Exposing Racial Dialect Bias in Abusive Language Detection: Can Explainability Play a Role?

Marta Marchiori Manerba[1(✉)] and Virginia Morini[1,2]

[1] Computer Science Department, University of Pisa, Pisa, Italy
{marta.marchiori,virginia.morini}@phd.unipi.it
[2] KDD Laboratory, ISTI, National Research Council, Pisa, Italy

Abstract. Biases can arise and be introduced during each phase of a supervised learning pipeline, eventually leading to harm. Within the task of automatic abusive language detection, this matter becomes particularly severe since unintended bias towards sensitive topics such as gender, sexual orientation, or ethnicity can harm underrepresented groups. The role of the datasets used to train these models is crucial to address these challenges. In this contribution, we investigate whether explainability methods can expose racial dialect bias attested within a popular dataset for abusive language detection. Through preliminary experiments, we found that pure explainability techniques cannot effectively uncover biases within the dataset under analysis: the rooted stereotypes are often more implicit and complex to retrieve.

Keywords: ML · NLP · Explainability · Interpretability · ML Evaluation · Fairness in ML · Algorithmic bias · Bias discovery · Algorithmic auditing · Data awareness · Discrimination

1 Introduction

Biases can arise and be introduced during each phase of a supervised learning pipeline, eventually leading to harm [17,41]. Within the task of automatic abusive language detection, this matter becomes particularly severe since unintended bias towards sensitive topics such as gender, sexual orientation, or ethnicity can harm underrepresented groups. The role of the datasets used to train these models is crucial. There might be multiple reasons why a dataset is biased, e.g., due to skewed sampling strategies or to the prevalence of a particular demographic group disproportionately associated with a class outcome [30], ultimately establishing conditions of privilege and discrimination. Concerning fairness and biases, in [24] is conducted an in-depth discussion on ethical issues and challenges in automatic abusive language detection. Among others, a perspective analyzed is the principle of non-discrimination throughout every stage of supervised machine learning pipelines. Several metrics, generic tools, and libraries such as [8,39] have

I. Koprinska et al. (Eds.): ECML PKDD 2022 Workshops, CCIS 1752, pp. 483–497, 2023.
https://doi.org/10.1007/978-3-031-23618-1_32

been proposed to investigate fairness in AI applications. Nevertheless, the solutions often remain fragmented, and it is difficult to reach a consensus on which are the standards, as underlined in a recent survey by [9], where the authors criticize the framing of *bias* within Natual Language Processing (NLP) systems, revealing inconsistency, lack of normativity and common rationale in several works.

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 where high-stakes choices are made. Inspecting non-discrimination of decisions and assessing that the knowledge autonomously learned conforms to human values also constitutes a real challenge. Indeed, in recent years working towards transparency and interpretability of black-box models has become a priority [11,21]. We refer the reader to the introduction conducted in [23], where authors cover selected explainability methods, offering an overall description of the state-of-the-art in this area.

Few approaches in the literature are at the intersection of fairness and explainability. In [1], through a user study, authors investigate the effects of explanations and fairness on human trust, finding that it increased when users were shown explanations of AI decisions. [6] develops a framework that evaluates systems' fairness through LIME [34] explanations and renders the models less discriminating, having identified and removed the sensitive attributes unfairly employed for classification. A model-agnostic strategy is proposed in [45]: from a biased black-box it aims at building a *fair surrogate* in the form of decision rules, guaranteeing fairness while maintaining performance. In [4] is described a Python package that allows for model investigation and development following a responsible ML pipeline, also performing bias auditing. We refer the reader to the review conducted in [2], where authors collect works that propose strategies to tackle fairness of NLP models through explainability techniques. Generally, authors found that, although one of the main reasons for applying explainability to NLP resides in bias detection, contributions at the intersection of these ethical AI principles are very few and often limited in the scope, e.g., w.r.t. biases and tasks addressed.

Given these evident socio-technical challenges, significant trust problems emerge, mainly regarding the robustness and quality of datasets and the related trustworthiness of models trained on these collections and their automated decisions. This work aims to investigate whether explainability methods can expose racial dialect bias attested within specific abusive language detection datasets. Racial dialect bias is described in [14] as the phenomenon whereby a comment belonging to African-American English (AAE) is more often classified as offensive than a text that aligns with White English (WE). For example, in [38], it is shown that annotators tend to label as offensive messages in Afro-American English more frequently than when annotating other messages, which could lead

to the training of a system reproducing the same kind of bias. Paradoxically, the systems learn to discriminate against the very demographic minorities they are supposed to protect against online hate, for whom it should help in creating a safe and inclusive digital environment.

To explore this issue, we chose the collection presented in [19] that gathers social media comments from Twitter manually annotated through crowdsourcing. The advantage of having data labelled by humans resides in the annotation's precision. However, it is a task that requires domain knowledge and can be very subjective [5] and time-consuming. We chose this dataset since it has been shown to contain racial dialect bias, introduced by the human annotator, who demonstrates a disparate treatment against certain dialect words [14]. For example, suppose terms belonging to the African-American language variant are used in the social media post. The instance is often more likely to be classified as abusive, even when, in fact, the content expressed is neutral, endorsing the importance of specific word variants rather than the offensive charge of the sentences. The focus of this work thus lies also in the impact on human annotation data, which can introduce different problems into the information formalized from the texts. As a result, the emerging biases propagate to the models drawn from these skewed collections. The quality of the annotation, and thus the models learned on these data, are significantly affected.

In this work, we adopt a qualitative definition of bias strongly contextual to abusive language detection and the type of unfairness we are investigating. We define as *bias* the sensitivity of an abusive language detection classifier concerning the presence in the record to be classified of terms belonging to the AAE dialect. Specifically, a classifier is considered biased or unfair if it tends to misclassify as abusive AAE records more often than those characterized by a white alignemnt linguistic variant. To understand whether these biases affect a model's outputs, we rely on explainability techniques, checking which aspects are relevant for the classification according to the model and the data on which it was trained. Suppose the explanation techniques give importance to misleading terms, not semantically or emotionally relevant. In that case, the explanation methods are effective for this debugging since they highlight how the knowledge learned from the model is neither reliable nor robust, revealing imbalances, possibly resulting from skewed and unrepresentative training data. Therefore, the question we try to answer is focused on testing if purely explanation techniques can identify biases in models' predictions inherited from problematic datasets. Specifically, according to our hypotheses, we would like to highlight those models demonstrate biases based on latent textual features, such as lexical and stylistic aspects, and not on the actual semantics or emotion of the text.

The rest of the paper is organized as follows. In Sect. 2 we briefly present necessary background knowledge. In Sect. 3, we conduct preliminary experiments to assess the effectiveness of explainability techniques application for evaluation and bias elicitation purposes. Finally, Sect. 4 discusses the limitations of our approach and indicates future research directions.

2 Setting the Stage

The following section reports the main methods and techniques leveraged in this contribution. We start by describing the AI-based text classifiers predicting the abusiveness, and then we proceed to the explanations algorithms used to interpret model outputs.

2.1 Text Classifiers

The task of detecting and predicting different kinds of abusive online content in written texts is typically formulated as a text-classification problem, where the textual content of a comment is encoded into a vector representation that is used to train a classifier to predict one of C classes.

Of course, when dealing with textual data, it is of utmost importance to consider both the suitable type of word representation and the proper type of classifier. Since traditional word representation (i.e., bag-of-words model) encode terms as discrete symbols not directly comparable to others [25], they are not fully able to model semantic relations between words. Instead, word embeddings like Word2vec [29], BERT Embeddings [16] and Glove [32] mapping words to a continuously valued low dimensional space, can capture their semantic and syntactic features. Also, their structure makes them suitable for deployment with Deep Learning models, fruitfully used to address NLP-related classification tasks. Among the available NLP classifiers (e.g., Recurrent Neural Networks like LSTM [22]), recently, in the literature have been introduced the so-called Transformer models that, differently from the previous ones, can process each word in a sentence simultaneously via the attention mechanism [44]. In particular, autoencoding transformer models such as Bidirectional Encoder Representations from Transformers (BERT) [16] and the many BERT-based models spawning from it (e.g., RoBERTa [26], DistilBERT [37]), has proven that leveraging a bidirectional multi-head self-attention scheme yields state-of-the-art performances when dealing with sentence-level classification.

Abusive Language Detection. Automatic abusive language detection is a task that emerged with the widespread use of social media [24]. Online discourse often assumes abusive and offensive connotations, especially towards sensitive minorities and young people. The exposition to these violent opinions can trigger polarization, isolation, depression, and other psychological trauma [24]. Therefore, online platforms have started to assume the role of examining and removing hateful posts. Since the large amount of data that flows across social media, hatred is typically flagged through automatic methods alongside human monitoring. Several approaches have been proposed to perform both coarse-grained, i.e., binary, and fine-grained classification. As noted, pre-trained embeddings such as contextualized Transformers [43], and ELMo [33] embeddings are among the most popular techniques [47]. For this reason, we adopt BERT in the experiments presented in the following sections.

2.2 Post-hoc Explanation Methods

Following recent surveys on Explainable AI [11,18,20,21,27,31,36], we briefly define the field to which the explainers we use in this contribution belong, i.e., post-hoc explainability methods. This branch pertains to the black-box explanation methods. The aim is to build explanations for a black-box model, i.e., a model that is not interpretable or transparent regarding the automatic decision process due to the complexity of its internal dynamics. Post-hoc strategies can be *global* if they target explaining the whole model, or *local* if they aim to explain a specific decision for a particular record. The validity of the local explanation depends on the particular instance chosen, and often the findings are not generalizable to describe the overall model logic. In addition, the explanation technique can be *(i) model-agnostic*, i.e., independent w.r.t. the type of black-box to be inspected (e.g., tree ensemble, neural networks, etc.), or *(ii) model-specific*, involving a strategy that has particular requirements and works only with precise types of models. Thus, given a black-box b and a dataset X, a local post-hoc explanation method ϵ takes as input b and X and returns an explanation e for each record $x \in X$. Returning to the general definition of post-hoc explainability, we now introduce more formally the objective of these methods. Given a black-box model b and an interpretable model g, post-hoc methods aim to approximate the local or global behaviour of b through g. In this sense, g becomes the transparent surrogate of b, which can mimic and account for its complex dynamics more intelligibly to humans. The approaches proposed in the literature differ in terms of the input data handled by b (textual, tabular); the type of b the interpretable technique can explain; the type of explanator g adopted (decision tree, saliency maps).

In the following, we briefly present the explanation techniques we chose to adopt. Specifically, Integrated Gradients and SHAP are used locally and globally, as described in Sect. 3.4.

Integrated Gradients. Integrated Gradients (IG) [40] is a post-hoc, model-specific explainability method for deep neural networks that attributes a model's prediction to its input features. In other words, it can compute how relevant a given input feature is for the output prediction. Differently from mostly attribution methods [7,42], IG satisfies both the attribution axioms *Sensitivity* (i.e., relevant features have not-zero attributions) and *Implementation Variance* (i.e. the attributions for two functionally equivalents models are identical). Indeed, IG aggregates the gradients of the input by interpolating in small steps along the straight line between a baseline and the input. Accordingly, a large positive or negative IG score indicates that the feature strongly increases or decreases the model output. In contrast, a score close to zero indicates that the feature is irrelevant to the output prediction. IG can be applied to any differentiable model and thus handle different kinds of data like images, texts, or tabular ones. Further, it is adopted for a wide range of goals like: *i)* understanding feature importance by extracting rules from the network; *ii)* debugging deep learning models perfor-

mance and *iii)* identifying data skew by understanding the important features contributing to the prediction.

SHAP. SHAP [28] is among the most widely adopted local post-hoc model-agnostic approaches [11]. It outputs *additive feature attribution methods*, a form of feature importance, exploiting the computation of Shapley values for its explanation process. High values indicate a stronger contribution to the classification outcome, while values close to or above zero indicate negligible or negative contribution. The importance is retrieved by unmasking each term and assessing the prediction change between the score when the whole input is masked versus the actual prediction for the original input. SHAP can also compute a global explanation over multiple instances and provides, in addition to the agnostic explanation model, the choice among different kernels, according to the specifics of the ML system under analysis.

3 Preliminary Experiments

In this section, we present the experiments[1] conducted to assess the effectiveness of explainability techniques application for evaluation and bias elicitation purposes.

3.1 Dataset Description

As dataset, we leverage the corpus proposed in [19], which collects posts from Twitter. The collection includes around 100K tweets annotated with four labels: HATEFUL, ABUSIVE, SPAM or NONE. Differently from the other datasets, it was not created starting from a set of predefined offensive terms or hashtags to reduce bias, which is a main issue in abusive language datasets [46]. This choice should make this dataset more challenging for classification. The strategy consisted of a bootstrapping approach to sampling tweets labelled by several crowdsource workers and then validated them. Specifically, the dataset was constructed through multiple rounds of annotations to assess raters' behavior and usage of the various labels. The authors then analyzed these preliminary annotations to understand which labels were most similar, i.e., related and co-occurring. The result consists of the labels to retain, i.e., the ones most representative and those to eliminate since they were redundant. From the derived annotation schema, labelling was conducted on the entire collection. For our experiments, we have used a preprocessed data version: retweets have been deleted, so the collection contains no duplicates; urls and mentions are replaced by '@USER' and 'URL,' and the order is randomised. We also removed the spam class, and we mapped both hateful and abusive tweets to the abusive class, based on the assumption that hateful messages are the most severe form of abusive language and that the

[1] The results of the experiments are available at https://github.com/MartaMarchiori/Exposing-Racial-Dialect-Bias.

term 'abusive' is more appropriate to cover the cases of interest for our study [12]. The dataset thus organized contains 49430 non-abusive instances and 23764 abusive ones. The number of abusive records is high since it results from the union of hateful and abusive tweets, as reported above. Besides, the class imbalance is typical of abusive language detection datasets: it reflects the dynamics of online discourse, where most content is not hateful. We do not introduce any other alterations to the dataset as the intention is precisely to examine the presence of bias in the collection as conceived and published by the data collectors.

We chose this dataset since in [3] is identified as a relevant source of racial dialect bias. As [3] claim, although this kind of bias is present in all of the collections investigated in their work, it is far more robust in the *Founta* dataset [19]. The authors trace this problem by making several assumptions. One reason may lie in the annotations not being conducted by domain experts. In addition, the platform used to collect and curate the collection may have had a significant impact. Therefore, a text classifier trained on this data will surely manifest a kind of racial bias, as the set is neither representative nor fair. Following such reasoning, the goal of this contribution focused on this collection is to assess via explanation methods if the trained model can correctly detect the comment's abusiveness or if it is predicting the grade of offensiveness based on dialect terms, i.e., manifesting an evident racial bias.

3.2 Methods Overview

Following the rationale in Sect. 2.1, we rely on a BERT-based model to predict the abusiveness. In the following paragraph, we explain the experimental setup and evaluation steps.

The dataset is split into $\sim 59,000$ records for training and $\sim 15,000$ for testing. As for the classifier architecture, we used the pre-trained implementation of BERT [15], i.e., BERT-BASE-UNCASED, available through the library Transformers[2]. We varied the learning rate between $[2e^{-5}, 3e^{-5}, 5e^{-5}]$. We trained the model for 5 epochs, finding that the best configuration was derived from the second iteration, reaching a weighted F1-score of 94.1% on the validation set. The performance achieved on the test set was also high (93.6% weighted F1-score).

Regarding the XAI techniques, IG's SEQUENCE CLASSIFICATION EXPLAINER was exploited, while for SHAP the LOGIT one, both with default parameters. Details on the subsets of instances for which explanations were calculated are provided in Sect. 3.4.

3.3 Local to Global Explanations Scaling

Before presenting the preliminary results, we briefly explain how we scale to a global explanation from the local ones for IG, attempting to represent the whole model. A straightforward way to accomplish this task consists of obtaining local predictions for many items and then averaging the scores assigned to each feature

[2] https://huggingface.co/bert-base-uncased.

across all the local explanations to produce a global one. Accordingly, for each record in the dataset, we store the local explanation, consisting of a key, i.e., the word present in the phrase, and a value, i.e., the feature importance. Then we average the obtained scores for each word. This process is repeated for each class predicted by the model in such a way to find what are the words that led the model to output a specific class.

3.4 Results

This section reports the experiments' results to test our hypotheses. We focus the analysis on the BERT-based abusive language detection classifier, adopting IG and SHAP as explanation techniques.

Global Explanations. We begin the analysis by illustrating the outcomes obtained by IG: the results are reported in Fig. 1 (a) as WordClouds. Among the most influential words for the predicted non-abusive class, we find *portrait* and *creativity*, followed by terms that belong to holidays, such as *passport, christmas*, and to a positive semantic sphere (*excitedly*). Interesting to note that the third most relevant non-abusive word is *bitch*. This behavior could be motivated by the fact that IG gives importance to this term in phrases that the classifier gets wrong, i.e., that it considers non-abusive when, in fact, they are. Another possible explanation could be found in the frequent use of this word informally with a friendly connotation in the African-American dialect, stripping this term of its derogatory meaning in specific linguistic contexts. As we would expect, among the most relevant terms for the predicted abusive class, we encounter insults, swear words, and imprecations, such as *fucked, shit, idiots, bastard, bitch, goddamn, crap, bullshit*. To note the presence of neutral words in this setting, which acquire a negative connotation in sentences with a strong toxic charge, such as *streets, clown, pigs, ska* (African-Jamaican folk music) and demographic groups like *homosexual, gay, lesbian, queer, jew*.

Sub-global Explanations. Although the most relevant patterns are primarily consistent with the related sentiment, e.g., toxic words for the abusive class, from this global overview, terms belonging to the African-American dialect did not clearly emerge. We, therefore, isolated from the test set the comments highly characterized by this slang, using a classifier[3] specifically trained to recognize texts belonging to the African-American English dialect [10]. The classifier works as follows: taking in input a text, such as *Wussup niggas*, it emits the probability that the instance belongs to AAE (0.87). Although authors suggest trusting the classifier prediction when the score is equal to or above 0.80, we relax this constraint by imposing 0.70 as bound to have a sufficiently populous subset to conduct preliminary sub-global analysis. We identified a cluster of only 74 AAE records, 65 abusive, and 9 non-abusive.

[3] https://github.com/slanglab/twitteraae.

(a) Whole test set.

(b) AAE subset.

Fig. 1. For each predicted class is shown a WordCloud representing the terms that obtained the higher global scores by IG for the whole test set and for the AAE subset respectively.

The results for IG, reported in 1 (b), are not remarkable, except for the importance of *ho* in the predicted non-abusive class. The hypothesis could be the same as that underlying the importance of *bitch*: *ho* is used informally in this slang. Among the words of lesser importance (with a score between 0.28 and 0.26) for the predicted abusive class, we find *em* and *gotta*, non-standard variants but not highly relevant to our bias detection. For comparison, we employ SHAP as additional explainer[4] (Fig. 2). SHAP already offers the possibility to compute explanations for multiple records; therefore, we do not have to perform the same local to global scaling applied to IG. For this predominantly abusive subset, the most important words identified by the logit explainer SHAP are *fucked, damn, fuck, bitch, fucking, dirty, shit, dick, ass.*

Since the findings concerning the evidence of racial dialect bias in this corpus are not as observable as we might have expected, we decide to narrow the investigation by focusing on local instances belonging to this subset to assess the classifier further.

[4] SHAP was not used on the entire test set (i.e., within the Global Explanations Section) due to the high computational costs of this explainability method. It was therefore preferred to apply it when analysing a narrower subset, i.e., in the sub-global setting.

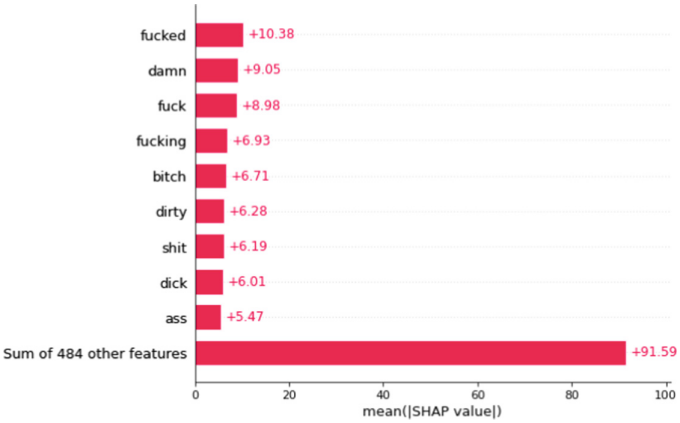


Fig. 2. Explanation for the AAE subset returned by the SHAP logit explainer, consisting of the average impact of each term for the abusive class.

Local Explanations. To further investigate possible racial dialect bias, we inspect local instances. Specifically, we focus the analysis on sentences belonging to the AAE subset according to different scenarios.

As a first exploration, we calculate the explanation for the three non-abusive instances misidentified as abusive by the classifier (specifically, with a probability > 0.5) precisely to assess whether there are AAE terms among the crucial words misleading the prediction. In Fig. 3, both IG and SHAP agree in finding *ass* as an important term, although in these contexts it is used with a neutral connotation, as is *hoes*, broken in both cases in *ho* and *es*. SHAP also gives importance to the contract negative form *ain'*, typically belonging to AAE writers.

Another aspect that we preliminarily investigate is the predicted abusive instances containing the most salient words (identified by the global IG scores). From both explanation methods, the locally most salient words in Figs. 4 and 5 turn out to be *ass, stupid ass, fuck, bitch*. Interestingly, both methods give importance to *nigga*, often split as *ni gga*. This kind of importance could be misleading if this term is used with a friendly informal connotation.

Summarizing, as first insights, we can easily assess that the global explanations highlight informative patterns, i.e., toxic terms for the predicted abusive class. By preliminary assessing certain local instances, we can gather additional findings regarding the influence of specific terms belonging to the AAE variant. Except in these isolated cases, the explainers, and therefore the classifier, do not seem to give importance to terms belonging to the AAE dialect. We can conclude that, in this setting, the pure explanation techniques cannot effectively highlight the racial bias instilled by the crowd-sourcing process, which, for this particular

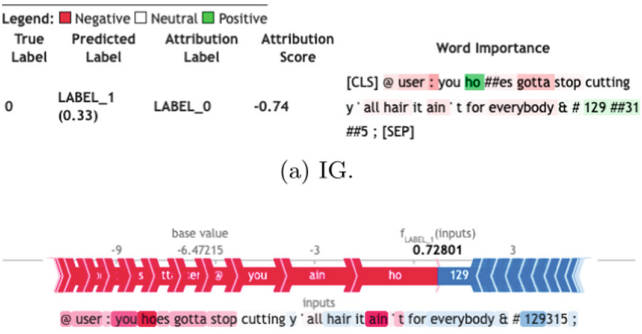


Fig. 3. Local explanation for the instance: *@USER: You hoes gotta stop cutting y'all hair it ain't for everybody&#129315.*

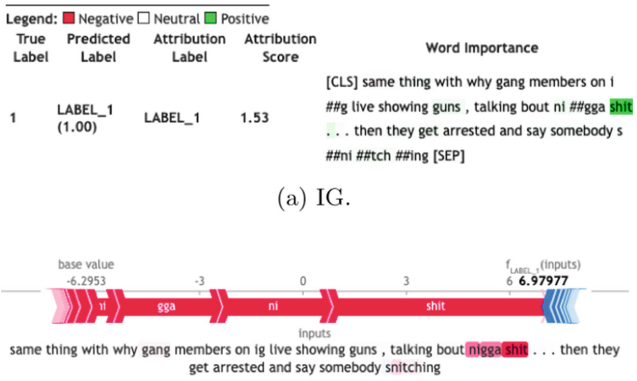
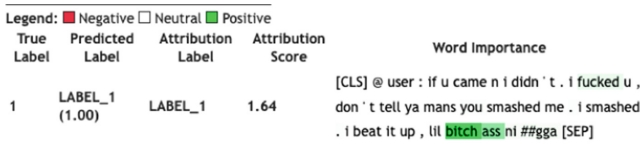


Fig. 4. Local explanation for the instance: *Same thing with why gang members on IG live showing guns, talking bout nigga shit...then they get arrested and say somebody snitching.*

dataset, is instead well documented in several works [3,38]. Since this stereotype is highly implicit, more specific and sophisticated bias checking techniques are needed to uncover it. Further, we see that the number of records belonging to the AAE variant in the test set is low. Further attempts by averaging the results from different subsets from cross-validation might yield more robust insights. Therefore, further experiments are needed to explore these preliminary hypotheses, involving individuals who speak AAE in everyday conversations and domain experts like linguists.



(a) IG.

(b) SHAP.

Fig. 5. Local explanation for the instance: *@USER: If u came n I didn't. I fucked u, don't tell ya mans you smashed me. I smashed. I beat it up, lil bitch ass nigga.*

4 Conclusion and Future Work

In this contribution, we investigated whether explainability methods can expose racial dialect bias attested within a popular dataset for abusive language detection, published in [19]. Although the experiment conducted is restricted to a single dataset and thus cannot directly lead to generalisable inferences, insights from the analysis of this specific collection are relevant to start discussing the limitations of applying explainability techniques for bias detection. The pure explainability techniques could not, in fact, effectively uncover the biases occurring in the *Founta* dataset: the rooted stereotypes are often more implicit and complex to retrieve. Possible reasons for this issue include the limited frequency of the AAE dialect identified in the test set and the shortages of explanation methods applicable to text but mainly developed for tabular data. In agreement with as pointed out in [2], current explainability methods applied to fairness detection within NLP suffer several limitations, such as relying on specific local explanations could foster misinterpretations, and it is challenging to combine them for scaling toward a global, more general level.

For future experiments, first, we want to explore other explanation techniques in addition to IG and SHAP, to compare whether other methods succeed bias discovery, e.g., testing Anchor[5] [35] and NeuroX[6] [13]. It would also be interesting to evaluate other transformer-based models to assess the impact of different pretraining techniques on bias elicitation.

Overall, labels gathered from crowd-sourced annotations can introduce noise signals from the annotators' human bias. Moreover, it is clear that when the labelling is performed on subjective tasks, such as online toxicity detection, it becomes even more relevant to explore agreement reports and preserve indi-

[5] https://github.com/marcotcr/anchor.
[6] https://github.com/fdalvi/NeuroX.

vidual and divergent opinions, as well as investigate the impact of annotators' social and cultural backgrounds on the produced labelled data. Having access to the disaggregated data annotations and being aware of the dataset's intended use can inform both models' outcome assessment and comprehension, including facilitating bias detection [41].

Acknowledgements. This work has been partially supported by the European Community Horizon 2020 programme under the funding schemes: H2020-INFRAIA-2019-1: Research Infrastructure G.A. 871042 *SoBigData++*, G.A. 952026 *HumanE AI Net*, ERC-2018-ADG G.A. 834756 *XAI: Science and technology for the eXplanation of AI decision making*), G.A. 952215 *TAILOR*.

References

1. Angerschmid, A., Zhou, J., Theuermann, K., Chen, F., Holzinger, A.: Fairness and explanation in AI-informed decision making. Mach. Learn. Knowl. Extraction **4**(2), 556–579 (2022)
2. Balkir, E., Kiritchenko, S., Nejadgholi, I., Fraser, K.C.: Challenges in applying explainability methods to improve the fairness of NLP models. arXiv preprint arXiv:2206.03945 (2022)
3. Ball-Burack, A., Lee, M.S.A., Cobbe, J., Singh, J.: Differential tweetment: mitigating racial dialect bias in harmful tweet detection. In: FAccT, pp. 116–128. ACM (2021)
4. Baniecki, H., Kretowicz, W., Piatyszek, P., Wisniewski, J., Biecek, P.: dalex: responsible machine learning with interactive explainability and fairness in python. arXiv preprint arXiv:2012.14406 (2020)
5. Basile, V., Cabitza, F., Campagner, A., Fell, M.: Toward a perspectivist turn in ground truthing for predictive computing. arXiv preprint arXiv:2109.04270 (2021)
6. Bhargava, V., Couceiro, M., Napoli, A.: LimeOut: an ensemble approach to improve process fairness. In: Koprinska, I., et al. (eds.) ECML PKDD 2020. CCIS, vol. 1323, pp. 475–491. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-65965-3_32
7. Binder, A., Montavon, G., Lapuschkin, S., Müller, K.-R., Samek, W.: Layer-wise relevance propagation for neural networks with local renormalization layers. In: Villa, A.E.P., Masulli, P., Pons Rivero, A.J. (eds.) ICANN 2016. LNCS, vol. 9887, pp. 63–71. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44781-0_8
8. Bird, S., et al.: Fairlearn: a toolkit for assessing and improving fairness in AI. Tech. Rep. MSR-TR-2020-32, Microsoft (2020)
9. Blodgett, S.L., Barocas, S., Daumé III, H., Wallach, H.: Language (technology) is power: a critical survey of "bias" in NLP. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5454–5476 (2020)
10. Blodgett, S.L., Green, L., O'Connor, B.: Demographic dialectal variation in social media: a case study of african-american english. In: Proceedings of EMNLP (2016)
11. Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D., Rinzivillo, S.: Benchmarking and survey of explanation methods for black box models. CoRR abs/2102.13076 (2021)

12. Caselli, T., Basile, V., Mitrović, J., Kartoziya, I., Granitzer, M.: I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In: Proceedings of the 12th Language Resources and Evaluation Conference, pp. 6193–6202. European Language Resources Association, Marseille, France (2020). https://www.aclweb.org/anthology/2020.lrec-1.760
13. Dalvi, F., et al.: Neurox: a toolkit for analyzing individual neurons in neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2019)
14. Davidson, T., Bhattacharya, D., Weber, I.: Racial bias in hate speech and abusive language detection datasets. arXiv preprint arXiv:1905.12516 (2019)
15. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
16. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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), pp. 4171–4186. Association for Computational Linguistics (2019)
17. Dixon, L., Li, J., Sorensen, J., Thain, N., Vasserman, L.: Measuring and mitigating unintended bias in text classification. In: AIES, pp. 67–73. ACM (2018)
18. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)
19. Founta, A., et al.: Large scale crowdsourcing and characterization of twitter abusive behavior. In: ICWSM, pp. 491–500. AAAI Press (2018)
20. Freitas, A.A.: Comprehensible classification models: a position paper. SIGKDD Explor. **15**(1), 1–10 (2013)
21. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5), 1–42 (2019)
22. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
23. Holzinger, A., Saranti, A., Molnar, C., Biecek, P., Samek, W.: Explainable AI methods-a brief overview. In: Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K.R., Samek, W. (eds.) xxAI - Beyond Explainable AI. xxAI 2020. LNCS, vol. 13200, pp. 13–38. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-04083-2_2
24. Kiritchenko, S., Nejadgholi, I., Fraser, K.C.: Confronting abusive language online: a survey from the ethical and human rights perspective. J. Artif. Intell. Res. **71**, 431–478 (2021)
25. Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., Brown, D.: Text classification algorithms: a survey. Information **10**(4), 150 (2019)
26. Liu, Y., et al.: Roberta: a robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
27. Longo, L., Goebel, R., Lecue, F., Kieseberg, P., Holzinger, A.: Explainable artificial intelligence: concepts, applications, research challenges and visions. In: Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds.) CD-MAKE 2020. LNCS, vol. 12279, pp. 1–16. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-57321-8_1
28. Lundberg, S.M., Lee, S.: A unified approach to interpreting model predictions. In: NIPS, pp. 4765–4774 (2017)

29. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, 2–4 May 2013, Workshop Track Proceedings (2013)
30. Ntoutsi, E., et al.: Bias in data-driven artificial intelligence systems - an introductory survey. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. **10**(3), e1356 (2020)
31. Pedreschi, D., et al.: Open the black box data-driven explanation of black box decision systems. CoRR abs/1806.09936 (2018)
32. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Empirical methods in natural language processing (EMNLP), pp. 1532–1543 (2014)
33. Peters, M.E., et al.: Deep contextualized word representations. In: NAACL-HLT, pp. 2227–2237. Association for Computational Linguistics (2018)
34. Ribeiro, M.T., Singh, S., Guestrin, C.: why should I trust you?: explaining the predictions of any classifier. In: KDD, pp. 1135–1144. ACM (2016)
35. Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: High-precision model-agnostic explanations. In: AAAI, pp. 1527–1535. AAAI Press (2018)
36. Samek, W., Montavon, G., Lapuschkin, S., Anders, C.J., Müller, K.: Toward interpretable machine learning: transparent deep neural networks and beyond. CoRR abs/2003.07631 (2020)
37. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of Bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
38. Sap, M., Card, D., Gabriel, S., Choi, Y., Smith, N.A.: The risk of racial bias in hate speech detection. In: ACL (1), pp. 1668–1678. Association for Computational Linguistics (2019)
39. Sokol, K., Hepburn, A., Poyiadzi, R., Clifford, M., Santos-Rodriguez, R., Flach, P.: FAT forensics: a python toolbox for implementing and deploying fairness, accountability and transparency algorithms in predictive systems. J. Open Source Softw. **5**(49), 1904 (2020)
40. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International Conference on Machine Learning, pp. 3319–3328. PMLR (2017)
41. Suresh, H., Guttag, J.V.: A framework for understanding unintended consequences of machine learning. CoRR abs/1901.10002 (2019)
42. Vashishth, S., Upadhyay, S., Tomar, G.S., Faruqui, M.: Attention interpretability across NLP tasks. arXiv preprint arXiv:1909.11218 (2019)
43. Vaswani, A., et al.: Attention is all you need. In: NIPS, pp. 5998–6008 (2017)
44. Vaswani, A., et al.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6000–6010. NIPS2017, Curran Associates Inc., Red Hook, NY, USA (2017)
45. Wang, T., Saar-Tsechansky, M.: Augmented fairness: an interpretable model augmenting decision-makers' fairness. arXiv preprint arXiv:2011.08398 (2020)
46. Wiegand, M., Ruppenhofer, J., Kleinbauer, T.: 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, pp. 602–608 (2019)
47. Zampieri, M., et al.: Semeval-2020 task 12: multilingual offensive language identification in social media (offenseval 2020). In: SemEval@COLING, pp. 1425–1447. International Committee for Computational Linguistics (2020)