On Gender Biases in Offensive Language Classification Models

Sanjana Marcé[1] **Adam Poliak**[2*]
[1]Columbia University
[2]Bryn Mawr College
apoliak@brynmawr.edu

Abstract

We explore whether neural Natural Language Processing models trained to identify offensive language in tweets contain gender biases. We add historically gendered and gender ambiguous American names to an existing offensive language evaluation set to determine whether models' predictions are sensitive or robust to gendered names. While we see some evidence that these models might be prone to biased stereotypes that men use more offensive language than women, our results indicate that these models' binary predictions might not greatly change based upon gendered names.

1 Introduction

Identifying offensive language in text is an increasingly important challenge that has sparked the release of datasets and advanced models focused on toxic language detection in multiple languages (Razavi et al., 2010; Pitenis et al., 2020; Sigurbergsson and Derczynski, 2020; Çöltekin, 2020; Founta et al., 2018). For these models to be trustworthy when deployed in sensitive, real-world contexts, they must perform equally well for text written by male, female, or non-binary authors.

However, based on known gender-based biases in NLP systems (Rudinger et al., 2018; Zhao et al., 2018; Sun et al., 2019; Gaut et al., 2020; Stanovsky et al., 2019; Savoldi et al., 2021), especially among models trained to detect abusive language (Park et al., 2018), we hypothesize that existing NLP systems that incorporate pre-trained word embeddings or transformer-based language models will perform differently given access to authors' names if those names are generally associated with a particular gender.[1] To test the hypothesis that offensive language identification models exhibit gender biases, we adopt the *Perturbation Sensitivity Analysis* framework (Prabhakaran et al., 2019). We perturb examples of an existing dataset by adding historically gendered or gender-ambiguous names to the original texts. We evaluate whether three classes of NLP models (bag of words, BiLSTM, and transformers) systematically change their predictions on our modified gendered examples.

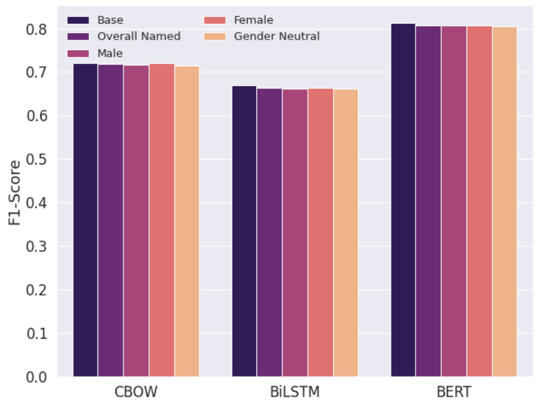


Figure 1: F1 Scores of the CBoW, BiLSTM, and BERT models isolated by each gender. The models' predictions do not noticeably change based on the gender of named examples.

Although we see statistically significant differences when comparing a bag of words model's and transformer model's predictions between male and female examples, we do not see convincingly strong evidence that the models' binary predictions for offensiveness consistently change with the addition of gendered names (Figure 1). Therefore, we compare how the model's predicted offensiveness probability changes for perturbed examples. We also explore if there are specific names for which the predicted class probability consistently changes. While we see some remnants of gendered

[*] Work performed while at Barnard College

[1] In this paper we use an author's name assigned at birth as a proxy for their gender. While we acknowledge the limitations associated with inferring gender from an individual's name, in doing so we recreate real-world circumstances in which NLP systems would make gendered associations based upon a speaker's or author's name even when their gender is not explicitly mentioned.

Proceedings of the The 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 174 - 183
July 15, 2022 ©2022 Association for Computational Linguistics

Example	CBoW	BiLSTM	BERT
▶ *@USER You are missing brains?*	0.741306	0.999869	0.839001
♀ *Vanessa tweeted @USER You are missing brains?*	0.859568	0.999833	0.755025
♂ *Matthew tweeted @USER You are missing brains?*	0.859568	0.999833	0.756230
○ *Oakley tweeted @USER You are missing brains?*	0.859568	0.999833	0.735549

Table 1: An example of an offensive tweet from the development set and the offensiveness probability each model (*CBoW, BiLSTM, BERT*) assigned to the unmodified (▶), female (♀), male (♂), and gender-neutral versions (○).

biases, our results offer encouraging evidence that downstream models using pre-trained representations that are known to encode gendered stereotypes (Bolukbasi et al., 2016; Garg et al., 2018; Zhao et al., 2018) might overcome these biases.

2 Motivation & Bias Statement

As user-generated content gradually dominates online spaces, offensive text has become more ubiquitous (Banks, 2010; Kumar et al., 2020). Unregulated inflammatory or hateful online discourse can have profound effects that extend beyond the web, from negative mental health impacts for targeted individuals to instigation of physical violence (Safi Samghabadi et al., 2020; Siegel, 2020). Hence, identifying and moderating toxic dialogue efficiently and accurately is a task that only grows more crucial, and developing automatic methods to detect and flag offensive language is critical.

Psychological studies spanning the past four decades conclude that, on average, "men use offensive language more than women" (although this gap has shrunk over time), likely as a result of how women are "socialized into subordinate roles and a less inflammatory manner of communicating" (Sapolsky and Kaye, 2005). Moreover, these observed patterns of offensive or abusive content authorship translate to online communities like Twitter (Mubarak et al., 2021).

Research into fairness in NLP indicates that systems trained on large corpora of human-written text tend to replicate existing stereotypes about gendered behavior (Sun et al., 2019; Babaeianjelodar et al., 2020). Thus, offensive language detection classifiers based on social-media data risk inheriting these underlying assumptions that male-authored tweets are more likely to utilize offensive language than text written by female individuals.

As it becomes more common for social media platforms to rely on NLP systems to detect and remove profane or hateful content online, it becomes increasingly vital that these classification models are robust to gender biases. While previous research has considered identity-based bias against a gendered *subject* in abusive language tasks (Park et al., 2018; Prabhakaran et al., 2019) and gender-based biases among annotations (Excell and Al Moubayed, 2021), how the perceived gender of a *speaker* or *author* affects output model classification remains understudied.

3 Experimental Setup

Our goal is to determine whether offensive language identification models are prone to gender biases. We train bag of word, BiLSTM, and transformer-based models on the Offensive Language Identification Dataset (OLID; Zampieri et al., 2019a). OLID is the official dataset used in the OffensEval shared tasks (Zampieri et al., 2019b, 2020), where tweets containing profanity, insults, threats, hate speech, etc, are labeled as offensive (Zampieri et al., 2019a). OLID contains 13,240 annotated English-language tweets (4400 offensive, 8840 not offensive) and 860 test examples (240 offensive, 620 not offensive). For model training, we split the original training set into 12,380 training and 860 dev examples.[2]

3.1 Gendered Test Set Creation

In order to evaluate whether the models' predictions change when the text explicitly mentions the author of the tweet, we modify the 860 test set examples using the following template:[3]

(1) *Name* tweeted *original tweet*

where *original tweet* is the original test example

[2] We provide all model implementation and hyperparameter tuning details in subsection 3.2.

[3] This template is similar to those previously used to evaluate natural language inference systems' abilities to capture different semantic phenomena (Poliak et al., 2018) and gender bias in named entity recognition systems (Mehrabi et al., 2020).

Hyper-parameter	Options	Models Used
Batch Size	16, 32, 64, 128	CBoW, BiLSTM, BERT*
Num. Hidden Features	1, 3, 5, 16, 64	CBoW, BiLSTM
Learning Rate	0.1, 0.01, 0.001, 0.0001	CBoW, BiLSTM
Dropout Rate	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9	CBoW

Table 2: Permutations of hyper-parameter configurations tested, with the models that use each hyper-parameter. *Due to machine memory constraints, only batch sizes of 16 and 32 were tested for BERT.

and *Name* is replaced with a name from an aggregated list of 212 historically gendered and gender ambiguous names in the United States to create a test set of 182,320 named tweets. Table 1 provides an example from our dataset.

Using standard practice (Vogel and Jurafsky, 2012; Bamman et al., 2014), we create a list of traditionally gendered names using publicly available government statistics. In particular, we compile data from the Social Security Administration's annual list of American baby names from 2000-2018.[4] We aggregate names with $p(gender|name) \geq 0.9$, filter out those names not recognized as singular tokens by the BERT and GloVe vocabularies, preventing OOV issues.[5] We select the top 100 most frequent names ascribed to newborns assigned female or male at birth.

While current research suggests that toxic language models may perform differentially on gendered input (Park et al., 2018), work remains to be done on how these models may misclassify text written by authors who do not conform to the gender binary. Therefore, we also include six gender-neutral names (*Justice, Milan, Lennon, Oakley, Marion,* and *Jackie*) that appear at approximately similar gender frequencies in the SSA data ($0.9 \geq \frac{p(male|name)}{p(female|name)} \geq 1.1$), are recognized by both pre-trained vocabularies, and were assigned to at least 4,000 newborns over the considered time-frame. We add one male (*he*), one female (*she*), and four gender-neutral pronouns (*one, they, someone* and *a person*).

3.2 Implementation Details

We explore classifiers based on three different classes of neural encoders. Each model was tested on a range of hyper-parameter configurations (Table 2), and the best configuration was chosen based on maximizing F1-Score on the validation set. Our trained models achieve comparable performance on the unnamed validation set to published results for similar classes of models on OLID (Ramakrishnan et al., 2019; Mahata et al., 2019; Zampieri et al., 2019a; Wu et al., 2019; Pavlopoulos et al., 2019; Aggarwal et al., 2019; Zhu et al., 2019).

Neural Bag of Words We trained a Continuous Bag of Words model (CBoW) to build classifiers for offensive and not offensive tweets and predict the output class of a new tweet based on the average vector representation of its tokens. To process the input examples, we use the NLTK tweet tokenizer and 100-dimensional GloVe embeddings (Pennington et al., 2014) pre-trained specifically for Twitter-sourced text.[6] Our CBoW model consists of a multi-layer perceptron (MLP) with a single hidden layer with one feature built on top of an embedding layer. The best performing model uses a batch size of 16 for training and validation, a learning rate of 0.001, and a dropout rate of 0.9 for regularization.

BiLSTM encoder The second type of encoder we consider is a Bidirectional LSTM (BiLSTM) (Schuster and Paliwal, 1997; Hochreiter and Schmidhuber, 1997). We process the input using the same tweet tokenizer and Twitter-trained GloVe embeddings as in the CBoW model. The best performing BiLSTM model architecture consists of a bidirectional LSTM layer with 128 output features and a MLP with 64 features in the hidden layer. For this model, weights are updated during training with a learning rate of 0.001 in an Adam optimizer and a training and validation batch size of 64.

[4] Prior research has similarly extracted gendered names from the Social Security Administration (Smith et al., 2013; Mohammad, 2019; Garg et al., 2018; HallMaudslay et al., 2019; Mehrabi et al., 2020; Shwartz et al., 2020)

[5] Filtering for names recognized by the BERT and GloVe vocabularies when collecting the top 100 gendered names recognized by pre-trained embeddings removed 11 more female names than male names. This might illustrate a bias against traditionally female names in these representations.

[6] Twitter GloVe embeddings downloaded from https://nlp.stanford.edu/projects/glove/

	CBoW				BiLSTM				BERT			
	TN	FN	TP	FP	TN	FN	TP	FP	TN	FN	TP	FP
-	62.79	12.44	15.47	09.30	58.95	13.37	14.53	13.14	65.12	07.91	20.00	06.98
♂	61.52	11.87	16.03	10.57	58.19	13.43	14.48	13.91	65.11	08.28	19.62	06.98
♀	61.92	11.93	15.97	10.17	58.47	13.48	14.42	13.62	65.35	08.55	19.36	06.75
○	61.51	12.05	15.86	10.58	58.09	13.41	14.50	14.00	65.13	08.43	19.48	06.97

Table 3: Aggregated confusion matrices of the CBoW, BiLSTM, and BERT models evaluated on the original, unmodified (-) test tweets and each named gender subgroup (male ♂, female ♀, and gender neutral ○). To enable easier comparisons, we normalized counts in the confusion matrices so that each cell represents the percentages of each type of prediction the models made across each gender.

Gender Model	-	♂	♀	○
CBoW	71.98	71.70	71.98	71.41
BiLSTM	66.97	66.21	66.37	66.16
BERT	81.31	80.75	80.60	80.55

Table 4: F1 scores for each model on the original unnamed (-) and male (♂), female (♀), and gender neutral (○) examples.

	t-stat	p-value
CBoW	2.1615	0.0153
BiLSTM	1.5833	0.0567
BERT	2.3691	0.0089

Table 5: Result of one-sided t-test comparing each models' predictions for male vs female authored-examples.

Transformers We fine-tune a HuggingFace pretrained BERT base-uncased model (Wolf et al., 2020) on our offensive training set using 2 epochs, 50 warm-up steps, a weight decay of 0.01, and a batch size of 16. We process the input examples using the BERT base-uncased tokenizer, the same tokenizer used when identifying OOV names.

3.3 Results

In our experiments, the models' F1-performances[7] slightly change on our examples modified with gendered or gender-neutral names (Figure 1 and Table 4). Compared to the original, unmodified test examples, the models' performance drops on the named examples and it seems that BERT's performance is most affected by the named examples compared to the other models. By adding the True Positives and False Positives rates in the confusion matrices Table 3, we notice an increase in offensive predictions across all genders for CBoW, a smaller increase for BiLSTM, and a slight decrease in offensive predictions for BERT.[8] In other words, just by adding a name or pronoun, the Glove-based models predict more examples as offensive and the BERT model predict fewer examples as offensive. However, across all models, the difference in predictions on the gendered and original examples is not statistically significant, as measured by t-tests.

Focusing just on the named examples, the models that do not use contextualized word representations (CBoW and BiLSTM) perform better on the female examples than the male or gender neutral examples, while the BERT model achieves a higher F1 score on the male examples than on the female or gender neutral examples. Turning towards our goal of identifying whether the models are prone to the stereotype that men use more offensive language than women, we notice that all models classify more male authored tweets as offensive than female authored tweets. Specifically the CBoW, BiLSTM, and BERT models respectively classify 0.46% (397), 0.35% (297), and 0.49% (435) more male authored-examples as offensive than female authored-examples.[9] While one-sided[10] t-tests (Table 5) comparing the models' predictions between

[7] We report F1-Score since both the training and test datasets are not balanced.

[8] CBoW classifies 24.77% of the unnamed test examples as offensive compared to 26.60% for male, 26.14% for female, and 26.44% for gender neutral examples. BiLSTM classifies 27.67% of the unnamed test examples as offensive compared to 28.39% for male, 28.04% for female, and 28.50% for gender neutral. BERT classifies 26.98% of the unnamed examples as offensive compared to 26.60% for male, 26.11% for female, and 26.45% for gender neutral examples.

[9] These are absolute differences between male TP + FP rates and female TP + FP rates.

[10] Specifically that the models categorize more male-authored than female-authored tweets as offensive.

male and female authored-examples indicate that these are statistically significant differences for CBoW and BERT, the small differences in magnitude might suggest that adding historically gendered names as speakers in our examples does not consistently or convincingly alter the models' class predictions for whether or not a tweet is offensive. The statistical significance for CBoW and BERT might be due to the large sample size in our study.

4 Further Analysis

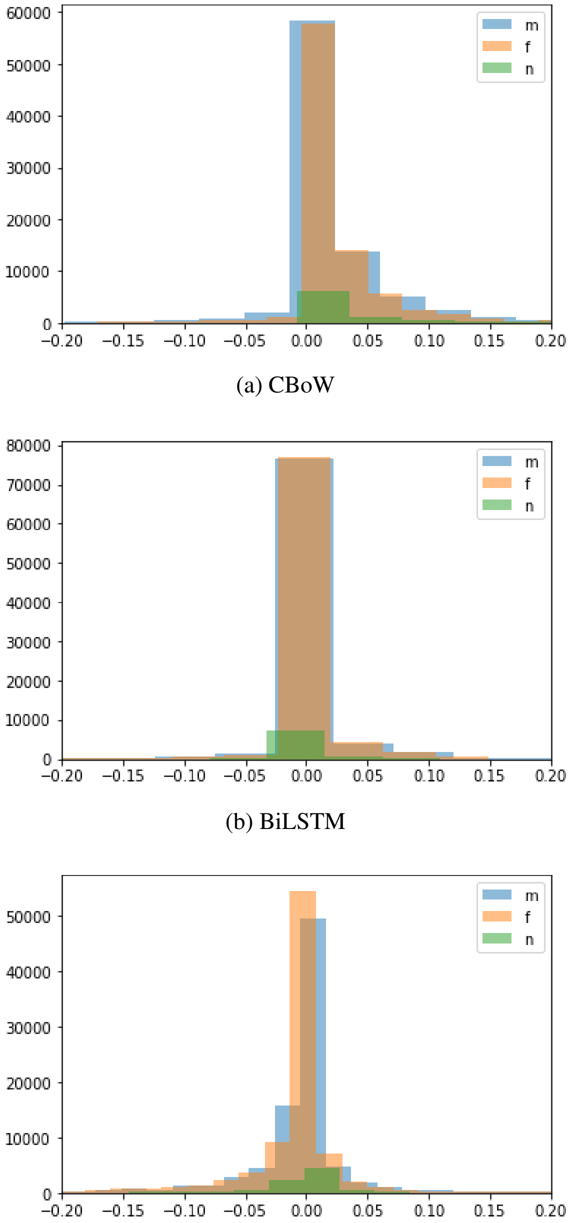
Since our results do not conclusively support our hypothesis that the models' binary predictions change for *all* considered models when explicitly adding gendered names to our test examples, we turn our attention towards exploring whether, and to what extent, the models' assigned probabilities change for our perturbed dataset. We also investigate whether these predicted probabilities consistently change for any specific names.

4.1 Offensiveness Probability Scores

Solely investigating whether a model's binary predictions change might mask gender biases should the model's predicted probabilities vary largely without crossing the label decision boundary. To explore whether this is the case, we compute the difference between a model's predicted offensiveness probability for every modified and corresponding unmodified example. The average differences are 0.021 ($\sigma = 0.059$) for CBoW, 0.007 ($\sigma = 0.059$) for BiLSTM, and -0.007 ($\sigma = 0.47$) for BERT.

Figure 2 plots the distribution of these differences grouped by gender for each model. These histograms illustrate that across all three models, for the majority of modified examples, the change in offensiveness probability is very small.

Additionally, these histograms further confirm our initial findings. For the CBoW model (Figure 2a), adding a gendered name seems to more likely lead to an increase in predicted offensive probability, and male names lead to larger increases. For the BiLSTM model (Figure 2b), the distributions of the differences for male and female examples almost match and a large majority of male (88.64%) and female (90.22%) examples have an absolute difference less than 0.025%. For the BERT model (Figure 2c), including gendered names in the examples lead to a decrease in predicted offensive probability, with more pronounced decreases for female names.



(a) CBoW

(b) BiLSTM

(c) BERT

Figure 2: Histograms plotting the change in offensive class probability between named and unnamed examples, grouped by gender (m: male, f: female, n: gender-neutral). A positive difference indicates that the model determined the named tweet to be more offensive than the base tweet.

These histograms demonstrates that there are very few examples where the model's predicted probabilities vary largely without crossing the label decision boundary. However, these histograms, specifically Figure 2a and Figure 2c, might reflect the stereotypes discussed by Sapolsky and Kaye (2005) that men use more offensive language than women.

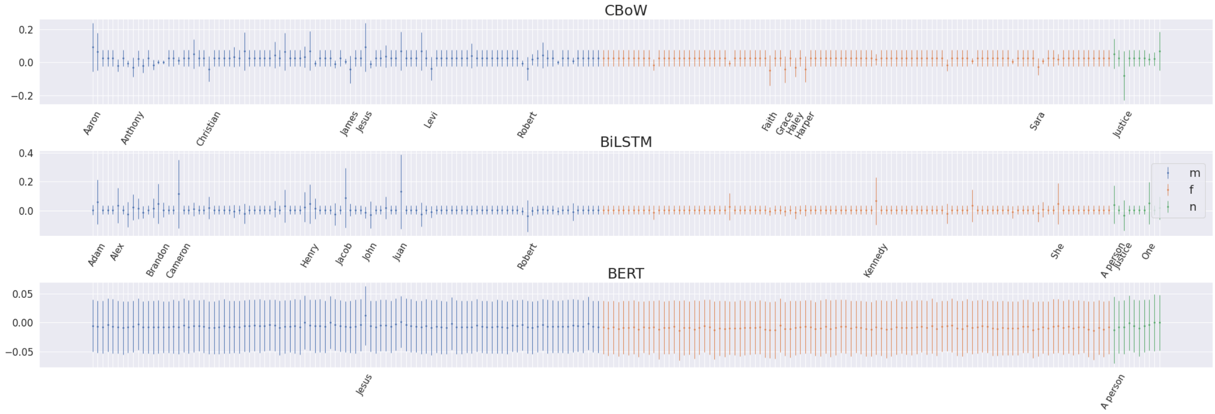


Figure 3: For each model, we plot how (average and standard deviation) the predicted offensiveness probability changed for each name. Y-axis indicates the difference. We label the names where the average difference was outside the typical standard deviation across the model.

4.2 Individual Name Impact

Prior work has shown that pre-trained representations might encode stereotypes about specific names (Shwartz et al., 2020). To test if these models similarly contain biases about specific names, we now group the difference between a model's predicted probability for each modified and corresponding unmodified example by prepended name.

Figure 3 plots how these differences vary for each name. The average differences in the BERT model's predictions consistently vary, but insignificantly. We notice just one name, *Jesus*, stands out as more offensive and one pronoun, *a person*, is uniquely less offensive. The mean change over all examples prepended with *Jesus* in the BERT model is 0.012, compared to an average change of -0.007 across all named examples. This finding is perhaps attributable to how *Jesus* is often used in colloquial English speech and on online platforms as a form of exclamation (Goddard, 2015).

For the GloVe-based models, we notice that the average and standard deviation of differences are identical for the same set of 158 names (and pronouns).[11] These models' predicted probabilities changed more for male than female names. Of the 54 names where the models' average probabilities differed from that of the 158 names, 36 are male, 13 are female, and 5 are gender-neutral. CBoW's average probability increased for 15 male, 0 female, and 2 gender-neutral names, and BiLSTM's increased for 21 male, 5 female, and 3 gender-neutral names. This suggests that the GloVe-based models might find male names to be more offensive than female names. However, there is little overlap between the male names that the CBoW and BiLSTM model usually predict as being more offensive (e.g. *Aaron*, *David*, and *Henry* for CBoW and *Adam*, *Brandon*. and *Jacob* for BiLSTM). For the name *Robert*, both models typically predict a lower offensive probability. The greater variations in the CBoW and BiLSTM predictions suggests that these models are more sensitive to the presence of specific gendered names compared to transformer-based models.

5 Conclusion

We asked whether there exists a measurable gender-based asymmetry in models' performances for predicting offensiveness when a tweet explicitly states the speaker's name. Our experimental results imply that a range of typical neural models might be robust to perceived author gender when classifying tweets as offensive though they might perceive male authored tweets to be slightly more offensive. Our work supports recent findings that intrinsic biases in the word embedding space may not correlate to extrinsic measures of bias in downstream applications (Goldfarb-Tarrant et al., 2021). While these findings on gender bias in offensive classification tasks are promising, we encourage further research to evaluate the extent to which these results generalize across more datasets and language phenomena as well as other social groups and intersectional identities, such as speaker race, age, and sexual orientation.

[11] The mean and standard deviation for these difference in CBoW's predictions for these names are respectively 0.025 and 0.049 and 0.005 and 0.031 for the BiLSTM.

6 Ethical Considerations

As noted in Antoniak and Mimno (2021), collecting gendered names from population-derived data has the limitation of centering the majority population, in this case US-born, white children. Moreover, while filtering for names not recognized by the GloVe or BERT vocabularies ensures our study only includes names that have pre-trained representations, this filtering might perpetuate biases in our tests since it disproportionately affected non-white names and female names.

Researchers have called on the NLP community to move beyond the gender binary (Larson, 2017; Prabhakaran et al., 2019). While our study included gender-neutral names and pronouns, we acknowledge that this set is drastically smaller than that of gendered names. We leave a deep study into the impact of gender-neutral names or pronouns as future work.

Using names as a proxy for gender is fraught with potential limitations and biases, particularly when an individual's gender identity does not match the gender historically associated with their name. However, NLP systems might make gendered associations based upon a speaker's name even when the speaker's gender is not explicitly mentioned. As discussed in footnote 1, we acknowledge these issues and strive to parallel the circumstances in which these systems may be deployed in the real world.

Acknowledgements

We are grateful to the anonymous reviewers who provided useful and insightful comments on this work. This work was partially supported by the Columbia University Data Science Institute's Data For Good Scholars program.

References

Piush Aggarwal, Tobias Horsmann, Michael Wojatzki, and Torsten Zesch. 2019. LTL-UDE at SemEval-2019 task 6: BERT and two-vote classification for categorizing offensiveness. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 678–682, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Maria Antoniak and David Mimno. 2021. Bad seeds: Evaluating lexical methods for bias measurement. 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 1889–1904, Online. Association for Computational Linguistics.

Marzieh Babaeianjelodar, Stephen Lorenz, Josh Gordon, Jeanna Matthews, and Evan Freitag. 2020. Quantifying gender bias in different corpora. In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 752–759, New York, NY, USA. Association for Computing Machinery.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

James Banks. 2010. Regulating hate speech online. *International Review of Law, Computers & Technology*, 24(3):233–239.

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

Çağrı Çöltekin. 2020. A corpus of Turkish offensive language on social media. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6174–6184, Marseille, France. European Language Resources Association.

Elizabeth Excell and Noura Al Moubayed. 2021. 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. Association for Computational Linguistics.

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 *Twelfth International AAAI Conference on Web and Social Media*.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Andrew Gaut, Tony Sun, Shirlyn Tang, Yuxin Huang, Jing Qian, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2020. Towards understanding gender bias in relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2943–2953, Online. Association for Computational Linguistics.

Cliff Goddard. 2015. "swear words" and "curse words" in australian (and american) english. at the crossroads of pragmatics, semantics and sociolinguistics. *Intercultural Pragmatics*, 12(2):189–218.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. 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 1926–1940, Online. Association for Computational Linguistics.

Rowan HallMaudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. 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 5267–5275, Hong Kong, China. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ritesh Kumar, Atul Kr. Ojha, Bornini Lahiri, Marcos Zampieri, Shervin Malmasi, Vanessa Murdock, and Daniel Kadar, editors. 2020. *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*. European Language Resources Association (ELRA), Marseille, France.

Brian Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain. Association for Computational Linguistics.

Debanjan Mahata, Haimin Zhang, Karan Uppal, Yaman Kumar, Rajiv Ratn Shah, Simra Shahid, Laiba Mehnaz, and Sarthak Anand. 2019. MIDAS at SemEval-2019 task 6: Identifying offensive posts and targeted offense from Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 683–690, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Ninareh Mehrabi, Thamme Gowda, Fred Morstatter, Nanyun Peng, and Aram Galstyan. 2020. *Man is to Person as Woman is to Location: Measuring Gender Bias in Named Entity Recognition*, page 231–232. Association for Computing Machinery, New York, NY, USA.

Saif M Mohammad. 2019. The state of nlp literature: A diachronic analysis of the acl anthology. *arXiv preprint arXiv:1911.03562*.

Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2021. Arabic offensive language on Twitter: Analysis and experiments. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 126–135, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

John Pavlopoulos, Nithum Thain, Lucas Dixon, and Ion Androutsopoulos. 2019. ConvAI at SemEval-2019 task 6: Offensive language identification and categorization with perspective and BERT. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 571–576, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in Greek. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5113–5119, Marseille, France. European Language Resources Association.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81, Brussels, Belgium. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation sensitivity analysis to detect unintended model biases. 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 5740–5745, Hong Kong, China. Association for Computational Linguistics.

Murugesan Ramakrishnan, Wlodek Zadrozny, and Narges Tabari. 2019. UVA wahoos at SemEval-2019 task 6: Hate speech identification using ensemble machine learning. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 806–811, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Amir H Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive language detection using multi-level classification. In *Canadian Conference on Artificial Intelligence*, pages 16–27. Springer.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. 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 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Niloofar Safi Samghabadi, Parth Patwa, Srinivas PYKL, Prerana Mukherjee, Amitava Das, and Thamar Solorio. 2020. Aggression and misogyny detection using BERT: A multi-task approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 126–131, Marseille, France. European Language Resources Association (ELRA).

Barry S. Sapolsky and Barbara K. Kaye. 2005. The use of offensive language by men and women in prime time television entertainment. *Atlantic Journal of Communication*, 13(4):292–303.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

M. Schuster and K. K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. 2020. “you are grounded!”: Latent name artifacts in pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6850–6861, Online. Association for Computational Linguistics.

Alexandra A Siegel. 2020. Online hate speech. *Social Media and Democracy: The State of the Field, Prospects for Reform*, pages 56–88.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. Offensive language and hate speech detection for Danish. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3498–3508, Marseille, France. European Language Resources Association.

Brittany N. Smith, Mamta Singh, and Vetle I. Torvik. 2013. A search engine approach to estimating temporal changes in gender orientation of first names. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL ’13, page 199–208, New York, NY, USA. Association for Computing Machinery.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Adam Vogel and Dan Jurafsky. 2012. He said, she said: Gender in the ACL Anthology. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 33–41, Jeju Island, Korea. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhenghao Wu, Hao Zheng, Jianming Wang, Weifeng Su, and Jefferson Fong. 2019. BNU-HKBU UIC NLP team 2 at SemEval-2019 task 6: Detecting offensive language using BERT model. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 551–555, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. 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 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. 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.

Jian Zhu, Zuoyu Tian, and Sandra Kübler. 2019. UM-IU@LING at SemEval-2019 task 6: Identifying offensive tweets using BERT and SVMs. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 788–795, Minneapolis, Minnesota, USA. Association for Computational Linguistics.