2021 IEEE Third International Conference on Cognitive Machine Intelligence (CogMI) | 978-1-6654-1621-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/COGMI52975.2021.00014

FairShades: Fairness Auditing via Explainability in Abusive Language Detection Systems

Marta Marchiori Manerba
University of Pisa
Pisa, Italy
marta.marchiori@phd.unipi.it

Riccardo Guidotti
University of Pisa
Pisa, Italy
riccardo.guidotti@unipi.it

***Abstract*—At every stage of a supervised learning process, harmful biases can arise and be inadvertently introduced, ultimately leading to marginalization, discrimination, and abuse towards minorities. This phenomenon becomes particularly impactful in the sensitive real-world context of abusive language detection systems, where non-discrimination is difficult to assess. In addition, given the opaqueness of their internal behavior, the dynamics leading a model to a certain decision are often not clear nor accountable, and significant problems of trust could emerge. A robust value-oriented evaluation of models' fairness is therefore necessary. In this paper, we present FairShades, a model-agnostic approach for auditing the outcomes of abusive language detection systems. Combining explainability and fairness evaluation, FairShades can identify unintended biases and sensitive categories towards which models are most discriminative. This objective is pursued through the auditing of meaningful counterfactuals generated within CheckList framework. We conduct several experiments on BERT-based models to demonstrate our proposal's novelty and effectiveness for unmasking biases.**

I. Introduction

Current abusive language detection models implemented with Artificial Intelligence (AI) technologies have been shown to inherit and perpetuate unintended bias towards specific demographic groups and protected attributes such as sexual orientation or religion [1], [2]. These skews pose a serious risk and limitation to online discourse, leading to the marginalization of minority voices. The authors of [3], for example, show that annotators tend to label as abusive more frequently the texts in Afro-American English than other messages, which could lead to the training of a system reproducing the same kind of prejudice. Given the sensitive context in which abusive language detection systems are deployed, a robust value-oriented evaluation of models' fairness is necessary. Furthermore, it is important to distinguish between different types of hatred, depending on the target group addressed. For example, misogynistic expressions show different linguistic peculiarities than racist ones. It is therefore crucial to conduct specialized and targeted analyses, addressing phenomena of abusive language towards different minorities, so that systems can be tuned to the complex and nuanced scenario of online speech. Indeed, fairness assessments are ultimately motivated by fundamental issues such as investigate whether there are social groups treated differently and in what linguistic contexts, whether conditions of privilege are confirmed, coupled with worsening for the disadvantaged, and the resulting drop in models' performance compared to other demographic group scores, as demonstrated in [4], [5], [6]. From this scenario, it becomes a priority the need to explore and disaggregate overall metrics. However, this process is complicated by the partial effectiveness of proposed methods that only work with certain definitions of fairness and bias, as well as by the limited availability of recognized benchmark datasets [7], and their focus on specific bias types, e.g., towards gender.

Besides fairness, another crucial aspect to consider lies in the opaqueness of models' internal behavior. In fact, if the dynamics leading to a certain automatic decision are not clear nor accountable, significant problems of trust for the reliability of outputs could emerge, especially in sensitive real-world contexts. Assessing that the knowledge autonomously learned conforms to human values and inspecting non-discrimination of decisions constitutes both a real challenge and a risk. Indeed, the objective of eXplainable Artificial Intelligence (XAI) is to propose strategies and methods to render AI systems and automatic decisions more intelligible to humans. In recent years, working towards transparency and interpretability of black box models has become a priority: multiple approaches and methods have been proposed [8], [9], [10].

To address these issues, in this paper we present *FairShades*, a model-agnostic approach for auditing abusive language detection classifiers that relies on explainability techniques. FairShades is a post-hoc explanation approach that performs a sub-global analysis composed of individual local interpretations unveiling the presence of biases. It combines explainability and fairness evaluation within a proactive pipeline to identify wrong correlations, unintended biases, and sensitive categories toward which the black box model under assessment is most discriminative. In particular, FairShades audits abusive language detection classifiers through synthetic but meaningful texts generated through the CheckList framework [11] and obtained by perturbing only sensitive identities present in the texts under analysis. Then, it learns a decision tree regressor used to identify unfair counterfactual terms, estimate their importance, and evaluate the degree of fairness of the abusive language detection system assessed.

The rest of the paper is organized as follows. In Section II we report a critical discussion of the literature and a review of the state of the art. In Section III we formalize the problem treated, and we briefly present necessary background

978-1-6654-1621-4/21/$31.00 ©2021 IEEE
DOI 10.1109/CogMI52975.2021.00014

knowledge. Section V presents our definition of *unfairness* and the proposed methodology. In Section VI we report the experiments demonstrating the novelty and effectiveness of our proposal on unmasking biases. Our evaluation shows that, although state of the art classifiers such as BERT-based models achieve high accuracy levels on a variety of natural language processing tasks, they demonstrate severe shortages on samples involving implicit stereotypes and protected attributes such as nationality or sexual orientation. Finally, Section VII discusses the limitations of our approach and indicates future research directions.

II. Related Works

Automatic abusive language detection is a task emerged with the widespread use of social media [12]. Often online discourse can assume abusive and offensive connotations, especially towards sensitive minorities and young people. The exposition to these violent opinions can trigger polarization, isolation, depression and other psychological trauma [12]. Therefore, online platforms have started to assume the role of examining and removing hateful posts. Since the large amount of data flowing across social media, hatred is typically flagged through automatic methods along with human monitoring. A number of approaches have been proposed to perform both coarse-grained, i.e. binary, and fine-grained classification: pre-trained embeddings such as contextualized Transformers [13] and ELMo [14] embeddings are among the most popular techniques [15].

Given the complexity of the internal dynamics of current deep learning models, it is crucial to understand and be able to account for the reasons of certain automatic decisions [8]. This need is further strengthened by their application in sensitive scenarios like health, legal practices, recruitment and automatic online content moderation [10]. Two main scenarios exist: the first is related to a transparency *by-design*, using methods such decision trees and decision rules; the second concerns the black box explanation. This latter branch can approach the problem with the aim of explaining the prediction for a specific instance, providing a local explanation [16], [17], [18], or aiming at explaining the whole internal logic of the model [19], [20]. A recent approach, related to XAI [21], suggests a local model-agnostic technique for sentiment classifiers leveraging on exemplar and counter-exemplar phrases generated through variational autoencoders. An interesting procedure, at the intersection of XAI and fairness, is *Model Cards* [22], a framework for Natural Language Processing (NLP) that establishes and encourages the responsible practice of “transparent model reporting”, to describe intended application scenarios, avoiding unintended harms[1]. Indeed, having access to the data on which the model was trained, and being explicitly aware of its intended and designed use, can inform both outcome comprehension and assessment, including facilitating bias detection [1].

Concerning the different definitions of fairness, they have been collected and organized in [1], [2] with the awareness that a single definition is not sufficient to address the multi-faceted problem in its entirety. Of particular interest is the concept of unconscious bias, which lies in the risk of a model generalizing a stereotyped conception of reality from unrepresentative and skewed data. In [25] is analyzed and criticized the formalization of *bias* within NLP systems, revealing inconsistency, lack of normativity, and common rationale in several papers. In [12] is conducted an in-depth discussion on NLP works dealing with ethical issues and challenges in automatic abusive language detection. Among others, a perspective analyzed is the principle of fairness and non-discrimination throughout every stage of supervised machine learning processes. In [26] is conducted a fine-grained fairness analysis of abusive language detection systems carried out with CheckList framework [11]. The authors, through the approach proposed, show the relevance played by training data in the treatment of certain phenomena and topics, such as body image stereotypes or feminism-related statements. Several metrics, generic tools and libraries such as [27], [28] have been proposed to investigate fairness in AI models, nevertheless, the solutions often remain fragmented and it is difficult to reach a consensus on which are the standards.

Since fairness and explainability are young disciplines and lack strong theoretical foundations, the most suitable strategy for exposing the bias is to build collaboratively at the intersection of these two AI ethics principles. This is precisely the insight on which FairShades is based and aims to operate. Compared to other explainers, such as the one presented in [16] or [21], the neighbors generation within FairShades is not random but is framed and controlled through CheckList within specific perturbations related to particular linguistic features, e.g., vocabulary, fairness, etc. Moreover, most works within XAI focus solely on explaining model logic without exploring or accounting for discrimination. Concerning bias assessment strategies, most techniques are often designed for tabular data and target a general concept of discrimination, without providing fine-grained analyses. Our contribution aims to combine and mutually enrich the approaches from these fields. The evaluation we provide is concretely relevant to a specific NLP application, i.e., abusive language detection. FairShades output, in fact, consisting of sensitive terms towards which the model exhibits discrimination, is immediately exploitable to become aware of disparate treatments and to design debiasing techniques precisely targeting the biases that emerged.

III. Setting The Stage

In this paper, we address the problem of fairness evaluation for abusive language detection systems. Let b be the abusive language detection system under examination considered as a black box classifier [8], x the textual instance to be classified belonging to the dataset X, $y = b(x)$ the decision of b for x as “abusive” or “non-abusive”. Besides, we name S the sensitive attribute, protected group or minority, i.e., *gender, race, nationality, etc.*, and we indicate with $v \in S$ a specific

[1] Similar documentation processes applied to data are proposed within the *Data Statements* [23] and *Datasheets for Datasets* [24].

sensitive value belonging to a protected group S (e.g. for "gender": *queer, transgender, non-binary, etc.*).

We keep our paper self-contained by summarizing here existing key concepts necessary to comprehend our proposal.

1) Counterfactual Fairness: We define *fairness* as b's behavior of producing similar decisions y for similar protected v mentioned, i.e., regardless of the specific value assumed, without disadvantaging minorities or amplifying pre-existing social prejudices. Recalling the definition of *counterfactual fairness* proposed in [6], we can reformulate that a $y = b(x)$ is fair if it does not change from the original text w.r.t. every other *counterfactual text* in which a different v appears, belonging to the same S. For instance, if *b("Paul is a brilliant academic, don't you think?")* = "$not\text{-}abusive$" but *b("Stephanie is a brilliant academic, don't you think?")* = "$abusive$", the model demonstrates a change in the prediction according to the perturbation of proper nouns, specifically whether the name belongs to a man or a woman. In fact, if y changes according to the perturbation of v into v' in the *counterfactual text*, then S turns out to be a discriminative concept because b shows that by changing v with v', y changes unfairly and unexpectedly.

2) CheckList Framework: CheckList [11] is the tool used in our proposal to stress the black box classifiers with potential counterfactual unfair texts. Usually, the generalization capability of NLP models is evaluated based on the performance obtained on a held-out dataset, by measuring F1 or accuracy. This process, although widely adopted by the NLP community as a way to compare systems performances and approaches, lacks informativeness since it does not provide insights into how to improve the models through the analysis of errors. To tackle this issue, CheckList [11] was developed as a comprehensive task-agnostic framework, inspired by behavioral testing, in order to encourage more robust checking and to facilitate the assessment of models' general linguistic capabilities, such as named entity recognition, negation and fairness. The package allows the generation of data through the construction of different ad-hoc tests by generalizations from templates and lexicons, general-purpose perturbations, tests expectations on the labels and context-aware suggestions using RoBERTa [29] fill-ins as prompter for specific masked tokens. The tests created can be saved, shared and utilized for different systems. CheckList includes three test types and a number of linguistic capabilities. The three types of tests are:

- Minimum Functionality Test (MFT): it involves classifying a record, as offensive or not. The test checks if the label assigned by the model is the same as the ground truth associated with the record. Each group of MFTs is designed to prove and explore how the model handles specific challenges related to a language capability, e.g. vocabulary, negation, etc..
- Invariance Test (INV): it verifies that model predictions with respect to a record do not change significantly on variants generated by altering the original sentence through the replacement of specific terms with similar expressions.
- Directional Expectation Test (DIR): it verifies that model predictions change as a result of the record perturbation, i.e., the score should raise or fall according to the modification applied.

3) Post-hoc Explainability: Following recent surveys on XAI [8], [30], [31], [32], [33], [34], [35], we briefly define the field to which FairShades contributes, i.e., post-hoc explainability methods. This branch belongs to the black box explanation methods, as opposed to the explanation by design techniques. The aim is to build explanations for a black box model, i.e., a model that, due to the complexity of its internal dynamics, is not interpretable nor transparent with regard to the automatic decision process. Post-hoc strategies can be *global* if they target explaining the whole model, or *local* if they aim to explain a specific decision for a particular record. The validity of the local explanation is therefore dependent on the particular instance chosen and often the findings by themselves are not generalizable to describe the overall model logic. In addition, the explanation technique can be *(i) model-agnostic*, i.e., independent w.r.t. the type of black box to be inspected (e.g., tree ensemble, neural networks, etc.), or *(ii) model-specific*, involving a strategy that has particular requirements and works only with precise types of models. Returning to the general definition of post-hoc explainability, we now introduce more formally the objective of these methods. Given a black box model b and an interpretable model g, post-hoc methods aim to approximate the local or global behavior of b through g. In this sense, g becomes the transparent surrogate of b, that can mimic and account for its complex dynamics in a way that is more intelligible to humans. The approaches proposed in the literature differ in terms of the input data handled by b (textual, tabular, etc.,); the type of b the interpretable technique can explain; the type of explanator g adopted (decision tree, saliency maps, etc.).

In our setting, FairShades aims at explaining b sub-globally, targeting to describe how b behaves specifically w.r.t. selected topics present in the corpus, i.e., sexism, racism or ableism. The approach is model-agnostic and therefore can be applied to any b, as its only requirement is a user-define function $y = b(x)$ that returns the prediction for the abusive class of a given instance $x \in X$. The input x is a textual record. The transparent technique g we adopt is a decision tree regressor.

IV. Fairness Estimation

Before presenting FairShades, we illustrate the $unfairness$, i.e., our proposal to estimate the level of fairness for an abusive language detection system on a given dataset. In line with [6], we propose an $unfairness$ definition that relies on the assessment of counterfactual worlds and unfair samples. In particular, we define $unfairness$ as the sensitivity of b with respect to the presence in the record to be classified of one or more entities $v \in S$. Specifically, b is considered unfair or biased if y changes according to the v present. For instance, in a given same sentence, the degree of "abusiveness" increases if terms such as *white* or *straight* are replaced by *black* or *non-binary*, revealing imbalances, possibly resulting from skewed and unrepresentative training data. We underline that framing the concept of (un)fairness in the specific scenario where the system is used is more effective in identifying biases and adopting the most suitable mitigation strategy. Thus, we adopt a definition strongly contextual to abusive language detection.

Definition 1 (Unfairness): Given a black box b, a corpus of texts X, and a sensitive category S, we define unfairness as

$$unfairness_S(b, X) = \frac{C}{T}$$

where $T = |X|$ is the number of records in X, and C is the number of records $x \in X$ for which exists at least a counterfactual text x' varying the prediction of b, i.e., $b(x) \neq b(x')$. Formally,

$$C = \sum_{x \in X} \mathbb{1}_{cond_Z(x)}$$

$$cond_Z(x) = \begin{cases} \textit{True iif } \exists x' {\in} Z \wedge v {\in} S \textit{ s.t. } b(x) {\neq} b(x') \wedge v {\in} x' \\ \textit{False otherwise} \end{cases}$$

where $\mathbb{1}_{cond_Z(x)}$ returns 1 when $cond_Z(x)$ is verified, 0 otherwise; and Z is a set of counterfactual texts obtained by perturbing x w.r.t. the set of sensitive values $v \in S$ attribute.

In other words, the measure of unfairness is calculated as the ratio of the records that have at least one unfair counterfactual (C in our formula) generated perturbing a sensitive value $v \in S$, over the number of records in the dataset grouped by the S present (T in our formula). For example, a model can be unfair at 0.48 w.r.t. samples involving sexism, i.e., $S =$ “$sexism''$, if $T = 27$ and the records involving discrimination, i.e., having at least one unfair counterfactual, are $C = 13$ (the ratio is therefore 13/27). The closer the value is to 1, the more the system is unfair, demonstrating biases. We acknowledge that the above definition is very narrow. As future work, we would develop a weighted version of the unfairness, to allow a more nuanced evaluation of models' performance.

In the following, we describe our proposal able to effectively measure the level of *unfairness* of an abusive language detection system adopting the formalization presented above.

V. FairShades

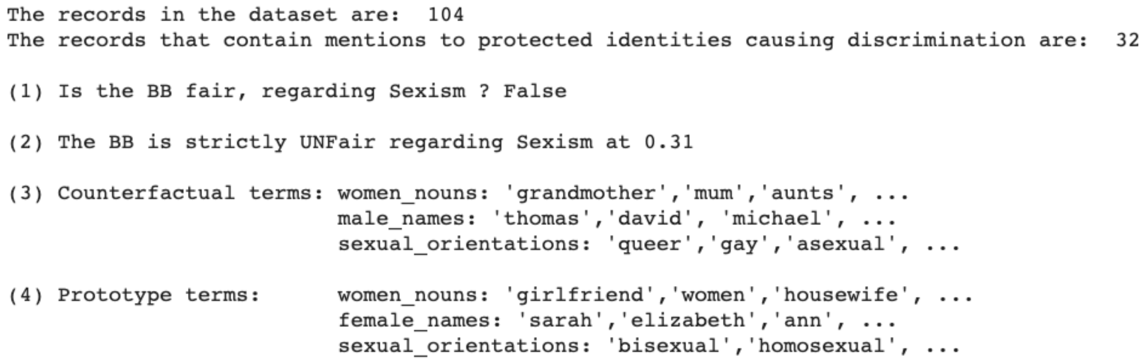
In this section, we describe FairShades, the model agnostic approach we design to conduct bias auditing in abusive language detection systems. Leveraging explainability techniques, following the formalism proposed in [8], FairShades can be characterized as a post-hoc sub-global explanation approach. FairShades is a task-specific approach for abusive language detection: it can be used to test the fairness for any abusive language detection system working on any textual dataset. However, its ideal application is on sentences that contain protected identities, i.e., expressions referring to nationality, gender, etc., as the primary scope is to uncover biases and not to explain the reasons for the prediction.

Inspired by post-hoc XAI approaches, FairShades trains an interpretable model (a decision tree regressor) on local synthetic neighborhoods generated through meaningful perturbations performed through the CheckList framework [11]. As in [36] for image classification, our idea is to benefit from counterfactuals to generalize and infer the classifier behavior, then analyze the prediction probability variation and correlate it with the record perturbation. Since the perturbations are performed through CheckList with ad-hoc testing and specialized lexicons, in the synthetically generated records we precisely vary the sensitive terms, i.e., the ones we are interested in observing system reaction, to infer potential inequalities.

FairShades output returns the fairness degree of the model under analysis w.r.t. a specific sensitive attribute S, e.g., gender or race, and the reasons associated with the result. Intuitively, the model is not fair if relies on confidential terms to perform the prediction. The *unfairness* measure, detailed in Section IV, is not binary (i.e., fair or unfair), but it is fuzzy, in the sense that a score between 0 and 1 is indicated within which the model behaves unfairly: the closer the value is to 1, the more it is unfair, demonstrating biases. In addition to the *unfairness* measure, the fairness explanation, returned in natural language through a message displayed to the user, consists of the most influential counterfactual sensitive terms identified during the black box auditing. Recalling the definition in [37] for “persistent counterfactual explanation”, *counterfactual terms* are those whose presence in the perturbed phrase cause a flip in the label, significantly influencing the prediction. Indeed, the result consists of the counterfactual terms within the binary classification, i.e., abusive or non-abusive class, discovering members' categories toward which the model is most biased. *Prototype terms* are also returned, i.e., those terms that cause an invariance behavior of the model for the records containing them, whose predictions do not significantly change and therefore, as stated in [38], the ones that share the same label as the original sample.

Lastly, we would like to stress that the evaluation is not performed on whole datasets, but on data subsets, obtained unifying records by the presence of certain sensitive terms. FairShades is therefore run on thematic groups of records, i.e., on posts about race, about women, etc. This approach allows a more in-depth exploration and distinction of bias nuances pertaining to each specific target of abuse.

As a practical example, we describe the explanation returned in Figure 1. FairShades takes as input 200 records from the *Automatic Misogyny Identification* dataset [39]. An abusive example of phrases in the corpus may be: *“@[mention] Which one of the two wins the prize for most hysterical woman?”*. In contrast, a non-abusive example is: *“Why do y'all blame the woman for what she wears when it comes to rape? That shit is WRONG.”*. From this input, as users we are interested in expressions related to the $bias = sexism$, expecting the topics addressed to be opinions on gender, sexual orientation or dialogues targeting women. Therefore, FairShades identifies and selects 104 sentences from the original 200. This filtering, resulting in a thematic subset on $bias = sexism$, is performed through hand-coded lexicons of protected terms related to the “topic” chosen. FairShades returns a *fairness explanation* composed by the following parts: *(i)* if the black box under



```
The records in the dataset are:  104
The records that contain mentions to protected identities causing discrimination are:  32

(1) Is the BB fair, regarding Sexism ? False

(2) The BB is strictly UNFair regarding Sexism at 0.31

(3) Counterfactual terms: women_nouns: 'grandmother','mum','aunts', ...
                          male_names: 'thomas','david', 'michael', ...
                          sexual_orientations: 'queer','gay','asexual', ...

(4) Prototype terms:      women_nouns: 'girlfriend','women','housewife', ...
                          female_names: 'sarah','elizabeth','ann', ...
                          sexual_orientations: 'bisexual','homosexual', ...
```

Fig. 1. FairShades fairness explanation for the BERT-based classifier on AMI subset grouped per "sexism" $bias$. We summarize and report: *(1)* whether the black box under consideration is fair w.r.t. the selected $bias$ in a binary direction, i.e., if the unfairness score is above 0 FairShades returns $False$, it returns $True$ otherwise; *(2)* the recorded value of the $unfairness$, still w.r.t. the same $bias$; *(3)* the counterfactual sensitive terms unfairly changing the classification outcome and *(4)* the prototypes sensitive terms that correctly do not change the classification outcome.

Algorithm 1: FairShades(X, b, $bias$)

```
Input  : $X$ - dataset or portion of a dataset, $b$ - black box,
         $bias$ - can be sexism, racism or ableism
Output: $F$ - fairness explanation
1  $X' \leftarrow SeparateCorpus(X, bias)$                // filter w.r.t. $bias$
2  $C \leftarrow 0$                                       // counter for $unfairness$
3  $\mathcal{L} \leftarrow \emptyset$                     // empty set to store local explanations
4  for $x \in X'$ do                                      // for each record
5      $L, q \leftarrow LITF(x, b(x))$                    // compute its local explanation
6      $\mathcal{L} \leftarrow \mathcal{L} \cup \{L\}$    // store each set of local $L$ in $\mathcal{L}$
7      if $q$ then                                        // if unfair counterf. term in $L$
8          $C \leftarrow C + 1$ // count that $x$ has at least one unfair counterf.
9  $I \leftarrow GroupInfluentialTerms(\mathcal{L})$      // group counterf. and prototypes
10 $U \leftarrow C/|X'|$                                  // calculate $unfairness$
11 $F \leftarrow BuildFairnessX(U, I)$                    // build fairness explanation
12 return $F$;
```

analysis is unfair[2] concerning the selected bias, i.e., *sexism* (1^{st} line of the explanation); *(ii)* specifically, it reports the $unfairness$ at 0.31, i.e., 32 records have at least one unfair counterfactual over the total of 104 (2^{nd} line of the explanation); *(iii)* "grandmother", "Thomas", "queer" and other terms are reported as counterfactuals, i.e., terms that unfairly change the decision outcome; *(iv)* a sets of prototype terms is also returned, including, for example, "girlfriend", "Sarah" and "bisexual"[3], that correctly does not affect the decision outcome (3^{rd} and 4^{th} line of the explanation, respectively).

In the following, we describe the details of FairShades, analyzing first the main method that retrieves the fairness explanation, and then presenting the approach applied to a single record that supports the identification of locally influential terms.

A. Fairness Explanation Extraction

The pseudo-code of FairShades is illustrated in Algorithm 1. FairShades takes as input X, i.e., a dataset or a portion of a dataset; b, the abusive language detection system working as a black box under examination; $bias$, identifying the type of prejudice to be investigated according to the target of abuse.

In line 1, X is divided into a subset X', obtained grouping the records according to $bias$, i.e., phrases having similar sensitive terms present. Thus, X' contains only records associated with the selected $bias$. This choice responds to the purpose of working on more similar neighbors and implicitly on similar content subjects. The $bias$ categories we currently identify, encoded in dictionaries of sensitive terms, are related to: *(i) sexism*, containing mentions to gender, sexual orientations, etc.; *(ii) racism*, containing expressions identifying nationalities, religions, etc.; *(iii) ableism*, i.e., terms referring to disabled and elderly people. We chose this approach because it conforms to the formalization and the process we have designed for fairness neighborhood perturbations (see later in Algorithm 2), but every other text-based clustering approach[4] would have worked, though leading to slightly different results.

In line 2, a counter C is set to 0: it is used to quantify records that have at least one unfair counterfactual. In line 3, FairShades creates an empty set $\mathcal{L}$ used to collect local fairness explanations. From line 4 to 6, for each record in the selection X', a local fairness explanation is computed (invoking Algorithm 2) and added to $\mathcal{L}$. A local fairness explanation consists of sets of: *(i)* a set L of "locally" influential terms, both counterfactuals and prototypes for each individual record; *(ii)* a boolean variable q that reveals if L contain at least one protected counterfactual term, thus stating that b is unfair or not on x. If this condition is verified, i.e., q is true, then C is incremented (line 7-8). In line 9, FairShades combines and groups the sets of terms collected

[2] A black box is *unfair* if the unfairness score is above 0.

[3] Prototypes terms are not to be regarded as "opposite" expressions to counterfactual terms. For example, we might expect to meet the term "man" in prototypes by seeing the term "woman" in counterfactuals: it is not necessarily the case. In fact, the term sets are neither related nor dependent in this current version of FairShades. The terms are generated according to the record perturbation method described in Section V-B.

[4] Applying a clustering algorithm or having sentences represented by word embeddings and applying a vector distance calculation on them to find the most similar instances are some of the possible alternatives.

Algorithm 2: LocalInfluentialTermsFinder(x, b)

```
Input  : $x$ - record, $b$ - black box
Output: $L$ - set of local counterfactuals and prototypes,
        $q$ - boolean value for $unfairness$
1 $y_{pred} \leftarrow b(x)$                        // get prediction of $b$ on $x$
2 $Z \leftarrow CheckListNeighGen(x)$               // perturb $x$ to generate $Z$
3 $DTR \leftarrow TrainDTR(Z, b(Z))$                // decision tree regressor
4 $L \leftarrow GetInfluentialTerms(DTR, y_{pred})$ // counterf. & proto.
5 $q \leftarrow cond_Z(x)$   // True if protected terms $\in Z$ cause pred. changes
6 return $L, q$;
```

in $\mathcal{L}$. For each record, the sets of local counterfactuals and prototypes $L \in \mathcal{L}$ are filtered to keep only those belonging to protected attributes, such as gender or nationality, selected through hand-coded lexicons collecting lists of sensitive terms. FairShades thus exhibits a "complete" set I, which summarizes the influential terms found in the subset of the dataset X' divided by counterfactuals and prototypes, further distinguished by protected attributes. Finally, in line 10 it computes the $unfairness$ measure U dividing C, i.e., the number of records with unfair counterfactuals, by the size of the dataset X'. The result is constructed in line 11, where the sub-global[5] fairness explanation F (an example in Figure 1) is composed of the $unfairness$ and the sets of counterfactuals and prototypes I.

B. Locally Influential Terms Finder

In Algorithm 2 we report the pseudo-code of the Local Influential Terms Finder[6] (LITF) which supports the main method in creating explanations of each individual record.

LITF takes as input a record $x \in X'$ and the black box b under examination. In line 1, b is applied to x, returning the prediction y_{pred} for the record. In line 2 it is applied the neighborhood generation on x, performed through the CheckList tool [11], partially following its framework and process. CheckList in fact is designed to complement the testing and evaluation phase of NLP models, but in this work it is used as neighborhood generator, deploying perturbation functions without framing them into test types (MFT, INV or DIR) or providing expectations, as in the standard use of it reported in Section III. In fact, we want to embrace its conceptual framework and the possibility of generating samples that test specific language skills within NLP models, but following the purpose of the explanation. Generally speaking, the automatic perturbations that we apply to the sentences could be considered as an INV test. In the following, we describe the linguistic capabilities identified as fundamental in [11] and report multiple examples generated from the diverse perturbation functions[7]:

- *Fairness* verifies that systems predictions do not change as a function of protected features: identification of any protected values present in the sentences and their replacement with similar ones.
  - You are too exceptional to be **lesbian**.
  - You are too exceptional to be **asexual**.
  - You are too exceptional to be **ace**.
  - You are too exceptional to be **bisexual**.
  - You are too exceptional to be **homosexual**.
- *Vocabulary* tests the ability of the model to recognize words or groups of words that are relevant for labelling: perturbation of neutral empty terms with similar ones, such as prepositions, articles, common auxiliary verbs.
  - I hate girls that are always mad and hysterical **in** the world for no reason. Why are they always so emotional?
  - I hate girls that are always mad and hysterical **towards** the world for no reason. Why are they always so emotional?
  - I hate girls that are always mad and hysterical **to** the world for no reason. Why are they always so emotional?
- *Robustness* investigates how the model deals with the addition of random and unrelated linguistic elements: (1) addition of irrelevant linguistic segments such as random strings, mentions, urls; (2) insertion of typos, neutral emojis, hashtags; (3) perturbation of punctuation and contractions, if present, otherwise are added.
  - **@5YqeBu** You are too exceptional to be gay.
  - You are too exceptional to be gay. **https://t.co/rqL**
  - You are too exceptional to be gay
  - You're too exceptional to be gay**.**
  - You are too exceptional to **b egay**.
  - You are too exceptional to be gay. **#gay.**
- *NER* analyses how the model reacts to the replacements of entities in the sentence: perturbations of locations and numbers, if present.
  - I love **Turkey**
  - I love **Uzbekistan**
  - I love **Madagascar**
  - I love my **7** dogs
  - I love my **4** dogs
  - I love my **8** dogs

The result of this phase is Z, the neighborhood which contains all the synthetic generated samples from the original x w.r.t. the linguistic capabilities just described.

Similarly to [17], [21], we train in line 3 a Decision Tree Regressor (DTR) on Z and $b(Z)$. Like in [16], Z is converted into a bag-of-words representation in order to obtain a human interpretable tree. This post-hoc explanation approach is performed to simulate and analyze locally the behavior, predictions and rationale of the black box b under consideration. In line with [40], we limit the depth of the DTR to 4 because *(i)* it is empirically shown that it is deep enough for being optimal and, *(ii)* a deeper tree would be more complex and less interpretable. An example of DTR is illustrated in Figure 2. From DTR and y_{pred}, in line 4, LITF identifies the set of influential terms L, encoded in the instances of the tree. The set L is composed of counterfactuals and prototypes. *Counterfactuals* are terms that cause the black box prediction $b(x)$ to change due to their replacement with similar expressions through ad-hoc perturbations $x' \in Z$. In these cases the label switches from the original prediction.

[5] We use the adjective sub-global because it refers to a subset of instance related to the same $bias$.

[6] This local functionality can be used independently of the main sub-global approach: it is also feasible to focus the analysis on a single record only.

[7] Compared to the methods and the lexicons provided in CheckList tutorials, we have automated some new perturbation functions, extending the use of the tool to the purpose of generating neighborhoods.

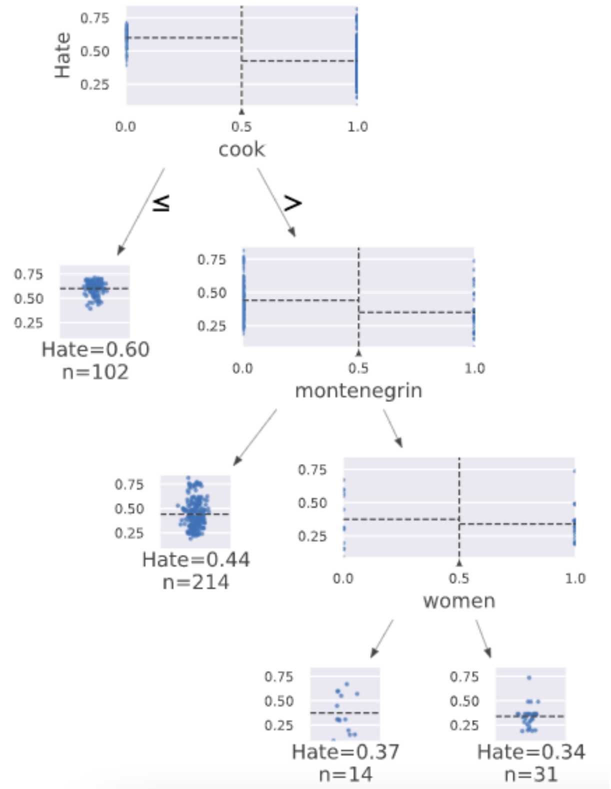
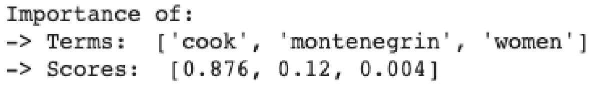


Fig. 2. Decision tree regressor visualization within local explanation for the abusive record *"I feel bad for Montenegrin women... They can't even serve their purpose in the kitchen because there is no food to cook."*. It reports on the leaves the probability scores for the hateful class w.r.t. specific conditions, i.e., the presence or absence of certain terms. This feature is expressed within each tree branch and results in bifurcations. The term under consideration, in the numerical bag-of-words representation, is characterized by either a value greater than 0.5 (i.e. 1, that means presence) or less than 0.5 (i.e. 0, absence).

For instance, if the original record x is classified as abusive by b, the record $x' \in Z$ contains a counterfactual term if x', obtained by perturbing x according to CheckList, i.e., having only a different word, it is classified as non-abusive, i.e, $b(x) \neq b(x')$. This process identifies the terms towards which the system is sensitive and discriminatory. On the other hand, *prototypes* are also recognized and returned. These terms are not particularly influential for the classification, as for these expressions $b(x')$ does not change w.r.t. $b(x)$, exhibiting an invariant behavior. From DTR we also get access to the local features importance (an example in Figure 3) that might be useful to better understand the reasons for the assigned outcome. Finally, in line 5, FairShades checks if Z contains at least one protected counterfactual term, thus classifying b as unfair on x. This information is stored in the boolean variable q: if the condition is verified, then q is True. LIFT therefore returns a local explanation composed of a tuple containing the influential terms L, both counterfactuals and prototypes, and the boolean variable q.



```
Importance of:
-> Terms:  ['cook', 'montenegrin', 'women']
-> Scores:  [0.876, 0.12, 0.004]
```

Fig. 3. Features importance within local explanation for the abusive record *"I feel bad for Montenegrin women... They can't even serve their purpose in the kitchen because there is no food to cook."*.

VI. Experiments

In this section, we report the experiments carried out to validate FairShades[8]. First we present the experimental settings and we describe the evaluation metrics used, then we examine the unfairnass of BERT-based models applied as abusive language detection systems.

A. Experimental Settings

We run our evaluation using a BERT-based classifier for English, a language representation model developed by Google Research, whose deep learning architecture obtained state-of-the-art results in several NLP tasks including sentiment analysis, natural language inference, textual entailment [41] and hate speech detection [29]. BERT can be fine-tuned and adapted to specific tasks by adding just one additional output layer to the neural network: this approach have been used by the vast majority of participants in the last Offenseval campaign reported in [15], yielding a very good performance on English (> 0.90 F1). For our experiments, we used two different pre-trained implementations of this model, available through the library Transformers[9]. The first system is a BERT model[10] [42], which was trained[11] on English benchmark hate speech datasets [43], [44], [45], [46], [47], [48] and fine-tuned on multilingual BERT model. Although the model was developed with the aim of testing new approaches for multilingual hate speech detection, especially for low-resource languages, our current exploration focuses on English only. However, we do not exclude a multilingual version of our tool at a later stage. The second system is a RoBERTa[12] based model [49], fine-tuned on TweetEval benchmark from [50], specifically on [46] for hate speech detection.

The datasets chosen gather mainly posts from Twitter. Two types of datasets are involved, for a total of six collections. The first are three synthetic datasets[13] created in [26] through CheckList templates covering different types of bias grouped by target, namely sexism, racism and ableism. The second type consists of three hate speech benchmark datasets: *(i) HatEval: Multilingual detection of hate speech against immigrants and women on Twitter* [46], part of the SemEval 2019 campaign, Task 5; *(ii) Automatic Misogyny Identification* [39], a new task

[8] FairShades is available at https://github.com/MartaMarchiori/FairShades, with an example notebook and the experiments realized.

[9] https://huggingface.co/transformers/

[10] huggingface.co/Hate-speech-CNERG/dehatebert-mono-english

[11] https://github.com/punyajoy/DE-LIMIT

[12] huggingface.co/cardiffnlp/twitter-roberta-base-hate

[13] The data are available are available at https://github.com/MartaMarchiori/Fairness-Analysis-with-CheckList.

TABLE I
ACCURACY AND F1 OF BERT AND ROBERTA BASED MODELS ON SUBSETS OF DATASETS FILTERED THROUGH HAND-CODED LEXICONS.

Dataset (Subset Size)	BERT		RoBERTa	
	Accuracy	F1	Accuracy	F1
Sexism (50)	0.36	0.30	0.60	0.67
Racism (50)	0.70	0.74	0.36	0.20
Ableism (50)	0.14	0.25	0.04	0.08
HatEval (200)	0.66	0.69	0.56	0.65
AMI (200)	0.77	0.75	0.77	0.73
Multilingual (200)	0.46	0.53	0.62	0.12

TABLE II
Unfairness MEASURE OF BERT AND ROBERTA BASED MODELS ON SELECTED DATASETS.

Dataset	$bias$	# sensitive-records	$\sum\|Z\|$	BERT	RoBERTa
Sexism	s	27	1575	**0.19**	0.48
Racism	r	31	7883	**0.19**	**0.58**
Ableism	a	19	627	0.21	**0**
HatEval	r	30	10108	**0.60**	0.53
AMI	s	104	5856	0.31	0.36
Multilingual	r	19	5711	0.42	0.11

part of the EVALITA 2018 campaign; *(iii) Multilingual and Multi-Aspect Hate Speech Analysis* [47]. We would like to point out that the use of these resources was limited to the general task of detecting whether a tweet is considered hateful or not; sub-tasks such as detecting the type of misogynistic attack (e.g., in AMI) or the specific target (e.g., in HatEval) were not considered, since FairShades itself allows a fine-grained analysis of biases and protected attributes present in the texts, following its own particular framework.

B. Evaluation Measures

Unfairness. As the main criterion to assess the black box under analysis, we propose to use, according to FS approach, the $unfairness$ measure as framed in Section IV.

Precision. To explore models fairness, there are also strategies to investigate it at group level, i.e. exploring and comparing classifier behavior w.r.t. each diverse race, gender or other sensitive attribute present in the data, to assess any disparate treatment. Therefore, in order to conduct a general fairness evaluation, in line with the literature, we deploy the python package `fairlearn`[14] [27] to compute the *precision* separately for each group-member. In particular, on the basis of the most frequently occurring sensitive attribute, the values belonging to it are analyzed and thus identified as main subgroup to be investigated. For example, if our texts deal more with opinions on gender, the most frequent attribute, based on the frequencies calculated and compared with the other attributes, will be *gender*: consequently, precision will be calculated separately for the values "non-binary", "trans", etc. Within FairShades, the computation of precision takes as input the neighborhoods Z. Specifically, the confusion matrix for each subgroup is computed from the ground truth value y_{real} associated with the original record and y_{pred}, i.e., those predicted for the neighborhoods by the black box b. Other required inputs may be the name of the sensitive attributes present and the sets of values belonging to them, obtainable through simple data processing.

C. Quantitative Evaluation

In the following, we outline a quantitative assessment[15] of the experiments conducted. In Table I we observe models performances on subsets of the presented datasets evaluated according to accuracy and F1. We recall that the versions of BERT and RoBERTa, described in the previous section, are pre-trained implementations available through the library Transformers[16]. We briefly comment that even just analyzing these results on selected subsets, both models demonstrate severe shortages. Low performances are reached for the synthetic dataset Ableism by both BERT and RoBERTa. The synthetic dataset Sexism also highlights drawbacks of BERT, while the synthetic Racism and the benchmark Multilingual by [47] point out limitations of RoBERTa through F1. A preliminary consideration surely relates to the fact that the lowest performances are reached for the synthetic data, which, in this context, would seem to be more challenging than real datasets or relating to topics and expressions not previously observed by the models. More experiments on a wider range of datasets are needed in order to broaden our analysis.

In Table II we report the $unfairness$ scores. The second column identifies the type of $bias$ as sexism (s), racism (r) or ableism (a). Besides, it is reported the number of identified "sensitive" records , i.e., those containing instances of sensitive attributes. Although it might seem a limited number of instances, the analysis is actually performed on $\sum|Z|$, reported in the following column, which represents the sum of the sizes of the neighborhoods created for all the identified sensitive records. We recall that the $unfairness$ is calculated through the ratio of the records that have at least one unfair neighbor over the number of records in the bias-grouped dataset: the closer the value is to 1, the more the system is unfair. The highest values of $unfairness$ are recorded for BERT on the HatEval dataset (0.60) and for RoBERTa on the synthetic dataset Racism (0.58), both grouped for the $bias$ $racism$. RoBERTa shows a close value for HatEval (0.53). It means that each of the models demonstrates unintended bias toward certain sensitive categories, likely derived from the training data they were exposed to, thus learning from collections that are neither balanced nor representative. The lowest values occur for BERT on the synthetic datasets Sexism and Racism (0.19). RoBERTa achieves 0 $unfairness$ on the synthetic examples Ableism. It could be motivated by the fact that the model did not frequently encounter abusive examples containing references to disabled, elderly, or homeless people. To conclude, we can see that, although RoBERTa is a variant of BERT, certain differences in the $unfairness$ scores are

[14] https://fairlearn.org/

[15] We limit the discussion to this type of evaluation but refer to the notebooks for more in-depth qualitative comparisons, e.g., concerning counterfactual and prototype terms identified by the models on the different datasets.

[16] https://huggingface.co/transformers/

TABLE III
PRECISION PER SUB-POPULATIONS OBTAINED BY THE BERT-BASED MODEL ON HATEVAL SUBSET GROUPED PER “RACISM” *bias*.

Country	Precision
- *mexico*	**0.08**
- *rwanda*	0.70
- *south korea, egypt, pakistan, germany, nigeria, kenya, russia, japan, italy, south africa, ethiopia, spain, myanmar, china, sudan, tanzania, algeria, vietman, argentina, iran, ukrain, united kingdom, turkey, india, bangladesh, indonesia, thailand, uganda, colombia*	0.75
- *uzbekistan*	0.78
- *philippines*	**0.86**
- ...	...

significant. BERT in our framework would appear to be more significantly sensitive than RoBERTa, as the values reached are on average higher. These preliminary hypotheses need to be confirmed with additional experiments, e.g., from the auditing of other diverse language models also with respect to a wider range of datasets. We leave these experiments as future works.

We report in Table III the precision[17] obtained by the BERT model on the HatEval dataset filtered per *bias racism*. From 200 randomly extracted records, FairShades selects 30 samples as pertaining to *racism*, generating a total of 10108 perturbed records on which to conduct the fairness analysis. The most frequent protected attribute is *country*, therefore the results on which we are focusing in this example analysis are the records that contain mentions or expressions related to countries. A great performance disparity is evident, demonstrating an unequal error distribution to different members of the same sensitive attribute. The lowest value, obtained for “Mexico”, amounts to 0.08, followed at a great distance by “Rwanda” at 0.7. Achieving poor precision means having a larger number of false positives, i.e., being incorrectly classified more often with the *hateful* label. The highest values, around 0.99, are obtained for “Poland”, “Brazil”, “United States” and “Sweden”. In general, while examining the results obtained on the other datasets, we have noticed that for some values within certain protected categories, diverse samples in datasets (and therefore in related synthetic neighborhoods) were missing. This situation constitutes a challenge because means that for one or more demographic few or no samples are available, as outlined in [51]. Therefore, it may happen that the denominator within precision formula is 0 (e.g. when true positive plus false negative is equal to 0 or true nor predicted samples are available), thus the metric is undefined. We tried to avoid this risk by selecting for these computations the most frequent sensitive attribute, conducting the analysis only on its values. We can hypothesize that metrics declined in this form are more suitable for tabular data, than for unstructured data such as text. For this reason and to further investigate, more experiments are needed. In general, the recognition of these sensitivities should lead the developer to quantitatively reassess the data used to train the model and to plan a second training or fine-tuning phase on more balanced data to account for the different minorities.

VII. CONCLUSION

In this paper, we have presented FairShades, a model-agnostic approach that relies on explainability techniques for auditing the outcomes of abusive language detection classifiers. Results on BERT-based models show that although these classifiers achieve high accuracy levels on a variety of natural language processing tasks, they demonstrate severe shortages on samples involving implicit stereotypes, expressions of hate towards minorities and sensitive attributes such as race or sexual orientation, in agreement with recent surveys.

A drawback of FairShades, closely related to CheckList, concerns the difficulty of including and dealing with contextual information [52]. Sensitive real-world statements often acquire a different connotation w.r.t. the degree of hatred if a certain race, gender, or nationality is present, due to historical or social references. Indeed, we will certainly have to reconsider and deepen the effectiveness of the invariance concept relating to fairness. With respect to other bias discovery works FairShades allows not limiting fairness evaluation to numerical metrics but offers also sets of related terms, pertaining for example to gender or race, for which the audited model demonstrate disparate treatments and, ultimately, unfair inequalities. However, the identification of these model skews is currently heavily dependent on the hand-coded lexicons we included within CheckList. This will certainly be an area for improvement, in order to make our result (i.e., the fairness explanation), more robust and in line with other works using established lexicons.

A future direction might include expanding the neighborhood generation process, e.g. deploying Polyjuice[18] [53], a general-purpose counterfactual generator trained by fine-tuning GPT-2 [54]. Moreover, implementing an option for exporting the explanations would be very useful, in order to allow users to save and share the results, revisiting them later. Significant aspects to explore with additional experiments would be testing commercial models like Google Perspective API[19] and evaluating FairShades with competitors explainers on the fairness dimension, i.e., the capability of identifying biases. Another priority consists in designing effective interactions with users and assessing the impacts on individuals from a human-centered ML perspective. Therefore, the opportunity to conduct robust user testing would also be extremely helpful, in order to collect human evaluation and improve FairShades quality of explanations.

ACKNOWLEDGMENT

This work has been partially supported by the EU Community H2020 programme under the funding schemes: G.A. 871042 *SoBigData++* (`sobigdata`), G.A. 952026 *HumanE-AI-Net* (`humane-ai`), G.A. 834756 *XAI* (`xai`), G.A. 860630 *NoBIAS* (`nobias`), and G.A. 952215 *TAILOR* (`tailor`).

[17] In Table III, we report only a portion of the results to support the discussion. The full version can be found in the published notebooks.

[18] https://huggingface.co/uw-hai/polyjuice

[19] https://www.perspectiveapi.com

REFERENCES

[1] H. Suresh and J. V. Guttag, "A framework for understanding unintended consequences of machine learning," *CoRR*, vol. abs/1901.10002, 2019.

[2] N. Mehrabi, F. Morstatter, N. Saxena *et al.*, "A survey on bias and fairness in machine learning," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 115:1–115:35, 2021.

[3] M. Sap, D. Card, S. Gabriel *et al.*, "The risk of racial bias in hate speech detection," in *ACL (1)*. Association for Computational Linguistics, 2019, pp. 1668–1678.

[4] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *FAT*, ser. Proceedings of Machine Learning Research, vol. 81. PMLR, 2018, pp. 77–91.

[5] L. Dixon, J. Li, J. Sorensen *et al.*, "Measuring and mitigating unintended bias in text classification," in *AIES*. ACM, 2018, pp. 67–73.

[6] M. Kusner, J. Loftus, C. Russell *et al.*, "Counterfactual fairness," *stat*, vol. 1050, p. 8, 2018.

[7] E. Ntoutsi, P. Fafalios, U. Gadiraju *et al.*, "Bias in data-driven artificial intelligence systems - an introductory survey," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 10, no. 3, 2020.

[8] R. Guidotti, A. Monreale, S. Ruggieri *et al.*, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 93:1–93:42, 2019.

[9] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[10] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, 2019.

[11] M. T. Ribeiro, T. Wu, C. Guestrin *et al.*, "Beyond accuracy: Behavioral testing of NLP models with checklist," in *ACL*. Association for Computational Linguistics, 2020, pp. 4902–4912.

[12] S. Kiritchenko, I. Nejadgholi, and K. C. Fraser, "Confronting abusive language online: A survey from the ethical and human rights perspective," *J. Artif. Intell. Res.*, vol. 71, pp. 431–478, 2021.

[13] A. Vaswani, N. Shazeer, N. Parmar *et al.*, "Attention is all you need," in *NIPS*, 2017, pp. 5998–6008.

[14] M. E. Peters, M. Neumann, M. Iyyer *et al.*, "Deep contextualized word representations," in *NAACL-HLT*. Association for Computational Linguistics, 2018, pp. 2227–2237.

[15] M. Zampieri, P. Nakov, S. Rosenthal *et al.*, "Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020)," in *SemEval@COLING*. International Committee for Computational Linguistics, 2020, pp. 1425–1447.

[16] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," in *KDD*. ACM, 2016, pp. 1135–1144.

[17] R. Guidotti, A. Monreale, F. Giannotti *et al.*, "Factual and counterfactual explanations for black box decision making," *IEEE Intell. Syst.*, vol. 34, no. 6, pp. 14–23, 2019.

[18] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *NIPS*, 2017, pp. 4765–4774.

[19] M. W. Craven and J. W. Shavlik, "Extracting tree-structured representations of trained networks," pp. 24–30, 1995.

[20] M. G. Augasta and T. Kathirvalavakumar, "Reverse engineering the neural networks for rule extraction in classification problems," *Neural Process. Lett.*, vol. 35, no. 2, pp. 131–150, 2012.

[21] O. Lampridis, R. Guidotti, and S. Ruggieri, "Explaining sentiment classification with synthetic exemplars and counter-exemplars," in *DS*, ser. Lecture Notes in Computer Science, vol. 12323. Springer, 2020, pp. 357–373.

[22] M. Mitchell, S. Wu, A. Zaldivar *et al.*, "Model cards for model reporting," in *FAT*. ACM, 2019, pp. 220–229.

[23] E. M. Bender and B. Friedman, "Data statements for natural language processing: Toward mitigating system bias and enabling better science," *Trans. Assoc. Comput. Linguistics*, vol. 6, pp. 587–604, 2018.

[24] T. Gebru, J. Morgenstern, B. Vecchione *et al.*, "Datasheets for datasets," 2018.

[25] S. L. Blodgett, S. Barocas, H. D. III *et al.*, "Language (technology) is power: A critical survey of "bias" in NLP," pp. 5454–5476, 2020.

[26] M. M. Manerba and S. Tonelli, "Fine-grained fairness analysis of abusive language detection systems with checklist," in *WOAH*, 2021, pp. 81–91.

[27] S. Bird, M. Dudík, R. Edgar *et al.*, "Fairlearn: A toolkit for assessing and improving fairness in AI," Microsoft, Tech. Rep. MSR-TR-2020-32, May 2020.

[28] K. Sokol, A. Hepburn, R. Poyiadzi *et al.*, "FAT Forensics: A Python Toolbox for Implementing and Deploying Fairness, Accountability and Transparency Algorithms in Predictive Systems," *Journal of Open Source Software*, vol. 5, no. 49, p. 1904, 2020.

[29] Y. Liu, M. Ott, N. Goyal *et al.*, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019.

[30] F. Bodria, F. Giannotti, R. Guidotti *et al.*, "Benchmarking and survey of explanation methods for black box models," *CoRR*, vol. abs/2102.13076, 2021.

[31] A. A. Freitas, "Comprehensible classification models: a position paper," *SIGKDD Explor.*, vol. 15, no. 1, pp. 1–10, 2013.

[32] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[33] D. Pedreschi, F. Giannotti, R. Guidotti *et al.*, "Open the black box data-driven explanation of black box decision systems," *CoRR*, vol. abs/1806.09936, 2018.

[34] L. Longo, R. Goebel, F. Lécué *et al.*, "Explainable artificial intelligence: Concepts, applications, research challenges and visions," in *CD-MAKE*, ser. Lecture Notes in Computer Science, vol. 12279. Springer, 2020, pp. 1–16.

[35] W. Samek, G. Montavon, S. Lapuschkin *et al.*, "Toward interpretable machine learning: Transparent deep neural networks and beyond," *CoRR*, vol. abs/2003.07631, 2020.

[36] E. Denton, B. Hutchinson, M. Mitchell *et al.*, "Detecting bias with generative counterfactual face attribute augmentation," *CoRR*, vol. abs/1906.06439, 2019.

[37] A. Artelt, F. Hinder, V. Vaquet *et al.*, "Contrastive explanations for explaining model adaptations," vol. 12861, pp. 101–112, 2021.

[38] C. Molnar, *Interpretable machine learning*. Lulu.com, 2020.

[39] E. Fersini, D. Nozza, and P. Rosso, "Overview of the evalita 2018 task on automatic misogyny identification (AMI)," vol. 2263, 2018.

[40] D. Bertsimas and J. Dunn, "Optimal classification trees," *Mach. Learn.*, vol. 106, no. 7, pp. 1039–1082, 2017.

[41] J. Devlin, M. Chang, K. Lee *et al.*, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT (1)*. Association for Computational Linguistics, 2019, pp. 4171–4186.

[42] S. S. Aluru, B. Mathew, P. Saha *et al.*, "Deep learning models for multilingual hate speech detection," *CoRR*, vol. abs/2004.06465, 2020.

[43] T. Davidson, D. Warmsley, M. Macy *et al.*, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ser. ICWSM '17, 2017, pp. 512–515.

[44] O. de Gibert, N. Perez, A. García-Pablos *et al.*, "Hate Speech Dataset from a White Supremacy Forum," in *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 11–20.

[45] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on twitter," in *SRW@HLT-NAACL*. The Association for Computational Linguistics, 2016, pp. 88–93.

[46] V. Basile, C. Bosco, E. Fersini *et al.*, "SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter," in *Proceedings of SEMEVAL 2019*, 2019.

[47] N. Ousidhoum, Z. Lin, H. Zhang *et al.*, "Multilingual and multi-aspect hate speech analysis," pp. 4674–4683, 2019.

[48] A. Founta, C. Djouvas, D. Chatzakou *et al.*, "Large scale crowdsourcing and characterization of twitter abusive behavior," in *ICWSM*. AAAI Press, 2018, pp. 491–500.

[49] G. Wiedemann, S. M. Yimam, and C. Biemann, "UHH-LT at semeval-2020 task 12: Fine-tuning of pre-trained transformer networks for offensive language detection," in *SemEval@COLING*. International Committee for Computational Linguistics, 2020, pp. 1638–1644.

[50] F. Barbieri, J. Camacho-Collados, L. E. Anke *et al.*, "Tweeteval: Unified benchmark and comparative evaluation for tweet classification," pp. 1644–1650, 2020.

[51] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," pp. 3315–3323, 2016.

[52] S. Menini, A. P. Aprosio, and S. Tonelli, "Abuse is contextual, what about nlp? the role of context in abusive language annotation and detection," *CoRR*, vol. abs/2103.14916, 2021.

[53] T. Wu, M. T. Ribeiro, J. Heer *et al.*, "Polyjuice: Automated, general-purpose counterfactual generation," *CoRR*, vol. abs/2101.00288, 2021.

[54] A. Radford, J. Wu, R. Child *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.