Root Cause Analysis of Bias Detection and Classification in Natural Language Processing

Ana Evans
ana.s.evans@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

May 2022

Abstract

Human biases have been shown to influence the performance of models and algorithms in various fields, including Natural Language Processing. While the study of this phenomenon is garnering focus in recent years, the available resources are still relatively scarce. The aim of our work is to determine if, and/or how, we can take advantage of these previously-available resources, namely publicly-available datasets, to train models in the task of Biased-language Detection and Classification. We analyse the performance of the developed models, first on the test set of our original data and then on the OpenSubtitles corpus. We find that the combination of datasets influences model testing and performance and, most notably, that while we obtain promising results in terms of Precision, Recall, and F1-score, those do not translate to the OpenSubtitles testing phase, resulting in a discrepancy between the results of both testing phases. We also analyse some issues with the field of Bias in NLP, such as scarcity of resources, reliance on non-persistent data and lack of attention given to downstream tasks. We discuss these issues in tandem with the development of our work.
Keywords: Bias, Bias Detection, Bias Classification, Hate Speech, NLP

***Warning**: This work contains examples of explicit and/or offensive language.*

1. Introduction

In recent years, we have become more aware of how human biases can affect our models and algorithms. This growing awareness is reflected in the fields dedicated to this research, such as Bias in NLP, which has seen more and more works developed in recent years and focusing on a variety of topics, from methods for Bias Detection to figuring out how Bias even finds its way into our models.

The presence of biases in training data, utilized across the field, seems to be the most notable culprit. Creating new, unbiased data to train our models with appears to be the obvious solution, but it is a highly costly process. Teaching models how to detect – and, thus, remove – biased content from existing datasets seems more achievable.

There are few benchmark datasets aimed at this task, and those datasets that do exist are relatively small, often do not focus on the same types of Bias, and are not even aimed at the same downstream tasks. Therefore, before we even concern ourselves with effectively removing Bias from training data, we must take a step back. Instead, we must ask: can we learn how to detect bias using these pre-existing resources? And, if so, how?

In order to answer these questions, we will be selecting pre-existing datasets, developed in the scope of Bias and/or Hate Speech Detection, and using them to train a model in the task of Biased-language Classification. We will evaluate the developed model using test set obtained from splitting our training data, and also a separate dataset, frequently used as training data for Dialogue Models. This will allow us to understand how our model actually performs in the downstream task we have aimed to tackle.

Throughout this work, we will find ourselves grappling with a number of issues currently befalling the field of Bias in NLP. These issues will confront us in every phase of our work, and become abundantly clear as we progress. Thus, in tandem with the aforementioned goal, we will also be describing and discussing these issues, as well as analysing how they influence our work and the future of this field of study.

1.1. Ethical Statement

Due to our reliance in pre-existing resources, we have made a number of concessions regarding the complexities of the phenomenons being studied, such as the reduction of “Gender” to the two binary

genders (and further exclusion of non-binary identities) or the uncritical approach to "Race", which, as a construct, is highly dependent of the sociocultural or national context it is discussed in [15].

Intersectionality is a term coined by Kimberlé Crenshaw in 1989 [8]. It refers to an analytical framework through which we can understand the ways that the dimensions of an individual's identity intersect and combine, thus producing a social and personal experience that cannot be fully described by either facet in isolation. Although we recognize the importance of adopting an Intersectionality framework in works such as ours, we were unable to adopt this approach due to our reliance on pre-existing resources.

The inclusion of this section in the current body of work arises due to the awareness that the study of Bias and Hate Speech is inherently a sensitive subject, which must be conducted with a degree of awareness and responsibility. As such, we must be critical in regards to the limitations we face in our work, as well as the limitations of Bias and Hate Speech Detection as fields of study.

2. Background

"Bias" refers to unequal treatment of a given subject due to preconceived notions regarding that very same subject, which necessarily influence our judgement. "Social bias", therefore, translates to unequal treatment of certain individuals or groups based on specific shared characteristics – namely, social constructs such as race, gender, gender identity, etc.

The definition of Bias in NLP must always be *task-specific* [4]; that is to say, it must always depend on the task being researched. In the scope of our work, we have restricted the definition of "Bias" to three distinct manifestations:

- The use of *derogatory terms* which specifically target an individual or a group based on the defined social characteristics (for example "bitch", "dyke", "tranny");
- The prevalence of *stereotypes*, which can also manifest through harmful beliefs (i.e. "All Muslims are terrorists."), stereotypical societal roles (i.e. "Women belong in the kitchen."), caricatures (i.e. "The Angry Black Woman"), or even apparently benevolent beliefs (i.e. "Asians are good at math.");
- Otherwise abusive language which specifically targets a group or an individual based on the defined social characteristics (i.e. "Gay people make me sick!", "I'd never date a black guy.").

We furthermore define that we will be researching the aforementioned manifestations when aimed at a pre-defined set of Target Categories, namely: Gender, Race, Profession, Religion, Disability, Sexual Orientation, Gender Identity, Nationality, and Age.

In works similar to ours, we find that a term which often approximates our definition of Bias is "Hate Speech". This is described by Founta et al. as "Language used to express hatred towards a targeted individual or group, or is intended to be derogatory, to humiliate, or to insult the members of the group, on the basis of attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender." (2018:495) [17]. Although Bias and Hate Speech share some similarities, they are not quite the same; while instances of Hate Speech will always be instances of Bias, the same cannot be said in reverse. However, since our proposed definition of Bias closely flirts with the concept of Hate Speech, we will be utilizing resources from both fields.

3. Related Work

3.1. Overview

When it comes to the study of bias in NLP, Bolukbasi et al. [6] is an almost obligatory mention, having conducted one of the earliest studies we could find on the topic, focusing on Gender Bias in Word Embeddings. While more studies on Bias in Word Embeddings have been released since this initial study [5, 20, 22, 23, 39], we have also seen researchers further widening the scope of Bias in NLP, pouring over models or tools frequently used in various NLP tasks and study them under the lens of bias – sometimes as tools for detection and mitigation, other times as sources or propagators of bias. There is work focused on Neural Networks [37], on state-of-the-art models such as BERT [29, 31], techniques such as Adversarial Learning [26, 45], and various NLP tasks, such as Coreference Resolution [46], Sentiment Analysis [24], Dialogue Generation [10], and even POS tagging and Dependency Parsing [18].

Another way in which models developed in the scope of NLP can perpetuate Bias is through their training data. A significant number of datasets is composed of non-curated content from the Web, due to the sheer amount of information that can easily be collected from online forums and platforms. While there are advantages to this approach (like the aforementioned ease in collecting large amounts of data, or the usage of casual, every day language instead of synthetic syntax), the fact remains that there is plenty of unsafe and offensive content on the Internet, which is uncritically collected to build these datasets. Luccioni and Viviano's [27] examine the Common Crawl Corpus[1],

[1] https://commoncrawl.org/

with a focus on finding instances of Hate Speech and sexually explicit content. The Common Crawl is a multilingual corpus, composed of 200 to 300 TB of text obtained from automatic web crawling, and with new versions being released monthly. After resorting to a series of different detection approaches, they found that 4.02% to 6.38% of their sample contained instances of Hate Speech, while 2.36% contained material deemed as sexually explicit. These percentages quickly become alarming when one considers the total size of the corpus in question and how easily harmful content can be taught to models learning from this data.

Rather than focusing on finding instances of Bias in NLP, there is also work focused on using NLP to detect and classify Bias in other in real-life applications, such as analysing the Case Law Access Project (CAP) dataset [2] in regards to Gender Bias [3], analysing how Wikipedia pages portray LGBTQ people across different languages [32], or even determining whether there are noticeable differences in the way book critics review the works of male and female authors [41].

Hate Speech Detection, as a field of study, follows a similar focus as the aforementioned works; namely, in utilizing state-of-the-art models and Machine Learning to detect and classify instances of Hate Speech. The detection of instances themselves might be simple, "yes-or-no" binary classification without specifying whom that phenomenon targets, simply whether or not it is present [2, 9, 11, 17, 19]. We refer to these as "Binary Classification" datasets. Other works also focus on a particular category or demographic, like sexism [14, 21, 38] or Islamophobia [7]. They might also focus on a simple "yes-or-no" classification (is the phenomenon present or not), or they might create their own subcategories for specific manifestations of the phenomenon in question. We refer to these as "Single-Target Classification" datasets. Lastly, some works consider several targets categories at the same time [4, 29, 31, 43, 44], which we shall name "Multi-Target Classification".

The growing relevance of this field can be attributed to the increased importance of monitoring language online platforms. This is why a significant part of the data utilized in this field is retrieved from social media platforms, with most works favouring Twitter as a platform and keyword-based retrieval of keywords with negative polarity [33], although there is also a growing focus on creating synthetic data [43].

The usage of NLP tools as a way to detect and classify both Bias and Hate Speech has a couple of consequences: namely, the need for testing approaches which evaluate models specifically in the scope of these fields [28, 34, 35] and the need for training data, annotated in regards to Bias and/or Hate Speech, which allows researchers to train models in the first place [4, 7, 9, 14, 17, 19, 21, 29, 31, 43, 44].

[2] https://case.law/

3.2. Critiques and Limitations

While Bias Detection and Hate Speech detection are not the same field, they intersect substantially and share common pitfalls. For those reasons, the commentary of this section refers to both fields interchangeably, unless otherwise specified.

The first issue in the current state-of-the-art is the lack of established taxonomies or centralized resources, whether in terms of terminology or benchmark datasets. While plenty of works use terms such as "Bias","Hate Speech", or "Abusive language", the definitions associated with these terms are rarely in agreement. The absence of concise and concrete criteria leads to a "sparsity of heterogeneous resources" [33]. However, one might also argue that there is no such thing as a set of pre-established criteria that could or should be applied, since there are also no objectively correct definitions to be constructed. Following this reasoning, we should instead strive for more clarity in the terminology used, as well as in the subtasks being studied [42].

The second limitation we would like to mention refers to the disproportionate focus given to certain target categories in these fields. We can find many examples of work done in regards to sexism or gender bias, and, to a lesser extent, racism or racial bias. However, we will be hard pressed to find significant data regarding ableism, transphobia, anti-semitism, and many, many other categories worthy of a similar focus [4, 15, 42]. Additionally, works with gender as a target category often fail to conduct their research under an intersectional lens, thus reducing the nuance and depth of the phenomenon they propose to research [15].

Furthermore, also in relation to uneven distribution of resources, there is the sheer amount of resources devoted to the English language in comparison to any other language. While this is, to a degree, understandable, due to how widely used English is in international contexts such as online spaces, it is not sustainable. The choice to center English-speaking internet users in this research, implicit or unintentional as it may be, creates its own form of data bias [15, 42]. While some works done in other languages do exist, these are few and far in between [16, 47].

Lastly, we would like to expand upon the issue of bias induced by dataset annotation. As humans, we are all prone to inherent biases. This is why, in general, datasets will be annotated by

more than one person, and why measures such as inter-annotator agreement exist. In theory, these measures should allow labels to be chosen with as little bias as possible, especially if researchers resort to a diverse pool of annotators.

However, we can still find instances of annotation bias. Sap et al. [36] found that entries of Hate Speech datasets which are written in AAE (African American English) are more likely to be annotated as toxic or offensive. In turn, models trained on this data propagate this bias and are more likely to classify tweets written in AAE english as more offensive than their Standard English counterparts. Excell and Al Moubayed [13] found that male annotators are more likely to rely on slurs and offensive language in the annotation process, and that a high inter-annotator agreement between male annotators (higher than between female annotators) leads to the final labels being those picked by male annotators. Models trained with this data have a tendency to prioritize slurs and offensive words in their classification. However, Excell and Al Moubayed reported an increase of 1.8% in performance once they train their model solely with female-annotated data.

In conclusion, the fields of Bias and Hate Speech detection in NLP are currently suffering from a series of pitfalls, from lack of centralized resources and agreed-upon taxonomies, to an unbalanced distribution of those very same resources. Furthermore, bias in dataset annotation is an issue that easily goes unnoticed unless researchers specifically seek to correct it and learn to account for it. While many of these problems can generously be attributed to the novelty of the fields in question, it stands to reason that an effort should be made to mitigate them, sooner rather than later.

4. Methodology

As mentioned in our introductory section, the objective of our work was to collect and combine pre-existing resources, namely datasets developed in the scope of Bias and/or Hate Speech Detection, and evaluate if these could be used to successfully train a model in Bias Detection and Classification. In order to do this, we evaluated the developed models in regards to precision, recall, and F1-score, using the training data test set, as well as run them over a corpus frequently used as training data in a pre-determined NLP task. We settled on OpenSubtitles [25], through the OPUS corpus [40], frequently used to train Dialogue Models [1]. This decision influenced the definition of Bias we have described in Section 2 and adopted in our work.

4.1. Data Retrieval and Treatment

After conducting our initial research, we settled on using the datasets depicted in Table 1. These are all datasets which have been made publicly available and were developed in the scope of Bias or Hate Detection.

Dataset	Twitter based?	Classification Type
CONAN [7]	No	Single Target
Davidson [9]	Yes	Binary
DynGen [43]	No	Multi Target
Founta [17]	Yes	Binary
Golbeck [19]	Yes	Binary
Hostile Sexism [21]	Yes	Single Target
MLMA [30]	Yes	Multi Target
StereoSet [29]	No	Multi Target
Waseem-Hovy [44]	Yes	Single Target

Table 1: Dataset Collection

Some of these datasets, namely Hostile Sexism [21] and Waseem-Hovy [44], were not made available with their original Twitter text, but rather with the Tweet IDs of each Tweet. A Tweet ID is an alphanumerical identifier of a Tweet and, through the functionalities offered by Twitter API [3], can be used to Look-Up Tweets, thus allowing us to retrieve these datasets in their entirety. However, we faced some issues regarding the non-persistent nature of Twitter data during the retrieval process, which led us to restrict our usage of the aforementioned datasets. Waseem-Hovy, which was originally a Multi-Target Classification dataset which targeted both “Gender” and “Race”, was now reduced to a Single-Target Dataset for Target “Gender”, since many of the tweets labeled for the “Race” category became unavailable. Hostile Sexism was originally a component of the Benevolent-Hostile Sexism dataset, but we were forced to dismiss the Benevolent component. Further discussion on this issue will be conducted in Section 5.

After retrieving the missing Twitter data, we proceeded to uniformise our dataset collections. We replaced Twitter-specific markers, such as usernames or hashtags, by specific text markers which would later be saved as special tokens; we selected only the relevant content from each dataset and saved it to identically structured CSV files; and, finally, we established label coherency through label mapping, thus guaranteeing that all datasets in our collection followed the same label schema [12].

4.2. Model Training

4.2.1 Experimental Setup

For this work, we used the Emotion-Transformer[4], developed in the scope of Emotion Detection but adaptable to our Bias Classification task. The

[3] https://developer.twitter.com/en/products/twitter-api

[4] https://github.com/HLT-MAIA/Emotion-Transformer

Group Name	Datasets	Questions
A	Davidson [9] + Founta [17] + Golbeck [19]	Baseline
B	Group A + Hostile Sexism [21] + Waseem-Hovy [44]	How do single-target datasets influence performance?
C	Group A + DynGen [43] + MLMA [30] + StereoSet [29]	How do synthetic and multi target datasets influence performance?
D	Group C + CONAN [7] + Hostile Sexism [21] + Waseem-Hovy [44]	Can we obtain better performance by using all of our resources together?

Table 2: Dataset Groups

Emotion-Transformer is built on top of a pretrained Transformer model. In this work, we chose the DistilBERT pretrained model from HuggingFace[5], which served as a necessary compromise between temporal efficiency and overall performance.

To establish the Emotion-Transformer's level of performance, we trained it with individual datasets of our collection and compared the obtained results against results reported in the publication of those same datasets. Any comparison of results for Benevolent-Hostile Sexism and Waseem-Hovy would be invalid, due to the alterations these datasets suffered, described in the previous section. Additionally, DynGen was evaluated in a multi-labeling task, which would make our evaluation of it as a single-labeling task irrelevant.

Out of the remaining datasets, only Davidson and MLMA reported performance results. Davidson originally reported an F1-score of *0.9*, using a Support Vector Machine with L2 regularization [9]. MLMA does not specify what type of methods were used in training and testing, but reports an F1-score *0.43* as its best result for the relevant classification task [30].

We obtained an F1-score of *0.8* for Davidson, training the Emotion-Transformer during 5 epochs, with Binary Cross-Entropy with Logits Loss and *max* pooling function; and an F1-score of *0,42* for MLMA, training the Emotion-Transformer during 4 epochs, with the same Loss and Pooling functions described for the previous experiment. While the F1-score obtained for Davidson is lower than originally reported, the values are still similar. Thus, we conclude that the Emotion-Transformer is able to perform at a similar level to those models used to test the original datasets.

We divided our datasets into four non-exclusive groups, named Group A, Group B, Group C, and Group D. Group A, as the smallest and most coherent of the groups, serves as our baseline for performance comparison. Groups B, C, and D each serve to answer a research question. These are described in Table 2.

We performed a non-deterministic split of each group's data, splitting it into training, testing, and validation sets (80% train and 10% for testing and validation each). In total, we conducted over 100 experiments, in which we trained the model with different parameters and training data combinations.

Some of these parameters remained unchanged throughout experiments, such as Seed Value (*12*), Patience (*1*), Gradient Accumulation Steps (*1*), Batch Size (*8*), Number of Frozen Epochs (*1*), Encoder Learning Rate (*1.0e-5*), Classification Head Learning Rate (*5.0e-5*), and Layerwise Decay (*0.95*). These were the default values set for the Emotion Transformer.

The tested parameters were: Number of Training Epochs, Loss Function, and Pooling Function. The best F1-score results for Groups B, C, and D trained in Single-Target Classification (for Group B, with Target Category "Gender", as well as "Unspecified Bias" and "Non-biased") and Multi-Target Classification (for Groups C and D, with Target Categories "Gender", "Race", "Profession", "Religion", "Disability", "Sexual Orientation", "Gender Identity", "Nationality", and "Age", as well as "Unspecified Bias" and "Non-biased"), were all obtained using the same Loss Function (Binary Cross Entropy with Logits Loss). We shall refer to them as Multi-B, Multi-C, and Multi-D, and are the following:

- **Multi-B**: *F1 = 0.8842*, trained during 6 epochs with avg Pooling Function;
- **Multi-C**: *F1 = 0.6046*, trained during 6 epochs with max Pooling Function;
- **Multi-D**: *F1 = 0.6132*, trained during 4 epochs with avg Pooling Function.

We also trained models in Binary-Target Classification. We will refer to these experiments as Group A, Binary-B, Binary-C, and Binary-D. The best results for this set were the following:

- **Group A**: *F1 = 0.8974*, trained during 4 epochs with avg Pooling Function;
- **Binary-B**: *F1 = 0.8909*, trained during 4 epochs with avg Pooling Function;
- **Binary-C**: *F1 = 0.8597*, trained during 4 epochs with max Pooling Function;

[5] https://huggingface.co/docs/transformers/model_doc/distilbert

- **Binary-D**: *F1 = 0.8515*, trained during 4 epochs with avg Pooling Function.

Lastly, we also used the model trained with Group A data, shown above, and tested its performance in Binary Classification on the testing sets of the other groups, to compare what or if the other models were truly learning. We will refer to these as Inter-B, Inter-C, and Inter-D. The best results for this set were the following:

- **Inter-B**: *F1 = 0.8780*;
- **Inter-C**: *F1 = 0.7840*;
- **Inter-D**: *F1 = 0.7650*.

When we tested Multi-C and Multi-D, we observed that both models obtained an F1-score of 0 for the "Age" category, which has very few entries. While we will discuss this topic further in Section 5, we re-trained the models with Group C and Group D data, but this time without the "Age" category. We will call these experiments NoAge-C and NoAge-D. The results for this set were the following:

- **NoAge-C**: *F1 = 0.6770*, trained during 6 epochs with max Pooling Function;
- **NoAge-D**: *F1 = 0.6728*, trained during 6 epochs with max Pooling Function.

We chose NoAge-D to test on the OpenSubtitles corpus. This decision was mostly motivated by the fact that we wanted to test one of the models trained in Multi-Target Classification. Since NoAge-C and NoAge-D obtained extremely similar results and we were facing temporal constraints, we chose the model that we already had access to and would not have to retrain, namely NoAge-D. To supplement our analysis, we decided we would also use the Binary-D and Group A models. The former because it was trained with the same training data as NoAge-D, but in a Binary Classification task, and comparing the performance of both models allows us to understand what this difference translates to in practice. The latter because it is our baseline and overall our best performing model.

4.3. Bias Detection in OpenSubtitles

We used the B-Subtle framework to select subtitles from two groups:

- Movies from the "Animation" genre, released from 2010 to 2017;
- Movies from the "Comedy" genre, released from 2010 to 2017.

To clarify, movies in the "Animation" genre are not necessarily family movies. Animation will includes, for example, shows such as "The Simpsons", "Family Guy", or "American Dad", which are notably not made for a younger audience. This selection was motivated by the fact that these two genres frequently host content that is irreverent or satirical, thus prone to exhibiting the type of language we mean to target with our work. The temporal selection was motivated by the sociocultural shifts observed in the decade of 2010 to 2020, characterized by a growing awareness of how Bias and Hate Speech can manifest, how that can or should impact the way we express ourselves, or the media we consume. Since OPUS only includes titles produced until 2018, and since there is only a small collection of available titles produced in that year, we decided to restrict our selection from 2010 to 2017.

We separated the subtitles belonging to each genre, treating them as different data groups. We ran the NoAge-D model over both the Animation and Comedy sets. Then, due to temporal constraints, we ran the Binary-D model over the Comedy set, and the Group A model over the Animation set. We obtained the following results:

	Total Entries	**Biased Entries**
Animation (NoAge-D)	2,645,479	38,852
Comedy (NoAge-D)	2,722,056	6,730
Animation (Group A)	2,645,479	41,156
Comedy (Binary-D)	2,722,056	8,075

Table 3: Entries classified as "biased" by the NoAge-D, Binary-D, and Group A models

We then compiled all the entries that the models classified as "biased". From the results yielded by each experiment, we randomly selected 75 from each year and genre. These were evenly distributed between 3 annotators. Each annotator was assigned 50 entries out of the aforementioned 75. These entries purposefully overlapped with the entries assigned to the other annotators, so that every entry would be annotated by 2 annotators. In total, each annotator would deal with a total of 400 entries. Annotators were also given an Annotation Guide and asked to review the sentences assigned to them and to classify them in accordance to the definition of Bias adopted in this work and described in Section 2.

We calculated Inter-Annotator Agreement (IAA) with Cohen-Kappa Coefficient, Pearson Correlation Coefficient, and Raw Agreement, which is nec-

essary because neither Cohen-Kappa nor Pearson Correlation can be calculated if annotators classify all their sentences as the same category. We average values of these three metrics, as obtained by the different annotator pairs. The results are depicted in Table 4.

	Cohen Kappa	Pearson Correlation	R.A.
Animation (NoAge-D)	0.6250	0.6504	0.9350
Comedy (NoAge-D)	0.3976	0.4369	0.9900
Animation (Group A)	0.7932	0.7959	0.9567
Comedy (Binary-D)	0.5151	0.5372	0.9917

Table 4: Average of IAA metrics results

The reason we resorted to annotator review was because the subtitle corpora was not annotated in regards to Bias. This means that there was no previous gold standard against which we could compare the results yielded by these models. Therefore, in order to calculate our model's Accuracy, we use our annotator's response in regards to this 75-entry sample as a gold standard. We consider "True Positives" only those entries which both annotators classified as biased. Results shown in Table 5.

	True Positives	Total	Accuracy
Animation (NoAge-D)	37	600	0.062
Comedy (NoAge-D)	3	600	0.005
Animation (Group A)	59	600	0.098
Comedy (Binary-D)	6	600	0.010

Table 5: Accuracy of Bias Classification for models NoAge-D, Group A, and Binary-D

As we can see, the best result was obtained by Group A on the Animation corpus, which rounds up to *0.1*. NoAge-D on the Animation corpus is the second best result, with *0.06*. The Binary-D model on the Comedy corpus, with *0.01*, doubles the result obtained with the NoAge-D model on the same corpus, which was a mere *0.005*. These are extremely low results, especially compared to those presented when we ran these models on their testing set data. We will be discussing these results in the next section.

5. Discussion

In this section, we will discuss both results obtained in previous sections as well as issues observed in the duration of this work, which largely relate to these results. Namely, the consequences of relying on non-persistent data to compose training datasets, the lack of coherence across available resources in regards to linguistic conventions, definitions, and skewed focuses, and the importance of testing developed models in the downstream task they were designed for.

5.1. Non-Persistent Data and Dataset Degradation

In Section 4, we briefly mentioned that due to privacy concerns, some Twitter-based datasets do not publicly share the textual content of the their collected tweets. Rather, they share Tweet IDs, which can be used to retrieve the text of the correspondent tweet.

Here is the catch: a tweet can only be retrieved *if that tweet still exists*. If we attempt to retrieve a tweet which no longer exists, or is no longer available, we will receive an error code and message. This means that some of this information is *non-recoverable* and, consequently, that Twitter-based datasets may be prone to *degradation*.

Once we realized this, we chose to not only analyse the results we had obtained in the scope of this issue, but also to repeat the retrieval process with the Founta dataset. Founta et al. [17] responded to privacy concerns by separating tweet identifiers and tweet text into separate files and then sharing both files, rather than withholding the text altogether. Ergo, we still possess the identifiers and are free to use them for our analysis.

Dataset	Total	Currently Available	Currently Unavailable
Benevolent Sexism	7,210	2,411	4,799
Hostile Sexism	3,378	2,718	661
Founta	99,996	53,857	46,139
Waseem-Hovy	16,907	10,370	6,537
Total	127,491	69,356	58,136
Total (%)	100.00%	54.40%	45.60%

Table 6: Unavailable Tweets Breakdown

The results of our analysis regarding unavailable tweets, across all four datasets, can be found in Table 6. Since Benevolent-Hostile Sexism separated the Benevolent and Hostile components into two files and their yielded results differed significantly, we chose to showcase them separately.

As can be seen in Table 6, *45.60%* of the tweets collected in these datasets had, at the time of retrieval, become unavailable. Additionally, we found that most unavailable tweets were either deleted or posted by deleted accounts (*46.61%* of unavailable tweets and *21.25%* of all the tweets in the

	A	A(%)	B	B (%)	C	C (%)	D	D (%)
Non-Biased	81,112	64.82%	88,754	64.22%	109,265	59.19%	120,851	59.79%
Biased (Unspecified)	44,016	35.18%	44,016	31.85%	51,947	28.14%	51,947	25.70%
Gender	-	-	5,433	3.93%	3,182	1.72%	8,615	4.26%
Race	-	-	-	-	10,613	5.75%	10,613	5.25%
Profession	-	-	-	-	1,855	1.00%	1,855	0.92%
Religion	-	-	-	-	2,632	1.43%	3,147	1.56%
Disability	-	-	-	-	1,575	0.85%	1,575	0.78%
Sexual Orientation	-	-	-	-	1,854	1.00%	1,854	0.92%
Gender Identity	-	-	-	-	1,132	0.61%	1,132	0.56%
Nationality	-	-	-	-	528	0.29%	528	0.26%
Age	-	-	-	-	23	0.01%	23	0.01%

Table 7: Breakdown of categories across data groups

datasets). A significant percentage was posted by accounts which were suspended at time of retrieval (*42.98%* of unavailable tweets and *10.60%* of all tweets).

This is not as surprising as it might appear at first. On one hand, deleting an account is not unusual. This fact alone means that the length of time between dataset creation and retrieval of a tweet ID contained in that dataset is proportional to the likelihood of that tweet becoming unavailable. On the other hand, and further exacerbating the previous point, Twitter allows users to flag or report content that they might find offensive. If the reported tweets are concluded to be so by Twitter's moderation team, accounts might find themselves suspended as a result. It is unsurprising that tweets belonging to a Hate Speech or Bias detection dataset might fall into this category, and thus that these datasets degrade over time.

However, unsurprising as it may be, it still warrants concern. Datasets are not only important resources, they are also inherently costly. That their value may deprecate over time due to reliance on non-persistent information presents a serious challenge, especially for a field as dependent on online-based resources as Hate Speech detection. Perhaps solutions such as Founta et al. [17], which still address privacy concerns while circumventing the issue of degradation, should be prioritized over simply sharing Tweet IDs with little to no regard as to the preservation of the data in question.

5.2. Diversity of Available Resources (or Lack Thereof)

Looking at the results obtained by our model training, and described in Section 4.2.1, we can clearly see a discrepancy between the results yielded in Binary Classification or Single-Target Classification experiments (namely, Group A, Multi-B, Binary-B, Binary-C, and Binary-D) and those yilded in Multi-Target Classification (namely, Multi-C, Multi-D, NoAge-C, and NoAge-D).

The main difference here is, at first glance, the difference of categories which the model is attempting to learn. Less obvious, perhaps, is the distribution of resources across those categories. All datasets roughly follow a one-third/two-thirds composition in regards to biased/non-biased entries, respectively. The decomposition of biased entries across categories, however, varies significantly. This can be seen in Table 7.

If we look at the results for Inter-B, Inter-C, and Inter-D, it is clear that not only does the Group A model not perform nearly as well on the other Groups' data as it does on its own, it also performs worse than experiments like Binary-B, Binary-C, and Binary-D. This leads us to conclude that these models are indeed learning from their training data how to identify forms of bias that the model trained solely on Group A data is unable to identify. Thus, using our resources conjointly does teach models new information.

However, these results fall apart once we try to teach them how to identify different categories of Bias. This is easily justifiable by the fact that, even with all these datasets, plenty of categories simply do not have enough content for the models to meaningfully learn how to identify them. "Gender", "Race", and "Religion" are the only categories that make up more than 1.00% of all available data for Groups C and D, with "Religion" never reaching 2.00%. The "Age" category is so insignificant that we removed it altogether and obtained NoAge-C and NoAge-D, which show an increased F1-score compared to Multi-C and Multi-D, respectively. Additionally, the most notable difference in performance between models trained with Group D and Group C data was their increased scores for the

"Gender" and "Religion" categories – rather noticeably, those targeted by the Single-Target datasets added to Group D.

This makes it rather clear that, while linguistic conventions across Twitter-based and synthetic datasets may impact model's performance, the more significant issue is the lack of available resources which we can use to teach our models how to recognize bias in regards to categories which are not "Gender" and "Race". This was an issue which was first introduced in Section 3.2 and which we can now see being reflected in practice.

We have learnt that models can learn to identify bias for a certain target category if trained with general/unspecified Bias/Hate Speech Detection datasets and a smaller number of entries labeled for a single category. This can be seen in the results obtained by Multi-B, Binary-B, and Inter-B. The quantity of entries necessary to obtain a satisfactory performance may or may not depend on whether these entries obey similar linguistic conventions as the general Bias/Hate Speech Detection datasets and/or if the utilized language is often found in the general datasets. However, the fact that it is possible at all is an extremely positive outcome, since it means that further research can focus on less costly strategies to teach models how to identify Bias for under explored categories.

5.3. Practical Accuracy vs. Theoretical F1: Result Discrepancy

We shall begin this section by analysing some of the results obtained in Section 4.3, and later by discussing the extreme discrepancy between the results obtained by testing our models with their testing data sets and testing them on the OpenSubtitles data. The relevant information for these discussions is depicted in Tables 3, 4, and 5.

There are some interesting insights to be garnered, such as the fact that the Group A model, tested on the Animation corpus, yields the best results on Accuracy, Cohen-Kappa, and Pearson Correlation. It is also the model with the highest count of biased entries. This combination of facts leads us to believe that this is the model that will overall accurately classify the largest amount of biased content. This is not a surprise, since we defined Group A as our testing baseline because we expected it would perform better than the rest. However, Group A was also never the main focus of our work – hence why it served only as a baseline.

More interesting is the difference in performance between NoAge-D and Binary-D on the Comedy corpus. Binary-D doubles the Accuracy score of NoAge-D, but these models were trained with the same training and validation data. The only significant difference between them is that NoAge-D was trained in the task of Multi-Target Classification and Binary-D was trained in the Binary Classification task, leading us to the conclusions discussed in previous section: namely, that the model's performance is definitely harmed when it attempts to learn the different Target Categories, which have a lot less available entries from which the model can learn from in the first place. Once more, this merely enforces our belief that there is urgent need to create more diverse and inclusive resources, rather than simply directing our attention towards one or two Target Categories which have already been more thoroughly invested in.

Additionally, we calculated the number of sentences which were labeled as biased by both experiment pairs (that is to say, by the pair of experiments conducted on each corpus). We found that there was a significantly higher overlap between the experiments conducted over the Comedy corpus in comparison to those in in the Animation corpus. This translates to *40.86%* and *34.06%* of all entries classified as biased by NoAge-D and Binary-D on the Comedy corpus, respectively, against a mere *12.87%* and *12.18%* of NoAge-D and Group A, respectively. A proper, sentence-by-sentence analysis of this overlap could yield illuminating results – we will have to, unfortunately, leave that to future work.

There is still more insight to be garnered from these experiments, more in regards to the content of the subtitles themselves. For example, the higher rate of Raw Agreement for both experiments ran over the Comedy corpus is a direct contrast to the Cohen-Kappa and Pearson Correlation Coefficients, but is also a relatively simple phenomenon. Since the Comedy corpus had a higher amount of non-biased sentences, or sentences in which the bias was less ambiguous, annotators reached an easier understanding than annotators of the Animation corpus. This supports the hypothesis that subtitles belonging to the Animation genre contain a higher amount of biased – or ambiguously biased – content than those of the Comedy genre.

Lastly, after observing and discussing the obtained results, we may now refer back to the research question motivating this work: "How can pre-existing resources, namely publicly available datasets, be used to train classifiers in the task of Bias Classification – if they can be used to this end at all?". We can now state that the answer to this question is: "They cannot – or, at least, not in this way."

Our models failed profusely, even our baseline, which featured a reasonably balanced split between classes, was composed solely of Twitter-based data and thus unlikely to fall prey to issues

resulting from being trained with different linguistic conventions, and composed by datasets which generally followed similar conventions and definitions. These were the problems we expected and prepared to tackle when we devised our dataset groups. Evidently, "extremely poor performance in the downstream task" was not one of those problems.

There are a number of concessions that can be made to partially justify this result. After all, we did not set out to build a highly specialized model, and the pre-trained model we did use, namely DistillBERT, is not as good as models such as BERT or RoBERTa. Either one of these changes could, and quite possibly would, have resulted in better performance of the developed models, as well as higher Accuracy.

There are other variables, however, that we can and should question. For example, the usefulness of the datasets we used in this work when used to train models in the sort of task we aimed for – or, even, in any downstream task. After all, we achieved very fair results in terms of precision, recall, and F1-score when we tested our models initially, which were not reflected in our downstream tasks.

The difference between those high scores and the extremely low Accuracy revealed in this work is, perhaps, the most significant conclusion that we can derive from this work. A notable majority of the datasets we collected, and even of those we found in later research, did not use their datasets in any sort of downstream task. After confronting the results of our work, we truly believe it is paramount for researchers to not only be clear in the downstream tasks they intend to tackle, but also, and most importantly, to take the extra step and properly test their work in the context of that very same task. This would allow researchers to obtain better understanding of their work and, consequently, bring significant advances to any field of study.

6. Conclusions

The field of Bias in NLP is growing quickly and garnering much needed attention. However, this field is also suffering from a number of significant pitfalls, such as skewed resources which tend to disproportionately target one or two types of biases while essentially ignoring others, incoherence in term usage and definitions across works, or even a lack of attention towards the downstream tasks being affected by the developed works.

In our work, we sought to discover if we could use pre-existing, publicly available resources to train a well-performing model in the task of Bias Detection and Classification. In order to determine this, we tested our model not just on the testing set of our training data, but also using subtitle corpora.

During our work, we ran into a series of issues, including some of the aforementioned ones. The reliance on non-persistent data leads to dataset degradation, which further sabotages whatever available resources exist. The disproportionate attention given to certain targets of bias means that there are not enough resources available to properly train models to identify those types of biases. And, lastly, that while we can obtain a satisfactory model performance when testing models with our test sets, this performance may not be reflected in the downstream task we aim to tackle.

These conclusions emphasize the need for clarity and diversity in further research in this field. It is paramount to diversify the focus of research, especially in an age in which social biases continue to grow in social importance. Technological advances must keep pace with societal ones, and that goal cannot be achieved if we remain stagnant and do not pay heed to recurring mistakes.

References

[1] D. Ameixa, L. Coheur, P. Fialho, and P. Quaresma. Luke, i am your father: dealing with out-of-domain requests by using movies subtitles. In *International Conference on Intelligent Virtual Agents*, pages 13–21. Springer, 2014.

[2] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760, 2017.

[3] N. Baker Gillis. Sexism in the judiciary: The importance of bias definition in NLP and in our courts. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 45–54, Online, Aug. 2021. Association for Computational Linguistics.

[4] S. Barikeri, A. Lauscher, I. Vulić, and G. Glavaš. RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1941–1955, Online, Aug. 2021. Association for Computational Linguistics.

[5] C. Basta, M. R. Costa-jussà, and N. Casas. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in*

Natural Language Processing, pages 33–39, 2019.

[6] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357, 2016.

[7] Y.-L. Chung, E. Kuzmenko, S. S. Tekiroglu, and M. Guerini. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy, July 2019. Association for Computational Linguistics.

[8] K. Crenshaw. *Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory, and antiracist politics [1989]*. Routledge, 2018.

[9] T. Davidson, D. Warmsley, M. Macy, and I. Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, 2017.

[10] E. Dinan, A. Fan, A. Williams, J. Urbanek, D. Kiela, and J. Weston. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online, Nov. 2020. Association for Computational Linguistics.

[11] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30, 2015.

[12] A. Evans. Root Cause Analysis of Bias Detection and Classification in Natural Language Processing (Masters Thesis). 2022.

[13] E. Excell and N. Al Moubayed. Towards equal gender representation in the annotations of toxic language detection. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 55–65, Online, Aug. 2021. Association for Computational Linguistics.

[14] E. Fersini, P. Rosso, and M. Anzovino. Overview of the task on automatic misogyny identification at ibereval 2018. *IberEval@ SEPLN*, 2150:214–228, 2018.

[15] A. Field, S. L. Blodgett, Z. Waseem, and Y. Tsvetkov. A survey of race, racism, and anti-racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1905–1925, Online, Aug. 2021. Association for Computational Linguistics.

[16] P. Fortuna, V. Cortez, M. Sozinho Ramalho, and L. Pérez-Mayos. MIN_PT: An European Portuguese lexicon for minorities related terms. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 76–80, Online, Aug. 2021. Association for Computational Linguistics.

[17] A. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 2018.

[18] A. Garimella, C. Banea, D. Hovy, and R. Mihalcea. Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498, 2019.

[19] J. Golbeck, Z. Ashktorab, R. O. Banjo, A. Berlinger, S. Bhagwan, C. Buntain, P. Cheakalos, A. A. Geller, R. K. Gnanasekaran, R. R. Gunasekaran, et al. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on web science conference*, pages 229–233, 2017.

[20] W. Guo and A. Caliskan. *Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases*, page 122–133. Association for Computing Machinery, New York, NY, USA, 2021.

[21] A. Jha and R. Mamidi. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the second workshop on NLP*

and computational social science, pages 7–16, 2017.

[22] M. Jiang and C. Fellbaum. Interdependencies of gender and race in contextualized word embeddings. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 17–25, Barcelona, Spain (Online), Dec. 2020. Association for Computational Linguistics.

[23] M. Kaneko and D. Bollegala. Gender-preserving debiasing for pre-trained word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1641–1650, 2019.

[24] S. Kiritchenko and S. Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[25] P. Lison and J. Tiedemann. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. 2016.

[26] H. Liu, W. Wang, Y. Wang, H. Liu, Z. Liu, and J. Tang. Mitigating gender bias for neural dialogue generation with adversarial learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 893–903, Online, Nov. 2020. Association for Computational Linguistics.

[27] A. Luccioni and J. Viviano. What's in the box? an analysis of undesirable content in the Common Crawl corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 182–189, Online, Aug. 2021. Association for Computational Linguistics.

[28] M. M. Manerba and S. Tonelli. Fine-grained fairness analysis of abusive language detection systems with CheckList. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 81–91, Online, Aug. 2021. Association for Computational Linguistics.

[29] M. Nadeem, A. Bethke, and S. Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online, Aug. 2021. Association for Computational Linguistics.

[30] N. Ousidhoum, Z. Lin, H. Zhang, Y. Song, and D.-Y. Yeung. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[31] P. Parikh, H. Abburi, P. Badjatiya, R. Krishnan, N. Chhaya, M. Gupta, and V. Varma. Multi-label categorization of accounts of sexism using a neural framework. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1642–1652, 2019.

[32] C. Y. Park, X. Yan, A. Field, and Y. Tsvetkov. Multilingual contextual affective analysis of lgbt people portrayals in wikipedia. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 479–490, 2021.

[33] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, and V. Patti. Resources and benchmark corpora for hate speech detection: a systematic review. *Lang. Resour. Evaluation*, 55:477–523, 2021.

[34] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh. 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, July 2020. Association for Computational Linguistics.

[35] P. Rottger, B. Vidgen, D. Nguyen, Z. Waseem, H. Margetts, and J. Pierrehumbert. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online, Aug. 2021. Association for Computational Linguistics.

[36] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith. 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, July 2019. Association for Computational Linguistics.

[37] S. Sharifirad, A. Jacovi, I. B. I. Univesity, and S. Matwin. Learning and understanding different categories of sexism using convolutional neural network's filters. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 21–23, 2019.

[38] A. Suvarna and G. Bhalla. # notawhore! a computational linguistic perspective of rape culture and victimization on social media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 328–335, 2020.

[39] Y. C. Tan and L. E. Celis. Assessing social and intersectional biases in contextualized word representations. *arXiv preprint arXiv:1911.01485*, 2019.

[40] J. Tiedemann. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer, 2012.

[41] S. Touileb, L. Øvrelid, and E. Velldal. Gender and sentiment, critics and authors: a dataset of norwegian book reviews. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 125–138, 2020.

[42] B. Vidgen, A. Harris, D. Nguyen, R. Tromble, S. Hale, and H. Margetts. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy, Aug. 2019. Association for Computational Linguistics.

[43] B. Vidgen, T. Thrush, Z. Waseem, and D. Kiela. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online, Aug. 2021. Association for Computational Linguistics.

[44] Z. Waseem and D. Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93, 2016.

[45] B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[46] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. 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, 2018.

[47] P. Zhou, W. Shi, J. Zhao, K.-H. Huang, M. Chen, R. Cotterell, and K.-W. Chang. Examining gender bias in languages with grammatical gender. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5276–5284, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.