Survey on Bias is Text Classification

Aida Mostafazadeh Davani
mostafaz@usc.edu

The rapid increase in online hate speech has led to a rise of profound research in defining and detecting this type of rhetoric. While some NLP studies seek to identify the abusive language in general (Wiegand et al., 2018; Zampieri et al., 2019), others restrict the definition of hate speech to the kind of abusive language that is directed to particular social groups and differentiate hate speech from offensive language (Davidson et al., 2017; Kennedy et al., 2018).

The development of hate speech studies includes creating typologies for hate speech (Waseem et al., 2017; Olteanu et al., 2018), annotating social media content (Davidson et al., 2017; Founta et al., 2018), and designing natural language processing models for detecting this type of language (Zhang et al., 2018b).

The awareness about fairness in machine learning have directed researchers to discover evidences of bias in detecting hate speech and abusive language (Hardt et al., 2016). The remainder of this document summarizes the methods that have been practiced to eliminate unintended bias from text classification, with a specific concern about hate speech detection.

1 Bias in Text Classification

Unintended bias in machine learning models has recently been introduced as an issue that can prevent a model from providing fair results (Hardt et al., 2016). An example of unintentional bias in NLP is when a model handles text documents regarding specific features that, according to the task description, should not be reflected in the model's behavior.

Recent studies of bias argue that this type of unintended bias is threatening the essential fairness in ML. Fairness is formally defined by three components: demographic parity, equality of odds, and equality of opportunity (Hardt et al., 2016). In this case, unintended bias is considered to be associated with equality of odds which necessitates the true positive and true negative rates to be in the same range for different groups (Hardt et al., 2016).

In the case of text classification, overly relying on over/under-represented terms in the training dataset for predicting the label is shown to be causing unintended bias (Dixon et al., 2018). Deep learning practices for text classification methods are supposed to reduce the model's dependence on specific words compared to word-level models. However, the sparsity of the positive labels, under-representation of specific word categories in the train set, or over-representation of a group in a subset of the dataset with a particular label can lead to unintended bias towards specific word categories.

Proposed methods for reducing the unintended bias, can be categorized into three classes based on their approach: balancing the dataset, restricting the model, adversarial training.

1.1 Balancing the Dataset

Dixon et al. (2018) introduce unintended bias in a model as observing different performance on subsets of the dataset that contain particular *identity terms* (words referring to *identity* groups). We would rather refer to *identity terms* as *social group terms*, since *identity* includes concepts that are not studied in the related work. We define *social group terms* (SGT) as term that refers to a specific social group (e.g. muslim, woman, russian)

By inserting additional data with negative labels that contain SGTs to balance the ratio of positive and negative labels for each *social group*, Dixon et al. (2018) try to make the training dataset less biased. Even though all sentences in the document are from comments in Wikipedia Talk Page, the added data is gathered from the Wikipedia articles, which makes the distribution of the new training set to be incongruous with the original dataset.

Another approach for generating a balanced dataset is to swap the SGT in sentences, to provide an equal representation of different groups Zhao et al. (2018); Park et al. (2018). In other words, these methods repeat the same sentence by substituting their identified SGTs with terms that refer to other social groups. As discussed by Wiegand et al. (2019), a disadvantage of this approach is that it ignores the other sentence-level sources of bias.

Wiegand et al. (2019) argues that datasets that are gathered by randomly sampling from a corpus are usually sparse regarding the positive labels. That leads to data gathering based on dictionaries or topics, which causes the trained models to be biased towards specific terms. Classifiers trained on biased datasets might achieve high accuracy, which should be evaluated by testing them on other datasets. It can be concluded from the analyses conducted by Wiegand et al. (2019) that the bias in the dataset is not restricted to the identity terms and can also involve the data collection method.

1.2 Restricting the Model

Since creating a balanced dataset faces issues such as ignoring higher-level causes of bias, approaches that penalize the existence of bias can more comprehensively address this issue. These approaches prevent the bias by defining and including it in the loss function of a model.

Garg et al. (2019) introduces the use of counterfactuals for examining unintended bias. Counterfactuals are sentences that are generated by substituting specific critical tokens (SGT) with other instances to test the fairness of the model. The modelś loss function is extended to minimize the difference among the error ranges for all counterfactuals generated for a specific document.

In some cases, the SGTs under study are considered fundamentally crucial for creating predictions (Garg et al., 2019). In other words, the sentence can only be regarded as a toxic comment when it is directed to a specific social group. These cases, which are called asymmetric counterfactuals, should be excluded when preserving fairness. This argument also implies that replacing the SGT with predefined tokens (such as ¡group_name¿) would not guarantee the issue with asymmetric examples.

In another proposed method, Liu and Avci (2019) assume that the bias in classification is due to the reliance on specific SGT and consequently use model interpretation to measure the importance of particular SGTs in predicting the label in classification tasks. The interpreted importance of these terms is considered for defining the loss function to prevent the adoption of SGT. Along with training a model that performs as accurate as of the biased models, the learned word embeddings are shown to include less bias.

The most important advantage of these approaches is that they can be applied in different contexts by modifying the formal specification of the bias.

1.3 Adversarial Models

Another class of approaches considers altering or extending the model by adding an adversarial network that minimizes the modelś bias. The adversarial network does not need to be as complicated as the prediction model, which presents this method as suitable for being combined with other models.

Zhang et al. (2018a) train an adversarial model by minimizing its capability for predicting the associated SGT mentioned in a text document while maximizing the classification accuracy. By having the adversarial network trained in parallel with the classifier, the loss function drives the hidden layers to acquire less information about the mentioned SGT.

In a similar approach, Madras et al. (2018) trains an autoencoder to learn a latent representation of the documents. The latent representation is then utilized by the classifier to predict the label and by the adversarial network to identify the SGT. The network is trained to minimize the autoencoder and classifier loss jointly with maximizing the adversary loss. Moreover, instead of cross-entropy loss, l_1 method is used for the adversary loss, which can more correctly guide the group-normalized loss in unbalanced datasets.

One advantage of the adversary models is their flexibility for defining the bias under study. By altering the loss function, it is possible to account for different components of fairness. However, due to the complexity of the network, these models appear to be challenging to train.

References

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

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

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*.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226. ACM.

Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323.

Brendan Kennedy, Drew Kogon, Kris Coombs, Joseph Hoover, Christina Park, Gwenyth Portillo-Wightman, Aida Mostafazadeh Davani, Mohammad Atari, and Morteza Dehghani. 2018. A typology and coding manual for the study of hate-based rhetoric.

Frederick Liu and Besim Avci. 2019. Incorporating priors with feature attribution on text classification. *arXiv preprint arXiv:1906.08286*.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*.

Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush R Varshney. 2018. The effect of extremist violence on hateful speech online. In *Twelfth International AAAI Conference on Web and Social Media*.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a lexicon of abusive words–a feature-based approach.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666*.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018a. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340. ACM.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018b. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European Semantic Web Conference*, pages 745–760. Springer.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.