Journal of the American Medical Informatics Association, 28(4), 2021, 839–849
doi: 10.1093/jamia/ocaa326
Advance Access Publication Date: 23 January 2021
Research and Applications

Research and Applications

The risk of racial bias while tracking influenza-related content on social media using machine learning

Brandon Lwowski and Anthony Rios

Department of Information Systems and Cyber Security, University of Texas at San Antonio, San Antonio, Texas, USA

Corresponding Author: Anthony Rios, Department of Information Systems and Cyber Security, University of Texas at San Antonio, San Antonio, TX USA; anthony.rios@utsa.edu

Received 20 June 2020; Editorial Decision 2 December 2020; Accepted 8 December 2020

ABSTRACT

Objective: Machine learning is used to understand and track influenza-related content on social media. Because these systems are used at scale, they have the potential to adversely impact the people they are built to help. In this study, we explore the biases of different machine learning methods for the specific task of detecting influenza-related content. We compare the performance of each model on tweets written in Standard American English (SAE) vs African American English (AAE).

Materials and Methods: Two influenza-related datasets are used to train 3 text classification models (support vector machine, convolutional neural network, bidirectional long short-term memory) with different feature sets. The datasets match real-world scenarios in which there is a large imbalance between SAE and AAE examples. The number of AAE examples for each class ranges from 2% to 5% in both datasets. We also evaluate each model's performance using a balanced dataset via undersampling.

Results: We find that all of the tested machine learning methods are biased on both datasets. The difference in false positive rates between SAE and AAE examples ranges from 0.01 to 0.35. The difference in the false negative rates ranges from 0.01 to 0.23. We also find that the neural network methods generally has more unfair results than the linear support vector machine on the chosen datasets.

Conclusions: The models that result in the most unfair predictions may vary from dataset to dataset. Practitioners should be aware of the potential harms related to applying machine learning to health-related social media data. At a minimum, we recommend evaluating fairness along with traditional evaluation metrics.

Key words: deep learning, classification, machine learning, social network, fairness

INTRODUCTION

Owing to the seasonal outbreaks of the influenza virus, there is an interest in digital tools and techniques for multiple tasks, including, but not limited to, digital contact tracing,[1,2] epidemiological studies,[3] and monitoring the prevalence of vaccinations.[4] The tools and techniques range from applications installed on user's personal phones to track the exact spread of a virus[2] to the development of machine learning–based techniques to study the spread of a virus using social media.[5–9] Similarly, machine learning–based methods have been developed to monitor the public's view on vaccines to combat the anti-vaccine narrative.[4] Here, we examine machine learning methods trained on social media data to track influenza-related content.

Current evidence suggests that there is a disproportionate incidence of disease and death among underrepresented minority groups. For example, there are significant racial disparities in influenza vaccinations.[10,11] Tse et al[12] report a nearly 10% difference in the influenza vaccination rate between non-Hispanic Black/African American adults over 50 years of age and non-Hispanic White adults. Fiscella et al[13] estimated that if influenza immunization rates

were equal for all races, nearly 2000 minority deaths could be prevented every year, saving more than 33 000 minority life-years.

In this article, we measure the fairness of machine learning–based tools for the specific task of detecting influenza-related messages on social media. Machine learning– and technology-based techniques have the potential to scale traditional public health tasks from a few hundred people at a time to millions (eg, digital contact tracing).[1] Therefore, digital tools have the potential to improve public health faster than ever before. Unfortunately, if there are even small differences in the performance of these tools across various demographic factors, then they have the potential to exacerbate the health disparities instead of improving them.

To understand bias in influenza tracking models, we ask the following questions:

- What is the relationship between overall classifier performance and fairness?
- Are the most (un)fair classifiers the same across different, but similar, influenza-related datasets?

Biases have been found in the machine learning methods developed for a wide variety of natural language processing tasks, including, but not limited to, text classification, learning word embeddings, and machine translation. For example, text classification models exhibit biases across gender and racial divides for tasks such as offensive language identification, resulting in differences in performance across groups.[14–17] Overall, much of the prior work has focused on traditionally nonbiomedical text classification tasks (eg, hate speech classification).

Word embeddings have also been shown to contain biases.[18–21] A word embedding is a learned representation or vector for text in which words with similar meanings have a similar representation, algorithmically capturing the meaning of words. Bolukbasi et al[18] show that the word embedding for *man* is similar to *doctor*, while *woman* is similar to *nurse*. Garg et al[22] developed a technique to study 100 years of gender and racial bias using word embeddings. Kurita et al[23] expanded on prior work to generalize bias measurement metrics for word embedding to contextual word embeddings (eg, BERT).[24,25] Machine translation systems have also been shown to exhibit biases.[26,27] Font and Costa-Jussà[26] showed that the sentence "She works in a hospital, my friend is a nurse" would correctly translate the word *friend* to *amiga*. However, the sentence "She works in a hospital, my friend is a doctor" tends to translate the word *friend* to *amigo*, implying that the friend is male. In general, many articles focus on testing whether bias exists in various models, or on developing techniques to remove bias from classification models for specific applications. In this article, we focus on measuring racial biases of machine learning methods in the biomedical natural language processing (NLP) domain.

Fairness can be defined in multiple ways. In this article, we focus on 2 specific definitions[28–30]: equality of opportunity and predictive equality. Simply, both definitions together are called equalized odds. Equality of opportunity assumes that the false negative rate (FNR) (see Evaluation for a complete definition) is equal between 2 groups. A high FNR could cause African Americans to potentially miss the opportunity to be identified. For instance, as a hypothetical scenario, if social media is mined to identify potential hotspots of the influenza virus, then a high FNR could lead to inadequate resources (eg, vaccinations) to fight the virus. Similarly, predictive equality is a measure of the difference between the false positive rates (FPRs) of 2 groups. A high FPR could be particularly harmful in the hypothetical scenario of the use of machine learning to detect vaccine-related misinformation. If information spread by African American communities is always (incorrectly) labeled as misinformation, then this could further exacerbate the disparities in the vaccination rate. It is also important to think about which is more important, predictive equality or equality of opportunity. This importance depends on the downstream application of the models. For the purpose of this article, we assume they are equally important.

This article focuses on measuring racial bias using the definition of equalized odds. Race is a complex construct, which is correlated with multiple facets such as dialect, socioeconomic class, and community.[31] Unfortunately, users do not generally self-report their race on social media—at least it is not common on Twitter. Instead, following the practice of prior researchers,[17,32–34] we rely on the correlation between dialect and race for our analysis. Specifically, we analyze the African American English (AAE) dialect. AAE has been shown to transfer from the use in face-to-face conversations to written text on social media.[33,35–37] AAE is a common dialect spoken by some, but not all, African Americans. It is important to emphasize that not all speakers of AAE are African American and not all African Americans are AAE speakers.[35] For more information about the correlation between AAE and racial constructs, please see Blodgett et al.[33]

As previously mentioned, having a machine learning model that is biased can have consequences. For instance, when using a machine learning model to predict potential epidemics, the model could correctly predict the spread of influenza for communities with a high-resource dialect like Standard American English (SAE), but, at the same time, have a high false negative rate for communities using low-resource dialects like AAE. In the computational linguistics community, "low resource" is used to simply mark languages or dialects that appear infrequently in the general population.[38] Thus, if we were to randomly sample English text from Twitter, then we would expect only a small fraction of the text to be AAE. As a real-world example on the impact of biased machine learning methods in the real world, Obermeyer et al[39] analyzed real world risk-prediction software that is applied to roughly 200 million people. The healthcare system relies on these algorithms to identify patients for "high-risk care management" programs. Their research shows that the algorithms were biased, causing differences in care between Black and White patients. Overall, it is important to understand how machine learning models will perform on a wide variety of tasks when applied to underrepresented populations.

Finally, we summarize our 3 major contributions as follows. First, we study the performance differences between SAE and AAE of machine learning models applied to various influenza-related tasks. Second, we explore the fairness of multiple machine learning algorithms including linear support vector machines (SVMs) and neural networks. Furthermore, we analyze the fairness of the neural networks using multiple pretrained vectors to understand the impact they have on the downstream performance of the model. Third, we provide a detailed discussion about the results presented in this article as well as this article's limitations.

MATERIALS AND METHODS

We provide an overview of our study in Figure 1. This article's methodology can be summarized in 4 steps. First, we train a convolutional neural network (CNN) to detect the dialect of individual tweets (ie, SAE vs AAE). Second, the model is used to classify the di-

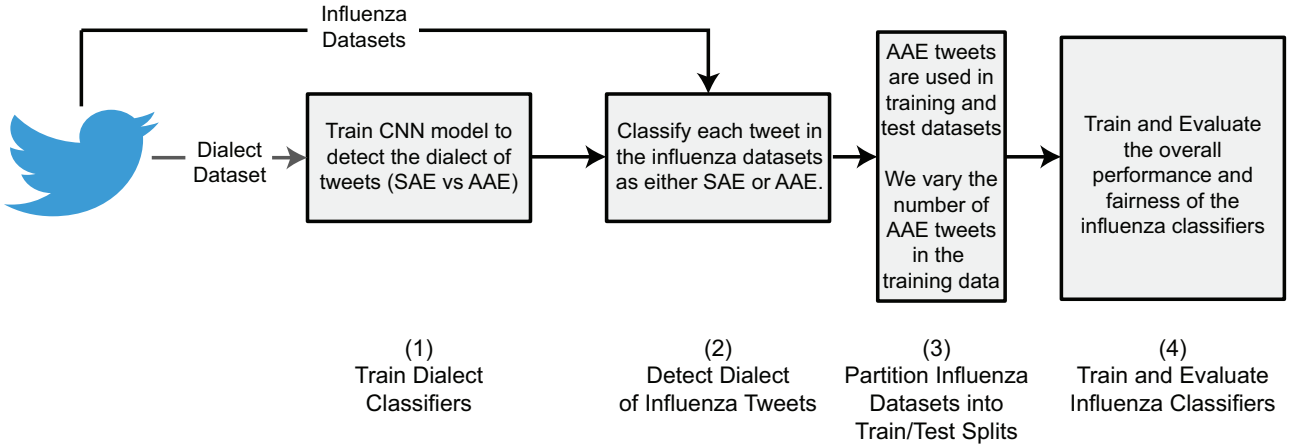

Figure 1. Overview of our data analysis pipeline. In summary, our pipeline has 4 major components: (1) training a dialect classifier to detect Standard American English (SAE) and African American English (AAE), (2) training multiple machine learning models on influenza datasets, (3) partitioning the influenza datasets to test fairness, and (4) the trained models are analyzed. CNN: convolutional neural network.

Table 1. Breakdown of total examples in each influenza-related dataset

FluTrack dataset summary					FluVacc dataset summary				
Class	Total	SAE	AAE	% SAE	Class	Total	SAE	AAE	% SAE
Related	2436	2334	102	95.81	**Vaccine related**	9517	9258	259	97.28
Not related	1900	1830	70	96.32	**Not related**	483	466	17	96.48
Awareness	1294	1242	52	95.98	**Intent**	3148	3027	121	96.16
Infection	1359	1303	56	95.88	**No intent**	6365	6228	137	97.85
Self	1392	1338	54	96.12	**Received**	3097	2981	116	96.25
Other	664	638	28	96.08	**Not received**	743	708	35	95.29

AAE: African American English; SAE: Standard American English.

alect of each tweet in various influenza-related datasets. Third, AAE tweets are partitioned in the training and testing datasets. In this experiment, we also subsample different numbers of AAE tweets in the training data to measure the impact of varying amounts of AAE training data on the fairness metrics. Fourth, we train and evaluate various models (ie, neural networks and linear models) on multiple influenza datasets to understand the biases in them and its relationship with their overall performance. In the following subsections, we describe the datasets we use for our experiments and each of our analysis steps in detail.

Datasets

In this section, we provide context on each dataset that we investigate. We also describe how they are used for training and evaluating the fairness of machine learning–based influenza classifiers. Specifically, we make use of 3 datasets: Dialect,[33] FluTrack,[5] and FluVacc.[4] Dialect is used for train a model to detect SAE or AAE text. FluTrack and FluVacc are used to train the influenza-related classifiers. The basic statistics of the influenza-related datasets are shown in Table 1. Overall, AAE tweets appear infrequently throughout every dataset used in our experiments, matching real-world conditions. We describe each dataset in detail subsequently.

Dialect dataset

Blodgett et al[33] developed a probabilistic model that combines geolocated tweets with the U.S. Census block group geographic areas to estimate message-level demographic information (block groups are the smallest geographical unit for which the U.S. Census bureau publishes sample data). Race and ethnicity information for each block group comes from the Census' 2013 American Community Survey. We make use of 59 million released tweets released by Blodgett et al[33] that contain message-level demographic estimates. Note that it is important to point out that message-level estimates are indicating whether the text similar to text written in areas with large African American or other communities (eg, Hispanic or White). The estimates are not indicative of the race or ethnicity of the individual users. Following the work by Elazar and Goldberg[40] and Rios,[17] we group the tweets into 2 linguistic styles: SAE and AAE. We limit our study to all tweets annotated with AAE and SAE with a confidence of at least 80%. This resulted in 1.6 million AAE tweets and millions of SAE tweets. To reduce the size of the SAE tweets, we randomly sample 5 million, resulting in a dataset of 6.6 million tweets. Finally, Dialect is used to train a CNN[41] to detect the dialect of each tweet. The CNN model is used in step 2 or our data analysis process, as shown in Figure 1.

FluTrack dataset

The FluTrack database[5] consists of 11 990 tweets collected from years 2009 to 2012 (because the dataset was released using Tweet IDs, only a subset of the dataset was available for our study, that is, some tweets and accounts were deleted since the original study). Each tweet is annotated with up to 3 labels (this is a multilabel clas-

sification task, not multiclass): related vs not related, awareness vs infection, and self vs other. It is important to note that there is a hierarchical structure between the labels. Specifically, only related tweets are annotated with the awareness vs infection and self vs other labels. During evaluation, to ensure the fairness estimates are easy to interpret, we only evaluate the awareness vs infection and self vs other classifiers on related test tweets, otherwise, we need to handle cascading errors. . The first class (related vs not related) categorizes each tweet based on whether it discusses an influenza-related topic or not. If a tweet is related to influenza, then it is categorized based on whether it is raising awareness to influenza or if it discusses a specific infection (awareness vs infection). Many tweets may simply raise awareness, instead of discussing an infection, meaning that tweets discuss beliefs related to influenza infections or preventative influenza measures are not useful for disease surveillance. Furthermore, each flu-related tweet is also labeled as self or other depending on whether it is about the user (self) or about another person (other). Both infection- and awareness-related tweets can be annotated as either self or other. For instance, many tweets discussing flu vaccines are annotated as awareness. So, the tweet "I am going to get the flu shot" would be labeled with both the awareness and self classes.

FluVacc dataset

Social media is not only useful for traditional disease surveillance tasks. For instance, social media can also be used to understand the public's view about potential treatments and vaccinations. This is important, especially if we want to combat potential misinformation campaigns at scale.[42] The FluVacc dataset is from Huang et al[4] and contains 10 000 annotated tweets. Each tweet is categorized with up to 3 major classes: vaccine related vs not related, which classifies whether a tweet is about influenza vaccines; received vs not received; and intent vs no intent. Similar to the FluTrack dataset, this is a multilabel task, and there is a hierarchical structure between the vaccine related vs not related class, and the others; at test time, to avoid handling cascading errors in our analysis, we only apply the received vs not received and intent vs no intent classifiers to vaccine-related tweets Received vs not received is used to detect whether a tweet discusses a user actually receiving a vaccine. Similarly, intent categorizes whether the user plans to receive the vaccine. It is important to note that a tweet may discuss receiving a vaccine and express the intent to receive it again.

Dialect detection with convolutional neural networks

As shown in step 2 of Figure 1, we train a CNN model[41] to predict the dialect of individual tweets using the Dialect dataset. The dialect dataset is split into 80% for training or validation and 20% for testing. Following Rios,[17] we use the CNN architecture from Kim.[41] The CNN model is trained with 900 filters that spans 3,4, and 5 words. For the AAE class, the final CNN has an F1 of 0.87, with a precision of 0.91 and a recall of 0.84. The precision, recall, and F1 for the SAE class were 0.97, 0.95, and 0.96, respectively. Once the model is trained, a new tweet can be passed through the CNN and the predicted dialect of the tweet is returned. This allows us to separate out data into different populations based on their dialects, which is important because these attributes are not provided in influenza datasets. See the Supplementary Appendix for a detailed evaluation of the dialect detection model on the influenza datasets.

Influenza classification models

We compare 3 models on each of the influenza datasets in step 4 (Figure 1): linear SVM, CNN, and bidirectional long short-term memory (BiLSTM). Furthermore, for both neural network models, we analyze the use of different pretrained word embeddings. We briefly describe each model subsequently.

Linear SVM

In biomedical research using social media, linear models have been shown to outperform neural networks for some tasks (eg, identifying adverse drug reactions).[43] We trained a linear SVM using term frequency–inverse document frequency weighting of unigrams and bigrams (ie, single words [eg, *vaccine*] and pairs of words like [eg, *flu vaccine*] are used as features) and L2 regularization. Term frequency–inverse document frequency weighting is a statistical measure that weights how important words are in a corpus. Furthermore, we searched for the best C value from the set $\{0.0001, 0.001, 0.01, 0.1, 1, 10\}$ using a validation dataset. The SVM is implemented using the LinearSVC classifier in scikit-learn.[44]

Convolutional neural network

The CNN architecture has shown success in text classification across many biomedical tasks.[45–47] For the CNN model implemented in this article, we use the architecture from Kim.[41] Essentially, the CNN can discover patterns and identify semantics found in different sized n-grams for the purpose of classification. Specifically, for each task, the Kim CNN models were trained with 512 filters for each span width of 3, 4, and 5 words. Because of the cost of training the model, hyperparameters were chosen manually following some of the best practices described in Zang and Wallace.[48] In general, we found the ngram ranges of 3, 4, and 5 words (similar to the Kim)[41] to perform the best with 512 filters with a dropout rate of 0.5. From our limited tests, further increasing the filters did not improve the CNN's performance. The model was trained with the Adam optimizer[49] for 30 epochs. The best epoch was chosen based on a held-out validation dataset. The model was implemented using the Keras Python package.[50]

Bidirectional LSTM

We trained a BiLSTM model, which has been shown to perform well across a wide variety of biomedical NLP tasks.[46,51] Unlike the CNNs, BiLSTM models are recurrent networks that are able to capture dependencies between words. BiLSTM units perform well with time series and sequence data since information can be kept across the entire sequence. By implementing a BiLSTM, dependencies of words are captured in both directions, forward and backward. The BiLSTM model is trained with a hidden state size of 512 for each direction. The model was trained with the Adam optimizer[49] for 30 epochs. The best epoch was chosen based on a held-out validation dataset. For the BiLSTM, we tried a few other hyperparameter configurations, eg, decreasing and increasing the size of the hidden state as well as the number of hidden layers. Overall, a hidden state size of 512 resulted in the best performance. As the number of layers increased, the training time grew exponentially for only a return of less than a fraction of a percentage point. The dropout rate was set to 0.5. The model was implemented using the Keras Python package.[50]

Pretrained word embeddings

Pretrained word embeddings have been shown to make a large impact on the overall performance of neural network-based text classification models.[41] In this article, we also explore the overall performance of the CNN and BiLSTM models trained with different pretrained embeddings. We evaluate several variations of GLOVE and Word2Vec.[52,53] Specifically, we test the pretrained Twitter-specific embeddings GLOVE 27B embeddings (http://nlp.stanford.edu/data/glove.twitter.27B.zip)with dimensions ranging from 50 to 200, GLOVE 6B embeddings (http://nlp.stanford.edu/data/glove.6B.zip)trained on Wikipedia 2014 and Gigaword 5 with 300 dimensions, and Word2Vec Skip-Gram-based embeddings trained on Google News (https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit)with 300 dimensions.

Evaluation

We evaluate the 3 influenza classifiers using both overall performance (ie, precision, recall, and F1) and fairness. Intuitively, based on our chosen evaluation metrics, we answer the following questions: Which classifier has the best overall performance on each influenza dataset? Which classifier is the fairest? Are fairness and overall performance related, that is, is the most accurate classifier the fairest?

To measure the fairness of the different models, we compare the absolute differences between the FPR and FNR calculated independently on SAE and AAE.[30] FPR and FNR are defined as

$$FPR = \frac{FP}{FP + TN} \quad \text{and} \quad FNR = \frac{FN}{FN + TP}$$

where *TP*, *FP*, *FN*, and *TN* represent the number of true positives, false positives, false negatives, and true negatives, respectively. Each score is calculated for the entire test dataset and the SAE and AAE test examples independently. The *FPR* and *FNR* scores for each group are combined using the false positive equality difference (FPED) and false negative equality difference (FNED).[14] Essentially, *FPED* is measuring the predictive equality, and FNED is measuring the equality of opportunity. *FPED* and *FNED* are defined as

$$\begin{aligned} FPED &= \sum_{t \in T} FPR - FPR_t \text{ and } FNED \\ &= \sum_{t \in T} FNR - FNR_t, \end{aligned}$$

respectively, where $T = \{\text{AAE, SAE}\}$. *FPR* and *FNR* represent the overall false positive and false negative rates, respectively. FPR_t and FNR_t represent the group-specific (ie, AAE or SAE) false positive and false negative rates. Smaller *FPED* and *FNED* scores represent fairer classifiers. Intuitively, if models have large false positive (or false negative) rates for certain underrepresented groups (eg, African Americans), then large absolute differences in *FPR*/*FNR* could potentially have unfair consequences if the model is used without this knowledge.

RESULTS

For evaluation, following prior work in methodological testing procedures of machine learning in the biomedical context,[54] we performed Monte Carlo cross-validation testing—sometimes referred to as repeated subsampling. Specifically, the dataset was split into 10 unique training, validation, and test splits. A total of 80% of the data was used for training and validation. A total of 20% of the data was used for testing. A total of 20% of each training data split was used as a validation dataset. Furthermore, because of the variance in performance produced by neural networks, on each data split, we repeatedly train each model 10 times (ie, each model was trained on each split 10 times using different random seeds). This procedure results in a total of 100 instances trained of each model. The results reported in this article are the average across both the data splits and multiple runs. Note that for some classes, there were not enough AAE tweets to evaluate using AAE tweets in both the training and testing datasets. Therefore, we report 2 sets of experiments. First, when we report the overall results in Tables 2 and 3, we did not use any AAE examples in the training data. Moreover, in the Supplementary Appendix we perform fairness experiments where all of the AAE examples are only used for testing. This is actually a likely scenario in which many dialects will not appear in a training dataset (eg, Chicano English or AAE variants such as urban or rural AAE). Therefore, these supplementary experiments will provide insight into how these models will perform for low-resource dialects that do not appear in the training dataset. Second, for the fairness results in Table 4 and Figures 2 and 3, up to 50% of the AAE tweets are used for training, while the other 50% are used for testing. For significance testing, we follow the strategy proposed in

Table 2. The mean P, R, and F1 scores for the 3 labels in the FluTrack dataset

	Related vs unrelated			Awareness vs infection			Self vs other		
	P	R	F1	P	R	F1	P	R	F1
Linear SVM	0.766	0.823	0.793	0.821	0.816	0.818	0.766	0.823	0.793
CNN GloVe 300	0.809[c]	**0.850**[b]	0.827[c]	0.903[c]	**0.906**[c]	**0.905**[c]	0.809[c]	0.847[b]	0.827[c]
CNN Twitter GloVe 50	0.813[c]	0.832	0.822[c]	0.850[c]	0.848[c]	0.849[c]	0.813[c]	0.832	0.823[c]
CNN Twitter GloVe 100	**0.816**[c]	**0.850**[b]	**0.832**[c]	**0.919**[c]	0.881[c]	0.900[c]	**0.816**[c]	**0.850**[b]	**0.832**[c]
CNN Twitter GloVe 200	0.800[c]	0.822	0.811[c]	0.866[c]	0.882[c]	0.874[c]	0.800[c]	0.822	0.811[c]
CNN Word2Vec 300	0.796[c]	0.839[a]	0.817[c]	0.902[c]	0.903[c]	0.903[c]	0.796[c]	0.839[a]	0.817[c]
BiLSTM GloVe 300	0.771	0.836	0.802[b]	0.857[c]	0.771	0.812	0.771[a]	0.836	0.802[b]
BiLSTM Twitter GloVe 50	0.759	0.845[a]	0.799[a]	0.748	0.760	0.754	0.759	0.845[a]	0.799[a]
BiLSTM Twitter GloVe 100	0.795[c]	0.794	0.794	0.821	0.752	0.785	0.795[c]	0.794	0.794
BiLSTM Twitter GloVe 200	0.767	0.837	0.800[a]	0.876[c]	0.737	0.800	0.767	0.837	0.800[a]
BiLSTM Word2Vec 300	0.788[c]	0.829	0.808[c]	0.833[a]	0.819	0.826	0.788[c]	0.829	0.808[c]

P: precision; R: recall. **Bold** font indicates the best result obtained in each column.

[a]P value (resulting from the Wilcoxon signed rank test) between .05 and .01.

[b]P value (resulting from the Wilcoxon signed rank test) between .01 and 0001.

[c]P value (resulting from the Wilcoxon signed rank test) that is $\leq$.001.

Table 3. The mean P, R, and F1 scores for the 3 labels in the FluVacc dataset

	Related vs unrelated			Received vs not received			Intent vs no intent`		
	P	R	F1	P	R	F1	P	R	F1
Linear SVM	0.987	0.994	0.991	0.886	0.939	0.911	0.829	0.828	0.828
CNN GloVe 300	**0.993**[c]	0.999[c]	**0.996**[c]	0.922[b]	**0.961**[b]	0.944[b]	**0.932**[c]	0.876[c]	0.903[c]
CNN Twitter GloVe 50	**0.993**[c]	0.999[c]	**0.996**[c]	0.917[c]	0.942	0.920[a]	0.900[c]	**0.904**[c]	0.902[c]
CNN Twitter GloVe 100	0.991[c]	0.999[c]	0.995[c]	0.926[c]	0.946	0.936[b]	0.931[c]	0.893[c]	**0.912**[c]
CNN Twitter GloVe 200	0.991[c]	**1.00**[c]	0.995[c]	**0.945**[c]	0.951[a]	**0.948**[b]	0.923[c]	**0.904**[c]	0.902[c]
CNN Word2Vec 300	0.992[c]	0.999[c]	**0.996**[c]	0.922[c]	0.949	0.935[b]	0.908[c]	0.876[c]	0.892[c]
BiLSTM GloVe 300	0.987	0.998[c]	0.992[c]	0.874	0.936	0.904	0.833	0.784	0.808
BiLSTM Twitter GloVe 50	0.987	0.996[b]	0.991	0.828	0.951	0.885	0.822	0.750	0.784
BiLSTM Twitter GloVe 100	0.985	0.997[c]	0.991	0.882	0.892	0.887	0.770	0.874[c]	0.818
BiLSTM Twitter GloVe 200	0.991[c]	0.998[c]	0.994[c]	0.902[a]	0.894	0.898	0.798	0.865[c]	0.830
BiLSTM Word2Vec 300	0.987	0.998[c]	0.993[c]	0.853	0.920	0.885	0.837	0.819	0.828

P: precision; R: recall. **Bold** font indicates the best result obtained in each column.

[a]Pvalue (resulting from the Wilcoxon signed rank test) between .05 and .01.

[b]P value (resulting from the Wilcoxon signed rank test) between .01 and 0001.

[c]P value (resulting from the Wilcoxon signed rank test) that is ≤.001.

Table 4. FluVacc results for the Intent class using a training dataset with a balanced number of AAE and SAE examples

	P	R	F1	FPED	FNED
Linear SVM	**0.786**	0.780	**0.783**	0.095	0.105
CNN GloVe 300	0.752	0.787	0.768	0.095	0.082
CNN Twitter GloVe 50	0.759	0.764	0.758	0.146[c]	0.093
CNN Twitter GloVe 100	0.760	**0.790**[a]	0.773	0.122[b]	0.086
CNN Twitter GloVe 200	0.777	0.784	0.779	0.096	0.081
CNN Word2Vec 300	0.766	0.741	0.752	0.080	0.064
BiLSTM GloVe 300	0.752	0.727	0.738	0.190[c]	0.104
BiLSTM Twitter GloVe 50	0.755	0.668	0.707	**0.257**[c]	**0.172**[c]
BiLSTM Twitter GloVe 100	0.755	0.717	0.734	0.250[c]	0.143[c]
BiLSTM Twitter GloVe 200	0.764	0.730	0.745	0.218[c]	0.142[c]
BiLSTM Word2Vec 300	0.772	0.644	0.699	0.172[c]	0.112[a]

This table shows theresults of undersampling the SAE examples to be equal to the number of Intent AAE examples. **Bold** font indicates the highest score in each column.

AAE: African American English; FNED: false negative equality difference, FPED: false positive equality difference; P: precision; R: recall; SAE: Standard American English.

[a]P value (resulting from the Wilcoxon signed rank test) between .05 and .01.

[b]P value (resulting from the Wilcoxon signed rank test) between .01 and 0001.

[c]P value (resulting from the Wilcoxon signed rank test) that is ≤.001.

prior biomedical studies[54] using the Wilcoxon signed rank test. Significance is calculated with respect to the linear SVM model (ie, we check if the neural network models are significantly better than the linear SVM for the overall results). For fairness metrics, we test if neural network models are significantly worse than the linear SVM.

FluTrack experiments

The overall performance results on the FluTrack[5] dataset is presented in Table 2. Both neural network-based models (ie, the CNN and BiLSTM) outperformed the baseline linear SVM. When comparing the CNN to the BiLSTM, the CNN outperformed the BiLSTM consistently across multiple word embeddings. This is an important factor to remember when discussing the fairness measurements. The best CNN model for related performed nearly 0.03 (3%) better than the best BiLSTM model. Similarly, the best awareness CNN model outperforms the best BiLSTM model by nearly 0.08 (8%). With regard to the best pretrained word embeddings for the CNN model, the Twitter GloVe 100 word embeddings outperformed the others for the Related and Self labels. Twitter GloVe 300 was the best for the Awareness label. For the BiLSTM, Word2Vec 300 generally had the best F1.

In Figure 2, we report the fairness results on the FluTrack dataset. The results are reported for the FluTrack class with the greatest number of AAE examples, related vs unrelated. In summary, we found that the neural network models generally had higher FPED and FNED scores than the linear SVM for the other classes. For the results in Figure 2, 50% of AAE related vs unrelated examples are used for training and the other 50% are used for testing. From the 50% of AAE examples used in the training dataset, we report the results of using different proportions of AAE examples in the training data: 0%, 20%, 40%, 60%, 80%, 100%. The scores do not vary substantially as more AAE examples are used in the training dataset. For instance, the CNN model, trained with the Word2Vec 300 embeddings, has similar FPED scores using 0% of the AAE

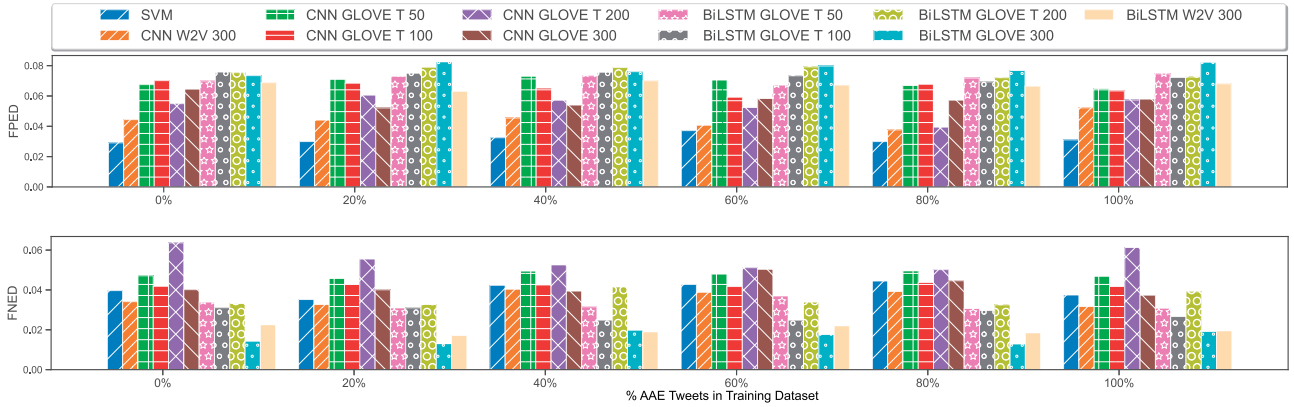

Figure 2. FluTrack's experimental results using African American English (AAE) tweets in both the training and test datasets. The false positive equality difference (FPED) and false negative equality difference (FNED) scores are plotted using different percentages of AAE tweets in the training dataset.

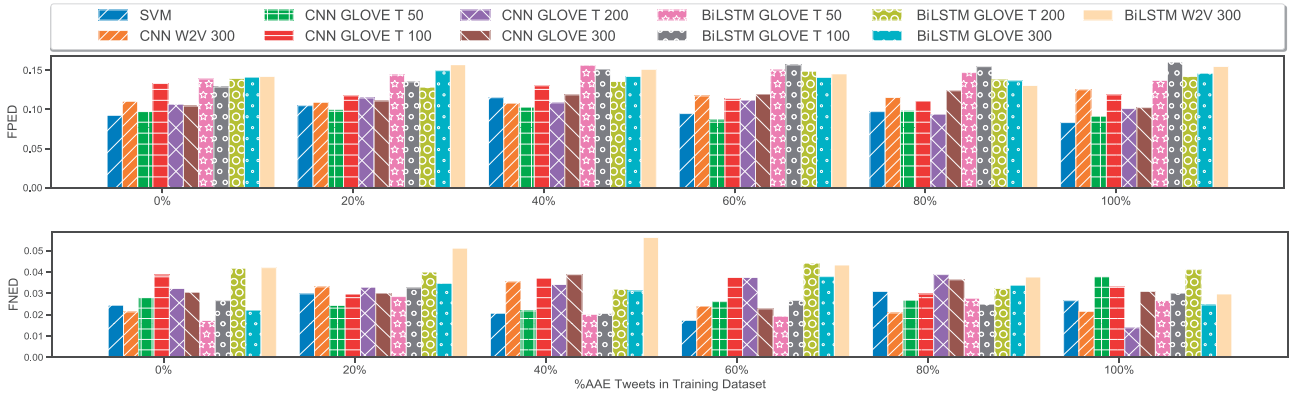

Figure 3. FluVacc's experimental results using African American English (AAE) tweets in both the training and test datasets. The false positive equality difference (FPED) and false negative equality difference (FNED) scores are plotted using different percentages of AAE tweets in the training dataset.

examples as it does use 100%. We find similar results with the FNED scores (eg, BiLSTM Word2Vec 300). See the Supplementary Appendix for FluTrack fairness experiments for more classes using all of the AAE examples in the test set.

FluVacc results

The overall performance results on the FluVacc[4] dataset is presented in Table 3. The results on FluVacc are similar to the findings on FluTrack. Specifically, we find that the CNN outperforms both the linear SVM and BiLSTM models across the precision, recall, and F1 metrics for each label. Specifically, the best CNN model for intent detection is 0.912, a nearly 10% absolute improvement over the linear SVM (0.828) and the best BiLSTM model (0.828). The best CNN model for the received label also outperformed the other methods by a large margin (eg, by more than a 4% absolute improvement over the next best BiLSTM model). Moreover, unlike the FluTrack results, the linear SVM model generally performs equivalent or better than the BiLSTM. For instance, the linear SVM's F1 score for the received label is 0.01 (1%) better than the best performing BiLSTM model. For the related label, while the CNN performed best overall, the results are similar across models. We found that the related label is relatively easy to classify because of certain keywords not appearing often in the not related label (eg, "vaccine"). We also find that the best pretrained word embeddings vary from model to model. For instance, the best embeddings for the CNN are generally GloVe 100 and GloVe 300, while the best BiLSTM embeddings are GloVe 300 and Word2Vec 300.

In Figure 3, we report the fairness results of using AAE tweets in the training set for the FluVacc dataset. Again, we used the class with the largest number of evenly distributed AAE examples, the intent vs no intent class. For a majority of the models for the task of intent classification, there is no consistent pattern of improvement of the FPED and FNED scores as we add more AAE tweets to the training set. On the contrary, adding AAE tweets seems to have little effect on the FPED and FNED scores. It is important to note that there is still an imbalance between SAE and AAE tweets in the training data. However, to the best of our knowledge, this is realistic because current research methodologies do not generally spend time collecting equal amounts of examples across all dialects. Therefore, our results paint a realistic picture about how current models perform. Finally, for the FPED scores, we also observe that the SVM generally results in the smallest score. See the Supplementary Appendix for FluVacc fairness experiments for more classes using all of the AAE examples in the test set.

How much does the imbalance between SAE and AAE examples in the training dataset affect the fairness metrics? In Table 4, we evaluate the performance of each model on the intent class using a balanced training dataset. We use the intent class because it has the largest number of AAE examples for both the intent and no intent cases. The SAE tweets are undersampled at random to match the total number of AAE examples. The results of this experiment are shown in Table 4. We make 2 major findings. First, while not directly comparable to Table 3, we find that undersampling results in a drop in F1 compared with using all of the dataset. Moreover, the linear SVM resulted in the most accurate method, which is expected given the smaller training dataset. Interestingly, we find that the BiLSTM method results in the most unfair results (ie, the highest FPED and FNED scores). The CNN models have higher FPED scores than the SVM and both the SVM and CNN have similar FNED scores.

Qualitative error analysis

In this section, we provide a couple of AAE examples (the examples have been slightly altered to preserve the privacy of users in the dataset) that resulted in incorrect predictions by the classifiers. We want to provide some insight into what aspects of AAE are potentially causing the problems. We found many examples in which the models had trouble classifying AAE text when they contain phonological variants of words. For instance, in the example from the FluVacc dataset

> "*Iont think my sister is making me go to school tomorrow since it's flu shot day at school*"

should be classified as "no intent." However, all of the classifiers classify it as "intent" instead. The suspected cause is the word *iont*—a well-known AAE phonological word variant[55]—which means "I don't." The likely cause of the errors is the limited number of AAE tweets. However, because it is not feasible to always collect enough AAE examples to handle these AAE word variants, how could this example be handled correctly? One potential solution would be to use models that operate at the character level, not the word level. Substrings of *iont* could correlate with *I don't*. The use of character information has been shown to be helpful in reducing bias in named entity recognition models.[56] Therefore, similar solutions could potentially help for influenza classification.

We found other AAE tweets that caused erroneous predictions for reasons not related to phonological word variants. For instance, the FluVacc example AAE tweet,

> "*I ain't donating sh^{a}t y'all kiss my a^{a}s already forced me to get the flu shot bullsh^{a}t*"

was correctly classified by the SVM classifier as received. However, most of the neural network methods incorrectly classified it has not received. In this example, we believe that the neural networks overfit the negation word *ain't*, and potentially, the curse words (expressing negative sentiment toward the vaccine). In this case, an obvious potential solution is to further regularize the neural networks to reduce overfitting (eg, with a larger dropout rate or L2 regularization). However, while more regularization may help, it is nontrivial to do when there are a small number of minority samples, or worse, minority samples do not appear in the dataset. It is important to note that it is likely not possible to collect large amounts of data for all dialects, so more data are not a solution for fair classifiers. Because it looks like the CNN and BiLSTM may rely on surface-level information, rather than on real natural language understanding, it may be beneficial to explore novel methods of training neural networks by augmenting the data using adversarial learning.[57]

DISCUSSION

Overall, the major finding of this article is that machine learning methods for influenza-related tasks using social media data are biased. We did not simply detect bias, we also quantified it across multiple machine learning models and datasets. With the interest of using social media to track the spread of viruses, these inaccuracies can cause a model to misrepresent certain neighborhoods as hot spots, or worse, identify communities with underrepresented populations as unlikely to develop a large number of infections. This can occur if the community, as a whole, uses a different dialect that is not consistent with the general population in which the data was collected. Note that the results of this experiment are specific to the datasets we evaluated. The models may be biased differently on other datasets and tasks.

Another interesting finding which generalizes across both the FluTrack and FluVacc datasets is that simple, ngram-based linear SVM models are competitive with some neural networks in terms of overall performance. We find that linear SVMs generally, but not always, result in fairer predictions then the best neural network methods on the 2 datasets we analyzed. Though neural network–based methods can achieve better performance compared with traditional statistical methods, interpretability is a major limitation for these deep learning methods. Therefore, in this case, linear SVMs provide a strong baseline while offering interpretability and fair results (as compared with the best neural network methods).

In summary, it is important to think about the potential impact the unfair results can have on minority communities. If statistics based on machine learning methods are used by policymakers, then unfair models could impact underrepresented group's access to certain over-the-counter medications, or worse, affect basic healthcare resources offered to their communities. For instance, if vaccines are limited, and a model incorrectly predicts that communities with certain large underrepresented populations will not be impacted by influenza (ie, high FPED), then they will be unfairly impacted. This could potentially increase health disparities that already exist because of economic disparities.

Limitations to this study

There are 4 limitations to this study. First, we rely on a SAE vs AAE dialect classifier to partition the datasets. The classifier is neither perfect nor is the classifier's training data. However, as was shown in prior work[17] and in our dialect evaluation in the Supplementary Appendix, the classifier does a good job at identifying tweets that contain common AAE syntactic and phonetic constructions.

Second, the number of AAE tweets is small. However, there is still evidence of bias in other classes with substantially more AAE data (eg, intent vs no intent which has more than 100 AAE tweets in each class). Furthermore, the bias is consistent across 2 datasets and multiple classes.

Third, we focus on dialect, which is directly related to neither race nor ethnicity. While there has been a wide array of research that predicts social identity (eg, race and sex) using text information, relying on text information alone to infer population-level statistics for race and ethnicity excludes people that do not write in a way that matches their group-identities "norm." Because race and ethnicity are impossible to fully detect automatically, we believe a more

inclusive way of obtaining social identity information is through optional self-reported surveys. The approach of asking rather than predicting (ie, relying on self-identified demographic information) is also recommended for studies about sex.[58,59] Overall, detecting social identity automatically can potentially lead to adverse outcomes. Hypothetically, predictions of social identity could be used to deprive people of opportunities. Yet, there are potential benefits of social identity detection methods in the field of biomedical informatics. For example, identity predictions could be used to measure potential health disparities. The decision process of choosing which applications will result in harm is complex. There have been recent proposals to introduce ethical review boards at the organizational level to help make such decisions—potentially extending the duties of current institutional review boards.[62] Currently, ethical issues in natural language processing applications are unlikely to raise the flags required to trigger an institutional review board approval process.[63]

Fourth, while we predict dialect, we do not make use of manually curated dialect annotations. Our evaluation strategy in the Supplementary Appendix relies on measuring well-known AAE phonetic and syntactic constructions. Moreover, our dialect classifier is trained using estimated dialect annotations. Why don't we manually annotate a small set of AAE tweets to evaluate or train the dialect classifier? Our answer to this question has 2 main points. First, it is difficult to decide a priori the "threshold" required for a tweet to be considered AAE. Is a tweet written in AAE if it contains a single AAE phonetic construct (eg, sumn)? Does it need more than 2 phonetic constructions? Does the tweet need to contain common syntactic patterns (eg, habitual be) to be AAE? Instead, we evaluate the dialect classifier in the Supplementary Appendix by comparing how likely well-known phonetic and syntactic constructions are in tweets labeled as SAE vs AAE by our classifier. We find that the well-known AAE constructions are more likely in tweets classified as AAE. We believe this evaluation strategy provides more flexibility than relying on an artificial threshold. Second, we only rely on a small number of well-known constructions. Thus, could manually annotation (eg, using Amazon Mechanical Turk) without relying on a few well-known constructions increase the variation of AAE text? Potentially, but, if we annotate tweets as AAE without relying on well-known constructions, we are at-risk of analyzing mock AAE rather than AAE itself.[60,61] Ronkin and Karn[61] defined mock AAE as "outgroup misappropriation of the language variety, which [indexes] racist stereotypes by reducing African Americans to stock outgroup images." Thus, at a minimum, we believe that such an annotation task would require self-identified AAE speakers, or new influenza-related data would need to be collected from self-identified AAE speakers on social media. But, beyond the minimum approach, more work is required to understand the best AAE annotation technique. Recent work in racial categorization for algorithm fairness suggests "various choices that go into the operationalization of race for the purposes of fairness-informed analysis or interventions significantly impact the result" and "measurement of race should be considered as an empirical problem in its own right."[64] Because dialect annotation strategies can have an impact on the outcome of algorithm fairness studies, we find that the empirical problem of annotating dialect should be carefully considered.

CONCLUSION

In this article, we used 2 influenza-related social media datasets to understand the potential biases in machine learning models trained on them. The major finding of this article is that the resulting models are biased. Therefore, practitioners should be aware of the potential harms related to biased methods. As future work, it is important to expand this study to other tasks, machine learning models (eg, BERT[24]) and demographic factors. Given the generalizability of the framework presented in this article, it can easily be applied to other datasets. Beyond measuring bias, we believe that it is also important to adapt recent methods to reduce the bias of state-of-the-art machine learning approaches[65] to the biomedical NLP domains.

FUNDING

This material is based on work supported by the National Science Foundation (Grant No. 1947697).

AUTHOR CONTRIBUTIONS

BL performed the experiments and drafted the initial manuscript. AR conceived of the study, oversaw the design, and reviewed and approved the manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGEMENTS

We thank the reviewers for their insightful comments and help improving this article.

DATA AVAILABILITY

The datasets underlying this article are available online. The Blodgett et al[33] demographic dataset is available online (http://slanglab.cs.umass.edu/TwitterAAE/). Both the FluTrack[5] and FluVacc[4] datasets are available online (http://www.cs.jhu.edu/~mdredze/data/).

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Ferretti L, Wymant C, Kendall M, *et al*. Quantifying sars-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science* 2020; 368 (6491): eabb6936.
2. Ekong I, Chukwu E, Chukwu M. COVID-19 mobile positioning data contact tracing and patient privacy regulations: exploratory search of global response strategies and the use of digital tools in Nigeria. *JMIR Mhealth Uhealth* 2020; 8 (4): e19139.
3. Salathé M, Freifeld CC, Mekaru SR, Tomasulo AF, Brownstein JS. Influenza a (h7n9) and the importance of digital epidemiology. *N Engl J Med* 2013; 369 (5): 401–4.
4. Huang X, Smith MC, Paul MJ, *et al*. Examining patterns of influenza vaccination in social media. In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence: AAAI-17*; 2017: 542–6.
5. Lamb A, Paul MJ, Dredze M. Separating fact from fear: tracking flu infections on Twitter. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; 2013: 789–95.
6. Corley C, Mikler AR, Singh KP, Cook DJ. Monitoring influenza trends through mining social media.*BIOCOMP* 2009; 2009: 340–6.

7. Corley C, Cook D, Mikler A, Singh K. Text and structural data mining of influenza mentions in web and social media. *Int J Environ Res Public Health* 2010; 7 (2): 596–615.
8. Santillana M, Nguyen AT, Dredze M, Paul MJ, Nsoesie EO, Brownstein JS. Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS Comput Biol* 2015; 11 (10): e1004513.
9. Ahmed N, Quinn SC, Hancock GR, Freimuth VS, Jamison A. Social media use and influenza vaccine uptake among white and African American adults. *Vaccine* 2018; 36 (49): 7556–61.
10. Fiscella K. Commentary—anatomy of racial disparity in influenza vaccination. *Health Serv Res* 2005; 40 (2): 539–50.
11. Bleser WK, Miranda PY, Jean-Jacques M. Racial/ethnic disparities in influenza vaccination of chronically-ill us adults: The mediating role of perceived discrimination in healthcare. *Med Care* 2016; 54 (6): 570–7.
12. Tse SC, Wyatt LC, Trinh-Shevrin C, Kwon SC. Racial/ethnic differences in influenza and pneumococcal vaccination rates among older adults in New York City and Los Angeles and orange counties. *Prev Chronic Dis* 2018; 15: E159–9.
13. Fiscella K, Dressler R, Meldrum S, Holt K. Impact of influenza vaccination disparities on elderly mortality in the united states. *Prevent Med* 2007; 45 (1): 83–7.
14. Dixon L, Li J, Sorensen J, Thain N, Vasserman L. Measuring and mitigating unintended bias in text classification. In: *AIES '18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*; 2018: 67–73.
15. Park JH, Shin J, Fung P. Reducing gender bias in abusive language detection. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*; 2018: 2799–804
16. Badjatiya P, Gupta M, Varma V. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In: *WWW '19: The World Wide Web Conference*; 2019: 49–59
17. Rios A. FuzzE: Fuzzy fairness evaluation of offensive language classifiers on African-American English. *Proc AAAI Conf Artif Intell* 2020; 34 (1): 881–9.
18. Bolukbasi T, Chang K-W, Zou JY, Saligrama V, Kalai AT. Man is to computer programmer as woman is to homemaker? *Debiasing word embeddings. Adv Neural Inf Process Syst* 2016; 4349–57.
19. Zhao J, Zhou Y, Li Z, Wang W, Chang K-W. Learning gender-neutral word embeddings. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*; 2018: 4847–53.
20. Zhao J, Wang T, Yatskar M, Cotterell R, Ordonez V, Chang K-W. Gender bias in contextualized word embeddings. 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)*; 2019: 629–34.
21. Rios A, Joshi R, Shin H. Quantifying 60 years of gender bias in biomedical research with word embeddings. In: *Proceedings of the 2020 SGIBioMed Workshop on Biomedical Language Processing*; 2020: 1–13.
22. Garg N, Schiebinger L, Jurafsky D, Zou J. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc Natl Acad Sci U S A* 2018; 115: E3635–44.
23. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. *Science* 2017; 356 (6334): 183–6.
24. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*; 2019: 4171–86.
25. Kurita K, Vyas N, Pareek A, Black AW, Tsvetkov Y. Quantifying social biases in contextual word representations. In: *1st ACL Workshop on Gender Bias for Natural Language Processing*; 2019.
26. Font JE, Costa-Jussà MR. Equalizing gender bias in neural machine translation with word embeddings techniques. In: *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*; 2019: 147–54.
27. Escudé FJ. *Determining Bias in Machine Translation with Deep Learning Techniques* [master's thesis]. Barcelona, Spain, Universitat Politècnica de Catalunya; 2019.
28. Verma S, Rubin J. Fairness definitions explained. In: *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*; 2018: 1–7.
29. Makhlouf K, Zhioua S, Palamidessi C. On the applicability of ml fairness notions. *arXiv*, doi: https://arxiv.org/abs/2006.16745, 19 Oct 2020, preprint: not peer reviewed.
30. Davidson T, Bhattacharya D, Weber I. Racial bias in hate speech and abusive language detection datasets. In: *Proceedings of the Third Workshop on Abusive Language Online*; 2019: 25–35.
31. Sen M, Wasow O. Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annu Rev Polit Sci* 2016; 19 (1): 499–522.
32. Sap M, Card D, Gabriel S, Choi Y, Smith NA. The risk of racial bias in hate speech detection. In: *Proceedings of the 57th Conference of the Association for Computational Linguistics*; 2019: 1668–78.
33. Blodgett SL, Green L, O'Connor B. Demographic dialectal variation in social media: a case study of African-American English. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*; 2016: 1119–30.
34. Blodgett SL, O'Connor B. Racial disparity in natural language processing: A case study of social media African-American English. *arXiv*, doi: https://arxiv.org/abs/1707.00061, 30 Jun 2017, preprint: not peer-reviewed.
35. Green LJ. *African American English: A Linguistic Introduction*. Cambridge, MA: Cambridge University Press; 2002.
36. Florini S. Tweets, tweeps, and signifyin' communication and cultural performance on "Black Twitter." *Television New Media* 2014; 15 (3): 223–37.
37. Eisenstein J. Identifying regional dialects in on-line social media. In: Boberg C, Nerbonne J, Watt D, eds. *The Handbook of Dialectology*. Hoboken, NJ: Wiley; 2017: 368–83.
38. Zalmout N, Habash N. Adversarial multitask learning for joint multi-feature and multi-dialect morphological modeling. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*; 2019: 1775–86.
39. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019; 366 (6464): 447–53.
40. Elazar Y, Goldberg Y. Adversarial removal of demographic attributes from text data. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*; 2018: 11–21.
41. Kim Y. Convolutional neural networks for sentence classification. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*; 2014: 1746–51.
42. Kouzy R, Abi JJ, Kraitem A, *et al.* Coronavirus goes viral: quantifying the covid-19 misinformation epidemic on Twitter. *Cureus* 2020; 12: e7255.
43. Sarker A, Belousov M, Friedrichs J, *et al.* Data and systems for medication-related text classification and concept normalization from twitter: insights from the social media mining for health (smm4h)-2017 shared task. *J Am Med Inform Assoc* 2018; 25 (10): 1274–83.
44. Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: machine learning in python. *J Mach Learn Res* 2011; 12: 2825–30.
45. Rios A, Kavuluru R. Convolutional neural networks for biomedical text classification: Application in indexing biomedical articles. In: *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*; 2015: 258–67.
46. Peng Y, Rios A, Kavuluru R, Lu Z. 2018 Extracting chemical–protein relations with ensembles of SVM and deep learning models. *Database (Oxford)* 2018; 2018: bay073.
47. Peng Y, Lu Z. Deep learning for extracting protein-protein interactions from biomedical literature. *BioNLP* 2017; 29–38.
48. Zhang Y, Wallace B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv*, doi: https://arxiv.org/abs/1510.03820, 6 Apr 2016, preprint: not peer reviewed.

49. Kingma DP, Ba J. Adam: A method for stochastic optimization. *arXiv*, doi: https://arxiv.org/abs/1412.6980, 30 Jan 2017, preprint: not peer reviewed.
50. Chollet F, *et al.* Keras. https://github.com/keras-team/keras Accessed March 1, 2020.
51. Kavuluru R, Rios A, Tran T. Extracting drug-drug interactions with word and character-level recurrent neural networks. In: 2017 *IEEE International Conference on Healthcare Informatics (ICHI)*; 2017: 5–12.
52. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. *Distributed representations of words and phrases and their compositionality. Adv Neural Inf Process Syst* 2013; 2: 3111–9.
53. Pennington J, Socher R, Manning C. Glove: global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*; 2014: 1532–43.
54. Mobadersany P, Yousefi S, Amgad M, *et al.* Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc Natl Acad Sci U S A* 2018; 115:E2970–9.
55. Jones T. Toward a description of African American vernacular English dialect regions using "Black Twitter." *Am Speech* 2015; 90 (4): 403–40.
56. Mishra S, He S, Belli L. Assessing demographic bias in named entity recognition. *arXiv*, doi: https://arxiv.org/abs/2008.03415, 8 Aug 2020, preprint: not peer reviewed.
57. Nie Y, Williams A, Dinan E, Bansal M, Weston J, Kiela D. Adversarial NLI: a new benchmark for natural language understanding. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; 2020: 4885–901
58. Scheuerman MK , Paul JM , Brubaker JR. How computers see gender: an evaluation of gender classification in commercial facial analysis services. In: *Proceedings of the ACM on Human-Computer Interaction*; 2019.
59. Keyes O. The misgendering machines: Trans/HCI implications of automatic gender recognition. In: *Proceedings of the ACM on Human-Computer Interaction*; 2018.
60. Smokoski HL, Voicing the Other: Mock AAVE on Social Media [master's thesis]. New York, NY, Graduate Center, City University of New York; 2016.
61. Ronkin M, Karn HE. Mock Ebonics: linguistic racism in parodies of Ebonics on the internet. *J Sociolinguist* 2002; 3 (3): 360–80.
62. Leidner JL, Plachouras V. Ethical by design: ethics best practices for natural language processing. In: *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*; 2017.
63. Hovy D, Spruit SL. The social impact of natural language processing. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*; 2016.
64. Hanna A, Denton E, Smart A, Smith-Loud J. Towards a critical race methodology in algorithmic fairness. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*; 2020: 501–12.
65. Jia S, Meng T, Zhao J, Chang K. Mitigating gender bias amplification in distribution by posterior regularization. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; 2020.