Challenges in Applying Explainability Methods to Improve the Fairness of NLP Models

Esma Balkır, Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C. Fraser
National Research Council Canada
Ottawa, Canada
{Esma.Balkir,Svetlana.Kiritchenko,Isar.Nejadgholi,Kathleen.Fraser}@nrc-cnrc.gc.ca

Abstract

Motivations for methods in explainable artificial intelligence (XAI) often include detecting, quantifying and mitigating bias, and contributing to making machine learning models fairer. However, exactly how an XAI method can help in combating biases is often left unspecified. In this paper, we briefly review trends in explainability and fairness in NLP research, identify the current practices in which explainability methods are applied to detect and mitigate bias, and investigate the barriers preventing XAI methods from being used more widely in tackling fairness issues.

1 Introduction

Trends in Natural Language Processing (NLP) mirror those in Machine Learning (ML): breakthroughs in deep neural network architectures, pretraining and fine-tuning methods, and a steady increase in the number of parameters led to impressive performance improvements for a wide variety of NLP tasks. However, these successes have been shadowed by the repeated discoveries that a high accuracy on the held-out test set does not always mean that the model is performing satisfactorily on other important criteria such as *fairness*, *robustness* and *safety*. These discoveries that models are adversarially manipulable (Zhang et al., 2020a), show biases against underprivileged groups (Chang et al., 2019), and leak sensitive user information (Carlini et al., 2021) inspired a plethora of declarations on Responsible/Ethical AI (Morley et al., 2021). Two of the common principles espoused in these documents are *fairness* and *transparency*.

Failures in fairness of models is often attributed, among other things, to the lack of transparency of modern AI models. The implicit argument is that, if biased predictions are due to faulty reasoning learned from biased data, then we need transparency in order to detect and understand this faulty reasoning. Hence, one approach to solving these problems is to develop methods that can peek inside the black-box, provide insights into the internal workings of the model, and identify whether the model is right for the right reasons.

As a result, ensuring the fairness of AI systems is frequently cited as one of the main motivations behind XAI research (Doshi-Velez and Kim, 2017; Das and Rad, 2020; Wallace et al., 2020). However, it is not always clear how these methods can be applied in order to achieve fairer, less biased models. In this paper, we briefly summarize some XAI methods that are common in NLP research, the conceptualization, sources and metrics for unintended biases in NLP models, and some works that apply XAI methods to identify or mitigate these biases. Our review of the literature in this intersection reveals that applications of XAI methods to fairness and bias issues in NLP are surprisingly few, concentrated on a limited number of tasks, and often applied only to a few examples in order to illustrate the particular bias being studied. Based on our findings, we discuss some barriers to more widespread and effective application of XAI methods for debiasing NLP models, and some research directions to bridge the gap between these two areas.

2 Explainable Natural Language Processing

With the success and widespread adaptation of black-box models for machine learning tasks, increasing research effort has been devoted to developing methods that might give human-comprehensible explanations for the behaviour of these models, helping developers and end-users to understand the reasoning behind the decisions of the model. Broadly speaking, explainability methods try to pinpoint the causes of a single prediction, a set of predictions, or all predictions of a model by identifying parts of the input, the model or the training data that have the most influence on the

Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022), pages 80 - 92

	Local	Global
Self-explaining	Gradients (Simonyan et al., 2014) Integrated Gradients (Sundararajan et al., 2017) SmoothGrad (Smilkov et al., 2017) DeepLIFT (Shrikumar et al., 2017) Attention (Xu et al., 2015; Choi et al., 2016) Representer Point Selection (Yeh et al., 2018)	Counterfactual LM (Feder et al., 2021b)
Post-hoc	LIME (Ribeiro et al., 2016) SHAP (Lundberg and Lee, 2017) Counterfactuals (Wu et al., 2021; Ross et al., 2021) Extractive rationales (DeYoung et al., 2020) Influence Functions (Koh and Liang, 2017; Han et al., 2020) Anchors (Ribeiro et al., 2018a)	TCAV (Kim et al., 2018; Nejadgholi et al., 2022) SEAR (Ribeiro et al., 2018b)

Table 1: Explainability methods from Sec. 2 categorized as local vs. global and self-explaining vs. post-hoc.

model outcome.

The line dividing XAI methods, and methods that are developed more generally for understanding, analysis and evaluation of NLP methods beyond the standard accuracy metrics is not always clear cut. Many popular approaches such as *probes* (Hewitt and Liang, 2019; Voita and Titov, 2020), *contrast sets* (Gardner et al., 2020) and *checklists* (Ribeiro et al., 2020) share many of their core motivations with XAI methods. Here, we present some of the most prominent works in XAI, and refer the reader to the survey by Danilevsky et al. (2020) for a more extensive overview of the field. We consider a method as an XAI method if the authors have framed it as such in the original presentation, and do not include others in our analysis.

A common categorization of explainability methods is whether they provide *local* or *global* explanations, and whether they are *self-explaining* or *post-hoc* (Guidotti et al., 2018; Adadi and Berrada, 2018). The first distinction captures whether the explanations are given for individual instances (*local*) or explain the model behaviour on any input (*global*). Due to the complex nature of the data and the tasks common in NLP, the bulk of the XAI methods developed for or applicable to NLP models are local rather than global (Danilevsky et al., 2020). The second distinction is related to how the explanations are generated. In *self-explaining* methods, the process of generating explanations is integrated into, or at least reliant on the internal structure of the model or the process of computing the model outcome. Because of this, self-explaining methods are often specific to the type of the model. On the other hand, *post-hoc* or *model-agnostic* methods only assume access to the input-output behaviour of the model, and construct explanations based on how changes to the different components of the prediction pipeline affect the outputs. Below, we outline some of the representative explainability methods used in NLP and categorize them along the two dimensions in Table 1.

Feature attribution methods, also referred to as *feature importance* or *saliency maps*, aim to determine the relative importance of each token in an input text for a given model prediction. The underlying assumption in each of these methods is that the more important a token is for a prediction, the more the output should change when this token is removed or changed. One way to estimate this is through the gradients of the output with respect to each input token as done by Simonyan et al. (2014). Other methods have been developed to address some of the issues with the original approach such as local consistency (Sundararajan et al., 2017; Smilkov et al., 2017; Selvaraju et al., 2017; Shrikumar et al., 2017).

Rather than estimating the effect of perturbations through gradients, an alternative approach is to perturb the input text directly and observe its effects on the model outcome. Two of the most common methods in this class are *LIME* (Ribeiro et al., 2016) and *SHAP* (Lundberg and Lee, 2017). LIME generates perturbations by dropping subsets of tokens from the input text, and then fitting a linear classifier on these local perturbations. SHAP is inspired by Shapely values from cooperative game theory, and calculates feature importance as the fair division of a "payoff" from a game where the features cooperate to obtain the given model outcome. *AllenNLP Interpret* toolkit (Wallace et al., 2019) provides an implementation for both types of feature attribution methods, gradient based and input perturbation based, for six core NLP tasks, including text classification, masked language modeling, named entity recognition, and others.

A third way to obtain feature attribution maps in architectures that use an attention mechanism

(Bahdanau et al., 2015) is to look at the relative attention scores for each token (Xu et al., 2015; Choi et al., 2016). Whether this approach provides valid explanations has been subject to heated debate (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019), however as Galassi et al. (2020) notes, the debate has mostly been centered around the use of attention scores as local explanations. There has also been some works that use attention scores for providing global explanations based on the syntactic structures that the model attends to (Clark et al., 2019).

Extractive rationales (DeYoung et al., 2020) are snippets of the input text that trigger the original prediction. They are similar in spirit to feature attribution methods, however in rationales the attribution is usually binary rather than a real-valued score, and continuous subsets of the text are chosen rather than each token being treated individually. Rationales can also be obtained from humans as explanations of human annotations rather than the model decisions, and used as an additional signal to guide the model.

Counterfactual explanations are new instances that are obtained by applying minimal changes to an input instance in order to change the model output. Counterfactuals are inspired by notions in causality, and aim to answer the question: "What would need to change for the outcome to be different?" Two examples of counterfactual explanations in NLP are *Polyjuice* (Wu et al., 2021) and *MiCE* (Ross et al., 2021). Polyjuice is model agnostic, and consists of a generative model trained on existing, human generated counterfactual data sets. It also allows finer control over the types of counterfactuals by allowing the user to choose which parts of the input to perturb, and how to perturb them with control codes such as “replace” or “negation”. MiCE uses model gradients to iteratively choose and mask the important tokens, and a generative model to change the chosen tokens so that the end prediction is flipped.

There are also methods that try to pinpoint which examples in the training data have the most influence on the prediction. The most common approach for this is *Influence Functions* (Koh and Liang, 2017; Han et al., 2020), where the goal is to efficiently estimate how much removing an example from the data set and retraining the model would change the prediction on a particular input. An alternative is *Representer Point Selection* (Yeh et al., 2018), which applies to a more limited set of architectures, and aims to express the logits of an input as a weighted sum of all the training data points.

Some explainability methods are designed to provide global explanations using higher level, semantic concepts. Feder et al. (2021b) use counterfactual language models to provide causal explanations based on high-level concepts. Their method contrasts the original model representations with alternative pre-trained representations that are adversarially trained not to capture the chosen high-level concept, so that the total causal effect of the concept on the classification decisions can be estimated. Nejadgholi et al. (2022) adapt Testing Concept Activation Vector (TCAV) method of Kim et al. (2018), originally developed for computer vision, to explain the generalization abilities of a hate speech classifier. In their approach, the concepts are defined through a small set of human chosen examples, and the method quantifies how strongly the concept is associated with a given label.

Finally, some methods produce explanations in the form of rules. One method in this category is Anchors (Ribeiro et al., 2018a), where the model searches for a set of tokens in a particular input text that predicts the given outcome with high precision. Although Anchors is a local explainability method in that it gives explanations on individual input instances, the generated explanations are globally applicable. SEAR (Ribeiro et al., 2018b), a global explainability method, finds universal replacement rules that, if applied to an input, adversarially change the prediction while keeping the semantics of the input the same.

3 Fairness and Bias in NLP Models

Unintended biases in NLP is a complex and multifaceted issue that spans various undesirable model behaviours that cause allocational and representational harms to certain demographic groups (Blodgett et al., 2020). When the demographic group is already marginalized and underprivileged in society, biases in NLP models can further contribute to the marginalization and the unfair allocation of resources. Examples include performance disparities between standard and African American English (Blodgett and O’Connor, 2017), stereotypical associations between gendered pronouns and occupations in coreference resolution (Rudinger et al., 2018) and machine translation (Stanovsky et al.,

2019), and false positives in hate speech detection on innocuous tweets mentioning demographic attributes (Röttger et al., 2021). In this section, we review some of the most popular methods and metrics to identify such biases. For a more comprehensive coverage, see recent surveys by Mehrabi et al. (2021) and Caton and Haas (2020).

Most works in ML fairness literature assume that biases in machine learning models originate from misrepresentations in training datasets and merely reflect the societal biases. However, as Hooker (2021) explains, design choices can amplify the societal biases, and automated data processing can lead to systematic un-precedented harms. Shah et al. (2020) identify five sources for bias in NLP models. *Selection bias* and *label bias* are biases that originate in the training data. The former refers to biases that are created when choosing which data points to annotate, and includes underrepresentation of some demographic groups as well as misrepresentation due to spurious correlations. The latter refers to biases introduced due to the annotation process, such as when annotators are less familiar with or biased against text generated by certain groups, causing more annotation errors for some groups than others. *Model bias* are biases that are due to model structure, and are responsible for the over-amplification of discrepancies that are observed in training data. *Semantic bias* refers to biases introduced from the pre-trained representations, and include representational harms such as stereotypical associations. Finally, *bias in research design* covers the larger issues of uneven allocation of research efforts across different groups, dialects, languages and geographic areas.

Research in fair ML has developed a number of metrics to quantify the biases in an ML model. These metrics are usually classified as *group fairness metrics* and *individual fairness metrics* (Castelnovo et al., 2022; Czarnowska et al., 2021). Group fairness metrics focus on quantifying the performance disparity between different demographic groups. Some examples are *demographic parity*, which measures the difference in the positive prediction rates across groups, *predictive parity*, which measures the difference in precision across groups, and *equality of odds*, which measures the differences between false positive and false negative rates across groups. Individual fairness metrics are based on the idea that the model should behave the same for similar examples regardless of the value of a protected attribute. A refinement to this approach is *counterfactual fairness*, where the criteria for fairness is that the model decision remains the same for a given individual in a counterfactual world where that individual belonged to a different demographic group. In NLP, this notion often appears as *counterfactual token fairness* (Garg et al., 2019), and is operationalized through test suites that include variations of the same text where some tokens associated with certain social groups are replaced with others, and the bias of the model is measured by the performance disparity between the pairs (Kiritchenko and Mohammad, 2018; Prabhakaran et al., 2019).

Both group fairness metrics and individual fairness metrics are instances of *outcome fairness*: whether a model is fair is determined solely on the outcomes with respect to various groups, regardless of how the algorithm produced those observed outcomes.[1] There is a complementary notion called *procedural fairness* that is often considered in organizational settings (Blader and Tyler, 2003), which aims to capture whether the *processes* that were followed to obtain the outcome are fair. In ML, this translates to whether the model's internal reasoning process is fair to different groups or individuals (Grgić-Hlača et al., 2018; Morse et al., 2021). For example, outcome fairness for a resume sorting system might be implemented as ensuring that the model has the same acceptance rates or the same precision and recall for groups defined by race, gender, or other demographic attributes. A procedural fairness approach, on the other hand, might aim to ensure that the decision making process of the system only relies on skill-related features, and not features that are strongly associated with demographic attributes, such as names and pronouns. The distinction between procedural and outcome fairness relates to different kinds of discrimination outlined in anti-discrimination laws, namely *disparate treatment* and *disparate impact* (Barocas and Selbst, 2016).

Fairness metrics have originally been developed for applications where the social group membership is known, for example in healthcare related tasks. An issue with applying these to NLP tasks is that either the demographic information is not available and needs to be estimated, or some auxiliary signal, such as the mention of a target group

[1] *Outcome fairness* is also referred to as *distributive fairness* in this literature.

or the gender of the pronoun, needs to be used. However, inferring people's social attributes from their data raises important ethical concerns in terms of privacy violations, lack of meaningful consent, and intersectional invisibility (Mohammad, 2022). Since determining whether the text is *about* a certain identity group is easier than whether it is *produced by* a certain identity group, there are more works investigating the former than the latter. An exception to this is the studies on disparate performance of models on certain dialects such as African American English (AAE) (Sap et al., 2019; Blodgett and O'Connor, 2017). This is possible due to the existence of a dialect identification tool for AAE, which was trained by pairing geo-located tweets with US census data on race (Blodgett et al., 2016).

One source of bias that the NLP community has devoted significant research effort to is word embeddings and pre-trained language models (Bolukbasi et al., 2016; Zhao et al., 2019), which Shah et al. (2020) characterizes as *semantic bias*. Although it is not framed as such, this can be seen as a particular global explanation for biases that the models demonstrate in downstream tasks. However, the effectiveness of these methods has recently been questioned by Goldfarb-Tarrant et al. (2021) who found that there is no correlation between intrinsic bias metrics obtained by embedding association tests, and extrinsic bias metrics on downstream tasks.

4 Applications of XAI in Fair NLP

To determine the uses of explainability methods in fair NLP, we search the ACL Anthology for papers that cite the explainability methods listed in Section 2, and that include keywords, "fair", "fairness", or "bias". We further exclude the papers that focus on other types of biases such as inductive bias, or bias terms in the description of the architecture. Our results show that although there are a number of papers that mention unintended or societal biases as wider motivations to contextualize the work (e.g., by Zylberajch et al. (2021)), only a handful of them apply explainability methods to uncover or investigate biases. All of the works we identify in this category use feature attribution methods, and except that of Aksenov et al. (2021), employ them for demonstration purposes on a few examples. Although our methodology excludes works that are published in venues other than ACL conferences and workshops, we believe that it gives a good indication of the status of XAI in fairness and bias research in NLP.

Mosca et al. (2021) use SHAP to demonstrate that adding user features to a hate speech detection model reduces biases that are due to spurious correlations in text, but introduces other biases based on user information. Wich et al. (2020) also apply SHAP to two example inputs in order to illustrate the political bias of a hate speech model. Aksenov et al. (2021) aggregate attention scores from BERT into global explanations in order to identify which words are most indicative of political bias.

Some works beyond the papers that our search methodology uncovered on the intersection of fairness for NLP and XAI are that of Kennedy et al. (2020), which uses Sampling and Occlusion algorithm of Jin et al. (2019) to detect bias toward identity terms in hate speech classifiers, and that of Mathew et al. (2021), which shows that using human rationales as an additional signal in training hate speech detection models reduces the bias of the model towards target communities. Prabhakaran et al. (2019) target individual fairness, and develop a framework to evaluate model bias against particular named entities with a perturbation based analysis. Although they do not frame their model as such, the automatically generated perturbations can be categorized as counterfactuals. Balkır et al. (2022) suggest the use of two metrics—necessity and sufficiency—as feature attribution scores, and apply their method to uncover different kinds of bias against protected group tokens in hate speech and abusive language detection models.

As summarized in Table 2, almost all these works focus exclusively on hate speech detection, and use local feature attribution methods. The range of bias types is also quite limited. This demonstrates the very narrow context in which explainability has been linked to fairness in NLP.

There are also some works beyond NLP that use XAI to improve fairness of ML models. Zhang and Bareinboim (2018), Parafita and Vitria (2021) and Grabowicz et al. (2022) leverage methods from causal inference to both model the causes of the given prediction and provide explanations, and to ensure that protected attributes are not influencing the model decisions through unacceptable causal chains. The disadvantage of these models is that they require an explicit model of the causal relations between features, which is a difficult task

Study	Overall Objective of the Study	Application	Bias Type	Explainability Method
Mosca et al. (2021)	Detecting classifier sensitivity towards identity terms vs. user tweet history	hate speech detection	social group bias	SHAP
Wich et al. (2020)	Measuring the effect of bias on classification performance	hate speech detection	political orientation	SHAP
Aksenov et al. (2021)	Classification of political bias in news	hate speech detection	political orientation	aggregated attention scores
Kennedy et al. (2020)	Reducing the classifier's oversensitivity to identity terms	hate speech detection	social group bias	feature importance (SOC)
Mathew et al. (2021)	Improving group fairness	hate speech detection	social group bias	LIME, attention
Prabhakaran et al. (2019)	Detecting biases related to named entities	sentiment analysis, toxicity detection	sensitivity to named entities	perturbation analysis
Balkır et al. (2022)	Detecting over- and under-sensitivity to identity tokens	hate speech and abusive language detection	social group bias	necessity and sufficiency

Table 2: Summary of the studies that apply explainability techniques to uncover unintended biases in NLP systems.

for textual data (Feder et al., 2021a). Pradhan et al. (2022) also suggest a causality inspired method that identifies subsets of data responsible for particular biases of the model. Begley et al. (2020) extend Shapely values to attribute the overall unfairness of an algorithm to individual input features. The main limitation of all these methods is that they are currently only applicable to low dimensional tabular data. How to extend these methods to explain the unfairness of NLP models remains an open research problem.

As abstract frameworks for connecting XAI to fair ML, P et al. (2021) outline potential synergies between the two research areas. Alikhademi et al. (2021) enumerate different sources of bias, and discuss how XAI methods can help identify and mitigate these.

5 XAI for Fair NLP through Causality and Robustness

The framework of *causality* (Pearl, 2009) is invoked both in fairness and explainability literature. The promise of causality is that it goes beyond correlations, and characterizes the causes behind observations. This is relevant to conceptualizing fairness since, as Loftus et al. (2018) argue, there are situations that are intuitively different from a fairness point of view, but that purely observational criteria cannot distinguish.

Causality tries to capture the notion of causes of an outcome in terms of hypothetical interventions: if something is a true cause of a given outcome, then intervening on this variable will change the outcome. This notion of intervention is useful for both detecting biases and for choosing mitigation strategies. Causal interventions are also the fundamental notion behind counterfactual examples in XAI. It is easier for humans to identify the cause of a prediction if they are shown minimally different instances that result in opposite predictions. Hence, causal explanations can serve as proofs of bias or other undesirable correlations to developers and to end-users.

Going beyond correlations in data and capturing causal relations is also an effective way to increase *robustness* and *generalization* in machine learning models. As Kaushik et al. (2020) argue, causal correlations are invariant to differing data distributions, while non-causal correlations are much more context and dataset specific. Hence, models that can differentiate between the two and rely solely on casual correlations while ignoring the non-causal ones will perform well beyond the strict i.i.d. setting.

Non-causal, surface level correlations are often referred to as *spurious correlations*, and a common use case of XAI methods for developers is to facilitate the identification of such patterns. A common motivating argument in XAI methods for debugging NLP models (Lertvittayakumjorn and Toni, 2021; Zylberajch et al., 2021), as well as counterfactual data augmentation methods (Kaushik et al., 2020; Balashankar et al., 2021; Yang et al., 2021), is that unintended biases are due to the model picking up such spurious associations, and XAI methods which can be used to improve the robustness of a model against these spurious patterns will also improve the fairness of a model as a side effect. There is indeed evidence that methods for robustness also reduce unintended bias in NLP models (Adragna et al., 2020; Pruksachatkun et al., 2021).

However, these methods are limited in that they can address unintended biases only insofar as the biases are present and identifiable as token-level spurious correlations.

6 Challenges and Future Directions

As we saw in Sec. 4 and 5, only a few studies to date have attempted to apply explainability techniques in order to uncover biases in NLP systems, to a limited extent. In this section, we discuss some possible reasons for a seeming lack of progress in this area and outline promising directions for future research.

Local explainability methods rely on the user to identify examples that might reveal bias. One issue in preventing wider adoption of XAI methods in fair NLP stems from the local nature of most explanation methods applicable to NLP models. An important step in identifying fairness problems within a model is identifying the data points where these issues might manifest. Since local explainability methods give explanations on particular data points, it is left to the user how to pick the instances to examine. This necessitates the user to first decide what biases to search for before employing XAI methods, limiting their usefulness for identifying unknown biases.

Local explanations are not easily generalizable. Even if an issue can be identified with a local XAI method, it is difficult to know to what extent the insight can be generalized. This is an issue because it is often essential to know what subsets of the input are affected by the identified biased behaviour in order to apply effective mitigation strategies. Some methods such as Anchors mitigate this problem by specifying the set of examples an explanation applies to. Other approaches use abstractions such as high-level concepts (Feder et al., 2021b; Nejadgholi et al., 2022) to provide more generalizable insights. Principled methods to aggregate local explanations into more global and actionable insights are needed to make local explainability methods better suited to identifying and mitigating unintended biases in NLP models. Also, future NLP research could explore global explainability methods that have been used to uncover unknown biases (Tan et al., 2018).

Not all undesirable biases are surface-level or non-causal. In the motivation for XAI methods, there is strong emphasis on identifying token-level correlations caused by sampling bias or label bias. Although methods that target these patterns are shown to also improve the fairness of models, not all sources of bias fit well into this characterization (Hooker, 2021), and hence might be difficult to detect with XAI methods that provide token-level explanations. For example, Bagdasaryan et al. (2019) show that the cost of differential privacy methods in decreasing the accuracy of deep learning NLP models, is much higher for underrepresented subgroups. A rigorous study of a model's structure and training process is required to discover such bias sources.

Another issue that is common in works that approach fairness through robustness is the characterization of unintended biases as non-causal associations in data (Kaushik et al., 2020; Adragna et al., 2020). In fact, it can be argued that many of the undesirable correlations observed in data are causal in nature, and will likely hold in a wide variety of different data distributions. For example, correlations between different genders and occupations—which arguably is the source of the occupational gender stereotypes picked up by NLP models (Rudinger et al., 2018)—are not due to unrepresentative samples or random correlations in the data, but rather underlying systemic biases in the distribution of occupations in the real world. To ensure a fair system, researchers must make a *normative decision* (Blodgett et al., 2020) that they do not want to reproduce this particular correlation in their model. This suggests that there may be inherent limitations to the ability of XAI methods to improve fairness of NLP methods through improving model robustness and generalization.

Some biases can be difficult for humans to recognize. Even for biases that could be characterized in terms of surface-level correlations, XAI methods rely on humans to recognize what an undesirable correlation is, but biased models are often biased in subtle ways. For example, if the dialect bias in a hate speech detection system is mostly mediated by false positives on the uses of reclaimed slurs, this might seem like a good justification to a user who is unfamiliar with this phenomenon (Sap et al., 2019). More studies with human subjects are needed to investigate whether humans can recognise unintended biases that cause fairness issues through explainability methods as well as they can recognise simpler data biases.

Explainability methods are susceptible to fairwashing. An issue that has repeatedly been raised with respect to XAI methods is the potential for "fairwashing" biased models. This refers to techniques that adversarially manipulate explanations in order to obscure the model's reliance on protected attributes. Fairwashing has been shown possible in rule lists (Aïvodji et al., 2019), and both gradient based and perturbation based feature attribution methods (Dimanov et al., 2020; Anders et al., 2020). This relates to the wider issue of the faithfulness of an explainability method: if there is no guarantee that the explanations reflect the actual inner workings of the model, the explanations are of little use. One solution to this problem would be to extend certifiable robustness (Cohen et al., 2019; Ma et al., 2021) beyond the model itself, and develop certifiably faithful explainability methods with proofs that a particular way of testing for bias cannot be adversarially manipulated. Another approach to mitigate this issue is to provide the levels of uncertainty in the explanations, giving the end-user more information on whether to trust the generated explanation (Zhang et al., 2019), or other ways to calibrate user trust to the quality of the provided explanations (Zhang et al., 2020b). However, the effectiveness of these methods depends substantially on whether the model's predicted probabilities are well-calibrated to the true outcome probabilities. Certain machine learning models do not meet this criterion. Specifically, the commonly used deep learning models have been shown to be over-confident in their predictions (Guo et al., 2017). Calibration of uncertainties is a necessary prerequisite, should they be used to calibrate user trust, as over-confident predictions can be themselves a source of mistrust.

Fair AI is focused on outcome fairness, but XAI is motivated by procedural fairness. Finally, it appears that there is a larger conceptual gap between the notions of fairness that the ethical AI community has developed, and the notion of fairness implicitly assumed in motivations for XAI methods. Namely, almost all the fairness metrics developed in Fair ML literature aim to formalize outcome fairness in that they are process-agnostic, and quantify the fairness of a model on its observed outcomes only. The type of fairness that motivates XAI, on the other hand, is closer to the concept of procedural fairness: XAI aims to elucidate the internal reasoning of a model, and make it transparent whether there are any parts of the decision process that could be deemed unfair.

We observe that due to the lack of better definitions of procedural fairness, the most common way XAI methods are applied to fairness issues is to check whether the model uses features that are explicitly associated with protected attributes (e.g., gendered pronouns). This practice promotes a similar ideal with "fairness through unawareness" in that it aims to place the veil of ignorance about the protected attributes not at the level of the data fed into the model, but into the model itself. In other words, the best one could do with these techniques seem to be to develop "colourblind" models which, even if they receive explicit information about protected attributes in their input, ignore this information when making their decisions. Although it is simple and intuitive, we suspect that such an approach has similar issues with the much criticized "fairness through unawareness" approach (Kusner et al., 2017; Morse et al., 2021). More clearly specified notions of procedural fairness, as well as precise quantitative metrics similar to those that have been developed for outcome fairness, are needed in order to guide the development of XAI methods that can make ML models fairer.

7 Conclusion

Publications in explainable NLP often cite *fairness* as a motivation for the work, but the exact relationship between the two concepts is typically left unspecified. Most current XAI methods provide explanations on a local level through post-hoc processing, leaving open questions about how to automatically identify fairness issues in individual explanations, and how to generalize from local explanations to infer systematic model bias. Although the two fields of explainability and fairness feel intuitively linked, a review of the literature revealed a surprisingly small amount of work at the intersection. We have discussed some of the conceptual underpinnings shared by both these fields as well as practical challenges to uniting them, and proposed areas for future research.

References

Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160.

Robert Adragna, Elliot Creager, David Madras, and Richard Zemel. 2020. Fairness and robustness in invariant learning: A case study in toxicity classification. In *Proceedings of the NeurIPS 2020 Workshop on Algorithmic Fairness through the Lens of Causality and Interpretability*.

Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. 2019. Fairwashing: the risk of rationalization. In *Proceedings of the International Conference on Machine Learning*, pages 161–170.

Dmitrii Aksenov, Peter Bourgonje, Karolina Zaczynska, Malte Ostendorff, Julian Moreno Schneider, and Georg Rehm. 2021. Fine-grained classification of political bias in German news: A data set and initial experiments. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 121–131.

Kiana Alikhademi, Brianna Richardson, Emma Drobina, and Juan E Gilbert. 2021. Can explainable AI explain unfairness? A framework for evaluating explainable AI. *arXiv preprint arXiv:2106.07483*.

Christopher Anders, Plamen Pasliev, Ann-Kathrin Dombrowski, Klaus-Robert Müller, and Pan Kessel. 2020. Fairwashing explanations with off-manifold detergent. In *Proceedings of the International Conference on Machine Learning*, pages 314–323.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. 2019. Differential privacy has disparate impact on model accuracy. *Advances in Neural Information Processing Systems*, 32.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*.

Ananth Balashankar, Xuezhi Wang, Ben Packer, Nithum Thain, Ed Chi, and Alex Beutel. 2021. Can we improve model robustness through secondary attribute counterfactuals? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4701–4712.

Esma Balkır, Isar Nejadgholi, Kathleen C Fraser, and Svetlana Kiritchenko. 2022. Necessity and sufficiency for explaining text classifiers: A case study in hate speech detection. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Seattle, WA, USA.

Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *CALIFORNIA LAW REVIEW*, 104:671.

Tom Begley, Tobias Schwedes, Christopher Frye, and Ilya Feige. 2020. Explainability for fair machine learning. *arXiv preprint arXiv:2010.07389*.

Steven L Blader and Tom R Tyler. 2003. What constitutes fairness in work settings? a four-component model of procedural justice. *Human Resource Management Review*, 13(1):107–126.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130.

Su Lin Blodgett and Brendan O'Connor. 2017. Racial disparity in natural language processing: A case study of social media African-American English. In *Proceedings of the 2017 Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in Neural Information Processing Systems*, 29.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *Proceedings of the 30th USENIX Security Symposium*, pages 2633–2650.

Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. 2022. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12.

Simon Caton and Christian Haas. 2020. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*.

Kai-Wei Chang, Vinod Prabhakaran, and Vicente Ordonez. 2019. Bias and fairness in natural language processing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts*.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in Neural Information Processing Systems*, 29.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does BERT look at? An analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. 2019. Certified adversarial robustness via randomized smoothing. In *Proceedings of the International Conference on Machine Learning*, pages 1310–1320.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. Association for Computational Linguistics.

Arun Das and Paul Rad. 2020. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2020. Eraser: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458.

Botty Dimanov, Umang Bhatt, Mateja Jamnik, and Adrian Weller. 2020. You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods. In *Proceedings of the AAAI Workshop on Artificial Intelligence Safety (SafeAI)*.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. 2021a. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *arXiv preprint arXiv:2109.00725*.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2021b. Causalm: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386.

Andrea Galassi, Marco Lippi, and Paolo Torroni. 2020. Attention in natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32(10):4291–4308.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940.

Przemyslaw Grabowicz, Nicholas Perello, and Aarshee Mishra. 2022. Marrying fairness and explainability in supervised learning. *arXiv preprint arXiv:2204.02947*.

Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2018. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 51(5):1–42.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning*, pages 1321–1330.

Xiaochuang Han, Byron C Wallace, and Yulia Tsvetkov. 2020. Explaining black box predictions and unveiling data artifacts through influence functions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5553–5563.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743.

Sara Hooker. 2021. Moving beyond "algorithmic bias is a data problem". *Patterns*, 2(4):100241.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556.

Xisen Jin, Zhongyu Wei, Junyi Du, Xiangyang Xue, and Xiang Ren. 2019. Towards hierarchical importance attribution: Explaining compositional semantics for neural sequence models. In *Proceedings of the International Conference on Learning Representations*.

Divyansh Kaushik, Amrith Setlur, Eduard H Hovy, and Zachary Chase Lipton. 2020. Explaining the efficacy of counterfactually augmented data. In *Proceedings of the International Conference on Learning Representations*.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *Proceedings of the International Conference on Machine Learning*, pages 2668–2677.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the International Conference on Machine Learning*, pages 1885–1894.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems*, 30.

Piyawat Lertvittayakumjorn and Francesca Toni. 2021. Explanation-based human debugging of NLP models: A survey. *Transactions of the Association for Computational Linguistics*, 9:1508–1528.

Joshua R Loftus, Chris Russell, Matt J Kusner, and Ricardo Silva. 2018. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Pingchuan Ma, Shuai Wang, and Jin Liu. 2021. Metamorphic testing and certified mitigation of fairness violations in NLP models. In *Proceedings of the Twenty-Ninth International Joint Conferences on Artificial Intelligence*, pages 458–465.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

Saif M Mohammad. 2022. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*.

Jessica Morley, Luciano Floridi, Libby Kinsey, and Anat Elhalal. 2021. From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. In *Ethics, Governance, and Policies in Artificial Intelligence*, pages 153–183. Springer.

Lily Morse, Mike Horia M Teodorescu, Yazeed Awwad, and Gerald C Kane. 2021. Do the ends justify the means? variation in the distributive and procedural fairness of machine learning algorithms. *Journal of Business Ethics*, pages 1–13.

Edoardo Mosca, Maximilian Wich, and Georg Groh. 2021. Understanding and interpreting the impact of user context in hate speech detection. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 91–102.

Isar Nejadgholi, Kathleen C Fraser, and Svetlana Kiritchenko. 2022. Improving generalizability in implicitly abusive language detection with concept activation vectors. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Dublin, Ireland.

Deepak P, Sanil V, and Joemon M. Jose. 2021. On fairness and interpretability. In *Proceedings of the IJCAI Workshop on AI for Social Good*.

Alvaro Parafita and Jordi Vitria. 2021. Deep causal graphs for causal inference, black-box explainability and fairness. In *Artificial Intelligence Research and Development: Proceedings of the 23rd International Conference of the Catalan Association for Artificial Intelligence*, volume 339, page 415. IOS Press.

Judea Pearl. 2009. *Causality*. Cambridge University Press.

Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation sensitivity analysis to detect unintended model biases. In *Proceedings*

of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5740–5745.

Romila Pradhan, Jiongli Zhu, Boris Glavic, and Babak Salimi. 2022. Interpretable data-based explanations for fairness debugging. In *Proceedings of the 2022 ACM SIGMOD International Conference on Management of Data*.

Yada Pruksachatkun, Satyapriya Krishna, Jwala Dhamala, Rahul Gupta, and Kai-Wei Chang. 2021. Does robustness improve fairness? approaching fairness with word substitution robustness methods for text classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3320–3331.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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*, pages 1135–1144.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018a. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018b. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912.

Alexis Ross, Ana Marasović, and Matthew E Peters. 2021. Explaining NLP models via minimal contrastive editing (MiCE). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3840–3852.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. Hatecheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626.

Deven Santosh Shah, H Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the International Conference on Machine Learning*, pages 3145–3153.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Proceedings of the 2nd International Conference on Learning Representations, Workshop Track*.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.

Gabriel Stanovsky, Noah A Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the International Conference on Machine Learning*, pages 3319–3328.

Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. 2018. Distill-and-compare: Auditing black-box models using transparent model distillation. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 303–310.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196.

Eric Wallace, Matt Gardner, and Sameer Singh. 2020. Interpreting predictions of NLP models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 20–23.

Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. 2019. AllenNLP Interpret: A framework for explaining predictions of NLP models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 7–12.

Maximilian Wich, Jan Bauer, and Georg Groh. 2020. Impact of politically biased data on hate speech classification. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 54–64.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6707–6723, Online. Association for Computational Linguistics.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning*, pages 2048–2057.

Linyi Yang, Jiazheng Li, Pádraig Cunningham, Yue Zhang, Barry Smyth, and Ruihai Dong. 2021. Exploring the efficacy of automatically generated counterfactuals for sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 306–316.

Chih-Kuan Yeh, Joon Kim, Ian En-Hsu Yen, and Pradeep K Ravikumar. 2018. Representer point selection for explaining deep neural networks. *Advances in Neural Information Processing Systems*, 31.

Junzhe Zhang and Elias Bareinboim. 2018. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. 2020a. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3):1–41.

Yujia Zhang, Kuangyan Song, Yiming Sun, Sarah Tan, and Madeleine Udell. 2019. "why should you trust my explanation?" understanding uncertainty in LIME explanations. In *Proceedings of the ICML Workshop AI for Social Good*.

Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020b. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 295–305.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1.

Hugo Zylberajch, Piyawat Lertvittayakumjorn, and Francesca Toni. 2021. Hildif: Interactive debugging of NLI models using influence functions. In *Proceedings of the First Workshop on Interactive Learning for Natural Language Processing*, pages 1–6.