Fine-Grained Fairness Analysis of Abusive Language Detection Systems with CheckList

Marta Marchiori Manerba
Department of Philology, Literature and Linguistics
University of Pisa, Italy
martamarchiori96@gmail.com

Sara Tonelli
Fondazione Bruno Kessler
Trento, Italy
satonelli@fbk.eu

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 paper, we create ad hoc tests through the CheckList tool (Ribeiro et al., 2020) to detect biases within abusive language classifiers for English. We compare the behaviour of two BERT-based models, one trained on a generic abusive language dataset and the other on a dataset for misogyny detection. Our evaluation shows that, although BERT-based classifiers achieve high accuracy levels on a variety of natural language processing tasks, they perform very poorly as regards fairness and bias, in particular on samples involving implicit stereotypes, expressions of hate towards minorities and protected attributes such as race or sexual orientation. We release both the notebooks implemented to extend the Fairness tests and the synthetic datasets usable to evaluate systems bias independently of CheckList.

1 Introduction

At every stage of a supervised learning process, biases can arise and be introduced in the pipeline, ultimately leading to harm (Suresh and Guttag, 2020; Dixon et al., 2018). 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. (2019), 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 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., 2019a), 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., 2019a). 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 create specialised datasets addressing different phenomena of abusive language, 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, in order to assess unintended biases and avoid, as far as possible, explicit harm or the amplification of pre-existing social biases. However, this bias-assessment 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).

Concerning the different definitions of fairness, they have been collected and organised both in (Suresh and Guttag, 2020) and (Mehrabi et al., 2019), with the awareness that a single definition is not sufficient to address the multi-faceted problem of fairness in its entirety. In this work, we adopt a definition for fairness that is strongly contextual to abusive language detection. We define

Proceedings of the Fifth Workshop on Online Abuse and Harms, pages 81–91
August 6, 2021. ©2021 Association for Computational Linguistics

unfairness as the sensitivity of an abusive language detection classifier with respect to the presence in the record to be classified of entities belonging to protected groups or minorities. Specifically, a classifier is considered unfair or biased if the prediction changes according to the identities present, i.e. in similar sentences, the degree of hate is increased if terms such as *white* or *straight* are replaced by adjectives such as *black* or *non-binary*, revealing imbalances, possibly resulting from skewed and unrepresentative training data. *Fairness*, on the other hand, is defined as the behaviour of producing similar predictions for similar protected mentions, i.e. regardless of the specific value assumed by sensitive attributes like race and gender, without disadvantaging minorities or amplifying pre-existing social prejudices.

We deploy the *CheckList* tool (Ribeiro et al., 2020), which was originally created to evaluate general linguistic capabilities of NLP models, extending it to test fairness of abusive language detection systems. Embracing CheckList systematic framework, we create tests from hand-coded templates, reproducing stereotyped opinions and social biases, such as sexism and racism. 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. As ultimate goal, the analysis of the failures could therefore lead to a general overview of the models' fairness: the ideal outcome would be to establish a proactive pipeline that allows the improvement of the systems, having highlighted the shortages through CheckList ad hoc synthetic testing. To the best of our knowledge, there has not yet been any work carried out with CheckList in this research direction.

2 Related work

Several tools and approaches have been proposed to identify the most frequent errors done by NLP tools. For example, Errudite (Wu et al., 2019) is a tool that allows interactive error analysis through counterfactuals generation, but it is limited to the tasks of Question Answering and Visual Question Answering.

TextAttack (Morris et al., 2020) – which, among other packages, deploys CheckList – is a model-agnostic framework useful for the expansion of the datasets and the increase of models robustness through adversarial attacks. Compared to CheckList, however, it is more complicated to handle and deploy for users with little NLP skills. An interesting aspect is that TextAttack includes in the package the so-called "recipes", i.e. attacks from the literature ready to run, that build a common ground for the assessment and comparison of models' performances.

As outlined in (Ribeiro et al., 2020), some methods to identify errors by NLP systems are task-specific, such as (Ribeiro et al., 2019) or (Belinkov and Bisk, 2018), while others focus on particular NLP components such as word embeddings, as in (Tsvetkov et al., 2016) or (Rogers et al., 2018). Compared to existing approaches, one of CheckList's major strengths lies in including the testing phase within a comprehensive framework. The evaluation, conducted through adaptable templates and a range of relevant linguistic capabilities, is on one hand more granular than overall measures such as accuracy; on the other hand it is more versatile, because it leaves liberty to the developer to enrich and expand the tests within new and more suitable capabilities, depending on the task and model under consideration.

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. Furthermore, the visibility reached by corporate tools, such as IBM AI Fairness 360 or Amazon SageMaker Clarify, which are designed and promoted by large IT companies, raises several questions: is self-regulation right? What would be the advantages and risks of conducting independent external auditing? Several metrics[1], generic tools and python packages[2] are available. Nevertheless, no consensus related to the above questions has

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

[2] Fairlearn, Dalex, InterpretML, FAT Forensics, Captum.

been reached yet among the involved players.

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) [3] also used in this work as a lexical resource, which pertains to the field of co-reference resolution. Our contribution instead aims to broaden fairness evaluation, specifically testing biases in abusive language detection systems through CheckList facilities.

Concerning abusive language detection, 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. The most popular Transformers were BERT (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 sections.

3 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 and negation. While we also evaluated these aspects, our main focus here 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 (Sap et al., 2020). The resulting samples exemplify several sexist, racist and ableist comments

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

and opinions: all of them are new aspects compared to the suites released by the authors (Ribeiro et al., 2020).

As described in the introduction, CheckList provides built-in tools to assist users in the creation of tests. Among others, WordNet allows the selection of synonyms, antonyms, hypernyms, etc. for a given expression. CheckList's templates take shape from these sets of semantically related words. We develop a further extension of the tool by integrating *SentiWordNet* (Baccianella et al., 2010), a lexical resource in which WordNet synsets have been associated with a sentiment score (negative, neutral or positive). In this way, CheckList can benefit from the sentiment-dimension of SentiWordNet. Indeed, during the development of templates and the perturbations of the records, SentiWordNet enables the selection of suitable linguistic substitutions for a given term, according to the label of the sentence to be created. An example: seeking a synonym that has a similar connotation as the adjective *happy* for the phrase "*The girl is happy*", the results returned include *glad*, with a positive denotations of 0.5. In this case, through SentiWordNet, it is possible to select a synonym term with a similar polarity, in order to create variants of the original sentence that preserve a similar semantic content and to assess how the model behaves with slightly different terms.

4 A Suite for Abusive Language Detection

Suites are objects designed by CheckList authors (Ribeiro et al., 2020) that enable users to organise, combine and save sets of tests, in order to reuse them several times and to aggregate results (i.e. failure rates) in a single run. Once a test is designed, it is added to the suite, specifying the test type (MFT, INV or DIR), a name, the language capability within which it is situated and a brief description. The suite will thus be composed of one or more capabilities, each of which is assessed through several tests. After the suite is created, it can be run to evaluate the output of a given classifier, provided that the system has been previously launched to label the records created for each test providing for each record a class and the respective probabilities. The results of the run of the suite are displayed through a visual and interactive summary, which reports misclassified samples and the various failure percentages obtained in each test (see Fig. 1 for an example).

The core of our work takes off from the notebooks released by CheckList authors (Ribeiro et al., 2020), specifically from the suite for the task of Sentiment Analysis[4], that builds a series of tests consisting in tweets about airline companies. 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 contribution is the extension of the Fairness capability, which we enrich with several tests addressing diverse abuse targets and dealing with different types of biases.

4.1 Fairness tests

The tests developed for analysing **Fairness** deal with social biases towards specific different targets and topics, such as sexism, racism and ableism. With respect to the three test types available within CheckList, we have only utilized MFT and INV, because 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.

The Fairness tests developed are grouped into categories which we have previously identified through the manual inspection of representative constructions and stereotypes annotated in the Social Bias Inference Corpus (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. The groups of biases are not exhaustive, but they are still representative, as resulting also by exploring the most frequently occurring abuse 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). We briefly describe the tests designed, grouping them by target and reporting between brackets the test type of each:

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";

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

- *Stereotyped female vs male work roles* and *Stereotyped male vs. female work roles* (INV): perturbing stereotyped professions connected to the "unconventional" gender[5], 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";

- *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.

The hand-coded templates about body image and toxic masculinity, belonging to the **MFT test type**, do not always use explicitly hateful terms: occasionally prejudices are expressed in a more subtle way that models are often not able to handle. Sample records for the body image topic are "Janet is blonde, that's why she is not smart" and "Never let a `[offensive woman expression]` with no make-up on tell you what to `[verb]`". Examples of toxic masculinity are found in "Harry, you're not a real man if you don't have muscles" and "John, you're not a real man if you don't have a job". Regarding examples of insults toward a specific nationality or religion, the examples crafted consist in sentences like "Nobody really like `[nationality]`, they are `[offensive expression]`" and "Stop givin' money to them, `[offensive expression] [nationality]`". Stereotypes about disability, homeless people and old people, are obtained from templates like "All `[offensive term towards homeless]` are a burden to the society", "All `[offensive term towards disabled]` aren't capable of doing anything" and "Being `[offensive term to old people]` is funny after all, what are they complaining about?".

The **INVariance test type** involves assessing whether changing the value of sensitive attributes causes also a change in the label predicted or in the probabilities (i.e. without reason, revealing biases). Race, gender, sexual orientation, religion and nationality are examples of sensitive features. From simple templates like "`[name]` is `[nationality]`" are generated records as "Jane is Chinese", "Jane is Egyptian" and similar (for each test, it is possible to specify the exact number of instances to produce). Using instead the keys *gender* and *sexual orientation* will result in "Jane is asexual", "Jane is queer" and more. Applying this same procedure for the other sensitive keys will result in simplified (because the sentences are very similar) but very targeted synthetic data (see Section 4.2). The tests involving the perturbation of race, nationality, religion, gender and sexual orientation are those developed by CheckList's authors; we extend them by adding "professions" as protected attribute, in order to assess whether predictions change if a male or a female assumes a specific job role.

4.2 Synthetic datasets generation

After constructing the tests[6], 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

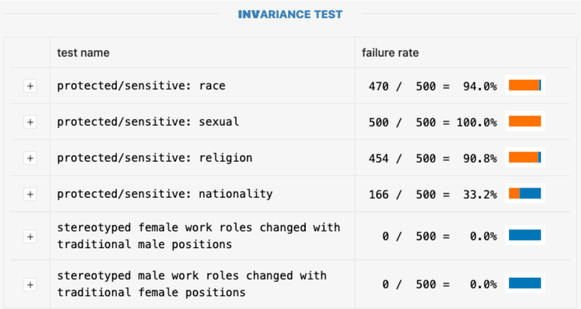
[5] The list used to identify the "swapped" professions is `https://github.com/uclanlp/corefBias/tree/master/WinoBias/wino`.

[6] Data Statement (Bender and Friedman, 2018): templates and related labels were manually defined by the first author, a non-native English speaker.

datasets, is separate from the standard CheckList process, which instead 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 (listed in 4.1), namely sexism, racism and ableism. The reason for distinguishing the records by abuse targets is due to the need for specialised datasets addressing different phenomena of abusive language with a fine-grained approach. The resulting data do not contain samples from datasets under license: the contents we release are therefore freely available[7].

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 abusive language 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.



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

5 System description

We run our evaluation using a standard BERT-based classifier for English, a language representation model developed by Google Research (Devlin et al., 2019), 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., 2019a). BERT can be fine-tuned and adapted to specific tasks by adding just one additional output layer to the neural network. We use this approach because language models like BERT, or variants like ALBERT and RoBERTa (Wiedemann et al., 2020), 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 the base model of BERT for English[8], trained on 3.3 billion words, which is made available on the project website (https://github.com/google-research/bert). We train two different classifiers in order to compare their behaviour w.r.t. biases. The first one is for generic abusive language detection, and is obtained by fine-tuning BERT on the (Founta et al., 2018) corpus. This dataset includes around 100K tweets annotated with four labels: hateful, abusive, spam or none. Differently from the other datasets, this 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 (Wiegand et al., 2019a). This should make this dataset more challenging for classification. For our experiments, we 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 serious form of abusive language and that the term 'abusive' is more appropriate to cover the cases of interest for our study (Caselli et al., 2020). The second model is trained with the AMI 2018 dataset (Fersini et al., 2018), which contains 4,000 tweets manually annotated as misogynistic or not. 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. Although BERT and similar language models may already encode biases (Bender et al., 2021), fine-tuning on different datasets may

[7] All the data and the Jupyter notebooks implemented to run the tests are available at https://github.com/MartaMarchiori/Fairness-Analysis-with-CheckList

[8] Uncased, 12-layer, 768-hidden, 12-heads, 110M parameters.

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).

indeed lead to a change in classification behaviour and therefore in its implicit biases.

6 Evaluation

In Table 1, we report a general overview of the performance of the two trained models on fairness tests. Each test involves 500 records randomly extracted from a larger subset, except for neutral statements feminism-related (200) and ableist stereotypes (220): the total number of records, considering all tests, amounts to 5,920. The metric computed by CheckList framework and reported in the table is the *failure rate*, i.e. the percentage of the records misclassified over the total number of records for that specific test[9]. Unlike metrics such as accuracy, the lower the failure rate (i.e. the closer to 0%) the better the model performs. In general, we notice that the overall failures are extremely high.

6.1 Fairness in Abusive Language Detection

Using the generic classifier trained on the dataset by (Founta et al., 2018), we observe that the hand-coded templates about body image and toxic masculinity, belonging to the MFT test type, are the most misclassified (respectively 92.8% and 99.2%). Regarding examples of insults toward a specific nationality or religion, the failure rate is of 30.2%. On stereotypes about disability, homeless people and old people, the model performs worse, reaching a failure rate of 43.2%.

With respect to the samples related to the perturbation of stereotyped professions connected to the "unconventional" gender, verified with the INVariance test type, the model shows zero failure. The issues arise when the sensitive features involved are *race*, *gender*, *sexual orientation* and *religion* (respectively 94%, 100% and 90.8% failures). This result means that overall the model is sensitive to alterations in these categories: probably this is caused by skewed training data, where e.g. the words "asexual" or "jew" in neutral, non-offensive contexts are not frequently attested. In addition, some sensitivity is demonstrated in changing the value of the protected attribute *nationality* (33.2% failure).

6.2 Fairness in Misogyny Detection

Using the model trained on the AMI dataset (Fersini et al., 2018), we observe some differences with respect to the generic abusive language model, as reported in in Table 1. The case where the change is most notable concerns stereotypes related to body image, for which the error drops from 92.8% to 8.6%. Analysing the perturbations of race, gender, sexual orientation and religion, we report a large decrease in errors: respectively from 94.0%, 100% and 90.8% for the first model to 14.8%, 54.0% and 1.6% for the second one. Surprisingly, comparing to the zero failures of the original model with respect to the perturbation of stereotyped professions, this last model reports 62% failures for stereotyped female work roles changed with "traditional" male positions. The same outcome is obtained for neutral identification statements related to feminism, where the first model reports

[9] Other significant metrics could be computed to strengthen the statistics obtained. Since this work is deeply rooted in CheckList framework, we focus our analysis on the options provided by the tool.

zero failures, while the second one achieves 76.5% failure.

This could be partially motivated by the fact that the Misogyny Classifier could have generalized a stereotyped conception of reality from skewed data on Misogyny Detection, e.g. learning to associate a high degree of toxicity with neutral posts containing terms such as *feminist* or negative correlation about women in positions of responsibility, since we can hypothesise that most of the examples the system was trained on contained references to these identities in offensive context.

7 Discussion and Conclusions

The approach that CheckList proposes should complement the evaluation of NLP models carried out by applying standard metrics such as F1 and accuracy. Indeed, in addition to the traditional held-out datasets, the creation of ad hoc examples, from the most basic ones to the most complex, contribute to highlight weaknesses that cannot be easily detected through large existing datasets. Furthermore, CheckList provides a way to explore the models' dynamics: through the analysis of the errors, we can infer which linguistic phenomena the system has not yet acquired from the data. However, in order to enable this fine-grained evaluation, several specific tests and templates should be created that, like in our case, may contain a small amount of examples because of the difficulty to create or retrieve a varied sample of records covering specific phenomena, e.g. feminist and ableist stereotypes.

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 (Sap et al., 2019). 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.

Although state-of-the-art models such as BERT-based 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., 2019b). 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. Moreover, the notebooks through which we built the suite are made available and the tests are easily editable and adaptable to specific data or linguistic aspects to be investigated.

A future direction of this work might be to expand the package integrating other linguistic resources, such as emotion or sentiment lexica. Concerning linguistic capabilities, for Fairness other stereotypes from a wider range of datasets could be more thoroughly explored and formalised into templates. It would be also interesting to analyse whether classification that takes into account the broader discourse context (Menini et al., 2021) is less prone to biases. Suites for other languages could be built as well, given that datasets for abusive language detection are available in many languages beyond English (Corazza et al., 2020).

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 hand-coded templates, that we generalized within the CheckList framework starting from real-user examples. The selection and the way in which the tests have been built certainly shaped the results.

Surely, this paper is not a complete or comprehensive work: for example, a direct interaction with the targeted users and the different stake-holders affected could have enriched the perspective and the insights retrieved. Furthermore, it is important to be aware that any solely technological solution 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).

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.

Acknowledgments

Part of this work has been funded by the KID_ACTIONS REC-AG project (n. 101005518) on "Kick-off preventIng and responDing to children and AdolesCenT cyberbullyIng through innovative mOnitoring and educatioNal technologieS", https://www.kidactions.eu/.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

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 the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation.

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*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

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*.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. 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*, pages 6193–6202, Marseille, France. European Language Resources Association.

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2020. A multilingual evaluation for online hate speech detection. *ACM Trans. Internet Techn.*, 20(2):10:1–10:22.

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.

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*, AIES '18, page 67–73, New York, NY, USA. Association for Computing Machinery.

Roel Dobbe, Sarah Dean, T. Gilbert, and Nitin Kohli. 2018. A broader view on bias in automated decision-making: Reflecting on epistemology and dynamics. *ArXiv*, abs/1807.00553.

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.

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.

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.

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*.

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*, pages 87–91, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

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. *CoRR*, abs/1907.11692.

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

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. *CoRR*, abs/2103.14916.

John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp.

Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended bias in misogyny detection. In *IEEE/WIC/ACM International Conference on Web Intelligence*, WI '19, page 149–155, New York, NY, USA. Association for Computing Machinery.

Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, Ioannis Kompatsiaris, Katharina Kinder-Kurlanda, Claudia Wagner, Fariba Karimi, Miriam Fernandez, Harith Alani, Bettina Berendt, Tina Kruegel, Christian Heinze, Klaus Broelemann, Gjergji Kasneci, Thanassis Tiropanis, and Steffen Staab. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *WIREs Data Mining and Knowledge Discovery*, 10(3):e1356.

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.

Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. 2019. Are red roses red? evaluating consistency of question-answering models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6174–6184, Florence, Italy. Association for Computational Linguistics.

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

Anna Rogers, Shashwath Hosur Ananthakrishna, and Anna Rumshisky. 2018. What's in your embedding, and how it predicts task performance. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2690–2703, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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

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*.

Harini Suresh and John V. Guttag. 2020. A framework for understanding unintended consequences of machine learning.

Yulia Tsvetkov, Manaal Faruqui, and Chris Dyer. 2016. Correlation-based intrinsic evaluation of word vector representations. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 111–115, Berlin, Germany. Association for Computational Linguistics.

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. 2019a. 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Michael Wiegand, Maximilian Wolf, and Josef Ruppenhofer. 2019b. Detecting Derogatory Compounds – An Unsupervised Approach. 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 2076–2081, Minneapolis, Minnesota. Association for Computational Linguistics.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2019. Errudite: Scalable, reproducible, and testable error analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 747–763, Florence, Italy. Association for Computational Linguistics.

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 the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

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.