inference from the researchers, since it does not show which specific features of the prototype were relevant for its selection.

This lack of causal explanation is addressed in the Bayesian case model (BCM) by first clustering the data and then generating prototypes as well as feature weights for these clusters (Kim et al., 2015), thus providing global explanations. However, if too many clusters are formed, both the computational time and the number of explanations are too high, which in turn no longer allows for interpretability. Guidotti et al. (2019) pointed out several improvements for

BCM: humans can interact with the model to improve the prototypes (Kim et al., 2015), or instances in which the classification does not fit well into the model can provide counterexamples (Kim et al., 2016). Subsequently, the overall strategies will be investigated with respect to how they may be leveraged to examine content validity in automated content analysis.

4 Using explanations to examine validity

Model-agnostic methods, such as LIME, Anchors, and MES, aim to explain individual classifications, whereas PALM identifies partial patterns, and Shapley values have the potential to trace the weight of features across the entire model, where it is not for the computational limits. However, given the fact that these solutions do not actually inspect or retrace the mechanics of the initial model, their usefulness regarding validity is limited. Whether the model has identified the same concepts technologically that have previously been defined for manual analysis remains unknown. While they may be useful for identifying discrepancies in classifications, they do not make use of but instead approximate the initial model (Rudin, 2019), thus creating an additional layer of uncertainty instead of alleviating it.

Model-specific methods provide a tool to partially inspect the model's validity; the fact that they create insight into the internal mechanics of a model suggests that they may be used to examine whether the theoretical concepts have been transferred to the technological operationalization. However, it is not sensible to infer how the model works as a whole from explanations for individual classifications (Mittelstadt et al., 2019), which would be an inductive fallacy. In fact, these methods also do not contribute to giving users a more comprehensive understanding of model behavior (Lertvittayakumjorn & Toni, 2019), may also give misleading explanations (Rudin, 2019), and should thus be used only with proper contextualization and caution.

Interpretable models provide "their own explanations, which are faithful to what the model actually computes" (Rudin, 2019, p. 1) and are thus especially interesting to researchers already competent in statistics. Their simplicity can offer insight into how the model has transferred theoretical operationalizations into technical features so that their explanations can actually act as indicators

for content validity. Note that some researchers have also critiqued Rudin's assumption that models can be inherently interpretable yet do not provide any data to substantiate their critique (Jacovi & Goldberg, 2020). Nevertheless, well-performing interpretable models also require time and effort, so the costs and benefits of the research project in question must be weighed. Due to the data structure and the inherent ambiguity of text, interpretable models for text classification currently do not receive much attention. Even Kim et al. (2015) who clearly advocated for interpretable models in 2015 and 2016, have moved on to developing model-specific methods by 2018. Although interpretable methods show the most promise for validity checks, interpretable methods in general are underrepresented in explainability research (Vilone & Longo, 2020). The aim of machine learning to develop generalized solutions (Fortuna et al., 2020) that can be applied to many problems does not necessarily overlap with that of the social sciences to consider problems in context.

In summary, existing strategies developed to explain the overall functioning or individual decisions of a text classification model offer limited help in examining a model's validity. Model-agnostic explanations may be used to gain some intuition when models are complex or proprietary but can be considered insufficient for a validity check. Similarly, model-specific explanations do not satisfy this use case either. While they access the model itself to provide explanations, they rarely explain it in its entirety, and local explanations should not be used to infer the functionality of the model as a whole. Interpretable models show the most promise for our use case. If trained carefully and with sufficient domain knowledge, they perform well and provide explanations that are appropriate for testing content validity. Nonetheless, since both explainability methods for text (as opposed to images or tabular data) and interpretable models are rare in the current body of research, an opportunity to collaborate beyond a simple splitting of tasks in automated content analysis emerges for communication scholars and computer scientists.

5 A call to develop methods to establish validity of automated content analysis

A critical rationalist perspective on automated content analysis substantiates the need to explain how a model works to examine its validity. Failure to provide adequate explanations creates opacity within the scientific process, preventing researchers from ensuring that their model has learned on informative features that sufficiently consider context instead of learning on artifacts or spurious correlations. Hence, only validated models should lead to inferences about the data's context. Whereas standardized content analysis has established strategies to strengthen and examine content validity, no such strategies have been established for supervised text classification. In creating a codebook informed by theory and empirical research, comprehensive coder training, feedback loops, and discussions in training sessions, as well as reliability scores, researchers gain confidence about the validity of their data and subsequent inferences. An automated model, however, is not involved in gaining a shared understanding of what is supposed to be coded; instead, it merely aims to mimic. The strategies to strengthen and examine validity thus look different for supervised text classification. As argued above, validity can be strengthened by using interpretable methods and examining whether the features that a model has learned preserve the context of meaning. Explainability methods partially enable such an examination; however, their current applications are not specific enough for scientific use.

The development of the explainability methods discussed above has mostly been motivated by the need to establish trust, identify bias or errors, and prevent damage by a faulty system. The quality of these methods, in line with a technocratic paradigm, tends to be evaluated *a posteriori*, for example, via a user's reaction, feedback, or subsequent performance (cf. Gilpin et al., 2019)—framing quality as the *plausibility* of the explanation. However, to verify the validity of a model, explanations cannot be measured with regard to their effects on users (see Herman, 2017). What matters in examining validity is not an explanation's effect on a user but that it explains a model concisely. Jacovi and Goldberg (2020) identified a difference between the plausibility and *faithfulness* of an explanation, which describes "how accurately it reflects the true reasoning process of a model" (p. 4198). In the context of using explanations as a tool to examine a model's validity, the

distinction between plausibility and faithfulness is especially valuable: here, faithfulness is not a measurement of explanation quality but a prerequisite.

To fully leverage the advantages of supervised text classification in automated content analysis, profound collaboration and innovation are needed from communication and computer scholars. Through the example of hate speech detection, this contribution has posited a need for quality control and has shown that adequate methods to establish the validity of a model are rare. While a concept such as hate speech presents a rather extreme example due to its complexity, it nonetheless illustrates the intricacies of two disciplines joining one method rather well. Since research on this specific topic is currently growing in both fields, the outlook of building better-performing and explainable models may motivate closer collaboration despite the additional effort. Scholars should collaborate on theoretical and empirical work to resolve epistemological differences, align research processes, develop joint measures for quality, and collect requirements for models that show what they actually compute in a way that is seminal to automated content analysis. This could, for example, result in the development of standardized strategies and criteria for validity in automated content analysis, specific interpretable models, and faithful explanations. As much as the research community and practitioners in the realm of hate speech will benefit from this work, we shall not underestimate how it may contribute to methodological improvements in computational communication studies in general.

Laura Laugwitz is a PhD candidate at the Institute for Journalism and Communication Studies at Universität Hamburg, Germany. https://orcid.org/0000-0001-8527-2504

References

Bien, J., & Tibshirani, R. (2011). Prototype selection for interpretable classification. *Annals of Applied Statistics*, *5*(4), 2403–2424. https://doi.org/10.1214/11-AOAS495

Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism*, *4*(1), 8–23. https://doi.org/10.1080/21670811.2015.1096598

boyd, danah, & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information Communication and Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Brosius, H.-B., Koschel, F., & Haas, A. (2009). *Methoden der empirischen Kommunikationsforschung [Methods of empirical research]* (5th ed.). VS Verlag für Sozialwissenschaften.

Chauviré, C. (2005). Peirce, Popper, abduction, and the idea of a logic of discovery. *Semiotica*, *4*(153), 209–221. https://doi.org/10.1515/semi.2005.2005.153-1-4.209

Chen, J., Song, L., Wainwright, M. J., & Jordan, M. I. (2019, May 6-9). *L-Shapley and C-Shapley: Efficient model interpretation for structured data* [Conference presentation]. 7th International Conference on Learning Representations, New Orleans, LA, Unites States.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the 11th International Conference on Web and Social Media*, 512–515. https://ojs.aaai.org/index.php/ICWSM/article/view/14955

Eden, A. H. (2007). Three paradigms of computer science. *Minds and Machines*, *17*(2), 135–167. https://doi.org/10.1007/s11023-007-9060-8

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, *51*(4), 1–30. https://doi.org/10.1145/3232676

Fortuna, P., Soler-Company, J., & Wanner, L. (2020). Toxic, hateful, offensive or abusive? What are we really classifying? An empirical analysis of hate speech datasets. *Proceedings of the 12th International Conference on Language Resources and Evaluation*, 6786–6794. http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.838.pdf

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2019). Explaining explanations: An overview of interpretability of machine learning. *Proceedings of the 5th International Conference on Data Science and Advanced Analytics*, 80–89. https://doi.org/10.1109/DSAA.2018.00018

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 1–42. https://doi.org/10.1145/3236009

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., & Meger, D. (2018). Deep reinforcement learning that matters. *32nd AAAI Conference on Artificial Intelligence*, 3207–3214. https://doi.org/10.48550/arXiv.1709.06560

Herman, B. (2017, December 7). *The promise and peril of human evaluation for model interpretability* [Poster presentation abstract]. Conference on Neural Information Processing Systems, Long Beach, CA, United States. https://arxiv.org/abs/1711.07414

Jacovi, A., & Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4198–4205). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.386

Kim, B., Glassman, E., Johnson, B., & Shah, J. (2015). *iBCM: Interactive Bayesian case model empowering humans via intuitive interaction.*(Report No. MIT-CSAIL-TR-2015-010). DSpace@MIT Computer Science and Artificial Intelligence Lab. https://dspace.mit.edu/handle/1721.1/96315

Kim, B., Khanna, R., & Koyejo, O. O. (2016). Examples are not enough, learn to criticize! Criticism for Interpretability. In D. D. Lee, M. Sugiyama, U. V Luxburg, I. Guyon, & R. Garnett (Eds.), *Proceedings of the 30th International Conference on Neural Information Processing Systems* (pp. 2280–2288). Curran Associates, Inc. https://dl.acm.org/doi/10.5555/3157096.3157352

Kim, B., Rudin, C., & Shah, J. (2015). The Bayesian Case Model: A generative approach for case-based reasoning and prototype classification. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Proceedings of the 27th International Conference on Neural Information Processing Systems* (vol. 2, pp. 1952–1960). MIT Press. https://dl.acm.org/doi/10.5555/2969033.2969045

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018): Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In J. Dy, & A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning: Vol. 80* (pp. 2668-2677). PMLR. http://proceedings.mlr.press/v80/kim18d.html

Krippendorff, K. (2019). *Content analysis: An introduction to its methodology* (4th ed.). Sage.

Krishnan, S., & Wu, E. (2017). PALM: Machine learning explanations for iterative debugging. *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics* 3(1), 1–6, ACM Press. https://doi.org/10.1145/3077257.3077271

Lacy, S., Watson, B. R., Riffe, D., & Lovejoy, J. (2015). Issues and best practices in content analysis. *Journalism & Mass Communication Quarterly*, *92*(4), 791–811. https://doi.org/10.1177/1077699015607338

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K.-R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, *10*(1), 1–8. https://doi.org/10.1038/s41467-019-08987-4

Lee, L. W., Dabirian, A., McCarthy, I. P., & Kietzmann, J. (2020). Making sense of text: Artificial intelligence-enabled content analysis. *European Journal of Marketing*, *54*(3), 615–644. https://doi.org/10.1108/EJM-02-2019-0219

Lei, T., Barzilay, R., & Jaakkola, T. (2016). Rationalizing neural predictions. In J. Su, K. Duh, & X. Carreras (Eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 107–117). Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1011

Lertvittayakumjorn, P., & Toni, F. (2019). Human-grounded evaluations of explanation methods for text classification. In K.Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 5194–5204). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1523

Lipton, Z. C., & Steinhardt, J. (2019). Troubling trends in machine learning scholarship. *Queue*, *17*(1), 1–15. https://doi.org/10.1145/3317287.3328534

Liu, F., & Avci, B. (2019): Incorporating priors with feature attribution on text classification. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 6274–6283). Association for Computational Linguistics. http://doi.org/10.18653/v1/P19-1631

Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., & Mukherjee, A. (2020). *HateXplain: A benchmark dataset for explainable hate speech detection*. ArXiv. http://arxiv.org/abs/2012.10289

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., & Terzopoulos, D. (2020). Image segmentation using deep learning: A survey. https://arxiv.org/abs/2001.05566v5

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). ACM Press. https://doi.org/10.1145/3287560.3287596

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data and Society*, *3*(2), 1–21. https://doi.org/10.1177/2053951716679679

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 279–288). ACM Press. https://doi.org/10.1145/3287560.3287574

Molnar, C. (2020). *Interpretable machine learning: A guide for making black box models explainable*. Lulu.

Nielsen, L. B. (2002). Subtle, pervasive, harmful: racist and sexist remarks in public as hate speech. *Journal of Social Issues*, *58*(2), 265–280. https://spssi.onlinelibrary.wiley.com/doi/pdf/10.1111/1540-4560.00260

Paasch-Colberg, S., Strippel, C., Laugwitz, L., Emmer, M., & Trebbe, J. (2020). Moderationsfaktoren: Ein Ansatz zur Analyse von Selektionsentscheidungen im Community Management [Moderation factors: an approach to analyzing selection decisions in community management]. In V. Gehrau, A. Waldherr, & A. Scholl (Eds.), *Integration durch Kommunikation: Jahrbuch der Publizistik- und Kommunikationswissenschaft 2019* (pp. 109–119). DGPuK. https://doi.org/10.21241/ssoar.67858

Pilny, A., McAninch, K., Slone, A., & Moore, K. (2019). Using supervised machine learning in automated content analysis: An example using relational uncertainty. *Communication Methods and Measures*, *13*(4), 287–304. https://doi.org/10.1080/19312458.2019.1650166

Pineau, J. (2020). The machine learning reproducibility checklist (Version 2.0). McGill. www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist-v2.0.pdf

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2020). Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation, 55*(1), 477–523. https://doi.org/10.1007/s10579-020-09502-8

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-agnostic interpretability of machine learning. In B. Kim, D. M. Malioutov, & K.R. Varshney (Eds.), *Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning* (pp. 91–95). arXiv. https://arxiv.org/abs/1606.05386

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1), 1527–1535. https://ojs.aaai.org/index.php/AAAI/article/view/11491

Risch, J., Ruff, R., & Krestel, R. (2020). Explaining offensive language detection. *Journal for Language Technology and Computational Linguistics*, *34*(1), 29-47. https://doi.org/10.21248/jlcl.34.2020.223

Rosenfeld, A., & Richardson, A. (2019). Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems*, *33*(6), 673–705. https://doi.org/10.1007/s10458-019-09408-y

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2017). Measuring the reliability of hate speech annotations: The case of the European refugee crisis. In M. Beißwenger, M. Wojatzki, & T. Zesch (Eds.), *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication* (vol. 17, pp. 6–9). Bochumer Linguistische Arbeitsberichte. https://doi.org/10.48550/arXiv.1701.08118

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Samek, W., & Müller, K. R. (2019). Towards explainable artificial intelligence. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11700 LNCS*, 5–22. https://doi.org/10.1007/978-3-030-28954-6_1

Scharkow, M. (2013). Thematic content analysis using supervised machine learning: An empirical evaluation using German online news. *Quality & Quantity*, *47*(2), 761–773. https://doi.org/10.1007/s11135-011-9545-7

Schmitt, J. B., Rieger, D., Rutkowski, O., & Ernst, J. (2018). Counter-messages as prevention or promotion of extremism?! The potential role of YouTube. *Journal of Communication*, *68*(4), 780–808. https://doi.org/10.1093/joc/jqy029

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. In D. Precup, & Y. W. The (Eds.), *Proceedings of the 34th International Conference on Machine Learning:* (vol. 70, pp. 4844–4866). PMLR. http://proceedings.mlr.press/v70/shrikumar17a.html

Stodden, V. (2010). The scientific method in practice: Reproducibility in the computational sciences. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1550193

Stoll, A., Ziegele, M., & Quiring, O. (2020). Detecting impoliteness and incivility in online discussions: Classification approaches for German user comments. *Computational Communication Research*, *2*(1), 109–134. https://doi.org/10.5117/CCR2020.1.005.KATH

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks In D. Precup, & Y. W. The (Eds.), *Proceedings of the 34th International Conference on Machine Learning* (vol. 70, pp. 5109–5118). PMLR. http://proceedings.mlr.press/v70/sundararajan17a.html

Trilling, D., & Jonkman, J. G. F. (2018). Scaling up content analysis. *Communication Methods and Measures*, *12*(2–3), 158–174. https://doi.org/10.1080/19312458.2018.1447655

Turner, R. (2016). *A model explanation system: Latest updates and extensions*. arXiv. https://arxiv.org/abs/1606.09517

van Aken, B., Risch, J., Krestel, R., & Löser, A. (2018). Challenges for toxic comment classification: An in-depth error analysis. In D. Fišer, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, & J. Wernimont (Eds.), *Proceedings of the 2nd Workshop on Abusive Language Online* (pp. 33–42). Association for Computational Linguistics. https://doi.org/10.18653/v1/w18-5105

van Aken, B., Winter, B., Löser, A., & Gers, F. A. (2020). VisBERT: Hidden-state visualizations for transformers. In A. El Fallah Seghrouchni, G. Sukthankar, T. Liu, & M. van Steen (Eds.), *Companion Proceedings of the Web Conference 2020* (pp. 207–211). Association for Computing Machinery. https://doi.org/10.1145/3366424.3383542

van Atteveldt, W., & Peng, T. Q. (2018). When communication meets computation: opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures*, *12*(2–3), 81–92. https://doi.org/10.1080/19312458.2018.1458084

Vilone, G., & Longo, L. (2020). *Explainable artificial intelligence: A systematic review*. arXiv. https://arxiv.org/abs/2006.00093

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In J. Andreas, E. Choi, & A. Lazaridou (Eds.), *Proceedings of the NAACL Student Research Workshop* (pp. 88–93). Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-2013

Yeh, C.-K., Hsieh, C.-Y., Suggala, A. S., Inouye, D. I., & Ravikumar, P. (2019). On the (in)fidelity and sensitivity for explanations. https://arxiv.org/abs/1901.09392v4