Measuring and Mitigating Bias in Sentiment Classification

12 May 2022

Introduction

With the ever-increasing use of machine learning models in real applications, a series of negative impacts are revealed, including prejudice and unfairness behind such models. A number of studies indicated that machine learning models tend to reflect a certain extent of systematic bias in terms of the specific demographic subsets (Dixon et al., 2018; Deshpande, 2021; Liu et al., 2019). Those models are generally trained upon the human-generated data under various scenarios and some are also tagged with labels marked by an individual. Thus the unfairness could originate from unintended bias in the early stage of training (Dixon et al., 2018).

Text classification is an important application of machine learning models. Such models could be utilised in hate speech filtering and sentiment analysis on social media and other settings. However, if a text classifier appears to hold preferences for a specific group in terms of its gender, race or socioeconomic status, it could result in implicit discrimination (Liu et al., 2021).

The bias in machine learning models and text classification is explored in this work. Various models are evaluated in terms of proposed evaluation metrics for unfairness in the context of classifying the sentiment of user-generated text from Twitter (Blodgett et al., 2016). The gap in performance between demographic groups is identified across all models. The underlying factors are being analysed from different perspectives and followed by an attempt in mitigating the existed bias.

Related Work

As machine learning models are proposed with better performance, the potential unfairness against groups with specific demographic attributes has become a concern and revealed by the existing studies. The bias is conceptualised as a system that tends to favour one conclusion over others (Deshpande, 2021). It is also claimed that every effective machine learning model is expected to reflect bias to a certain extent due to the nature of "learning" attributes for accurate prediction (Dixon et al., 2018). For example, a text classification model for hate speech detection is supposed to pick out texts expressing hate or the intention of violence against others. However, the bias in machine learning models is generally perceived as the preference for irrelevant demographic features which may affect the performance across specific groups. Such bias is also denoted as unintended bias

(Dixon et al., 2018), which will be the focus of this report.

The study conducted by Dixon et al. (2018) explored the bias in classifying toxic comments on Wikipedia Talk Pages. It demonstrated that the imbalance of training data with certain identity terms could result in unintended bias in the model and therefore proposed a mitigation approach by augmenting training data. The work by Deshpande (2021) discussed the sentiment analysis for user-generated data from various social media platforms. The bias in sentiment classification was identified when attempting to generalise to new platforms. It addressed the importance of mitigating unfairness in order to train more generalised machine learning models which can be applied to various settings. It was pointed out that factors influencing the unintended bias may include "unfair representation in training data, insufficient training data, or faulty feature selection" (Deshpande, 2021).

In terms of feature selection, Liu et al. (2019) looked into the effect of specific attribute words in the training data for a model in the study. A list of race words was attached with each in a pair of common phrases in African American English and its corresponding phrase in standard English. It was found that African Americans would receive more negative responses compared to White Americans. Besides, it was surprising that the demographic information of authors could also lead to bias in deep text classification (Liu et al., 2021). Even though specific demographic attributes didn't appear in the text, the content may implicitly expose certain language styles including the habit of wording and the tone. Thus, this implicit demographic information about the author was captured by machine learning models to make unfair predictions against specific groups.

Methodology

Data Sets

This study investigates the unintended bias in classifying the sentiment of text across multiple machine learning models. The data set is a collection of Tweets with labelled sentiment as either *Positive* or *Negative* (Blodgett et al., 2016). Demographic labels are attached to each Tweet as well but are not considered as one of the features. The demographic label specifies the variety of English for the content including African American English (AAE) and Standard American English (SAE).

The data set is separated into training and test sets with 40,000 records in the training set and 4,000 records in the test set. The distribution of the labelled sentiment for the training set is balanced and each sentiment

class contains 20,000 Tweets. Furthermore, the ratio of the demographic set is verified to be balanced for both labels as well. A summary of sentiment/demographic distribution is presented below 1:

		Demographic		**Total**
		AAE	SAE	
Sentiment	Positive	1,0000	1,0000	2,0000
	Negative	1,0000	1,0000	2,0000
Total		2,0000	2,0000	4,0000

Table 1: Distribution of records

Feature Representations

The raw content of each Tweet is mapped to a 384-dimensional list as the sentence embedding. Sentence embeddings are generated with a pre-trained sentence transformer: *Sentence-BERT* (Reimers and Gurevych, 2019). *Sentence-BERT* is also known as *SBERT*, which is based on the state-of-the-art language representation model *BERT*. *BERT* deals with the bidirectional representation of languages with a multi-layer architecture and adapts the fine-tuning approach to train the parameters (Devlin et al., 2018). Similar to *BERT*, *SBERT* is structured with a siamese and triplet network to update the weights and derive fix-sized sentence embeddings such that semantically similar sentences are close to each other. The similarity in semantics can be quantified using techniques like cosine-similarity (Reimers and Gurevych, 2019).

Machine Learning Models

To investigate the unintended bias in sentiment classification, various machine learning models are tested and compared. The following four classifiers are selected and implemented using Scikit-learn library (Pedregosa et al., 2011):

- K-Nearest Neighbour Classifier (*KNN*)
- Gaussian Naïve Bayes Classifier (*NB*)
- Logistic Regression Classifier (*LR*)
- Multilayer Perceptron Classifier (*MLP*)

These models are trained on the training set with the majority of the parameters set as the default, except the number of neighbours for *KNN* is set to be 5 and the maximum iteration set to 500 for both *LR* and *MLP* to ensure the result will converge. Trained models are used to predict the sentiment label for the test set with results collected and analysed in the following sections.

Bias Evaluation Metrics

Apart from the usual metrics used to assess the accuracy rate as the indicator of the overall model performance, the bias in machine learning models is measured and evaluated with some specialised metrics.

The *selection rate* metrics is one of the metrics used to evaluate the bias in each demographic group. The *selection rate* denotes the proportion of instances predicted as *Positive* for groups labelled as *SAE* or *AAE*. It's a probabilistic concept and can be denoted as the following mathematical expression below 1. If the model does not prefer any group, the *selection rate* for each group should be roughly the same. The difference in selection rate across demographic groups can be referred to as *demographic parity difference*. Both *selection rate* and *demographic parity difference* can be easily used via *Fairlearn* library and are implemented to quantify bias in this work (Bird et al., 2020).

$$\Pr(\hat{y} = Positive) = \frac{TP + FP}{T + N} \tag{1}$$

Another bias evaluation is mentioned in the work by Dixon et al. (2018) and referred to as *error rate equality difference*. *Error rate equality difference* metrics denotes the variation of false positive rates and false negative rates between demographic groups. It measures the unintended bias in the model by calculating the sum of differences in terms of false positive rates and false negative rates for each term of demographic labels. Higher *error rate equality difference* indicates a larger gap in performance across groups. The formal definition can be expressed in the following formula 2.

$$\sum_{t \in T} (|FPR - FPR_t| + |FNR - FNR_t|) \tag{2}$$

Bias Mitigations

It is pointed out that the imbalance in the data set could lead to bias in machine learning models by other studies (Dixon et al., 2018; Liu et al.,

2021). Thus, the data set used in this study is checked to ensure its balance in terms of both sentiment label and demographic label as explained in the above section.

However, it is noticed that certain phrases frequently appear in the content of Tweets while they seem irrelevant to the meaning in the context. Such irrelevant phrases are usually regarded as noise in the data collected from users. By examining the raw content of Tweets in the data set, phrases such as *"_TWITTER-ENTITY_"* and leading " \ " are picked out. Such examples are presented below in table 2 and noise phrases are marked as bold.

Text	Sentiment	Demographic
"\We're staying at the Hilton , as in Paris Hilton !!! " **_TWITTER-ENTITY_**"	positive	SAE
"\ **_TWITTER-ENTITY_** : **_TWITTER-ENTITY_** **_TWITTER-ENTITY_** yall broke " when you have money ?"	positive	AAE
"**_TWITTER-ENTITY_** I let it all out last night . I need you here	negative"	AAE

Table 2: Examples of noise phrases in text

To mitigate the effect of noise in the data set and also test the impact on the measured bias across machine learning models, a Python script is written to filter the Tweets in both training and test sets by removing such irrelevant phrases.

Results

Accuracy

Even though the overall accuracy of the data set across models is not the focus of this study, the variation of accuracy between demographic groups reflects the existence of a gap. The accuracy rate for each demographic group predicted using different classifiers is presented below in table 3. It suggests that all these classifiers tend to show better performance for instances using SAE.

	Accuracy			Gap
	Overall	AAE	SAE	
KNN	0.663	0.636	0.690	5.35%
NB	0.615	0.578	0.652	7.45%
LR	0.698	0.665	0.732	6.75%
MLP	0.635	0.600	0.670	7.05%

Table 3: Accuracy across different models

The gap in accuracy between SAE and AAE is demonstrated in figure 1. *KNN* performs the best with the lowest gap at 5.35% and the rest three have similar performance at around 7%.

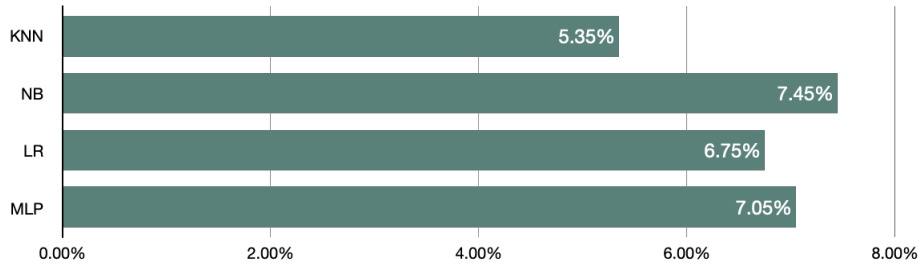

Figure 1: Gap in accuracy rate across classifers

Selection Rate & Demographic Parity Difference

The statistics collected about *selection rate* are presented below for both demographic groups. The *demographic parity difference (DPD)* is calculated as the difference of *selection rate* between Tweets labelled as *AAE* and *SAE*. Table 4 contains the statistics for original data set and table 5 is based on the data set with certain noise phrases already removed.

	Selection Rate			**DPD**
	Overall	AAE	SAE	
KNN	0.504	0.534	0.474	-6.05%
NB	0.416	0.461	0.371	-8.95%
LR	0.519	0.566	0.473	-9.25%
MLP	0.497	0.529	0.466	-6.25%

Table 4: Selection rate across different models

	Selection Rate (Mitigated)			**DPD**
	Overall	AAE	SAE	
KNN	0.447	0.498	0.397	-10.10%
NB	0.547	0.680	0.414	-26.60%
LR	0.513	0.574	0.453	-12.15%
MLP	0.506	0.537	0.475	-6.15%

Table 5: Selection rate with noise removed

Bar chart 2 is created to better illustrate the *demographic parity difference* across different models and also reflects the effect of removing noise phrases in the data set. *KNN* and *MLP* classifiers are the two which performed the best with relative low *demographic parity difference*. However, it is surprising that all models tend to show a higher level of unintended bias except for *MLP*. *NB* reaches a 26.6% gap between *AAE* and *SAE* while *MLP* roughly stays the same with around 6%. But the direction is consistent as Tweets with *AAE* have a higher probability to be predicted as *Positive*.

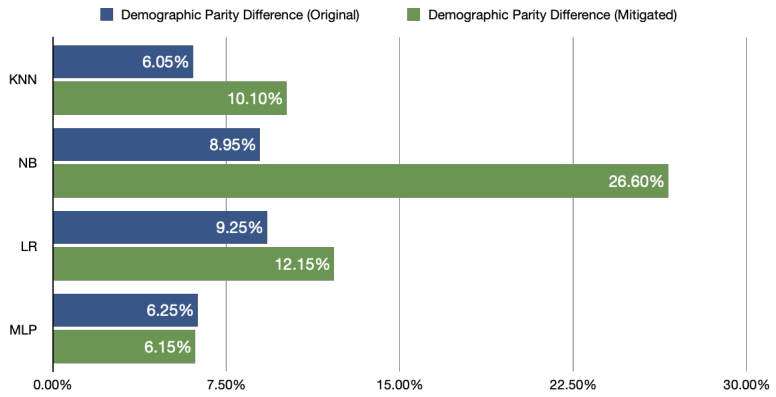

Figure 2: Demographic Parity Difference before and after noise removed

Error Rate Equality Difference

Figure 3 represents the *error rate equality difference* using various classifiers for sentiment prediction. It includes the statistics before and after removing noise phrases in the raw texts and the overall trend is similar to *demographic parity difference*. *NB* seems to be affected most with its sum of errors reaching more than 0.5.

The sum of *error rate equality difference* will be the key indicator of the bias in machine learning models in the following sections and the variation across models will be discussed along with the analysis of the impact of the attempted mitigation technique.

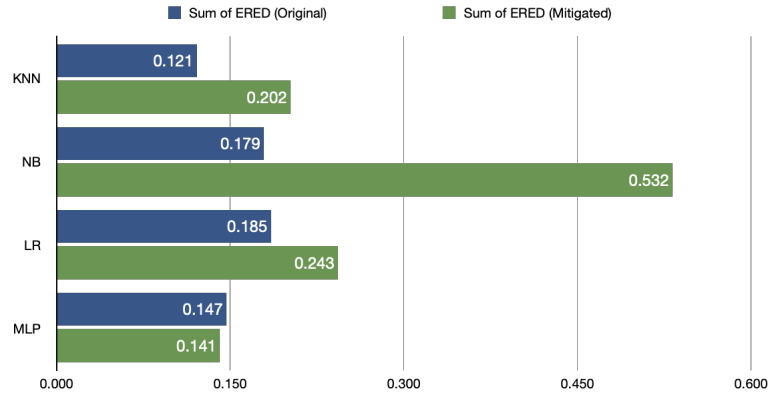

Figure 3: Error Rate Equality Difference before and after noise removed

Discussion

This study investigates the unintended bias in sentiment classification for a collection of Twitter contents and evaluates a series of machine learning models. According to the discrepancy between demographic groups captured by bias evaluation metrics such as *demographic parity difference* and *error rate equality difference*, the existence of unintended bias is identified across all models that have been evaluated. It is claimed that the imbalance of training data, unfair representation of data set or biased feature selection could be the potential underlying factors (Dixon et al., 2018; Deshpande, 2021). Therefore, the following sections will address these issues with analysis based on the statistics gathered.

Imbalance in Data Set

The imbalance in training data is proved to have a negative impact on the disparity between demographic groups in the work by Dixon et al. (2018), which may lead to the unintended bias in machine learning models. A simple technique to mitigate the impact of imbalanced data is proposed to strategically add complementary data to the training set. Thus, the machine learning models would not be biased due to the disproportionate distribution in the training set.

Fortunately, the training data used in this study is balanced both for sentiment and demographic labels (Blodgett et al., 2016). The data set is fairly distributed as illustrated in table 1. Therefore, this study will analyse the potential reasons for unintended bias from other perspectives.

Noise Phrases

As mentioned earlier, certain irrelevant phrases frequently appear in the raw content of Tweets. Examples of such Tweets can be found in table 2. The data set is further examined and is noticed that the noise phrase *"_TWITTER-ENTITY_"* takes the top in the 10 most frequent words in both *AAE* and *SAE*. The frequency of words is calculated based on its corresponding TFIDF value and the top 10 frequent words are shown in the table 6.

To eliminate the effect of noise phrases in the data set, both training and test data are processed to remove the words classified as noise. However, even though the overall performance in terms of the accuracy is noticed to be improved for all models, both *demographic parity difference* and *error*

AAE		SAE	
Word	**TFIDF Value**	**Word**	**TFIDF Value**
entity_	1794.854	**entity_**	1851.360
_twitter	1794.854	**_twitter**	1851.360
amp	450.186	**just**	610.969
got	446.814	**like**	399.704
like	429.866	**don**	356.531
shit	426.811	**want**	317.264
lol	425.855	**right**	264.159
ass	382.262	**know**	256.267
just	318.571	**work**	237.803
don	308.800	**hate**	233.880

Table 6: Top 10 most frequent word in AAE and SAE

rate equality difference are observed to grow as indicated in figure 2 and 3. These two metrics are indicators of the level of unintended bias in machine learning models.

From another perspective, the rising bias in results suggests that the existence of noise phrases in texts may mitigate the degree of unfairness across demographic groups to some extent. It is pointed out that the unintended bias could emerge due to the overfitting of the given training data (Dixon et al., 2018). While the embedded noise affects the performance of models in terms of accuracy, it prevents models to overfit certain features which may be associated with demographic information in the training data. From the TFIDF value in table 6, it suggests roughly the same amount of noise is added to each of the demographic groups. Therefore the distribution of noise phrases is balanced as well which does not bring a negative impact on the bias. In other words, due to the existence of noise fairly distributed to *SAE* and *AAE*, the possibility of overfitting is reduced and thus mitigated a certain level of bias in machine learning models.

The captured bias provides another approach to evaluating machine learning models. It is surprising that *KNN* classifier tends to perform the best among the four classifiers. Its lazy learning mechanism could be one of the key factors behind the statistics. Unlike other models, *KNN* classifier doesn't actually "learn" on the given training set since it only calculates and compares the distance to the existing instances. It's similar to looking up a word in a dictionary while the dictionary only returns the response from the database where the dictionary itself is not involved in the processing and

prediction.

If compare the evaluated bias before and after removing noise phrases, the *NB* classifier is observed to be impacted the most with nearly tripled the *error rate equality difference*. Considering the assumption of conditional independence for features, the huge difference can be explained. In a task to classify the sentiment of text, the models are expected to capture the overall semantics of the sentence rather than the existence of a single word. Besides, the values in sentence embeddings generated with *SBERT* are certainly associated with each other so the assumption of conditional independence is obviously not supported.

The *MLP* classifier is considered the most resilient model in terms of bias. The sophisticated model reflects the capability to generalise the training set to predict the sentiment of Tweets in unseen cases. It is not affected by the noise and presents excellent performance both before and after the mitigation while keeping the bias on a relatively low level.

Feature Selection

The work by Liu et al. (2021) reveals the unfair outcome could be produced due to specific identity terms included in the context. A collection of identity terms refer to the phrases that may be associated with or relevant to specific demographic groups. A similar conclusion is supported by Dixon et al. (2018) in the study to claim that identity terms such as "gay" and "Muslim" are explicitly connected to the demographic information. Classifiers are supposed to make predictions based on language features that are related to the demographic attributes of groups (Liu et al., 2021). Such terms are captured by the models as important features and thus produce biased classifications.

	Word	Total Count	AAE Ratio	SAE Ratio
Going - Goin	Going	708	0.366	**0.634**
	Goin	77	**0.831**	0.169
Relax - Chill	Relax	15	0.467	0.533
	Chill	119	**0.849**	0.151
Hello - Yo	Hello	4	0.0	**1.0**
	Yo	351	**0.972**	0.028
Friend - Homie	Friend	144	0.375	**0.625**
	Homie	18	**0.944**	0.056

Table 7: Distribution of sensitive word pairs in AAE and SAE

In previous work by Liu et al. in 2019, a list of attribute word pairs of a common phrase in African American English and corresponding phrase in Standard American English is identified to bias the results produced by the models. A few word pairs are selected and checked for distribution in the training set. The result is presented in table 7 which indicates a disproportionate distribution across demographic groups. Some extremely skewed ratios are marked in bold in the table.

Figure 4 illustrates the imbalance of certain attribute terms in AAE and SAE more clearly. Terms including "Goin", "Chill" are more frequently used in African American English while terms like "Going", "Relax" are prevalent in Standard American English. If such attribute terms in the Tweets are captured by the models as the main feature to produce outcomes, unfair predictions are possible to be produced and potentially lead to unfair applications Dixon et al., 2018.

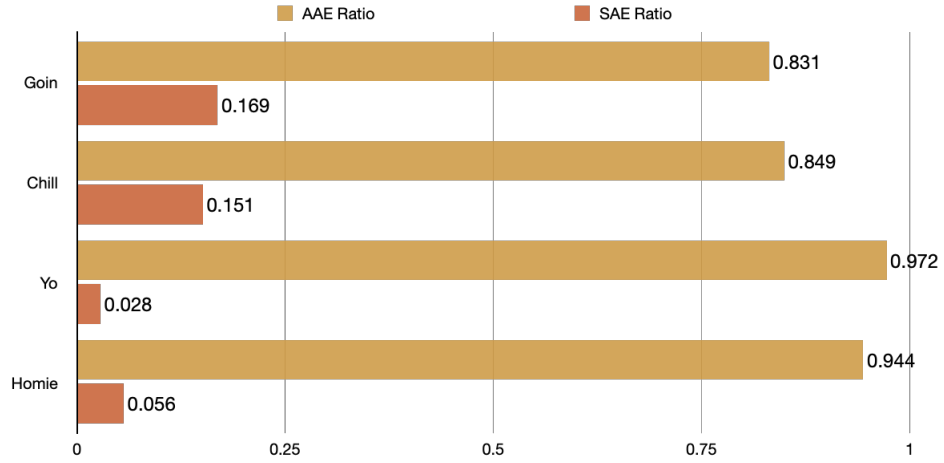

Figure 4: Disproportionate distribution of selected attribute terms

Conclusion

In this study, the bias produced in sentiment classification for a collection of Tweets is investigated across various machine learning models. Evaluation metrics such as *demographic parity difference* and *error rate equality difference* are used to quantify the level of bias. Factors including disproportionate distribution of data in the training set, the existence of noise phrases and specific attribute terms associated with demographic groups

are considered to affect the unfairness in models.

It is attempted to mitigate bias by removing noise phrases in the given data set but found it results in increased level of bias. The potential cause behind this is discussed and the discrepancy in performance in terms of bias among classifiers is analysed based on the feature of classifiers.

Bibliography

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., & Walker, K. (2020). *Fairlearn: A toolkit for assessing and improving fairness in AI* (tech. rep. MSR-TR-2020-32). Microsoft. https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

Blodgett, S. L., Green, L., & O'Connor, B. (2016). Demographic dialectal variation in social media: A case study of african-american english. *arXiv preprint arXiv:1608.08868*.

Deshpande, A. (2021). *Sentiment classification bias in user generated content* (Doctoral dissertation). Syracuse University.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 67–73.

Liu, H., Dacon, J., Fan, W., Liu, H., Liu, Z., & Tang, J. (2019). Does gender matter? towards fairness in dialogue systems. *arXiv preprint arXiv:1910.10486*.

Liu, H., Jin, W., Karimi, H., Liu, Z., & Tang, J. (2021). The authors matter: Understanding and mitigating implicit bias in deep text classification. *arXiv preprint arXiv:2105.02778*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.