UNIVERSITÀ DI PISA

DIPARTIMENTO DI FILOLOGIA, LETTERATURA E LINGUISTICA

CORSO DI LAUREA IN INFORMATICA UMANISTICA MAGISTRALE

TESI DI LAUREA

FairShades - Fairness Auditing in Abusive Language Detection Systems

Candidata: Marta Marchiori Manerba

Relatore: Dott. Riccardo Guidotti

Controrelatore: Prof. Alessandro Lenci

ANNO ACCADEMICO 2020/2021

Abstract

Current abusive language detection systems have demonstrated unintended bias towards sensitive features such as nationality or gender. This is a crucial issue, which may harm minorities and underrepresented groups if such systems were integrated in real-world applications. In this thesis, we present *FairShades*, a model-agnostic approach for auditing the outcomes of Abusive Language Detection Systems. Combining Explainability and Fairness evaluation, the tool is able to identify wrong correlations, unintended biases and sensitive categories toward which the models are most discriminative. This objective is pursued through the auditing of meaningful counterfactuals generated by CheckList framework, obtained perturbing sensitive identities present in the texts to be classified. A Decision Tree Regressor is trained on the synthetic neighbourhood and used to simulate and analyse the behaviour, predictions and rationale applied by the black box under consideration. Our approach performs both local and sub-global analysis, combining the individual interpretations. We conduct several experiments on research BERT-based models in order to demonstrate the novelty and effectiveness of our proposal on unmasking biases. Although these classifiers 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.

Keywords: *NLP; XAI; Fairness in ML; Algorithmic bias; Algorithmic Auditing; Digital Discrimination; Intersectionality; Hate Speech Detection; Abusive Language Detection Systems*

Contents

1 Introduction

In December 2020, the Court of Bologna issued a ruling[1] against the algorithm used by Deliveroo, defining it as illegitimate and discriminatory, as the ranking calculation disfavoured riders who requested exemptions from deliveries for health reasons, without guaranteeing the fundamental rights of all workers. Another paradigmatic example emerges from an algorithm used in Amazon's human resources[2]: the machine is provided with large amounts of data on past recruitment so that it can learn the company's policy and automate the process for future hiring. The system may unintentionally learn to discard a CV because it belongs to a woman, or a person with a foreign name or from a certain region of the world. Making these automated decisions based on sensitive dimensions, such as gender or nationality, creates an unfair model that discriminates against certain social groups because of the bias emerging from historical data. Fairness of models is one of the core values for the development and ethical use of AI systems. The principle is recurrent in many guidelines, published for example by the European Commission[3] (*"Diversity, non-discrimination and fairness"*) and by tech companies such as Google[4], within the goal *"Avoid creating or reinforcing unfair bias"*. The publication of ethical principles of this kind, however, runs the risk of falling into *"ethics-washing"*, i.e., merely drawing up a set of ethically acceptable values to build a positive image of the policy of the public or private actor, without following up with a real change in practices and models used. At every stage of a supervised learning process, biases can arise and be introduced in the pipeline, ultimately leading to harm[5] (Suresh and Guttag (2019); Mehrabi et al. (2019)). When it comes to systems whose goal is to automatically detect abusive language, this issue becomes particularly serious, since unintended bias towards sensitive attributes such as gender, sexual orientation or nationality can harm underrepresented groups. Sap et al. (2019a), for example, show that annotators tend to label messages in Afro-American English more frequently than when annotating other messages, which could lead to the training of a

[1] http://www.bollettinoadapt.it/wp-content/uploads/2021/01/Ordinanza-Bologna.pdf

[2] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[3] In https://digital-strategy.ec.europa.eu/en/library/excellence-and-trust-ai-brochure. Consider also the most recent regulatory framework proposal on Artificial Intelligence, published at https://tinyurl.com/EUR-Lex-AI.

[4] https://ai.google/principles

[5] An interesting curated list of harmful AI usages and outcomes is published at https://github.com/daviddao/awful-ai.

system reproducing the same kind of bias.

The role of the datasets used to train these models is crucial: as pointed out by Wiegand et al. (2019), there may be multiple reasons why a dataset is biased, e.g. due to skewed sampling strategies, prevalence of a specific subject (*topic bias*) or of content written by a specific author (*author bias*). Mitigation strategies may involve assessing which terms are frequent in the presence of certain labels and implementing techniques to balance the data by including neutral samples containing those same terms to prevent the model from learning inaccurate correlations (Wiegand et al. (2019)). 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 specialised 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.

Given the sensitive context in which abusive language detection systems are deployed, a robust value-oriented evaluation of the model's fairness is necessary: the risks otherwise might be to contribute and lead to the marginalisation of voices in online discourse belonging to certain demographic groups. These bias assessments are therefore 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 social group scores. From these questions, it becomes a priority the need to explore and disaggregate overall metrics emerges. However, this process is complicated by the partial effectiveness of proposed methods that only work with certain definitions of bias and fairness, as well as by the limited availability of recognised benchmark datasets (Ntoutsi et al. (2020)) and their focus on specific bias types such as gender, when available. Another crucial challenge emerges from the requirement to balance bias identification (and the related need to access sensitive pieces of information) with privacy protection, as pointed out by Gebru et al. (2018).

In addition to Fairness, another crucial aspect to consider, related to these complex models used on high-dimensional data, lies in the opaqueness of their internal behaviour. In fact, if the dynamics leading a model 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. Inspecting non-discrimination of decisions and assessing that the knowledge autonomously learnt conforms to human values also constitutes both a real challenge and a risk. Indeed in recent years working towards transparency and interpretability of black box models has become a priority: multiple approaches and methods have been proposed to face these matters (Guidotti et al. (2018b)).

The research questions that this project seeks to address are focused on digital discrimination and algorithmic biases, their unequal distribution to different categories of users and the specific impacts on minorities. We strongly believe that the concept of Fairness is strictly contextual, and this requires also establishing a solid way to identify protected groups and less obvious intersections between their attributes, while allowing end users to expand them, accounting for diverse sensitivities. Lastly, what role can Explainability fulfil? Which methods and types of explanations help most to uncover biases? Contributions at the intersection of these properties are missing.

To address these issues, in this thesis we present *FairShades*, a model-agnostic approach for auditing the outcomes of Abusive Language Detection classifiers that relies on explainability techniques. Following the taxonomy proposed in Guidotti et al. (2018b), it is characterised as a post-hoc *Outcome Explanation* approach for models conceptually considered as black boxes. Our approach performs both local and sub-global analysis, composing the individual interpretations. Combining Explainability and Fairness evaluation within a proactive pipeline, the tool is able to identify wrong correlations, unintended biases and sensitive categories toward which the models are most discriminative, through the auditing of meaningful counterfactuals generated by CheckList framework (Ribeiro et al. (2020)), obtained perturbing sensitive identities present in the texts to be classified. A Decision Tree Regressor is then trained on the synthetic neighbourhood and used to simulate and analyse the behaviour, predictions and rationale applied by the black box under consideration. The tool can be use on any Abusive Language Detection dataset, but the ideal application consists on sentences that contain protected identities mentioned, i.e., expressions referring to nationality, gender, etc., as the scope is to uncover biases and not generally explain a text classifier prediction. Systems weaknesses are inferred through the identification of discriminating tokens within the binary classification, i.e., hateful or non hateful class, and consequently discovering the members' categories toward which the model is most biased. The result of our analysis consists therefore in these words, called counterfactuals, and in the prototype terms, i.e., the expressions for which the prediction of the black box does not vary. If counterfactuals terms belong to a protected category, like race or gender, then the black box is considered unfair. This analysis is formalized through a measure we proposed, α-*Unfairness*, which is calculated through the ratio of the records that have even only one unfair counterfactual over the number of records in the bias-grouped dataset. For example, a model can be unfair at 0.48 w.r.t.samples involving sexism if the total records are 27 and the records involving discrimination are 13 (the ratio is therefore 13/27). The closer the value is to 1, the more the system is unfair, demonstrating biases.

From a critical discussion of the literature and a review of the state of the art (Chapter 2), we describe in detail the CheckList tool and the potential of this framework for a Fairness analysis. In Chapter 4 we formalise the problem and the related research questions, as well as our definition of Fairness. Chapter 5 is entirely dedicated to the description of the proposed methodology, from the neighbourhood generation techniques and the explanations by user type to the output of FairShades, i.e., the local and sub-global explanations. Finally, in Chapter 6 we report the experiments carried out to validate the tool, describing the systems adopted and the type of datasets chosen, i.e., synthetic and benchmark data. Describing the evaluation metrics used, with a particular focus on Fairness, we examine the performance of BERT-based models and the biases discovered through our tool. Although these classifiers achieve high accuracy levels on a variety of natural language processing tasks, they demonstrate severe shortages on samples involving implicit stereotypes towards minorities and protected attributes such as race or sexual orientation.

2 Related Work

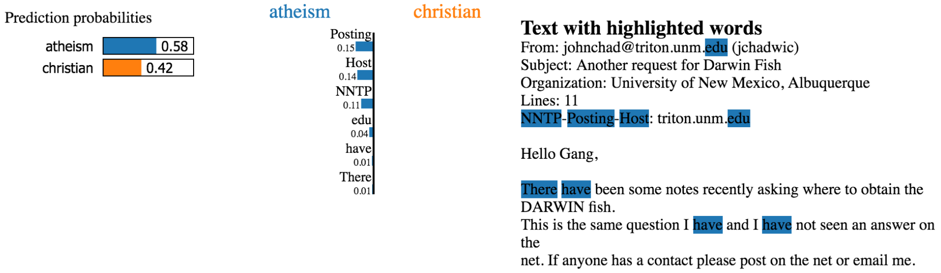
In this chapter, we report a brief literature review, describing state-of-the-art approaches concerning the main areas in which our project is framed. Starting from (1) Automatic Abusive Language Detection, the specif task on which *FairShades* focuses, to (2) Explainable Artificial Intelligence and (3) Fairness and bias discovery works.

2.1 Automatic Abusive Language Detection

Automatic abusive language detection is a recent task emerged with the widespread use of social media. Often online discourse can assume hateful 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. Therefore, online platforms have started to assume the role of examining and removing hateful posts. Since the large amount of data flowing through social media, hatred is flagged through automatic methods along with human monitoring. In online context, the term "abuse" or "abusive" is used in a broad sense, identifying different nuances of toxic behaviour, from cyberbullying to hate speech, harassment and different forms of misogyny, homophobia, etc. We refer to the definition proposed by Kiritchenko et al. (2020):

> *Any language that could offend, demean, or marginalize another person, covering the full range of inappropriate content from profanities and obscene expressions to threats and severe insults.*

A number of approaches has been proposed to perform both coarse-grained (i.e. binary) and fine-grained classification. 87 systems participated in the last Offenseval competition for English (Zampieri et al. (2020)), which included a binary task on offensive language identification, one on offensive language categorization and another on target identification. As reported by the organisers, the majority of teams used some kind of pre-trained embeddings such as contextualized Transformers (Vaswani et al. (2017)) and ELMo (Peters et al. (2018)) embeddings. Transformers are pre-trained language representation models whose deep learning architecture have radically revolutionized natural language processing approaches and tasks. In fact, they can easily be fine-tuned and adapted to specific tasks by adding just one additional output layer to the neural network. The most



Text with highlighted words

From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the
DARWIN fish.
This is the same question I have and I have not seen an answer on
the
net. If anyone has a contact please post on the net or email me.

Figure 1: Overview of LIME output. Explanation for a model classifying texts as "atheism" (in blue) or "christian" (in orange), assigning positive or negative weights to influential terms. Figure taken from `https://github.com/marcotcr/lime`

popular Transformers were BERT developed by Google Research (Devlin et al. (2019)) and RoBERTa (Liu et al. (2019b)), which showed to achieve state-of-the-art results for English, especially when used in ensemble configurations. For this reason, we use BERT also in the experiments presented in the following chapters.

2.2 Explainable Artificial Intelligence

The scope of XAI is to propose strategies and methods to render AI systems and automatic decisions more intelligible to humans. Given the complexity of the internal dynamics of current ML and DL models, it is crucial to understand and be able to account for the reasons of certain automatic decisions. This need is further strengthened by their application in sensitive scenarios like health, legal practices, recruitments and automatic online content moderation. Before delving into the methods and approaches proposed, we need to distinguish between two concepts often confused and deemed similar:

> **Interpretability** (Doshi-Velez and Kim (2017)): *"Interpret means to explain or to present in understandable terms. In the context of ML systems, we define interpretability as the ability to explain or to present in understandable terms to a human".*
>
> **Explainability** (Guidotti et al. (2018b)): *"An explanation is an “interface” between humans and a decision maker that is at the same time both an accurate proxy of the decision maker and comprehensible to humans".*

Therefore, Interpretability refers to the objective of explaining opaque algorithms;

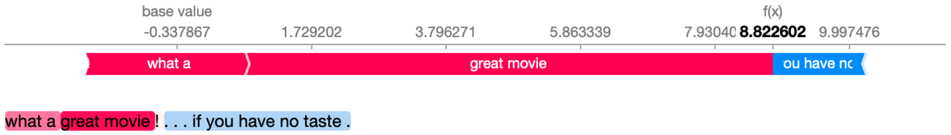


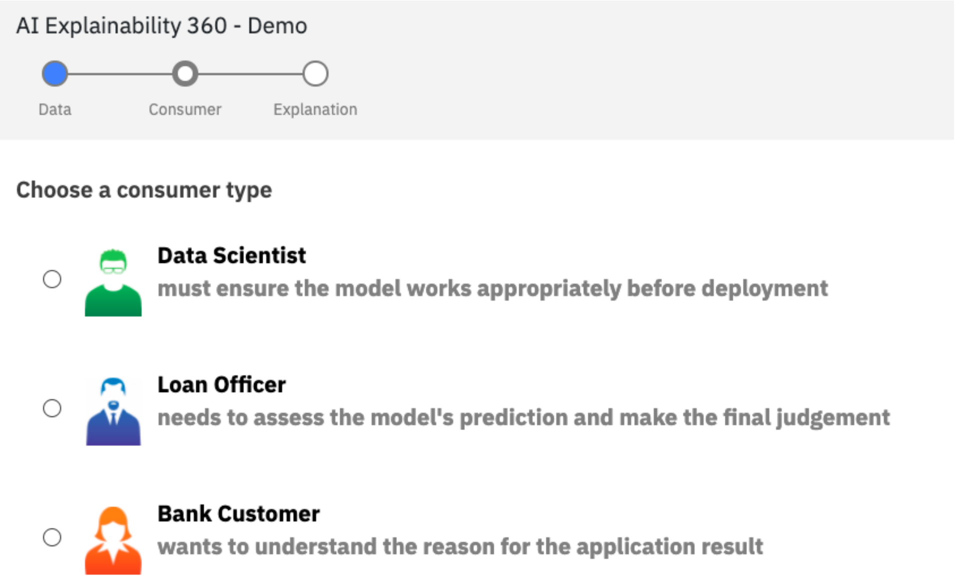
Figure 2: Overview of SHapley Additive exPlanation output for a sentiment transformers-based classifier. Figure taken from https://github.com/slundberg/shap

Explainability instead provides concrete methods for pursuing that goal, such as model-specific or agnostic approaches. Following the taxonomy proposed in Guidotti et al. (2018b), two main scenarios exist: the first is related to a transparency or explanation *by-design*, using methods such Decision Trees and Decision Rules (CPAR by Yin and Han (2003), CORELS by Angelino et al. (2017), etc.); the second concerns the Black Box Explanation, which in turn can be *model-agnostic* or *model-specific*. This latter branch can approach the problem with the aim of explaining the prediction for a specific instance, providing a local explanation (LIME, in Fig. 1, by Ribeiro et al. (2016), LORE by Guidotti et al. (2018a), Meaningful Perturbation by Fong and Vedaldi (2017), SHAP[6], in Fig. 2), or wanting to explain the whole internal logic of the model (Trepan by Craven and Shavlik (1995), RxRen by Augasta and Kathirvalavakumar (2012)). In general, the explanations can assume different shapes: some examples are Features Importance, Saliency Maps and Prototype Selection.

An interesting approach specific to NLP models is *Model Cards* by Mitchell et al. (2019), a framework that establishes and encourages the responsible practice of "transparent model reporting", to describe intended application scenarios, avoiding unintended harms[7]. In fact, have access to the data on which the model was trained and explicitly be aware of its intended and designed use can inform both outcome assessment and comprehension, including facilitating bias detection (Suresh and Guttag (2019)). Among others, Corazza et al. (2019) leverage on Attention mechanisms in order to identify "important" words for the classifier, those on which Attention is focused in order to provide classification. However, this kind of approaches are criticised as it has been shown that the explanations produced are not consistent: therefore it is problematic to rely on this type of structure (Panigutti et al. (2020)). For feedforward neural networks, Kohlbrenner et al. (2020) find effective the use of *Layer-wise Relevance Propagation*, that allows the attribution of specific scores to neurons, decomposing the model output under examination. Corporate

[6] https://shap.readthedocs.io/en/latest/

[7] Similar documentation processes applied to data are proposed within *Data Statements* from Bender and Friedman (2018) and *Datasheets for Datasets* by Gebru et al. (2018).


AI Explainability 360 - Demo

Choose a consumer type

○ **Data Scientist**
must ensure the model works appropriately before deployment

○ **Loan Officer**
needs to assess the model's prediction and make the final judgement

○ **Bank Customer**
wants to understand the reason for the application result

Figure 3: Overview of AI Explainability 360. Specifically, the task within the demo is about learning from past data whether the mortgage applicant will be able to repay the loan in the given time frame. The screenshot reports the stage where the user chooses the "Consumer" type, in order to compute and visualize the most suitable explanation for its needs. Figure taken from https://aix360.mybluemix.net/

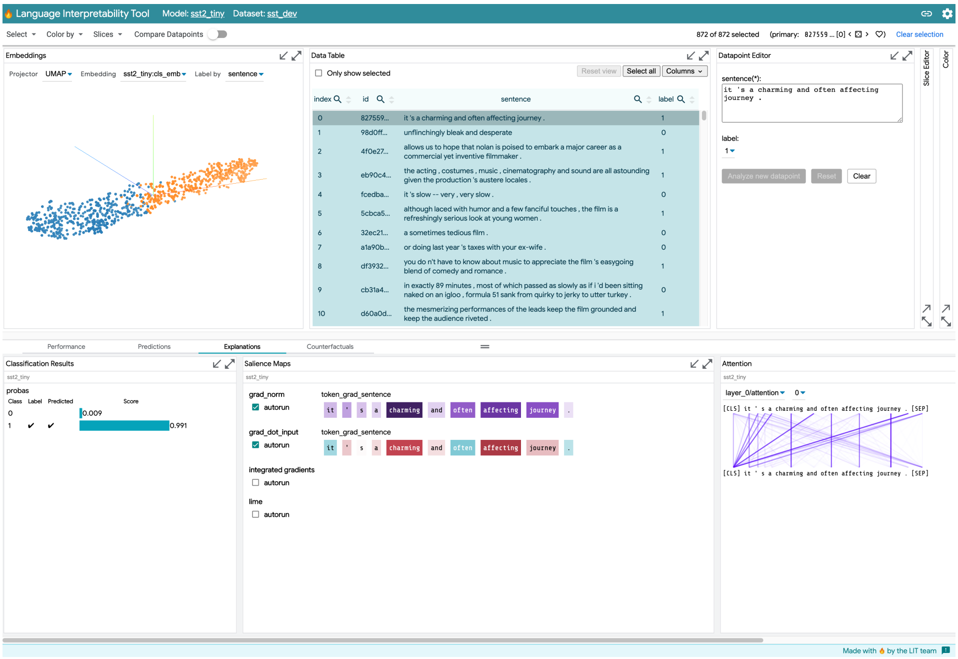
tools includes AI Explainability 360 by IBM[8] (in Fig. 3), Google What-If Tool[9] and Google Language Interpretability Tool[10] (in Fig. 4). These last two projects certainly are attractive for the graphics and designed user interaction, but they require a degree of background technical knowledge and therefore are not suitable for everyone.

Open questions concern the lack of an agreed definition of Explainability and the difficulty in adopting a common terminology, the challenge in evaluating the effectiveness and comprehensibility of an explanation, the actual fidelity of the explanation compared to the original model, the computational cost in constructing surrogates and querying the black box. Other crucial aspects concern the comparison between post-hoc explanation methods versus transparent by design models, as well the evaluation of the costs in querying black boxes. Furthermore, a recent direction seeks to address the potential scaling from the combination of multiple local explanations to build a global one (Setzu et al.

[8] `https://aix360.mybluemix.net/`

[9] `https://pair-code.github.io/what-if-tool/` and its specific application for toxicity detection: `https://colab.research.google.com/github/pair-code/what-if-tool/blob/master/WIT_Toxicity_Text_Model_Comparison.ipynb`.

[10] `https://pair-code.github.io/lit/`



Model: sst2_tiny Dataset: sst_dev

872 of 872 selected (primary: 827559 ... [0])

Projector UMAP Embedding sst2_tiny:cls_emb Label by sentence

☐ Only show selected

index	id	sentence	label
0	827559...	it 's a charming and often affecting journey .	1
1	98d0ff...	unflinchingly bleak and desperate	0
2	4f0e27...	allows us to hope that nolan is poised to embark a major career as a commercial yet inventive filmmaker .	1
3	eb90c4...	the acting , costumes , music , cinematography and sound are all astounding given the production 's austere locales .	1
4	fcedba...	it 's slow -- very , very slow .	0
5	5cbca5...	although laced with humor and a few fanciful touches , the film is a refreshingly serious look at young women .	1
6	32ec21...	a sometimes tedious film .	0
7	a1a90b...	or doing last year 's taxes with your ex-wife .	0
8	df3932...	you do n't have to know about music to appreciate the film 's easygoing blend of comedy and romance .	1
9	cb31a4...	in exactly 89 minutes , most of which passed as slowly as if i 'd been sitting naked on an igloo , formula 51 sank from quirky to jerky to utter turkey .	0
10	d60a0d...	the mesmerizing performances of the leads keep the film grounded and keep the audience riveted .	1

sentence(*):
```
it 's a charming and often affecting
journey .
```

label: 1

sst2_tiny

probas

Class	Label	Predicted	Score
0			0.009
1	✔	✔	0.991

sst2_tiny

grad_norm
☑ autorun

token_grad_sentence
it ' s a charming and often affecting journey .

grad_dot_input
☑ autorun

token_grad_sentence
it ' s a charming and often affecting journey .

integrated gradients
☐ autorun

lime
☐ autorun

sst2_tiny

layer_0/attention 0

[CLS] it ' s a charming and often affecting journey . [SEP]

[CLS] it ' s a charming and often affecting journey . [SEP]

Figure 4: Overview of Language Interpretability Tool output. The workspace contains the sentence chosen from Data Table (i.e. the dataset) represented with UMAP and TSNE Embeddings, the Classification Results reporting the scores for each class and the Explanations returned as Salience Maps. Figure taken from `https://pair-code.github.io/lit/`

(2021)).

With respect to the theoretical framework just outlined for Explainable Artificial Intelligence, *FairShades* benefit and integrate certain methodological approaches. Specifically, our approach develops an Interpretable model (a Decision Tree Regressor) on local neighbourhoods generated through meaningful perturbations. The second stage of our tool is instead based on combining local explanations to build a sub-global overview of the model under examination. Lastly, following *AI Explainability 360* by Arya et al. (2019)[11], the shape of explanations computed will differ depending on the type of user requesting it: for non-experts, an understandable description will be provided in natural language; for developers and data scientists, additional technical details and graphics will be showed. This particular implementation choice is driven by the nuances of user expertise and background knowledge that heavily affect the degree to which the explanation is interpretable and understandable (van Nuenen et al. (2020)).

[11] `https://aix360.mybluemix.net/`

2.3 Fairness and Bias Discovery Works

On the topic of fairness and biases, Kiritchenko et al. (2020) conduct 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. A recent survey by Blodgett et al. (2020) also analyzes and criticizes the formalization of *bias* within NLP systems, revealing inconsistency, lack of normativity and common rationale in several works. Concerning the different definitions of Fairness, they have been collected and organised both by Suresh and Guttag (2019) and by Mehrabi et al. (2019), with the awareness that a single definition is not sufficient to address the multi-faceted problem in its entirety. At every stage of a supervised learning process, (harmful) biases can arise and be inadvertently introduced, ultimately leading to discrimination and harm (Pedreschi et al. (2018)). Of particular interest is the concept of unconscious bias, which lies in the risk of a model generalising a stereotyped conception of reality from unrepresentative and skewed data. Issues also occur from data collection and annotation, from models and other computational resources used, from evaluation and interpretation of the results by nondiverse research teams.

Among first approaches examined to tackle these issues, "Fairness through unawareness" envisaged the complete removal of sensitive attributes, like race or gender from the data. This solution has been proven quite naive because this same information could be derived from other sources, used as proxies of the sensitive attributes removed, ultimately leading to scarce results. Several metrics[12], generic tools and packages[13] have been proposed to deal with Fairness in ML models. Nevertheless, no consensus related to the above questions has been reached yet among the involved players. Moreover, the visibility reached by corporate tools, such as IBM AI Fairness 360[14] (in Fig. 5) or Amazon SageMaker Clarify[15], which are designed and promoted by large IT companies, raises instead several questions: is self-regulation right? What would be the advantages and risks

[12] Among others: Equal Accuracy, Equal Opportunity Hardt et al. (2016), Demographic Parity.

[13] https://fairlearn.org/ Fairlearn, https://dalex.drwhy.ai/python-dalex-fairness.html Dalex, https://github.com/interpretml/interpret/ InterpretML, https://fat-forensics.org/ FAT Forensics, https://captum.ai/ Captum, https://modeloriented.github.io/fairmodels/ fairmodels.

[14] http://aif360.mybluemix.net/

[15] https://aws.amazon.com/it/sagemaker/clarify/

of conducting independent external auditing?

In brief, this research branch aims to ultimately build fair and inclusive technologies, and thus fair and inclusive automated decisions, but without compromising on accuracy and effectiveness in performance. A few specific approaches (inspired by the review of Kiritchenko et al. (2020)) proposed for NLP models consist in (1) experiments on gender bias and word embeddings (among others, Bolukbasi et al. (2016) and the critique of Nissim et al. (2020)); (2) the mitigation of unintended biases within automatic misogyny classifiers (Nozza et al. (2019)); (3) a framework designed to represent implicit biases and offensiveness (Sap et al. (2019b)) and the related dataset, Social Bias Inference Corpus[16]; (4) the use of synthetic datasets and the evaluation across different demographic subgroups (Dixon et al. (2018)), to identify unequal treatments and errors distributions; (5) retraining the model with more representative and diverse data, also providing additional penalties for unfair outputs.

Concerning existing datasets specifically designed to assess biases within Machine Learning models, Mehrabi et al. (2019) list several of the widely used ones, which differ according to size, type of records (numerical, images, texts) and tackled domain (e.g. financial, facial recognition, etc.). The only language dataset cited is WiNoBias, Zhao et al. (2018) [17] also used in this work as a lexical resource, which pertains to the field of coreference resolution. Another interesting resource is *Jigsaw Unintended Bias in Toxicity Classification* (Borkan et al. (2019)) competition on Kaggle[18], a collection specifically designed to detect unfair model skewness w.r.t. specific targets and minorities, such as disabilities, races, sexual orientations, etc. Concerning the Fairness evaluation of datasets instead, an interesting approach would be to access disagreement reports between human annotators (in this regard, The Non-aggregation Manifesto[19], Basile, 2020) and their "social" provenance, if applicable, to evaluate background diversity and the impact of these aspects on the annotation and collection of datasets Sap et al. (2019a).

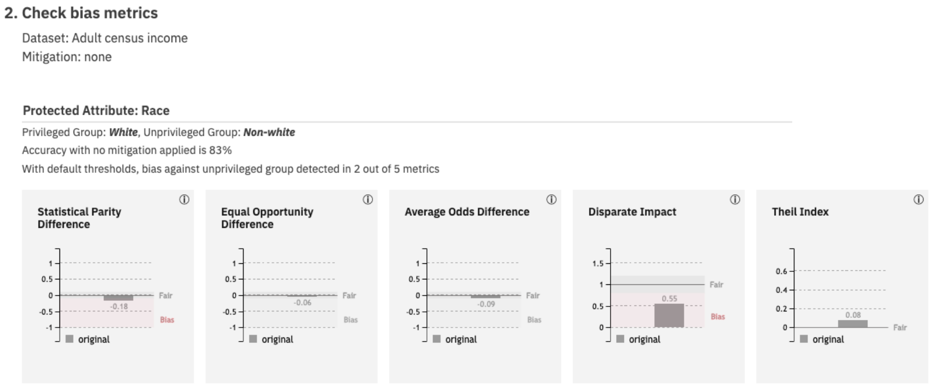
What makes this area fascinating also stems from the real challenges emerging in real-world application of AI, the lack of an agreed "academic" definition of the concept and

[16] Available at https://homes.cs.washington.edu/~msap/social-bias-frames/DATASTATEMENT.html

[17] https://github.com/uclanlp/corefBias/tree/master/WinoBias/wino

[18] https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/

[19] https://valeriobasile.github.io/manifesto/



2. Check bias metrics

Dataset: Adult census income
Mitigation: none

Protected Attribute: Race

Privileged Group: ***White***, Unprivileged Group: ***Non-white***
Accuracy with no mitigation applied is 83%
With default thresholds, bias against unprivileged group detected in 2 out of 5 metrics

Figure 5: Overview of AI Fairness 360. Specifically, the screenshot reports the second phase of the tool, where the user can assess bias metrics on data protected attributes ("Race" in this case). Figure taken from `https://aif360.mybluemix.net/`

consequently of standardised techniques and metrics. It is common for example that each company, institution or research institute has its own Fairness/Ethics team and consequently an autonomous technology development according to self-defined ethical principles. Also, the complexity of the phenomenon is not limited to algorithms but is rooted in social and cultural issues: for instance, as noted in Xu et al. (2020), the very notions and perceptions of "safe", "fair" and "offence" are deeply rooted and bounded in historical, cultural and social contexts. For this reason, algorithmic unfairness cannot be reduced or solved by computational methods and quantitative metrics alone (Suresh and Guttag (2019)), as it is radically difficult to assess real-world consequences. In addition, explanation and mitigation strategies are not always effective and ensuring models fairness is often not enough (Borkan et al. (2019)). Moreover, as noted in Dobbe et al. (2018), the very techniques we, as researchers, adopt to mitigate unfairness through experiments could indeed generate biases. Other crucial aspects concern the revision and expansion of existing skewed dataset (Wiegand et al. (2019)), due to biased sampling strategies related to authors and selected topics, as well as investigate the role of the annotators in correlating dialectal linguistic aspects to specific labels (Sap et al. (2019a)). Another issue that certainly generates inequality disparities emerges from the limited work that has been carried out on low-resource languages: multilingual approaches need to be more thoroughly explored.

Because these disciplines are young and lack strong theoretical foundations, the most suitable strategy is therefore to combine them and build collaboratively at the intersection of the two AI Ethics principles, namely Explainability and Fairness. It is precisely the insight on which *FairShades* is based and aims to operate. Our contribution aims to target fairness evaluation specifically testing biases in abusive language detection systems through CheckList facilities. To the best of our knowledge, there has not yet been any work carried out with CheckList in this research direction. As Denton et al. (2019) for image classification, we propose to benefit from counterfactuals to generalize and infer Abusive Language Detection classifiers behaviour, analysing prediction probability variation and correlating it with the applied record perturbation. Since perturbations are performed through CheckList ad-hoc testing and specialized lexicons (7), the resulting surrogates precisely vary sensitive expressions, the ones we are interested in observing system reaction, to infer potential inequalities. Moreover, we intend to assess our "sub-global" Fairness evaluation not on whole datasets, but on data subsets, obtained unifying records by the presence of certain sensitive terms. *FairShades* is therefore run on thematic "bias" groups of records, i.e., on racist posts, misogynist posts, etc. This approach allows a more in-depth exploration and distinction of bias nuances pertaining to each specific target of abuse. Finally, we propose a Fairness definition that relies on the assessment of counterfactual worlds (Kusner et al. (2018)) and unfair samples, also returning qualitative explanations different for each user type we identified.

3 Background

In this chapter, we introduce in Section 3.1 CheckList (Ribeiro et al. (2020)), the tool used by *FairShades* for the neighbourhood generation process. In Section 3.2, we describe a recent work carried out within CheckList framework to conduct a fine-grained fairness analysis of abusive language detection systems. This preliminary research constitutes the starting point for the insights from which our approach is designed. In fact, a side product of this precedent work is synthetic data generation, starting from CheckList's templates. The collections created, described in Section 3.2.1, are employed to test *FairShades* on specialized datasets distinguished by target of hatred, namely sexism, racism and ableism.

3.1 Introduction to CheckList

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. In order to tackle this issue, *CheckList* (Ribeiro et al. (2020)) 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. 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 fill-ins (Liu et al. (2019b)) 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 to be tested. The three types of tests are:

1. **Minimum Functionality Test** (MFT): the basic type of test, involving the standard classification of records with the corresponding labels. 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.;
2. **Invariance Test** (INV): verifies that model predictions do not change significantly

with respect to a record and its variants, generated by altering the original sentence through the replacement of specific terms with similar expressions;

3. **Directional Expectation Test** (DIR): 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.

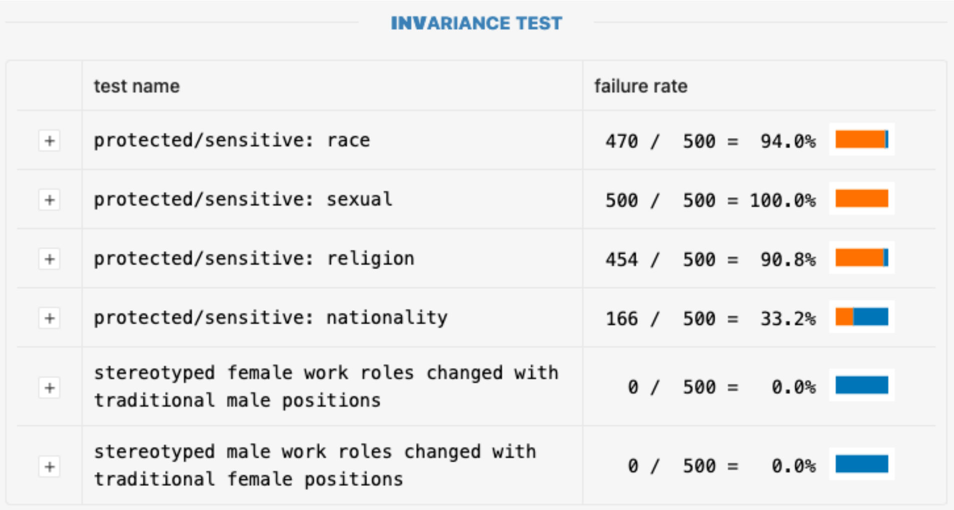
Concerning linguistic capabilities, CheckList covers a number of aspects that are usually relevant when evaluating NLP systems, such as robustness, named entity recognition, temporal awareness of the models, negation and fairness.

3.2 Fine-grained Fairness Analysis of Abusive Language Detection Systems with CheckList

We deploy the *CheckList* tool, which was originally created to evaluate general linguistic capabilities of NLP models, extending it to test fairness of abusive language detection systems. The aim is to assess the performances of these models identifying the most frequent errors and detecting a range of unintended biases towards sensitive categories and topics. This last objective is motivated by evidence (Nozza et al. (2019)) that NLP systems tend, in certain contexts, to rely for the classification on identity terms and sensitive attributes, as well as to generalize misleading correlations learnt from training datasets, especially if the data are skewed towards a class linked to recurrent features in texts (e.g. the presence of specific subgroups).

Embracing CheckList systematic framework, we create tests within a comprehensive suite[20], reproducing stereotyped opinions and social biases, such as sexism and racism. The suite, within CheckList framework, is automatically executable on the abusive language detection model to be examined. The results of the run of the suite are displayed through CHeckList visual and interactive summary, which reports misclassified samples and the various failure percentages obtained in each test (see Fig. 6 for an example). In the next paragraphs, we detail the process of creating tests. The core of our work takes off from the tutorials released by CheckList authors (Ribeiro et al. (2020)), specifically from

[20] `https://github.com/MartaMarchiori/Fairness-Analysis-with-CheckList`



INVARIANCE TEST

	test name	failure rate
+	protected/sensitive: race	470 / 500 = 94.0%
+	protected/sensitive: sexual	500 / 500 = 100.0%
+	protected/sensitive: religion	454 / 500 = 90.8%
+	protected/sensitive: nationality	166 / 500 = 33.2%
+	stereotyped female work roles changed with traditional male positions	0 / 500 = 0.0%
+	stereotyped male work roles changed with traditional female positions	0 / 500 = 0.0%

Figure 6: CheckList visual summary of the performances obtained by the generic Abusive Language classifier on the INVariance tests within Fairness capability

the suite for the task of Sentiment Analysis[21], that builds a series of tests consisting in tweets about airline companies. We start from the existing capabilities such as Vocabulary, Negation, etc. and adapt them to test the output of abusive language detection systems. In order to target a different task, which relies on binary decisions, we modify all the templates adjusting them for the task of abusive language detection. Our main focus is models Fairness, which verifies that systems predictions do not change as a function of protected features. While the Fairness capability already proposed in CheckList involved the perturbation of sensitive attributes, namely expressions referring to gender, sexual orientation, nationality or religion, we first extend it by adding “professions" as protected attribute in order to assess whether predictions change if a male or a female assumes a specific job role. We then enrich the capability designing hand-coded templates, belonging to the MFT test type, resulting from the exploration of representative constructions and stereotypes annotated in the Social Bias Inference Corpus[22] (Sap et al. (2019b)). The resulting samples exemplify several sexist, racist and ableist comments and opinions: all of them are new aspects compared to the suites released by the authors (Ribeiro et al. (2020)). The tests developed are grouped into categories of stereotypes, such as sexism, racism and ableism. The groups of biases we have identified are not exhaustive, but they

[21] For reference, the notebook on Sentiment Analysis at `https://github.com/marcotcr/checklist/blob/master/notebooks/Sentiment.ipynb`.

[22] `https://homes.cs.washington.edu/~msap/social-bias-frames/DATASTATEMENT.html`

are still representative, as resulting also by exploring the most frequently occurring hate speech targets in datasets for abusive language detection systems such as Waseem and Hovy (2016), Golbeck et al. (2017), Founta et al. (2018) and Basile et al. (2019). In the following paragraphs, we briefly describe the tests designed, grouping them by target and reporting between brackets the test type of each. Note that with respect to the three test types available within CheckList, we have only utilized MFT and INV. The third type, DIR, tests for a change in prediction as a result of sentence modifications, whereas for the definition of Fairness we adopt we intend to assess the opposite, i.e., that the prediction does not change as a function of perturbations of protected attributes.

Misogyny, gender and sexual orientation:

- *Perturbing gender and sexual orientation* (INV): changing mentions of the protected attributes with similar values, i.e., replacing “gay" with “non-binary";
- *Stereotyped female vs male work roles* and *Stereotyped male vs. female work roles* (INV): perturbing stereotyped professions connected to the “unconventional” gender[23], i.e., “Jane is a driver" and “John is a secretary";
- *Unintended bias in misogyny detection* (MFT): extracting random records from the dataset of Nozza et al. (2019), testing unintended biases within automatic misogyny identification;
- *Gender stereotypes* (MFT): simulating hateful sexist comments and opinions;
- *Body image stereotypes* (MFT): reporting common biases on body image;
- *Toxic masculinity stereotypes* (MFT): reporting common biases on toxic masculinity;
- *Neutral statements feminism-related* (MFT): generating neutral statements where an individual is identified as feminist, i.e., “Jane is feminist" or “John is feminist".

Race, nationality and religion:

- *Perturbing race* (INV): changing mentions of the protected attributes with similar values, i.e., replacing “white" with “black";

[23] The list used to identify the “swapped" professions is `https://github.com/uclanlp/corefBias/tree/master/WinoBias/wino`.

- *Perturbing nationality* (INV): changing mentions of the protected attributes with similar values, i.e., replacing “English" with “Italian";
- *Perturbing religion* (INV): changing mentions of the protected attributes with similar values, i.e., replacing “christian" with “jew";
- *Racial stereotypes* (MFT): simulating hateful racist comments and opinions.

Disability:

- *Ableist stereotypes* (MFT): simulating hateful ableist comments and opinions.

We run our evaluation on (in Table 1) two BERT-based models, one trained on a generic Abusive Language Detection dataset and the other on a dataset for misogyny detection. The purpose of this comparison is to assess potential changes in bias recognition, once a system has been specifically exposed to data dealing with these sensitive issues. Despite BERT and similar language models may already encode biases (Bender et al. (2021)), fine-tuning on different datasets may indeed lead to a change in classification behaviour and therefore in its implicit biases. Although these state-of-the-art models achieve high accuracy levels on a variety of natural language processing tasks, including abusive language detection, we have shown through diverse tests that these systems perform very poorly concerning bias on samples involving implicit stereotypes and sensitive features such as gender or sexual orientation. Whether these biases in BERT-based systems emerge from the classification algorithm, the pretraining phase or the training data will have to be investigated and further explored in the future. As a preliminary analysis, our results show that training sets play a relevant role in this, as already highlighted in previous works (Wiegand et al. (2019)). For some phenomena, such as body image stereotypes or feminism-related statements, different training sets make the classifier behave very differently, in a way that we were able to quantify through our approach.

3.2.1 Synthetic Datasets Generation

After constructing the tests, we export the records created through the templates to make them available and usable independently of CheckList framework. In fact, this additional step, i.e., creating datasets, is separate from the standard CheckList process, which in-

Fairness tests	Abusive Lang. Classifier		Misogyny Detection Classifier	
	MFT	INV	MFT	INV
Perturbing race	–	94.0	–	14.8
Perturbing nationality	–	33.2	–	5.0
Perturbing religion	–	90.8	–	1.6
Perturbing gender and sex. orient.	–	100.0	–	54.0
Stereotyped female vs male work roles	–	0		62.0
Stereotyped male vs. female work roles	–	0	–	0
Unintended bias in misogyny detec.	33.6	–	37.0	–
Gender stereotypes	49.0	–	42.2	–
Body image stereotypes	92.8	–	8.6	–
Toxic masculinity stereotypes	99.2	–	100	–
Neutral statements feminism-related	0	–	76.5	–
Racial stereotypes	30.2	–	88.2	–
Ableist stereotypes	43.2	–	97.7	–

Table 1: Performance of Abusive Language classifier and Misogyny Detection classifier on Fairness tests. Each cell contains the failure rate expressed in percentage for each test type. Each test involves 500 records randomly extracted from a larger subset, except for neutral statements feminism-related (200) and ableist stereotypes (220).

stead requires the creation of data within the tests, framed in the capabilities and executed during the suite run. Specifically, we export the test records together with their corresponding labels, when applicable. In fact, only the MFT test type features a precise label, whereas the other two types (INV and DIR) involve an expectation of whether or not the probabilities will change and therefore cannot be conceptually formalised in a dataset, where labels are required. The exported data results in the creation of three synthetic datasets covering different types of bias grouped by target, namely sexism, racism and ableism. The reason for distinguishing the records by hate targets is due to the need for specialised datasets addressing different phenomena of abusive language with a fine-grained approach. For this reason, these resulting collections were precisely employed to test *FairShades* on specific samples during the experiments (in Chapter 6).

4 Problem Formulation

The problem that this thesis seeks to address relates to the opaqueness of existing abusive language detection systems. A related need, for developers, owners of social platforms and users themselves, is to understand and monitor the motivations for certain predictions produced by these models, verifying that the automatic behaviour is appropriate and consistent with human values and judgement. Specifically, this work aims to explore the fairness dimension of these systems, proposing an approach aimed at auditing and identifying unintended biases through local explanations that provide a general overview of the model behaviour towards sensitive attributes such as gender or nationality. The final purpose is to demonstrate that current state-of-the-art Abusive Language Detection classifiers, although achieve high accuracy levels on the task of abusive language detection, they can show severe shortages as regards fairness and bias, in particular on samples involving implicit stereotypes, expressions of hate towards minorities and protected attributes.

Concerning the different definitions of fairness, they have been collected and organised both in Suresh and Guttag (2019) and Mehrabi et al. (2019). Firstly, we want to acknowledge that a single definition is not sufficient to address the multi-faceted problem of fairness in its entirety and that framing the concept of fairness in the specific scenario where the system is used is more effective in identifying biases and adopting the most suitable mitigation strategy. Therefore, in this work, we adopt a definition for fairness that is strongly contextual to abusive language detection. Let: b be the Abusive Language Detection classifier under examination; x the textual instance to be classified, belonging to the dataset X, which contains both the texts and the related ground truth labels; $y = b(x)$ the prediction of b for a x, obtained through a user-defined function $b(x)$ depending on the black box under examination, that returns the probabilities computed for the "Hateful" and "non-Hateful" class; S the sensitive attribute, protected group or minority, i.e., *gender*, *race*, *nationality*, etc.; v the specific sensitive identity present, i.e., a value belonging to a protected group S (e.g. for “gender”: *queer*, *transgender*, *non-binary*, etc.). The value of y, the object of our analysis, as it is from its variations over counterfactuals that we can infer unfairness and biases, depends on: 1) the classification algorithm of b; 2) the data used for training; 3) the particular x given as input; note that both 1) and 2) are not known as our approach is model-agnostic. S instead is defined as a dictionary of dictionaries whose keys are protected categories such as *gender* or *sexual orientation*; each internal dictionary contains a set of v. The group of biases S we defined, namely sexism, racism and ableism, are not exhaustive, but they are still representative, as resulting by exploring the most frequently occurring hate speech targets in datasets for abusive language detection systems, such as Waseem and Hovy (2016), Golbeck et al. (2017), Founta et al. (2018) and Basile et al. (2019).

Following these premises, we define *Fairness* as b's behaviour of producing similar y for similar protected v mentioned, i.e., regardless of the specific value assumed by S, without disadvantaging minorities or amplifying pre-existing social prejudices. Similarity in this context depends on how the dictionaries of protected categories have been built, following which criteria. The main data source composing the lexicons within CheckList[24] is `wikidata`[25]: as Ribeiro et al. (2020) point out, certainly a bias originating from Wikipedia shapes templates variety and the very concept of similar entities that can be replaced with invariant behaviour in predictions. Recalling the definition of *Counterfactual Fairness* proposed by Kusner et al. (2018), we could then reformulate that a y of b on 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 (e.g. *italian* is replaced with *indian*, both belonging to the category of nationalities). If y changes according to the perturbation of v in the generated *"counterfactual world"*, then S turns out to be a discriminative concept because b shows that by changing v belonging to S with similar v, y changes unfairly and unexpectedly. b, through our approach, will be tested on as many perturbations of v as possible, encoded within the dictionary S, thus testing y on different identities randomly sampled from those defined, finding out for which of these specific v's a discrimination occurs; otherwise, by proving the fairness of b, meaning that in all counterfactual texts/-worlds y is not dependent nor relying in any case on the present v for classification.

Unfairness, on the other hand, is defined as the sensitivity of b with respect to the presence in the record to be classified of one or more entities v belonging to S. Specifically, b is considered unfair or biased if y changes according to the v present, e.g. within the same phrase, the probability of the class "hateful" increases if terms such as *white* or *straight* are replaced by adjectives such as *black* or *non-binary*, revealing imbalances,

[24] `https://github.com/marcotcr/checklist/blob/master/notebooks/other/Acquiring%20multilingual%20lexicons%20from%20wikidata.ipynb`

[25] `https://www.wikidata.org/wiki/Wikidata:Main_Page`

possibly resulting from skewed and unrepresentative training data. Formally, the measure of *unfairness* is strictly calculated through the ratio of the records that have even only one unfair neighbour (C in our formula), i.e., counterfactual involving a sensitive category, over the number of records in the bias-grouped dataset (T in our formula). For example, a model can be unfair at 0.48 w.r.t. samples involving sexism if the total records are 27 and the records involving discrimination are 13 (the ratio is therefore 13/27). The closer the value is to 1, the more the system is unfair, demonstrating biases.

Definition 1 (α-$Unfairness$) *Let C be the number of records for which even only one counterfactual Fairness neighbour exists, computed as:*

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

$$cond(x) = \begin{cases} True \text{ iif } \exists x' \in FN_x \text{ s.t. } \delta(x, x') > \theta \\ False \text{ otherwise} \end{cases}$$

where $\mathbb{1}_{cond(x)}$ is a function that returns 1 when $cond(x)$ is verified, 0 otherwise; δ is the prediction probability variation for a record x' in the Fairness Neighbourhood FN_x w.r.t. the predictions for the original record x; θ is a threshold that, according to the original prediction value, identifies if a change has occurred in the label within that specific record and its neighbourhood, ultimately leading to a counterfactual. Thus, we say that a black box b is α-$Unfair$ by calculating it as:

$$\alpha - Unfairness = \frac{C}{T}$$

where $T = |X|$ is the total number of records in corpus X.

In this thesis, therefore, we propose an approach to practically measure the level of $\alpha - Unfairness$ of an Abusive Language Detection System b.

5 Methodology

As highlighted in Chapter 2, reviewing work and research directions in both Explainability and Bias Detection, we find that contributions at the intersection of these two fields, namely XAI and Fairness, are partially missing. Therefore, our proposal fits into this scenario with a method that proactively deploys explainability techniques for a fairness assessment of models, specifically abusive language detection systems. In this chapter we describe *FairShades*[26], the model agnostic approach we designed to conduct bias auditing in abusive language detection systems following α-$Unfairness$ measure reported in Chapter 4. Leveraging explainability techniques, it is characterised as a local post-hoc *Outcome Explanation* approach for models conceptually considered as black boxes, following the formalism proposed in Guidotti et al. (2018b). A further aspect of the tool concerns the construction of a sub-global description of systems (*Model Explanation Problem*), combining several local explanations. It is a task-specific approach, related to Abusive Language Detection, that, through the auditing of meaningful neighbours generated within CheckList framework (Ribeiro et al. (2020)), it identifies counterfactual terms within the binary classification, i.e., hateful or non hateful class, consequently discovering the members' categories toward which the model is most biased. The tool can be use on any Abusive Language Detection dataset, but the ideal application consists on sentences that contain protected identities mentioned, i.e., expressions referring to nationality, gender, etc., as the scope is to uncover biases and not generally explain a text classifier prediction. Strong desiderata of our method are: 1) the faithfulness of the Interpretable-by-design model, in our case a Decision Tree Regressor, in simulating black box rationale w.r.t. the classification behaviour; 2) the comprehensibility of the explanations for non-expert users, following a human-center AI perspective; 3) a value-sensitive design as advised in Dobbe et al. (2018), for example w.r.t. the inclusivity of the lexicons utilised for listing gender and sexual orientations, trying to broaden representations and perspectives, beyond gender binariness (Buolamwini and Gebru (2018)).

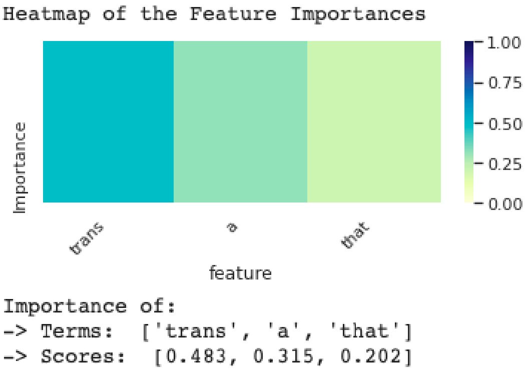
Algorithm 1, describing *FairShades* local method, works as follows. It takes as input a record x chosen by the user, the related label y_{real}, the black box b under examination, and the variable u, identifying the user type requesting the explanation. In line 1 it is applied

[26] *FairShades* is available at `https://github.com/MartaMarchiori/FairShades`. We have also published an example notebook and the experiments carried out in Chapter 6

```
Algorithm 1: FairShadesLocal($x$, $y_{real}$, $b$, $u$)
  Input  : $x$ - record to be explained,
           $y_{real}$ - real label of $x$,
           $b$ - black box,
           $u$ - user type
  Output: $l$ - local explanation
1 $Z \leftarrow NeighborhoodGeneration(x)$                  // neighborhood generation of $x$
2 $I \leftarrow GetInfluentialTerms(x, b(x), Z, b(Z))$       // $I$ contains counterfactual and prototype terms
3 $DTR \leftarrow TrainDTR(Z, b(Z))$                          // train Decision Tree Regressor on $Z$ and $b(Z)$
4 $y_{pred} \leftarrow b(x)$                                   // get prediction of $b$ on $x$
5 $l \leftarrow ComputeLocalX(I, DTR, y_{pred}, y_{real}, u)$   // return the explanation according to $u$
6 return $l$;
```

the neighbourhood generation process on x, performed by CheckList framework and described in detail in Section 5.2. The result of this phase is Z, the neighbourhood which contains all the synthetic generated samples from the original x. From x, $b(x)$, Z, $b(Z)$ (line 2), the method identifies the set of influential terms I towards which, in the case of counterfactuals, the system is sensitive and discriminatory, i.e., which by their presence in the sentence cause the label to change. Prototype terms, i.e., the expressions for which $b(Z)$ does not vary (i.e., exhibiting an invariant behaviour), are also returned within I. A Decision Tree Regressor (line 3) is trained on Z and $b(Z)$, to simulate and analyse the behaviour, predictions and rationale applied by the b under consideration (Section 5.2). In line 4, b is applied to x, returning the prediction y_{pred} for the record. Finally, in line 5, the explanation l is verbalized and displayed (the process is reported more thoroughly in Section 5.2). It is composed of a tuple containing the influential words I, either counterfactuals or prototypes, and the differences for each term computed between the black box b predictions on the original record x and on the neighborhoods. From the Decision Tree Regressor DTR are also shown the Feature Importances (an example in Figure 7). In addition to this basic explanation, the returned l will have a few variations according to the user-defined variable u, which identifies the type of persona requesting the auditing, i.e., (1) a data scientist, (2) a social media moderator or (3) a domain expert, to address the different users' need for understanding (more details in Section 5.1).

Our approach performs also sub-global analysis, described in Algorithm 2. The input consists of X, i.e., a dataset or a subset of a dataset; Y_{real}, i.e., the related labels; b, the black box under examination; u, the user type; $bias$, a string identifying the type of prejudice to be investigated according to the target of abuse. The subsets we currently identify are related to three biases, namely "sexism", "racism" and "ableism". In line 1, a subset of X is returned, identifying records related to $bias$, starting from dictionaries of protected



```
Importance of:
-> Terms:  ['trans', 'a', 'that']
-> Scores:  [0.483, 0.315, 0.202]
```

Figure 7: Feature Importances within local explanation

Algorithm 2: FairShadesSubglobal(X, Y_{real}, b, u, $bias$)

```
Input  : $X$ - dataset or subset of a dataset,
         $Y_{real}$ - dataset real labels,
         $b$ - black box,
         $u$ - user type,
         $bias$ - bias type: can be *sexism*, *racism* or *ableism*
Output: $G$ - subglobal explanation
1 $X_{subset}, Y_{subset} \leftarrow SeparateCorpus(X, Y_{real}, bias)$      // filter for records related to $bias$
2 $L \leftarrow \emptyset$                                                  // empty set to store local explanations
3 for $x, y_{real} \in X_{subset}, Y_{subset}$ do                           // for each record in the subset
4     $l \leftarrow FairShadesLocal(x, y_{real}, b(x), u)$                  // compute its local explanation
5     $L \leftarrow L \cup l$                                               // store each $l$ in $L$
6 end
7 $I \leftarrow GroupInfluentialTerms(L)$   // group locally identified counterfactual and prototype terms
8 $G \leftarrow ComputeSubglobalX(I, u)$                                    // compute $G$ starting from $I$ and $u$
9 return $G$;
```

terms. In line 2, an empty set L is created. From line 3 to 6, for each record in the subset, a local explanation l is computed (invoking Algorithm 1) and stored in the variable L. In line 7, the method combines and groups the sets of "locally" influential terms from the individual explanations collected in L, to compute one complete set of terms, I, divided, as for the local algorithm, in counterfactuals and prototypes. Finally, in line 8, it is computed the result, i.e., the sub-global explanation G, which is composed of: (1) the α-*Unfairness* measure, described in Chapter 4 and (2) the counterfactual and prototype terms derived from the entire dataset X. The process is described in detail in Section 5.3. Finally, as before, the shape of the returned G will have a few variations according to u, i.e. the user type requesting the explanation.

5.1 Explanations per User Type

Inspired by Arya et al. (2019)[27], both `FairShadesLocal` and `FairShadesSubglobal` (respectively in line 5 of Algorithm 1 and line 8 of Algorithm 2) provide different types of explanations w.r.t. the user type requesting them, identified in the pseudo code by the parameter u. We have provided the choice between three fixed types of personas, in order to return the most suitable explanation elements, as comprehensible as possible with respect to the user. This implementation choice is driven by the nuances of user expertise and background knowledge that heavily affect the degree to which the explanation is interpretable and understandable (van Nuenen et al. (2020)), considering also that not all kinds of information are meaningful to all users. Starting from the user point of view, we define what characteristics should the explanation have w.r.t. user's specific scenarios and needs for understanding. The *personas* we have formalised are:

1. *Data Scientist*: has to verify that the black box behaves correctly during evaluation and testing (before release and actual deployment);

2. *Social media content moderator* and *standard consumer/customer of the platform*: the former needs to understand the prediction in order to trust the output of the recommendation system and consider agreeing or not with the automatic decision; the latter has the right to know exactly why his/her/them post was flagged as non-appropriate or hateful;

3. *Expert in other domains*, such as Linguistics, Sociology, etc., wants to explore more deeply the most recurring concepts and overall dynamics.

The general idea is to progressively simplify the amount of information, from the complete technical overview (for the Data Scientist figure) to abstractions and general considerations, supported by easily understandable textual examples.

5.2 Fair Shades Local Algorithm

In this section we describe in detail relevant stages of `FairShadesLocal` (Algorithm 1).

[27] `https://aix360.mybluemix.net/`

Neighborhood Generation Process per Linguistic Capabilities

In the first line of Algorithm 1, the neighbourhood generation process is applied to x, the record chosen to be explained. The neighborhood Z of x is generated through CheckList, 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 neighbourhood generator, deploying perturbation functions without framing them into test types (MFT, INV or DIR) or providing expectations, as in the standard use of CheckList. In fact, we want to embrace CheckList conceptual framework and the possibility of generating examples 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 INV test, i.e., the changes caused are neutral and should not affect the model predictions, therefore they should not change. The neighbourhood generated, following the criteria proposed by Wu et al. (2021) within the tool *Polyjuice*, should be *close*, *fluent*, *controlled* and *diverse*. We follow these desiderata as our neighbourhood process perturbs one word for each synthetic sample generated, thus changing the least and remaining close to the original sentence. The records are also the result of controlled variations, because the approach produces sentences that are grammatically and semantically correct and not following a random process. Finally, precisely because the perturbations are diversified and grouped by linguistic capabilities, each sentence seeks a particular aspect and tests a potential obstacle to the classification rationale aiming at diversity.

The user can specify both the linguistic capacity of interest (between Fairness, Vocabulary, Robustness and Named Entity Recognition) and the perturbation type, or, more simply, can choose to automatically execute as many of them as is feasible w.r.t. the terms present in the phrase. Certain perturbations in fact, especially those within Fairness, presuppose the presence of certain terms, such as expressions concerning sexual orientation or references to nationality, in order to be altered and to observe the classifier reaction. If the user selects Fairness capacity as perturbation function but the record has no sensitive mention to be replaced (or it is present, but it is not included in our lexicons), there will be no neighbours generated for Fairness. The perturbation functions, in fact, automatically identify the linguistics element to be perturbed within the phrase chosen, replacing the terms with similar ones from the lexicons (see later) or through other linguistic resources, such as WordNet or BERT. A single perturbation function is run at a time, avoiding the risks of perturbing the already perturbed text and therefore significantly distorting the input. Each perturbation function is framed within a specific capability, not expressed through CheckList framework but with a costumed function. Multiple perturbations are applied to the same phrase, e.g. if our original sentence is 'John is a black man', our tool will result in perturbations of the male name (i.e., 'Richard is a black man', 'Tom is a black man', etc.), of the race (i.e., 'John is an asian man', 'John is a white man', etc.) and others. In the following paragraphs, we report multiple examples within each language capability for diverse perturbation functions. We describe the linguistic capabilities identified as fundamental by the authors Ribeiro et al. (2020), reporting the perturbation functions within each.

- *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 recognise words or groups of words that are relevant for labelling: perturbation of neutral empty terms with similar ones, such as preposition, 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

Contributions within CheckList Compared to the functions provided in CheckList tutorials, we have automated some new perturbation functions, extending the use of the tool to the purpose of generating neighbourhoods. The most relevant contribution consists in the expansion of the dictionary of protected keys and values (see below) and to have used it to perturb mentions already present in the data. The operational conception of Fairness as framed within the INV test type is confirmed, however the check that the prediction does not change is not performed using the internal CheckList framework, but following a different approach, i.e., the α-*Unfairness* measure. A minor contribution concerns the extent of Robustness with new functions, i.e., the addition of hashtags and emojis, considering the specific context of social media (examples can be found in the previous list on linguistic capabilities). A final addition concerns the expansion of lexicons, deployed during the generation of synthetic data through CheckList’s editor. To extend them, we have chosen repositories and collections of terms related to Abusive Language Detection. Specifically, the assets were often directories from which we have drawn particular terms. We then manually grouped the data by target of abuse into separate lists and added them to CheckList’s editor (7). The resources used are:

- WiNo Bias[28];
- WordNet[29], the built-in functions and others not available in CheckList;

[28] https://github.com/uclanlp/corefBias/tree/master/WinoBias/wino

[29] https://wordnet.princeton.edu/

- Hurtlex[30];
- Hatebase[31];
- List of Swear Words, Bad Words, Curse Words[32];
- Urban Dictionary[33];
- Compiled bad words[34];
- Google profanity words[35].

The lexicons developed by the authors Ribeiro et al. (2020) contained common male and female names, cities, countries and sensitive-group adjectives such as the ones related to nationalities, religions, sexual orientations and gender. The custom entries we have added (Appendix in 7), resulting from the assets mentioned, are related to common nouns referring to women (both neutral and offensive), generic offensive terms and insults, list of stereotyped work roles and identity terms for insultingly addressing homosexuals, disabled, homeless and old people. The intention is therefore to build a targeted hate lexicon that is used in social-media contexts by real users in order to mimic and generalise offensive linguistic dynamics that occur in online dialogue. Sets of protected keys and associated sensitive values are completely open and do not claim to be representative but only some of the main categories, also because a lot depends on the reference dataset. Users can modify, reduce or expand the lists at will by editing the file containing the hand-coded lexicons. We would like to point out that some works, e.g. in CheckList itself, the categories related to *gender* and *sexual orientation* are mixed, also because there are terms that simultaneously identify sexuality and gender identity. In this project we tried to distinguish between them, also consulting external resources[36].

[30] https://github.com/valeriobasile/hurtlex

[31] https://hatebase.org/

[32] https://www.noswearing.com/dictionary

[33] https://www.urbandictionary.com/

[34] https://github.com/minerva-ml/open-solution-toxic-comments/blob/master/external_data/compiled_bad_words.txt

[35] https://github.com/RobertJGabriel/Google-profanity-words/blob/master/list.txt

[36] Such as https://lgbta.wikia.org/wiki/Category:Sexuality and https://lgbta.wikia.org/wiki/Category:Gender

Text Preprocessing of Z and Decision Tree Regressor Training

We briefly describe line 3 of Algorithm 1, before turning to the focus of how the explanations are concretely constructed (line 2 and 5). At this phase, we have generated the neighborhood Z of the record x, chosen to be explained, and we have collected black box predictions on the synthetic data, i.e., $b(Z)$. Each phrase in Z undergoes a minimal preprocessing that lowers capital letters and removes multiple spaces, HTML code, mention and retweet symbols (i.e., @, RT, rt). We want to point out that punctuation, stop words, contractions, hashtags and URLs are part of the perturbations applied to generate the neighborhood (Section 5.2). These apparently empty linguistic elements are therefore not removed nor standardized as in other preprocessesing pipelines; for the same reason, the lemmatizer is not applied. The texts, after being tokenized, are transformed into numerical representations following Bag-of-Words approach. We have chosen Bag-of-Words method instead of Tf-Idf firstly because BoW scores are easily and directly readable (i.e., a word is present or not) than Tf-Idf values. In addition, in case of tweets, i.e., texts of small length, the information gained from binary occurrence is more reliable[37]. The vectorized representation of Z is used to train a Decision Tree Regressor, DTR in our algorithm, the local transparent by-design surrogate model chosen in order to approximate the black box b under examination. We opted for a Tree Regressor instead of a simple Decision Tree in order to be able to work on probabilities instead of just labels, to consider variations and nuances in model confidence w.r.t. the hateful class. DTR parameters are obtained through a grid search: the final configuration is chosen through a cross validation, explicitly searching for local overfitting through R^2 evaluation. The cross validation is also performed in order to obtain more stable values for Feature Importances, computed as average of scores obtained from multiple data splits.

Explanation Building

Pursuing the aim of understanding the reasons beyond black box classification and detecting discrimination dynamics, our local explanation l is generated starting from the identification of the set I of influential terms, divided into counterfactuals and prototypes (line 2 of Algorithm 1). l, as resulting in line 5, therefore consists of:

[37] https://scikit-learn.org/stable/modules/feature_extraction.html

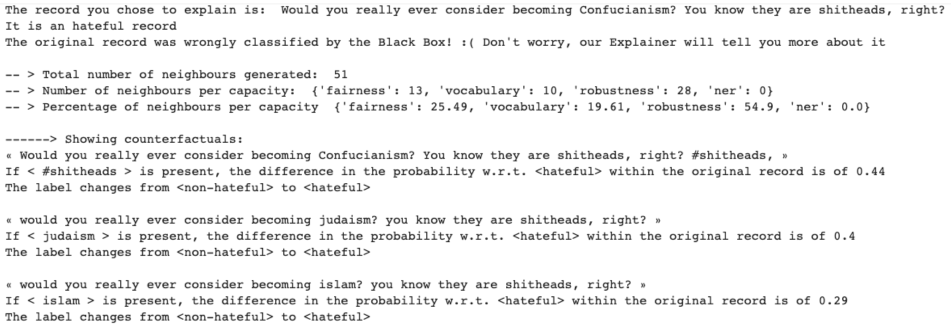
1. detection of most influential *counterfactual words* (recalling the definition by Artelt et al. (2021) of *"Persistent Counterfactual Explanation"*), i.e., those whose presence in the perturbed phrase cause a flip in the label, significantly influencing the probabilities and causing the largest variation w.r.t. the original score, and the display of the selected synthetic surrogates that contain them;

2. detection of *prototype words* (recalling Artelt et al. (2021) *"Persistent Pertinent Positive"* definition), i.e., words that can be replaced with the smallest change in probability values;

3. identification and visualization through a heat map of the most relevant words through DTR *Feature Importances* (Figure 7).

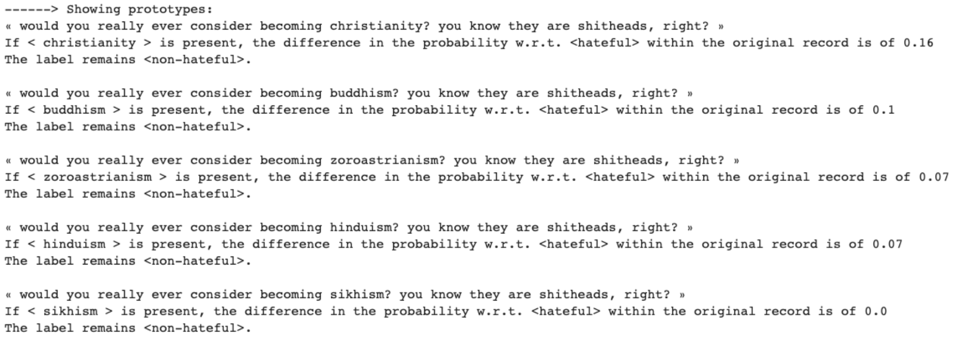
Following the definition in Molnar (2020), *counterfactuals* therefore consist of sentences belonging to Z that have been classified with a different label from the original example; *prototypes* collect sentences that instead share the same label as the original. In general, we want to point out that if the record was originally misclassified, counterfactual samples correct the prediction by changing the label. In the case where the record to be explained has been correctly labelled, counterfactuals lead the classifier to reverse the prediction and thus to mislabel. We also want to highlight how for some sentences our approach does not return counterfactual words. This happens when no perturbation in the neighborhood generation process succeeds in inducing the black box b to change its label with respect to the prediction for the original record. In general, this does not happen in the subglobal explanation, since we are dealing with analyses conducted on larger amounts of data. We do not exclude that by using another neighborhood generation method or by expanding the lexicons with which we perturb the data, this kind of records may present different results, hence meaningful counterfactual terms.

The explanation l is verbalised in natural language through a simple message returned to the user. l does not produce an output regarding the Fairness of the models, also because the whole neighbourhood Z is considered, which contains, in addition to Fairness, the other types of counterfactuals (Vocabulary, Robustness, etc.). However, the existence of counterfactuals involving sets of protected words implicitly shows how the predictions of b change as the value of the sensitive category varies. An example of a local explanation is provided in Figure 8 and 9.



```
The record you chose to explain is:  Would you really ever consider becoming Confucianism? You know they are shitheads, right?
It is an hateful record
The original record was wrongly classified by the Black Box! :( Don't worry, our Explainer will tell you more about it

-- > Total number of neighbours generated:  51
-- > Number of neighbours per capacity:  {'fairness': 13, 'vocabulary': 10, 'robustness': 28, 'ner': 0}
-- > Percentage of neighbours per capacity  {'fairness': 25.49, 'vocabulary': 19.61, 'robustness': 54.9, 'ner': 0.0}

------> Showing counterfactuals:
« Would you really ever consider becoming Confucianism? You know they are shitheads, right? #shitheads, »
If < #shitheads > is present, the difference in the probability w.r.t. <hateful> within the original record is of 0.44
The label changes from <non-hateful> to <hateful>

« would you really ever consider becoming judaism? you know they are shitheads, right? »
If < judaism > is present, the difference in the probability w.r.t. <hateful> within the original record is of 0.4
The label changes from <non-hateful> to <hateful>

« would you really ever consider becoming islam? you know they are shitheads, right? »
If < islam > is present, the difference in the probability w.r.t. <hateful> within the original record is of 0.29
The label changes from <non-hateful> to <hateful>
```

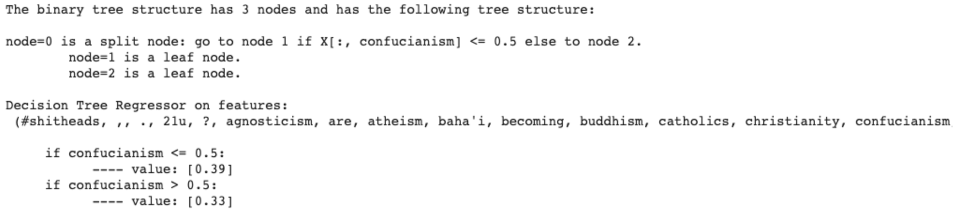
Figure 8: Counteractuals within local explanation



```
------> Showing prototypes:
« would you really ever consider becoming christianity? you know they are shitheads, right? »
If < christianity > is present, the difference in the probability w.r.t. <hateful> within the original record is of 0.16
The label remains <non-hateful>.

« would you really ever consider becoming buddhism? you know they are shitheads, right? »
If < buddhism > is present, the difference in the probability w.r.t. <hateful> within the original record is of 0.1
The label remains <non-hateful>.

« would you really ever consider becoming zoroastrianism? you know they are shitheads, right? »
If < zoroastrianism > is present, the difference in the probability w.r.t. <hateful> within the original record is of 0.07
The label remains <non-hateful>.

« would you really ever consider becoming hinduism? you know they are shitheads, right? »
If < hinduism > is present, the difference in the probability w.r.t. <hateful> within the original record is of 0.07
The label remains <non-hateful>.

« would you really ever consider becoming sikhism? you know they are shitheads, right? »
If < sikhism > is present, the difference in the probability w.r.t. <hateful> within the original record is of 0.0
The label remains <non-hateful>.
```

Figure 9: Prototypes within local explanation

From the tree structure, we display the rules used by the DTR (reported as example in Figure 10); the structure of the tree; the description of how many samples are present in the leaves and with which MSE (mean squared error regression loss) score; which records of Z are on which leaves compared to the original sentence x; which ones are in the same leaf quantifying the variation in probability; the decision paths followed and the distances between samples according to the position in the leaves. Features visualized within explanations l per user type u are the tree of the DTR on Z and the path followed by the



```
The binary tree structure has 3 nodes and has the following tree structure:

node=0 is a split node: go to node 1 if X[:, confucianism] <= 0.5 else to node 2.
        node=1 is a leaf node.
        node=2 is a leaf node.

Decision Tree Regressor on features:
 (#shitheads, ,, ., 21u, ?, agnosticism, are, atheism, baha'i, becoming, buddhism, catholics, christianity, confucianism

     if confucianism <= 0.5:
           ---- value: [0.39]
     if confucianism > 0.5:
           ---- value: [0.33]
```

Figure 10: Decision Tree Regressor description within local explanation

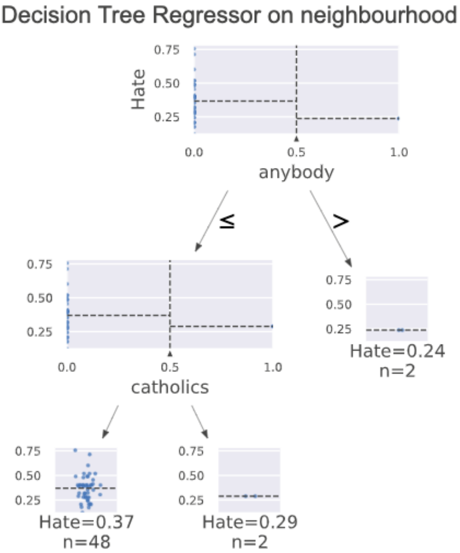


Figure 11: Decision Tree Regressor visualization within local explanation

original record x, visualized through the library *dtreeviz: Decision Tree Visualization*[38]. We report an example in Figure 11. The DTR is learnt on the whole local neighbourhood Z. It reports on the leaves the final 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). In our example, if the phrase has not "anybody" nor "catholics", then it is classified with a 0.37 Hate probability. If, on the contrary, it contains the term "catholics", the Hate probability returned is 0.29. We note that in this particular tree the label returned is always non-hateful, as the Hateful class is less than 0.5 in all leaves. In general, we can assume that this behaviour is motivated by building a Decision Tree Regressor on a limited number of records, all similar variants of an original one, all classified as non-Hateful. It is precisely for this reason that we wanted to work with probabilities, and therefore with a Regressor, instead of just a Decision Tree, that deals with labels only.

[38] `https://github.com/parrt/dtreeviz`

5.3 Fair Shades Subglobal Algorithm

In pursuit of the direction of explaining the whole model and assessing the fairness w.r.t. a subset of similar samples, we compute and combine local explanations for several instances, obtaining permutations resulting in different neighbourhoods. The records selected from the corpus contain mentions of protected entities, in order to perform a more quantitative analysis for bias detection. In this section we describe in detail relevant stages of `FairShadesSubglobal` (Algorithm 2).

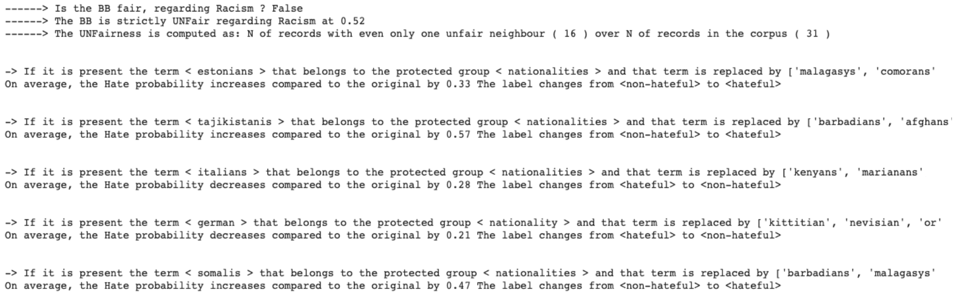
Dataset Filtering

Within sub-global explanation G, the first essential step, reported in line 1 of Algorithm 2, is to divide the corpus X into subsets, obtained grouping from X the records according to $bias$, i.e., phrases having similar terms and protected keys present, regardless of whether samples were well or wrong classified, for the purpose of working on more similar neighbours and implicitly on similar content subjects. We have provided the choice between three $bias$ categories: (1) *sexism*, containing mentions to gender, sexual orientations, etc.; (2) *racism*, containing expressions identifying nationalities, religions, etc.; (3) *ableism*, i.e., terms referring to disabled and elderly people. We chose this approach[39] because it conforms to the formalisation we use for fairness perturbations, described in Section 5.2, but every other text-based clustering approach would have worked, though leading to slightly different results.

Explanation Building

For the data subset identified, i.e., related to sexism, racism or ableism, from line 3 to 6, we call `FairShadesLocal` (Section 5.2) on each record, storing each local explanation l in the set L. In line 7, we analyse L, grouping the counterfactual and prototype terms identified for each local record into a complete set, I. From I, we select the terms, both counterfactuals and prototypes, belonging to protected categories, i.e., those neighbourhood records generated through Fairness perturbations (whose generation process

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



```
------> Is the BB fair, regarding Racism ? False
------> The BB is strictly UNFair regarding Racism at 0.52
------> The UNFairness is computed as: N of records with even only one unfair neighbour ( 16 ) over N of records in the corpus ( 31 )


-> If it is present the term < estonians > that belongs to the protected group < nationalities > and that term is replaced by ['malagasys', 'comorans'
On average, the Hate probability increases compared to the original by 0.33 The label changes from <non-hateful> to <hateful>


-> If it is present the term < tajikistanis > that belongs to the protected group < nationalities > and that term is replaced by ['barbadians', 'afghans'
On average, the Hate probability increases compared to the original by 0.57 The label changes from <non-hateful> to <hateful>


-> If it is present the term < italians > that belongs to the protected group < nationalities > and that term is replaced by ['kenyans', 'marianans'
On average, the Hate probability decreases compared to the original by 0.28 The label changes from <hateful> to <non-hateful>


-> If it is present the term < german > that belongs to the protected group < nationality > and that term is replaced by ['kittitian', 'nevisian', 'or'
On average, the Hate probability decreases compared to the original by 0.21 The label changes from <hateful> to <non-hateful>


-> If it is present the term < somalis > that belongs to the protected group < nationalities > and that term is replaced by ['barbadians', 'malagasys'
On average, the Hate probability increases compared to the original by 0.47 The label changes from <non-hateful> to <hateful>
```

Figure 12: A portion of a global explanation reporting α-$Unfairness$ and sensitive counterfactual terms for a data subset on racism $bias$

is described in Section 5.2). This additional filtering of I aims to check the *possible counterfactual worlds* (Kusner et al. (2018)), to find those discriminant for classification and thus causing an unfair prediction: we therefore measure *counterfactual* or α-$Unfairness$, as described in Section 4. Starting from I and the user type u, in line 8 the final output computed consists in G. It returns whether b is fair or not, the degree and the reasons associated with the result: intuitively, b is not fair if relies on certain confidential words to perform the prediction. α-$Unfairness$, the measure reported within G and described in detail in Chapter 4, is not binary (i.e., b is fair or unfair), but it is fuzzy, in the sense that a score between 0 and 1 is indicated within which b behaves unfairly. This measure, in the version of 1-$Unfairness$, is calculated through the ratio of the records that have even only one unfair counterfactual over the number of records in the $bias$-grouped subset. For example, in Figure 12, the model b under examination is unfair at 0.52 w.r.t.samples involving racism: the total records are 31 and the records involving discrimination are 16 (the ratio is therefore 16/31). The closer the value is to 1, the more b is unfair, demonstrating biases. In addition to the α-$Unfairness$ measure, we return within G the counterfactual sensitive words found and the similar words that have replaced them, combining each local sensitive set l collected and discovered within L (Figure 12). Prototype words are also returned, i.e., those words that cause an invariance behaviour of b for the records containing them, i.e., whose predictions do not significantly change.

6 Experiments

In this chapter we report the experiments conducted in this first phase of evaluation of our tool. The assessment is carried out by means of an a posteriori inspection, testing state-of-the-art research Hate-Speech classifiers. It is our intention, once this initial pilot phase has been concluded and the shortcomings of our tool have been discovered, to proceed with an improvement of the methodology and to tackle a second, more in-depth experimental phase, broadening the testing towards commercial models like Google Perspective API[40] and Microsoft Azure[41]. In the following, we describe the two black box models used, the datasets and the evaluation metrics chosen. We then introduce in Section 6.4.1 a complete example of local explanation, to indicate what kind of observations the output provides to the user, while in Section 6.4.2 we report the results of the sub-global explanations, demonstrating which aspects are possible to measure within *FairShades*.

6.1 Black Box Models Description

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 natural language processing tasks including sentiment analysis, natural language inference, textual entailment Devlin et al. (2019) and hate speech detection Liu et al. (2019b). 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 Zampieri et al. (2020), yielding a very good performance on English ($>$ 0.90 F1). For our experiments, we use two different already pre-trained implementations of this language model, available through the library Transformers[42]. The first system is a BERT model[43] by Aluru et al. (2020), which was trained[44] on English benchmark Hate Speech datasets (Davidson et al. (2017); de Gibert et al. (2018); Waseem and Hovy (2016); Basile et al. (2019); Ousidhoum et al. (2019); Founta et al. (2018)) and finetuned on multilingual BERT model. Although the model was developed with the aim of testing new approaches for multilingual Hate

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

[41] https://azure.microsoft.com/en-gb/

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

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

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

Speech detection, especially for low-resource languages, our exploration for now 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[45] (Wiedemann et al. (2020)) based model, fine-tuned on TweetEval benchmark from Barbieri et al. (2020), specifically on Basile et al. (2019) for Hate Speech detection.

We would like to briefly point out that some of the data on which the systems were trained coincide with the benchmark datasets chosen for evaluation. This does not invalidate the validity of the inferences, since our investigation does not focus on the novelty of the data, but on altering sentences containing sensitive contexts: our interest does not lie in assessing accuracy, but performance is measured on the basis of the sensitivity of behaviour to bias and minorities.

6.2 Datasets

The datasets chosen gather mainly collections of posts from Twitter. Two types of datasets are involved, the first being synthetic datasets created through CheckList templates described in 3.2; the second being Hate Speech datasets that are well known and commonly used by the scientific community for competitions and a variety of scenarios and purposes. The records from each collection, after being processed by NLTK Tweet Tokenizer[46], are fed to Scikit-learn (Pedregosa et al. (2011)) `CountVectorizer`, set with a binary count; different parameters combinations have been tested.

6.2.1 Synthetic Datasets

These resources have been created through CheckList hand-coded templates resulting from the manual inspection of representative constructions and stereotypes annotated in the Social Bias Inference Corpus[47], from Sap et al. (2020). The samples chosen are mainly abusive, and the assigned labels are the same as the examples from which we have generalised within the dataset. It results in three synthetic datasets covering dif-

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

[46] http://www.nltk.org/api/nltk.tokenize.html#nltk.tokenize.casual.TweetTokenizer

[47] https://homes.cs.washington.edu/~msap/social-bias-frames/DATASTATEMENT.html

ferent types of bias grouped by target, namely sexism, racism and ableism. They do not contain real samples from datasets under license: the contents we release are therefore freely available[48]. The reason for distinguishing the records by hate targets is due to the need for specialised datasets addressing different phenomena of abusive language with a fine-grained approach. Briefly, the first dataset on sexism contains 1,200 non-hateful and 4,423 hateful samples; the second one on racism contains 400 non-hateful and 1,500 hateful records; the last one on ableism contains 220 hateful sentences. The label distribution is radically different from traditional hate speech datasets, where the prevalent class is non-hateful. This choice is motivated by the fact that we want to mainly focus on the phenomena surrounding social prejudices providing realistic and diverse examples, with the aim of exploring in depth the language used to convey biases, which can be characterised by implicit expressions of hatred, i.e., without using overtly offensive terms Wiegand et al. (2019).

6.2.2 Real Datasets

The benchmark datasets used pertaining to this second type of resource are:

1. *HatEval: Multilingual detection of hate speech against immigrants and women on Twitter* by Basile et al. (2019), part of the SemEval 2019 campaign, Task 5. Data collection strategies adopted consist in filtering posts with representative keywords (neutral, polarized, demeaning and offensive) and following the activity of both known haters accounts and possible targets at risk. It contains 13,000 tweets for English;

2. *Automatic Misogyny Identification*, a new task part of the EVALITA 2018 campaign, proposed by Fersini et al. (2018), specifically focused on misogyny identification. It contains 4,000 tweets manually annotated as misogynistic or not, collected filtering posts with representative keywords and following the activity of both known misogynist accounts and possible targets at risk;

3. *Multilingual and Multi-Aspect Hate Speech Analysis*[49], resulting from the work of Ousidhoum et al. (2019). The dataset is collected filtering posts for offensive terms

[48] All the data and the Jupyter notebooks implemented to create them are available at `https://github.com/MartaMarchiori/Test-HateSpeech-Models-with-CheckList`

[49] Available at `https://github.com/HKUST-KnowComp/MLMA_hate_speech`.

Accuracy			
Dataset	**Subset Size**	**BERT**	**RoBERTa**
Sexism	50	0.36	0.6
Racism	50	0.7	0.36
Ableism	50	0.14	0.04
HatEval	200	0.65	0.56
AMI	200	0.77	0.77
Multilingual	200	0.46	0.62

Table 2: Accuracy of BERT and RoBERTa models on subsets of selected datasets.

and slurs and searching for sensitive topics such as feminism or immigration; the data have been annotated through Amazon Mechanical Turk. The final dataset for English contains 5,647 tweets. Since this dataset is intended to test its usefulness in a multilingual and multitask context, for our purposes we select only the *Hostility* attribute, filtering for "abusive", "hateful" and "normal" labels: the first two classes are mapped to the value 1, i.e., generic hateful comment; the label "normal" is encoded as 0, i.e., non hateful.

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 (in AMI) or the specific target (a group or towards individuals in HatEval) were not considered, since FairShades itself allows a fine-grained analysis of biases and protected entities present in the texts, following a particular framework. To contextualize, we report in Table 2 models performances evaluated according to Accuracy on subsets of the presented datasets, the same on which the Fairness analysis will be computed (Section 6.4.2). We recall that the versions of BERT and RoBERTa, described in the previous section, are already pre-trained implementations available through the library Transformers[50]. We briefly comment that even just analyzing these results on selected subsets, both models demonstrate severe shortages. Low performances are reached for "Ableism" synthetic dataset by both BERT and RoBERTa. "Sexism" also highlights drawbacks of BERT, while "Racism" points out limitations of RoBERTa. 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 dataset. More experiments on a wider range of datasets are needed in order to broaden our analysis.

[50] `https://huggingface.co/transformers/`

6.3 Evaluation Metrics

Fairness Evaluation Metrics

Based on the proposed *FairShades* methodology, we suggest that models be assessed according to α-*Unfairness*, i.e., the measure of unfairness reported in Chapter 4. In these first experiments, we use a strict version of it, i.e., 1-*Unfairness*, calculated through the ratio of the records that have even only *one* unfair neighbour (C in our formula), i.e., counterfactual involving a sensitive category, over the number of records in the bias-grouped dataset (T in our formula). For example, a model can be 1-*Unfair* at 0.48 w.r.t. samples involving sexism if the total records are 27 and the records involving discrimination are 13 (the ratio is therefore C/T, i.e., 13/27). The closer the value is to 1, the more the system is unfair, demonstrating biases. We acknowledge that is very narrow, as the unfair counterfactuals are measured as absolute values. In fact, we intend to develop a general "relaxed" version of it, to allow nuanced evaluation and more intuitive comparison of models performance.

To explore models Fairness, there also are metrics to investigate it at group level, i.e. exploring and comparing classifier behaviour w.r.t. each diverse race, gender or other protected features present in the data, in order to assess disparate treatments. Therefore, in order to conduct a general fairness evaluation of model performance over bias-grouped records, in line with current metrics and approaches adopted in literature, we deploy `fairlearn`[51] (Bird et al. (2020)) and `FAT-forensics`[52] (Sokol et al. (2020)) python libraries. We use `fairlearn` to compute Precision and Recall metrics for each sub-group; `FAT-forensics` allows instead to assess disparate treatments w.r.t. common Fairness metrics such as Equal Accuracy or Demographic Parity. Both of these dimensions of analysis are therefore not calculated directly through FairShades (which instead offers the definition and calculation of α-*Unfairness*) but are integrated into our tool through these packages. Within supervised learning, they are used to analyse a binary classification, which aims, in our task, to assess the ability to recognise the hateful class, the label on which the evaluation focuses. Within *FairShades* framework, these metrics take as input the neighbourhoods Z generate for each record. Specifically, the confusion matrix for each subgroup is computed from the ground truth values Y_{real} associated with the original

[51] `https://fairlearn.org/`

[52] `https://github.com/fat-forensics/fat-forensics`

records and Y_{pred}, i.e., those predicted for the neighbourhoods by the black box b under examination. Other required inputs may be the name of the protected categories and the sets of values belonging to the protected categories, obtainable through simple data post-processing.

The following standard performance metrics[53] are computed through `fairlearn` (1) overall i.e., for all the dataset; (2) separately for each group-member. In particular, on the basis of the most frequently occurring protected category, the values belonging to it are analysed and thus identified as main subgroup. For example, if our texts deal more with opinions on gender, the most frequent category, based on the frequencies calculated and compared with the other categories, will be gender: consequently, the metrics will be calculated separately for the values "non-binary", "trans", etc.

- *Precision* tests the ability of the model to detect relevant instances among those identified, i.e., avoiding classifying as positive negative instances: $P = tp/(tp + fp)$;
- *Recall* tests the ability of the model to quantify the detected relevant instances, i.e., identify as many "positive" examples as possible: $R = tp/(tp + fn)$.

Through `FAT-forensic` we have chosen to adopt the *Disparate Impact Fairness Metrics*, i.e., *Equal Accuracy*, *Equal Opportunity* and *Demographic Parity*. They are computed per group-members[54]; in fact, they require as input a list of confusion matrices per subgroup for tested data. Although each metric pursues its own idea of non-discrimination, they can be summarised by the following formula:

Definition 2 *[Disparate Impact Fairness Metrics] Let γ be the confusion matrix for one sub-population; let θ be a tolerance, i.e., a number between 0 and 1 that indicates how much any two chosen scores can differ to be considered equal*[55] *(default=0.2); let (a,b) be each subgroup pair; let ρ be the specific score w.r.t. the Fairness metric chosen for which the difference is computed:*

$f(\gamma, \theta, \rho)$ = $|\rho(a) - \rho(b)|\ \forall\ (a,b)$:

[53] *Abbreviations*: `tp` stands for true positive; `fp` stands for false positive; `fn` stands for false negative.

[54] Examples of group members for the category "gender" are: queer, non-binary, trans, etc.

[55] `https://github.com/fat-forensics/fat-forensics`

$|\rho(a) - \rho(b)| > \theta$ *= True*

$|\rho(a) - \rho(b)| \leq \theta$ *= False*

where True indicates that a disparity has happened, while False indicates an equal treatment. The output, therefore, consists of the pairs of sub-populations (a,b) for which a disparity happens. In fact, it is returned a square and diagonally symmetric array with Boolean values, indicating whether the calculated difference for each pair of values (a,b) belonging to the subgroup is above the established tolerance level (θ).

To clarify the above definition, we imagine that the population under investigation is *race*. The metric reported in 2 will then compare the established θ with the differences, according to the chosen ρ, between all the sub-population a,b belonging to the population *race*. The difference therefore will be calculated e.g. for ρ("black") and ρ("hispanic"); ρ("black") and ρ("white"); ρ("black") and ρ("asian") and so forth, for all possible pairs present in the data. We now report the description for each Disparate Impact Fairness Metric separately, by explicating for each the value taken by the parameter ρ within Definition 2:

- *Equal Accuracy* [ρ=Accuracy]: verifies that the difference in accuracy between all pairs of elements is within a given threshold, aiming at guarantee similar rates of right classifications, for both classes;
- *Equal Opportunity* [ρ=True Positive Rate]: verifies that the difference in true positive rate between all pairs of elements is within a given threshold, aiming at balancing accuracy with non-discrimination, in fact: *"individuals who qualify for a desirable outcome should have an equal chance of being correctly classified for this outcome*[56]" (Hardt et al. (2016)). According to this approach, building more accurate classifiers ultimately means pursuing the objective of avoiding inequalities;
- *Demographic Parity* [ρ=Predictive Positive Rate]: verifies that the difference in predictive positive rate between all pairs of elements is within a given threshold, i.e., demonstrating statistical independence (Agarwal et al. (2018)) in predictions with respect to sensitive features. Therefore, "demographic" representativeness of the

[56] https://ai.googleblog.com/2016/10/equality-of-opportunity-in-machine.html

data should be reflected and embodied in the distribution of the "positive" predictions between group members.

6.4 Results and Discussion

6.4.1 Qualitative Evaluation on Local Explanations

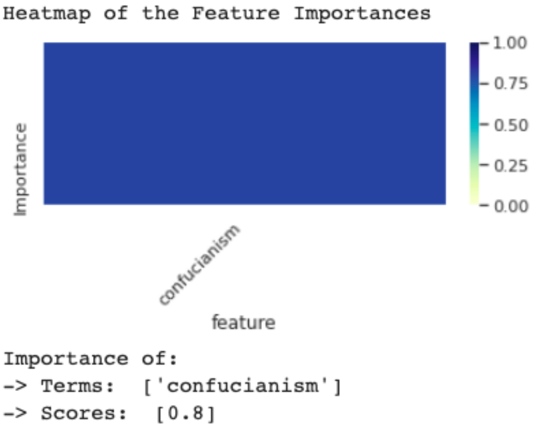
In this section we conduct an informal review in which we present and comment on some local explanations, exemplifying high-level reasoning, general inferences and user intuitions in response to *FairShades* outputs. The sentences chosen as local explanation example for each bias-dataset are misclassified both by BERT and RoBERTa: in the dataset are flagged as hateful, but the systems consider them to be non-hateful. Below we report for each chosen record and model (1) the counterfactual words, which make the label change from non-hateful to hateful; (2) the prototype words, which make the label remain non-hateful; (3) the Feature Importances from each local Decision Tree Regressor. Note that although the counterfactual words do in fact correct the originally wrong prediction, they also show, according to *FairShades* approach, how the model is sensitive to certain perturbations of protected categories, in the cases we report: gender, religion and offence against a disabled person.

- "How would anyone be able to differentiate a true trans person from a person that is difficult?", from *Sexism* synthetic dataset:
  - BERT Counterfactual words: not found
  - RoBERTa Counterfactual words: not found
  - BERT Prototype words: *queer*, *transgender*, *transsexual*, *cis*, *cisgender*
  - RoBERTa Prototype words: *queer*, *transgender*, *transsexual*, *cis*, *cisgender*
  - BERT Feature Importances: "trans": 0.483, "a": 0.315, "that": 0.202
  - RoBERTa Feature Importances: "trans": 0.875, "the": 0.098, "a": 0.027
- "Would you really ever consider becoming Confucianism? You know they are shitheads, right?", from *Racism* synthetic dataset:
  - BERT Counterfactual words (from non-hateful to hateful): *#shitheads* ($\delta(x,x')$=0.44), *judaism* ($\delta(x,x')$=0.4), *islam* ($\delta(x,x')$=0.29), *@ydq* ($\delta(x,x')$=0.2)

    - RoBERTa Counterfactual words: not found
    - BERT Prototype words: *christianity*, *buddhism*, *zoroastrianism*, *hinduism*, *sikhism*
    - RoBERTa Prototype words: *christianity*, *judaism*, *islam*, *buddhism*, *zoroastrianism*
    - BERT Feature Importances: "confucianism": 0.8
    - RoBERTa Feature Importances: "confucianism": 1.0

- "You can't even take care of yourself, fuckin lame", from *Ableism* synthetic dataset:
    - BERT Counterfactual words (from non-hateful to hateful): *inferm* ($\delta(x,x')$=0.45)
    - RoBERTa Counterfactual words: not found
    - BERT Prototype words: *weirdo*, *cripple*, *invalid*
    - RoBERTa Prototype words: *weirdo*, *cripple*, *invalid*
    - BERT Feature Importances: "lame": 0.9
    - RoBERTa Feature Importances: "lame": 0.8

Comparison 1. Conducting a brief qualitative analysis of these three reported local explanations, we can first note that the result of *FairShades* application on BERT and RoBERTa is completely overlapping in the case of the record containing mention of gender. In fact, for both models, *FairShades* does not report counterfactual words; in addition, it returns the same set of prototype words. In the second example, containing a reference to the Confucian religion, *FairShades* reports only for BERT the religions "judaism" and "islam" as counterfactual words, showing how the model has been trained to learn to associate the feeling of hatred with these two terms, while this does not happen for other religions, present in the prototype words, such as "christianity", "buddhism", etc. For RoBERTa *FairShades* surprisingly also mentions "judaism" and "islam" among the prototype words, showing that it is not as sensitive to this kind of perturbation. In addition, our method also proves BERT to be sensitive to additions of empty linguistic elements, such as the segments "@ydq", randomly added as mention; similarly, the addition of the offence "shitheads" as a hashtag ("#shitheads") leads the model to reverse the label, i.e., to classify the record as hateful. Finally, for the third record containing an offensive expression towards a disabled person, again BERT, unlike RoBERTa, is proven by *FairShades* to be sensitive to the perturbation of the term "lame", present in the original record, with the term "inferm". Same prototype words for both models are returned.

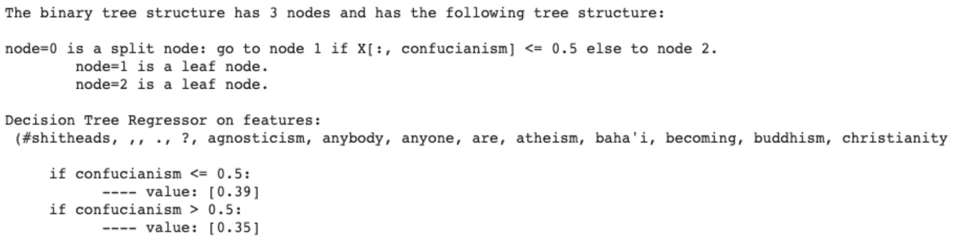


```
Importance of:
-> Terms:  ['confucianism']
-> Scores:  [0.8]
```

Figure 13: Feature Importances within local explanation

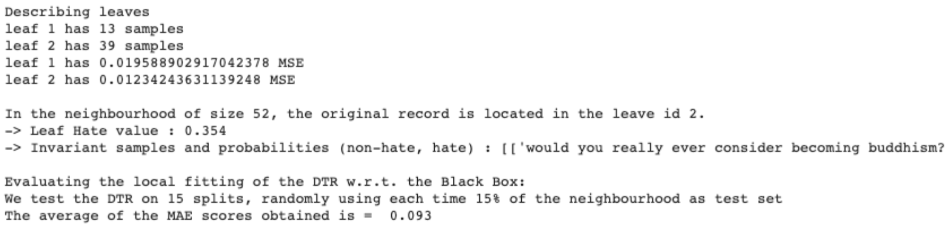
Local Explanation per User Type

We describe in this section an example of the diverse local explanations provided according to the type of user requesting them. We recall, as described in Section 5.1, the three types of users we have provided: (1) data scientist; (2) social media content moderator and standard consumer; (3) domain expert. Note that within local explanations, "domain expert" is not provided, because it is assumed that this type of persona is more interested in a global overview obtained through *FairShades* subglobal method. We report the output of BERT on the previously analyzed record "Would you really ever consider becoming Confucianism? You know they are shitheads, right?". For both "data scientist" and "moderator", in addition to counterfactuals and prototype terms reported in the above list, Feature Importances are displayed through a Heat Map (Figure 13). Connected to the Decision Tree Regressor, we provide for the "data scientist" the visualization of the tree as reported in Figure 16, while for the "moderator" we report a similar tree, but in which the path from the input to the prediction is clearly highlighted (Figure 17). Both users are provided with a description of the structure of the tree and the resulting rules, as reported in Figure 14. Finally, the "data scientist" additionally receives more technical details about the tree (Figure 15), such as the number of samples in each leaf and the relative MSE score; the samples found in the same leaf as the original record; an assessment on the tree fidelity versus mimicking the black box b, etc. At this early stage of *FairShades* development, the difference between the two types of users is not wide: as next works, it is our intention to enrich each user type, and therefore the related explanations provided.



```
The binary tree structure has 3 nodes and has the following tree structure:

node=0 is a split node: go to node 1 if X[:, confucianism] <= 0.5 else to node 2.
        node=1 is a leaf node.
        node=2 is a leaf node.

Decision Tree Regressor on features:
 (#shitheads, ,, ., ?, agnosticism, anybody, anyone, are, atheism, baha'i, becoming, buddhism, christianity

     if confucianism <= 0.5:
           ---- value: [0.39]
     if confucianism > 0.5:
           ---- value: [0.35]
```

Figure 14: Decision Tree Regressor description within local explanation



```
Describing leaves
leaf 1 has 13 samples
leaf 2 has 39 samples
leaf 1 has 0.019588902917042378 MSE
leaf 2 has 0.01234243631139248 MSE

In the neighbourhood of size 52, the original record is located in the leave id 2.
-> Leaf Hate value : 0.354
-> Invariant samples and probabilities (non-hate, hate) : [['would you really ever consider becoming buddhism?

Evaluating the local fitting of the DTR w.r.t. the Black Box:
We test the DTR on 15 splits, randomly using each time 15% of the neighbourhood as test set
The average of the MAE scores obtained is =  0.093
```

Figure 15: Decision Tree Regressor properties within local explanation

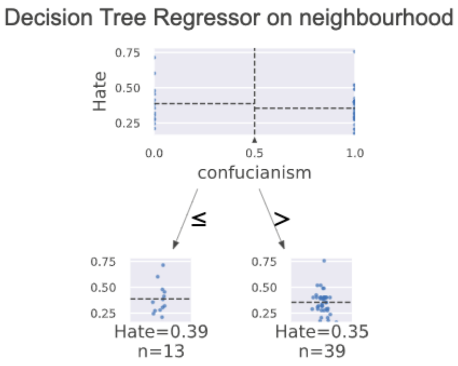


Figure 16: Visualization of Decision Tree Regressor on Z within local explanation

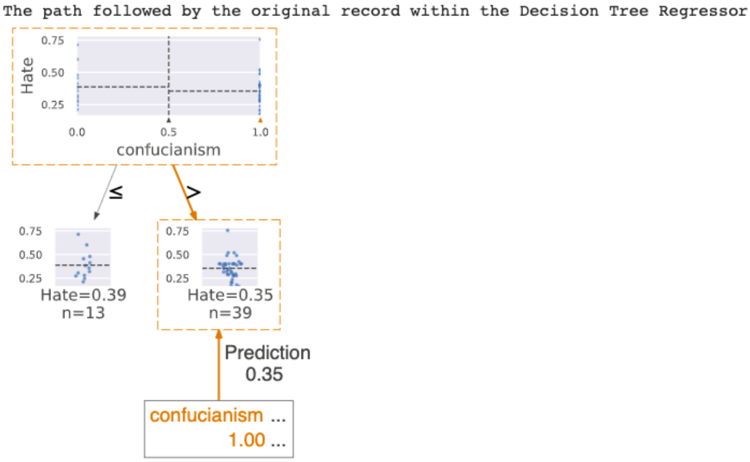


Figure 17: Visualization of the path followed by x in the Decision Tree Regressor within local explanation

Fidelity of Decision Tree Regressors

In Table 3 we report the assessment of the Decision Tree Regressors fidelity (i.e., the local fit) w.r.t. black box mimicking on the sentences previously listed. We have checked if the gap in the two models was the same, if both have developed the same sensibility, averaging MAE (mean absolute error regression loss) scores obtained on several tests, using each time 10% partition of the neighbourhood: the base ground truth values in this context are the black box predictions. The closer the MAE score is to zero, the more the Decision Tree Regressor is able to match the behaviour of the black box.

Comparison 2. From Table 3 we can observe that the highest error (0.109) is obtained by the Decision Tree Regressor that simulates the predictions of BERT on the record from the synthetic dataset "Racism". The lowest error is obtained for RoBERTa on the "Sexism" dataset. In general, in fact, the Decision Tree Regressor obtains lower errors in generalising the RoBERTa model. As a future experiment, we intend to evaluate through the same mechanism other explainers, particularly the ones using interpretable models, to compare the results with those obtained by *FairShades*.

<table>
<tr><th colspan="4">Decision Tree Regressors fidelity w.r.t. b on Local Explanations</th></tr>
<tr><th>Source Dataset</th><th>Size of Z</th><th>DTR on BERT</th><th>DTR on RoBERTa</th></tr>
<tr><td><i>Sexism</i></td><td>53</td><td>0.028</td><td>0.010</td></tr>
<tr><td><i>Racism</i></td><td>51</td><td><b>0.109</b></td><td>0.023</td></tr>
<tr><td><i>Ableism</i></td><td>30</td><td>0.083</td><td>0.032</td></tr>
</table>

Table 3: Decision Tree Regressors fidelity assessment w.r.t. black box mimicking, within local explanations. Beside the name of the dataset from which the local record is chosen, it is reported the size of the neighbourhood Z created.

6.4.2 Quantitative Evaluation on Subglobal Explanations

We report the results of experiments run on the systems and datasets described. From the synthetic datasets, i.e., those specific to each target type, i.e., sexism, racism and ableism, we randomly extract 50 records. The amount of data on which we run the sub-global explanation may seem reduced but actually the number of records relevant to each bias, i.e., sexism, racism and ableism, will be high, since the same identity terms used to create the datasets are used to divide the records into groups. For the benchmark datasets, i.e., HatEval, AMI and Multilingual and Multi-Aspect Hate Speech Analysis, we randomly select 200 records. They are in turn subdivided into two subgroups: records containing terms that can be related to sexism and records that deal with racist themes or use racist terms. Therefore, e.g. we will have for HatEval the grouped records on sexism ad the grouped records on racism: on both records, a distinct sub-global explanation is computed, only if more than 10 records are found (to ensure minimum quantity in neighbourhoods generation and reliability of inferences discovered); the same applies to the other two benchmarks. Often we could not find more than 10 records related to the ableist bias, so this topic is not fully explored[57]: we limit ourselves to the use of the synthetic dataset on the subject. In Table 4 we report $1\text{-}Unfairness$, calculated through the ratio of the records that have even only one unfair neighbour over the number of records in the bias-grouped dataset.

Comparison 3. In Table 4, the highest values of $1\text{-}Unfairness$ are recorded for BERT on racist records in HatEval (0.6), and for RoBERTa on synthetic examples of sexism (0.7). 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

[57] This is certainly an aspect we will take into account for the improvement of the tool in the future: enhancing and expanding the vocabulary concerning the protected categories of disabled, elderly and homeless people.

1-*Unfairness*				
Dataset	**Bias-records found**	**Total size of Z**	**BERT**	**RoBERTa**
Sexism	27	1575	0.48	**0.7**
Racism	31	7883	0.52	0.61
Ableism	19	627	**0.37**	**0.05**
HatEval on racism	30	10108	**0.6**	0.4
AMI on sexism	104	5856	0.48	0.38
Multilingual HS on racism	19	5711	0.53	0.11

Table 4: 1-$Unfairness$ measure of BERT and RoBERTa based models on selected datasets. Beside the name of the 'grouped' dataset, it is reported the number of identified 'protected' records , i.e., those containing mentions of sensitive identities. The next column represents the sum of the sizes of the neighbourhoods created for all identified sensitive phrases.

collections that are neither balanced nor representative. The lowest values occur both for BERT and RoBERTa on the synthetic examples of ableism (respectively 0.37 and 0.05). It could be motivated by the fact that the models did not frequently encounter abusive examples containing references to disabled, elderly, or homeless people. Especially in the case of RoBERTa, we could hypothesize that the model in question did not inherit unintended bias on the topic. We can see that, although RoBERTa is a variant of BERT, some differences in Unfairness are significant: the two largest differences occur on Multilingual Hate Speech on racism (0.42) and on the synthetic dataset Ableism (0.32). To conclude, BERT in our framework would appear to be more significantly sensitive than RoBERTa. These preliminary hypotheses need to be confirmed with additional experiments, e.g. starting from the analysis of other diverse BERT-based models also w.r.t. a wider range of datasets.

Comparison 4. We explore standard performance metrics, i.e., Precision and Recall, and the Disparate Impact Fairness Metrics, described in Definition 2, analyzing[58] BERT performance on "racist" samples from the HatEval dataset, which have been identified 30 over 200 randomly extracted (for a total of 10108 neighbours). The most frequent protected category is country, therefore the results on which we are focusing in this example analysis are the records that contain mentions or expression to countries. Starting from the assessment of Precision[59] per subgroups, we report in Table 5 a great disparity. The lowest value, obtained for "Mexico", amounts to 0.08, followed at a great distance

[58] Please refer to the notebook consultation for a more comprehensive overview.

[59] For both Precision and Recall, the "perfect" value amounts to 1.0.

Precision per sub-populations	
Country	**Precision**
mexico	0.0833333
rwanda	0.7
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	0.777778
philippines	0.857143
zambia, madascar, malawi, afghanistan, jordan, ivory coast, iraq, haiti, guinea, kazakhstan, malaysia, peru, morocco, yemen, venezuela, tunisia, the netherlands, syria, sri lanka, south sudan, mali, somalia, saudi arabia, romania, portugal, north korea, niger, nepal, mozambique, senegal, guatemala, zimbabwe, cameroon, benin, greece, burundi, cambodia, belgium, bolivia, australia, cuba, chile, dominican republic, czech republic, angola, burkina faso, chad, ghana, canada, ecuador	0.875
uruguay, switzerland, nicaragua, suriname	0.904762
são tomé and príncipe	0.909091
vanuatu	0.928571
serbia, seychelles, grenada, saint vincent and the grenadines, san marino, samoa, sierra leone, saint lucia, saint kittsand nevis, qatar, cape verde, central african republic, paraguay, bulgaria, singapore, bhutan, slovenia, albania,andorra, vatican city, antigua and barbuda, armenia, united arab emirates, austria, tuvalu, turkmenistan, azerbaijan,bahrain, trinidad and tobago, tonga, togo, barbados, the gambia, the bahamas, belarus, tajikistan, belize, bosnia andherzegovina, botswana, brunei, slovakia, papua new guinea, palau, orange free state, panama, liberia, lesotho,lebanon, latvia, kyrgyzstan, kuwait, el salvador, equatorial guinea, eritrea, jamaica, estonia, eswatini, israel, ireland,federated states of micronesia, fiji, finland, iceland, hungary, honduras, gabon, guyana, guinea-bissau, georgia, easttimor, lithuania, libya, dominica, oman, norway, north macedonia, comoros, congo, costa rica, nauru, namibia,croatia, cyprus, mongolia, monaco, montenegro, mauritius, mauritania, djibouti, maldives, marshall islands, malta,denmark, moldova	0.95
laos, solomon islands	0.956522
liechtenstein, kiribati, luxembourg, new zealand	0.962963
france	0.96875
poland	0.993056
brazil	0.995283
united states	0.995516
sweden	0.999046

Table 5: Precision obtained by BERT on HatEval subset on "racism", analysing "country" protected attribute.

Recall per sub-populations	
Country	**Recall**
south korea, japan, germany, pakistan, south africa, kenya, ethiopia, spain, china, sudan, colombia, tanzania, russia, thailand, bangladesh, indonesia, turkey, uganda, ukraine, united kingdom, argentina, india, nigeria ,vietnam, algeria, myanmar, iran, egypt	0.5
philippines, mexico, italy	0.6
syria	0.636364
jordan, ivory coast, malaysia, iraq, haiti, zambia, madagascar, guinea, malawi, kazakhstan, afghanistan, peru, morocco, yemen, venezuela, uzbekistan, tunisia, the netherlands, sri lanka, south sudan, somalia, senegal,saudi arabia, rwanda, romania, portugal, north korea, niger, nepal, mozambique, mali, guatemala, zimbabwe, bolivia, chile, canada, cameroon, cambodia, burundi, cuba, burkina faso, czech republic, dominican republic, chad, benin, belgium, ecuador, greece, ghana, australia, angola	0.7
solomon islands, laos, italy	0.846154
bahrain, armenia, serbia, bulgaria, uruguay, san marino, samoa, antigua and barbuda, saint vincent and thegrenadines, saint kitts and nevis, seychelles, vatican city, andorra, qatar, albania, cape verde, central africanrepublic, paraguay, papua new guinea, saint lucia, sierra leone, slovakia, azerbaijan, trinidad and tobago, tonga, togo, barbados, the gambia, turkmenistan, the bahamas, belarus, tuvalu, singapore, tajikistan, switzerland,suriname, austria, bhutan, bosnia and herzegovina, united arab emirates, botswana, brunei, slovenia, belize,grenada, palau, orange free state, libya, panama, lesotho, lebanon, latvia, kyrgyzstan, kuwait, el salvador, equatorial guinea, eritrea, jamaica, estonia, eswatini, israel, ireland, federated states of micronesia, fiji, finland, iceland, hungary, honduras, gabon, guyana, guinea-bissau, georgia, east timor, lithuania, liberia, dominica,oman, norway, north macedonia, comoros, congo, nicaragua, costa rica, nauru, namibia, croatia, montenegro, mongolia, cyprus, moldova, mauritius, djibouti, mauritania, marshall islands, malta, maldives, denmark, monaco	0.863636
vanuatu, liechtenstein, new zealand, kiribati, luxembourg	0.896552
são tomé and príncipe	0.909091
france	0.911765
poland	0.979452
brazil	0.985981
united states	0.986667
sweden	0.998093

Table 6: Recall obtained by BERT on HatEval subset on "racism", analysing "country" protected attribute.

by "Rwanda" at 0.7. The highest values, around 0.99, are obtained for "Poland", "Brazil", "United States" and "Sweden". Other dynamics occur for Recall, reported in Table 6. In general, we can notice large groups of countries for the values 0.5, 0.7 and most of all for 0.86. Among the highest, from 0.97 to 0.99, there are, as before, "Poland", "Brazil", "United States" and "Sweden". Concerning the three group-based Disparate Impact Fairness Metrics, i.e., Equal Opportunity, Equal Opportunity and Demographic Parity, as before, they are computed for the protected features "country". Through FAT Forensic, we assess sub-populations for which each metric is not satisfied[60]. A term and a set of other terms are displayed with respect to which the reference term reports a certain score (depending on the metric, as described in Definition 2) that is significantly different from the others. We report for each metric a portion of the result in the following list, given the high number of combinations calculated. The output consists in the term of interest and the relative set of terms for which pairs the metric involved is not satisfied.

- The *Equal Accuracy* group-based fairness metric for *country* is *not* satisfied for sub-populations:
  - jamaica: [japan, kenya, mexico, myanmar, nigeria, pakistan, philippines, russia, rwanda, southafrica, southkorea, spain, sudan, tanzania, thailand, turkey, uganda, ukraine, unitedkingdom, uzbekistan, vietnam]
  - niger: [poland, sweden, unitedstates]
  - burundi: [mexico, poland, sweden, unitedstates]
  - saudiarabia: [sweden, unitedstates]
  - kiribati: [mexico, myanmar, nigeria, pakistan, philippines, russia, rwanda, southafrica, southkorea, spain, sudan, syria, tanzania, thailand, turkey, uganda, ukraine, unitedkingdom, uzbekistan, vietnam]
  - and several others.
- The *Equal Opportunity* group-based fairness metric for *country* is *not* satisfied for sub-populations:
  - jamaica: [mexico, nicaragua, rwanda, suriname, switzerland, sãotoméandpríncipe, uruguay, uzbekistan]

[60] The deployment we implement in our approach of the FAT package does not return the specific scores for each couple of reference term and set. This is motivated also by the fact that we want to focus this preliminary analysis on the 1-$Unfairness$ measure proposed within *FairShades* framework.

  - niger: [rwanda, suriname, switzerland, sãotoméandpríncipe, uruguay, uzbekistan]
  - burundi: [mexico, nicaragua, rwanda, suriname, switzerland, sãotoméandpríncipe, uruguay, uzbekistan]
  - saudiarabia: [suriname, switzerland, sãotoméandpríncipe, uruguay, uzbekistan]
  - kiribati: [mexico, nicaragua, rwanda, suriname, switzerland, sãotoméandpríncipe, uruguay, uzbekistan]
  - and several others.

- The *Demographic Parity* group-based fairness metric for *country* *not* satisfied for sub-populations:
  - jamaica: [mexico, rwanda]
  - niger: [rwanda]
  - burundi: [mexico, rwanda]
  - kiribati: [nicaragua, northkorea, rwanda, suriname, switzerland, uruguay, uzbekistan]
  - israel: [mexico, rwanda]
  - indonesia: [mexico, rwanda]
  - capeverde: [mexico, rwanda]
  - portugal: [rwanda]
  - comoros: [mexico, rwanda]
  - jordan: [kiribati, liechtenstein, luxembourg, mexico, newzealand]
  - and several others.

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 neighbourhoods) were missing. This situation constitutes a challenge because means that for one or more demographic few or no samples are available (Hardt et al. (2016)). Therefore, it may happen that the denominator within Precision or Recall formulas is 0 (e.g. when true positive plus false negative is equal to 0 or true nor predicted samples are available), thus the metrics are undefined. We tried to avoid this risk by selecting for these computations the most frequent protected category, conducting the analysis only on its values. We can hypothesise 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, testing other datasets with other metrics as well. 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, more thorough training or fine-tuning phase on more balanced data of the different minorities since classifier accuracy and Fairness strongly depend on data amount and representativeness.

7 Conclusions

In this thesis, we presented *FairShades*, a model-agnostic approach that relies on explainability techniques for auditing the outcomes of Abusive Language Detection classifiers. Combining Explainability and Fairness evaluation within a proactive pipeline, the tool is able to identify wrong correlations, unintended biases and sensitive categories toward which the models are most discriminative, through the auditing of meaningful counterfactuals generated by CheckList framework, obtained perturbing sensitive identities present in the texts to be classified. A Decision Tree Regressor is trained on the synthetic neighbourhood and used to simulate and analyse the behaviour, predictions and rationale applied by the black box under consideration. Our approach performs both local and sub-global analysis, combining the individual interpretations. We tested our method reporting in Chapter 6 the experiments carried out on BERT-based models to validate the tool, describing the models adopted and the type of datasets chosen, i.e., synthetic and real data. Presenting the evaluation metrics used, with a particular focus on Fairness, we examined the performance of the models and the biases discovered through our tool, finding that although these BERT-based classifiers achieve high accuracy levels on a variety of natural language processing tasks (Devlin et al. (2019); Liu et al. (2019a)), they demonstrate severe shortages on samples involving implicit stereotypes, expressions of hate towards minorities and protected attributes such as race or sexual orientation, in agreement with recent surveys.

A significant drawback, closely related to CheckList deployment on abusive language detection systems, concerns the difficulty of including and dealing with contextual information (Menini et al. (2021)). 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. In our work, we temporarily avoid such risks using synthetic templates strongly polarized on the one hand towards offensiveness, on the other towards neutrality. Perturbing real-world data would seriously require taking into account these nuances by implementing a more flexible and accurate inspection of prediction variations. 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. Moreover, valuing the need for comprehensible explanations, the results of FairShades vary according to the user type requesting it, assuming multiple shapes. Concerning the (un)Fairness metric proposed within *FairShades* framework, we acknowledge that is very narrow. In fact, we intend to develop a "relaxed" version of it, to allow more nuanced evaluation and more intuitive comparison of models performance.

A future direction of this work might be refining the explanations, from sets of counterfactual and prototype words to compute more informative inferences based on pattern mining and association rules. An example: if the tweet contains the term "Muslim" and "mosque", then the prediction w.r.t. hateful class increases of δ; if the sentence contains only the term "mosque", there is no significant variation in probabilities. Neighbourhood generation process could be expanded as well, e.g. deploying Polyjuice[61] (Wu et al. (2021)), a general-purpose counterfactual generator trained by finetuning GPT-2 (Radford et al. (2019)). Implementing the option for exporting the explanations would be very useful, in order to allow users to save and share the results, revisiting them later. Another improvement could regard designing effective interactions with users, drawing inspiration from the Fairness Dashboard[62] provided by FairLearn (Bird et al. (2020)). Exploring how to analyze the predictions from an intersectional point of view, as in the work of Buolamwini and Gebru (2018), would also be valuable. Unstructured data like texts challenges this investigation in a different way w.r.t. tabular sources, where e.g. protected attributes like race and gender can easily be cross-referenced and verified. It would be also interesting to take into account the impact of latent, not directly observable features like dialects and other linguistic variations, which indeed are proven to be a source of bias (Sap et al. (2019a)).

Significant aspects to explore with additional experiments would be testing commercial models like Google Perspective API[63] and Microsoft's Text Analytics API[64] and evaluating *FairShades* with competitors explainers on the Fairness dimension, i.e., the capability of identifying biases, as well as comparing the neighbourhood generation process with perturbations of other similar tools, e.g. LIME (Ribeiro et al. (2016)), LORE (Guidotti et al. (2018a)), X-SPELLS (Lampridis et al. (2020)) and ANCHORS (Ribeiro et al. (2018)).

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

[62] https://pypi.org/project/raiwidgets/#fairness-dashboard

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

[64] https://docs.microsoft.com/en-us/azure/cognitive-services/Text-Analytics/

This last experiment could be performed with a classifier trained to detect synthetic from human-produced texts: the generation process more able to trick the classifier is the most effective in simulating realistic sentences.

Finally, given the complexity of the problem, an interdisciplinary approach is a must (Romei and Ruggieri (2014)). The key to enrich and further this project lies in the prospect of engaging in interdisciplinary ways to develop ethical use of AI: legal and privacy experts, social scientists, ethical philosophers, UX designers and digital humanists are all needed. Another priority consists in raising new and complex questions within human-centered ML, assessing the impacts on individuals. 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.

The broader goal of this project was to explore socio-cultural implications regarding the concrete impacts and consequences that AI exerts on our everyday life. The specific focus concerned contexts where AI systems can cause harm, amplifying discriminations against minorities. The concrete aim of this research was therefore performing a strong value-oriented evaluation, in order to guarantee beneficial and just automated decisions for everyone, with the explicit objective of preventing unequal treatment of legally protected groups and intersections between them. This perspective also enabled the pursuit of broader goals, such as moving beyond the binariness associated with gender, encouraging a pluralism of views and voices as a fundamental element of these systems and, finally, cultivating an awareness that data is never neutral, but is always the expression of a prevailing context and perspective (D'Ignazio and Klein (2020)). Finally, a secondary focus, yet to be undertaken further, was the prospect of seeking ways of raising awareness and disseminating both effects of human traits on algorithmic response (Zarsky (2016)) and impacts of automated decision-making systems, increasing public understanding of being able to hold IT companies accountable, ultimately empowering citizens (Criado et al. (2020)).

In fact, from *FairShades* output, i.e., biases detected towards counterfactual terms, automated debiasing techniques could be adopted. A recent contribution by Zhou et al. (2021) analyses and highlights the main limitations of the current strategies applied to the debiasing of abusive language detection models, concluding that, in the sensitive context of online hate speech, they are not always effective. Further studies and experiments will be necessary to explore the most relevant approaches.

As suggested in Dobbe et al. (2018), proposing a contribution within the Machine Learning domain responsibly and consciously means foremost acknowledging our own biases. In particular, we are referring to the implementation of perturbation functions and hand-coded lexicons, that we encoded within CheckList framework: the selection and the way in which processes have been built certainly shaped the results.

What has guided our commitment are the pressing user need in demanding transparency and the intent of developing truly inclusive tools that can meet the needs of diverse experiences of minorities in online spaces, understanding social disparities and language nuances in expressing opinions (Saha et al. (2019)). A further criterion we have tried to pursue is contextuality, i.e., adopting the effectiveness of domain-specific ethical frameworks and context-specific applications, developing a Fairness definition for the outcomes of abusive language detection systems. Bias identification must also be contextual since the notion of harm is culturally dependent and minority perspectives could be missed in the generalization process: for this reason, we embraced protected lexicons as open, evolving sets. Surely, this project is not a complete or comprehensive work: for example, in addition to future works, a direct interaction with the targeted users and the different stake-holders affected could have enriched the perspective and the insights retrieved. In fact, as long as these research branches lack perspectives and minorities, solutions will never be beneficial or just for everyone (Xu et al. (2020)). Furthermore, it is important to be aware that any solely technological solutions will be partial, as not considering the broader social issue that is the source of these biases means simplifying and “fixing" only on the surface (Ntoutsi et al. (2020)). Therefore, the complexity of the phenomenon is not limited to algorithms but is rooted in socio-cultural issues: it cannot be solved by computational methods alone nor with mathematical explanations understandable only to data scientists, just as algorithmic fairness is not enough to effectively counteract certain types of harms (Suresh and Guttag (2019)). Regardless, we strongly believe that abusive language classifiers need a robust value-sensitive evaluation, in order to assess unintended biases and avoid, as far as possible, explicit harm or the amplification of pre-existing social biases, trying to ultimately build systems that contributes in a beneficial way to the society and all its citizens.

List of Figures

List of Tables

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR.

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep learning models for multilingual hate speech detection.

Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. 2017. Learning certifiably optimal rule lists for categorical data. *arXiv preprint arXiv:1704.01701*.

André Artelt, Fabian Hinder, Valerie Vaquet, Robert Feldhans, and Barbara Hammer. 2021. Contrastive explanations for explaining model adaptations. *arXiv e-prints*, pages arXiv–2104.

Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012*.

M Gethsiyal Augasta and Thangairulappan Kathirvalavakumar. 2012. Reverse engineering the neural networks for rule extraction in classification problems. *Neural processing letters*, 35(2):131–150.

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of SEMEVAL 2019*.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32, Microsoft.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of bias in nlp. *arXiv preprint arXiv:2005.14050*.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *arXiv preprint arXiv:1607.06520*.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, Serena Villata, and Fondazione Bruno Kessler. 2019. Inriafbk drawing attention to offensive language at germeval2019. In *KONVENS*.

Mark Craven and Jude Shavlik. 1995. Extracting tree-structured representations of trained networks. *Advances in neural information processing systems*, 8:24–30.

Natalia Criado, Xavier Ferrer-Aran, and Jose M Such. 2020. Is my program sexist? using norms to attest digital discrimination. *IEEE Technology and Society Magazine*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515.

Emily Denton, Ben Hutchinson, Margaret Mitchell, and Timnit Gebru. 2019. Detecting bias with generative counterfactual face attribute augmentation. *arXiv preprint arXiv:1906.06439*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Catherine D'Ignazio and Lauren F Klein. 2020. *Data feminism*. Mit Press.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Roel Dobbe, Sarah Dean, Thomas Gilbert, and Nitin Kohli. 2018. A broader view on bias in automated decision-making: Reflecting on epistemology and dynamics. *arXiv preprint arXiv:1807.00553*.

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

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the evalita 2018 task on automatic misogyny identification (ami). *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12:59.

Ruth C Fong and Andrea Vedaldi. 2017. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3429–3437.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *11th International Conference on Web and Social Media, ICWSM 2018*. AAAI Press.

T Gebru, J Morgenstern, B Vecchione, JW Vaughan, HD Wallach III, and K Crawford. 2018. Datasheets for datasets. arxiv.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Jennifer Golbeck, Zahra Ashktorab, Rashad O Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, et al. 2017. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 229–233. ACM.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. 2018a. Local rule-based explanations of black box decision systems. *arXiv preprint arXiv:1805.10820*.

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

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413*.

Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C Fraser. 2020. Confronting abusive language online: A survey from the ethical and human rights perspective. *arXiv preprint arXiv:2012.12305*.

Maximilian Kohlbrenner, Alexander Bauer, Shinichi Nakajima, Alexander Binder, Wojciech Samek, and Sebastian Lapuschkin. 2020. Towards best practice in explaining neural network decisions with lrp. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.

Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2018. Counterfactual fairness. *stat*, 1050:8.

Orestis Lampridis, Riccardo Guidotti, and Salvatore Ruggieri. 2020. Explaining sentiment classification with synthetic exemplars and counter-exemplars. In *International Conference on Discovery Science*, pages 357–373. Springer.

Ping Liu, Wen Li, and Liang Zou. 2019a. NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, Minneapolis, Minnesota, USA.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

Stefano Menini, Alessio Palmero Aprosio, and Sara Tonelli. 2021. Abuse is contextual, what about nlp? the role of context in abusive language annotation and detection. *arXiv preprint arXiv:2103.14916*.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

Christoph Molnar. 2020. *Interpretable machine learning*. Lulu. com.

Malvina Nissim, Rik van Noord, and Rob van der Goot. 2020. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, 46(2):487–497.

Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended bias in misogyny detection. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 149–155.

Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. 2020. Bias in data-driven artificial intelligence systems—an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3):e1356.

Tom van Nuenen, Xavier Ferrer, Jose M Such, and Mark Cote. 2020. Transparency for whom? assessing discriminatory artificial intelligence. *Computer*, 53(11):36–44.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. *CoRR*, abs/1908.11049.

Cecilia Panigutti, Alan Perotti, and Dino Pedreschi. 2020. Doctor xai: an ontology-based approach to black-box sequential data classification explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 629–639.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Dino Pedreschi, Fosca Giannotti, Riccardo Guidotti, Anna Monreale, Luca Pappalardo, Salvatore Ruggieri, and Franco Turini. 2018. Open the black box data-driven explanation of black box decision systems. *arXiv preprint arXiv:1806.09936*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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. 2018. Anchors: High-precision model-agnostic explanations. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*.

Andrea Romei and Salvatore Ruggieri. 2014. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(5):582–638.

Koustuv Saha, Sang Chan Kim, Manikanta D Reddy, Albert J Carter, Eva Sharma, Oliver L Haimson, and Munmun De Choudhury. 2019. The language of lgbtq+ minority stress experiences on social media. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–22.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019a. 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.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2019b. Social bias frames: Reasoning about social and power implications of language. *arXiv preprint arXiv:1911.03891*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *ACL*.

Mattia Setzu, Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi, and Fosca Giannotti. 2021. Glocalx-from local to global explanations of black box ai models. *Artificial Intelligence*, 294:103457.

Kacper Sokol, Alexander Hepburn, Rafael Poyiadzi, Matthew Clifford, Raul Santos-Rodriguez, and Peter Flach. 2020. FAT Forensics: A Python Toolbox for Implementing and Deploying Fairness, Accountability and Transparency Algorithms in Predictive Systems. *Journal of Open Source Software*, 5(49):1904.

Harini Suresh and John V Guttag. 2019. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California.

Gregor Wiedemann, Seid Muhie Yimam, and Chris Biemann. 2020. UHH-LT at SemEval-2020 task 12: Fine-tuning of pre-trained transformer networks for offensive language detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1638–1644, Barcelona (online). International Committee for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. 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 602–608.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2021. Polyjuice: Automated, general-purpose counterfactual generation. *arXiv preprint arXiv:2101.00288*.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2020. Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*.

Xiaoxin Yin and Jiawei Han. 2003. Cpar: Classification based on predictive association rules. In *Proceedings of the 2003 SIAM International Conference on Data Mining*, pages 331–335. SIAM.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of SemEval*.

Tal Zarsky. 2016. The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values*, 41(1):118–132.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. 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 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2021. Challenges in automated debiasing for toxic language detection.

Appendix 1: Lexicons

The lexicons available in CheckList's editor and developed by the authors (Ribeiro et al. (2020)) contained common male and female names, cities, countries and sensitive-group adjectives such as the ones related to nationalities, religions, sexual orientations and gender. The custom entries we have added, resulting from the assets mentioned in 5.2, are related to common nouns referring to women (both neutral and offensive), generic offensive terms and insults, list of stereotyped work roles and identity terms for insultingly addressing homosexuals, disabled, homeless and old people. The intention is therefore to build a targeted hate lexicon that is used in social-media contexts by real users in order to mimic and generalise offensive linguistic dynamics that occur in online dialogue. Sets of protected keys and associated sensitive values are completely open and do not claim to be representative but only some of the main categories, also because a lot depends on the reference dataset. Users can modify, reduce or expand the lists at will by editing the file containing the hand-coded lexicons. We would like to point out that some works, e.g. in CheckList itself, the categories related to *gender* and *sexual orientation* are mixed, also because there are terms that simultaneously identify sexuality and gender identity. In this project we tried to distinguish between them, also consulting external resources[65]. For reference, we report in the following the protected keys, that summarizing are: *sexuality; gender identity; male names; female names; neutral expressions referring to women*; *work roles*; *race*; *nationality; country; city; religion* and finally *identity terms*, often offensive, related to *disabled*, *elderly* and *homeless*. First we list the protected terms, then we create from these lists the dictionary that contains them.

```
############## terms addressing women
women_noun = [woman , sister , daughter , girl , wife , young lady ,
    lady , bride , grandmother , mum , mother , housewife ,
    girlfriend , partner]
women_noun_plural = [women , sisters , aunts , daughters , girls ,
    mothers , wives , young ladies , ladies , grandmothers , mums ,
    mothers , housewives , girlfriends , partners]


```

[65] Such as https://lgbta.wikia.org/wiki/Category:Sexuality and https://lgbta.wikia.org/wiki/Category:Gender

```
############## offensive
offensive_women_noun = [slut, bitch, whore, cunt, spinster,
    ditsy, hussy, frigid, witch, wench, harpy, trollop, dyke,
    skank, butch]
offensive_women_noun_plural = [sluts, bitches, whores, cunts]
offensive_generic = [twat, prick, motherfucker, shit, shithead,
     dick, dickhead, cock, cockface, cockhead, cocksucker,
    pussy, asshole, arsehole, assfucker, fag, faggot, bastard,
    douche, bugger]
dis = [weirdo, lame, cripple, invalid, inferm]
homeless = [tramp, hobo, drifter, beggar, vagrant]
old = [codger, grandpa, oldie, geezer, elder]


############## work-related
work_role = [academic, quarterback, fireman, baker,
    entrepreneur, interpreter, attendant, secretary, laborer,
    writer, student, designer, poet, mechanic, investigator,
    dentist, tailor, ranger, foreman, lifeguard, construction
    worker, professor, architect, nurse, analyst, assistant,
    driver, footballer, editor, supervisor, hitter, farmer,
    librarian, chief, guard, administrator, organizer, teacher,
     gardener, economist, janitor, hairdresser, journalist,
    executive, educator, priest, manager, clown, scientist,
    builder, doorman, police, cook, cashier, shooter, counselor
    , magician, sheriff, singer, carpenter, detective, salesman
    , reporter, accountant, repairman, officer, fighter,
    investor, advisor, astronaut, paramedic, consultant,
    electrician, housekeeper, photographer, engineer, physician
    , doctor, soldier, waiter, developer, activist, attorney,
    mover, agent, businessman, lawyer, artistboxer, player,
    historian, intern, pilot, CEO, painter, receptionist, clerk
    , salesperson, cleaner, postman, surgeon, cameraman, nun,
    auditor]

# these sterotyped lists are taken from the file
    male_occupations.txt and female_occupations.txt from the
    WiNo dataset
fem_work_role = [attendant, cashier, teacher, nurse, assistant,
     secretary, auditor, cleaner, receptionist, clerk,
    counselor, designer, hairdresser, writer, housekeeper,
    baker, accountant, editor, librarian, tailor]
male_work_role = [driver, supervisor, janitor, cook, mover,
    laborer, construction worker, chief, developer, carpenter,
    manager, lawyer, farmer, salesperson, physician, guard,
    analyst, mechanic, sheriff, CEO]


############## final dictionary of protected categories

protected = {

    sexuality: [gay, lesbian, asexual, ace, bisexual, bi,
    homosexual, straight, cishet, heterosexual, pansexual, pan,
     demisexual, polysexual, bicurious, pancurious, polyamorous
    , aromantic, aro, biromantic, panromantic],

    gender_identity: [queer, trans, transgender, transsexual,
    cis, cisgender, cissexual, nonbinary, non-binary, enby, NB,
     genderqueer, genderfluid, genderflux, agender, bigender],

    race: [black,hispanic, white, asian, european, latino,
    middle eastern, african, african american, american],

    religion: list(map(lambda x: x.lower(), editor.template({
    religion}).data)),

    nationality: list(map(lambda x: x.lower(), editor.template
    ({nationality}).data)),

    country: list(map(lambda x: x.lower(), editor.template({
    country}).data)),

    city: list(map(lambda x: x.lower(), editor.template({city}).data)),

    male: list(map(lambda x: x.lower(), editor.template({male}).data[:10])),

    female: list(map(lambda x: x.lower(), editor.template({female}).data[:10])),

    women_noun: list(map(lambda x: x.lower(), editor.template({women_noun}).data)),

    women_noun_plural: list(map(lambda x: x.lower(), editor.template({women_noun_plural}).data)),

    work_role: list(map(lambda x: x.lower(), editor.template({work_role}).data)),

    dis: list(map(lambda x: x.lower(), editor.template({dis}).data)),

    homeless: list(map(lambda x: x.lower(), editor.template({homeless}).data)),

    old: list(map(lambda x: x.lower(), editor.template({old}).data))

}
```