The Risk of Racial Bias while Tracking Influenza-Related Content on Social Media using Machine Learning

Authors

Brandon Lwowski

Department of Information Systems and Cyber Security,

University of Texas at San Antonio, USA

Anthony Rios

Department of Information Systems and Cyber Security,

University of Texas at San Antonio, USA

***Corresponding Author**: Anthony Rios*

Email: anthony.rios@utsa.edu

***Key Words**: Deep Learning, Classification, Machine Learning ,Social Network, Fairness*

***Word Count**: 5900*

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, USA

Abstract

Objective. With the seasonal outbreaks of the influenza virus, social media can be used to understand and track them using machine learning. 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 potential biases of machine learning methods developed to monitor and track the spread of viruses on social media.

Materials and Methods. Two influenza-related datasets are used to train various text classification models with multiple pre-trained word embeddings. We measure the fairness of influenza classification models by comparing the results on tweets written in Standard American English and African American English.

Results. We find that all of the tested machine learning methods are biased. We also find that the best performing neural network methods generally result in more unfair results than linear models.

Discussion. The ad-hoc use of machine learning can be harmful for certain sub-populations if fairness is neither measured nor taken into consideration. Neural network-based methods achieve better performance compared with traditional statistical methods, but interpretability is a limitation for deep learning. In contrast, linear models still provide a strong baseline while also being more interpretable and generally resulting in fairer predictions.

Conclusion. The major finding of this paper is that the resulting models built using social media data are biased. Therefore, practitioners should be aware of the potential harms related to them.

BACKGROUND AND SIGNIFICANCE

Due 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]. This paper examines machine learning methods using social media 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. (2018) [12] report a nearly 10% difference in the influenza vaccination rate between non-Hispanic Black/African American adults over 50 and non-Hispanic whites. Fiscella et al. (2007) estimated that if influenza immunization rates were equal for all races, nearly two thousand minority deaths could be prevented every year, saving more than 33 thousand minority life years [13].

In this paper, we look at the potential impact machine learning-based tools can have on health disparities in the context 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 (e.g., 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 questions such as, What is the relationship between overall classifier performance and fairness, and 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 non-biomedical text classification tasks (e.g., hate speech classification).

Word embeddings have also been shown to contain biases [18–21]. A word embedding is a learned representation/vector for text where words with similar meanings have a similar representation, algorithmically capturing the words meaning. , Bolukbasi et al. (2016) show that the word embedding for "man" is similar to "doctor", while "woman" is similar to "nurse" [18]. Garg et al. (2018) developed a technique to study 100 years of gender and racial

bias using word embeddings [22]. Kurita et al. (2019) expanded on prior work [23] to generalize bias measurement metrics for word embedding to contextual word embeddings (e.g., BERT [24]) [25]. Machine translation systems have also been shown to exhibit biases [26, 27]. Font and Costa-jussá (2019) show 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 papers focus on testing whether bias exists in various models, or on developing techniques to remove bias from classification models for specific applications. In this paper, we focus on measuring racial biases of machine learning methods in the biomedical NLP domain.

Fairness can be defined in multiple ways. In this paper, we focus on two 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 the Evaluation Section for a complete definition) is equal between two 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 hot-spots of the influenza virus, then a high FNR could lead to inadequate resources (e.g., vaccinations) to fight the virus. Similarly, Predictive Equality is a measure of the difference between the False Positive Rates (FPR) of two 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. Which is more important, Predictive Equality or Equality of Opportunity? This depends on the downstream application of the models. For the purpose of this paper, we assume they are equally important.

This paper focuses on measuring *racial* bias using definitions 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 analyse the African American English *dialect* (AAE). AAE is a common dialect spoken by some, but not all, African Americans [*]. Furthermore, AAE has been shown to transfer from the use in face-to-face conversations to written text on social media [33, 35–37].

As previously mentioned, having a machine learning model that is biased can have huge 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,

[*]It is important to note that not all speakers of AAE are African American and not all African American's are AAE speakers [35]. For more information about the correlation between AAE and racial constructs, please see Blodgett et al. (2016) for a complete discussion [33].

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. A high false negative rate for such communities could reduce the supply of medical equipment and vaccinations if statistics based on these models are used by policy makers, further increasing potential health disparities among minorities. Moreover, a high false positive rate could have a substantial impact on the economic conditions in neighborhoods with large minority populations, further expanding the existing unemployment and pay disparities they experience [39]. Overall, if policy-related decisions are made using unfair models, then this can impact the governing bodies choices on where to intervene to stop the spread of influenza, as well as expanding potential economic harms. As an real-world example on the impact of biased machine learning methods in the real world, Obermeyer et al. (2019) [40], 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. Essentially, our work is an extension of their study. But, we look at biases in machine learning applied to social media. 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 three major contributions as follows: First, we study the performance differences between SAE and AAE of machine learning models applied to various influenza-related tasks, including identifying influenza-related tweets, detecting whether a tweet is about an infection or simply raising awareness, detecting whether a user is discussing themselves or someone else, and identifying vaccine-related tweets. Second, we explore the fairness of the influenza classifiers across multiple machine learning algorithms including linear support vector machines 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 paper as well as this paper's limitations.

MATERIALS AND METHODS

We provide an overview of our study in Figure 1. This paper's methodology can be summarized in four steps. **(1.)** We train a convolutional neural network (CNN) to detect the dialect of individual tweets (i.e., SAE vs AAE). **(2.)** Next, the model is used to classify the dialect of each tweet in various influenza-related datasets. **(3.)** We use different partitions of the data for two separate experiments. First, in Experiment 1, the influenza-related datasets are split into a 80% training/development split and 20% test split. Because of the limited number of AAE tweets in

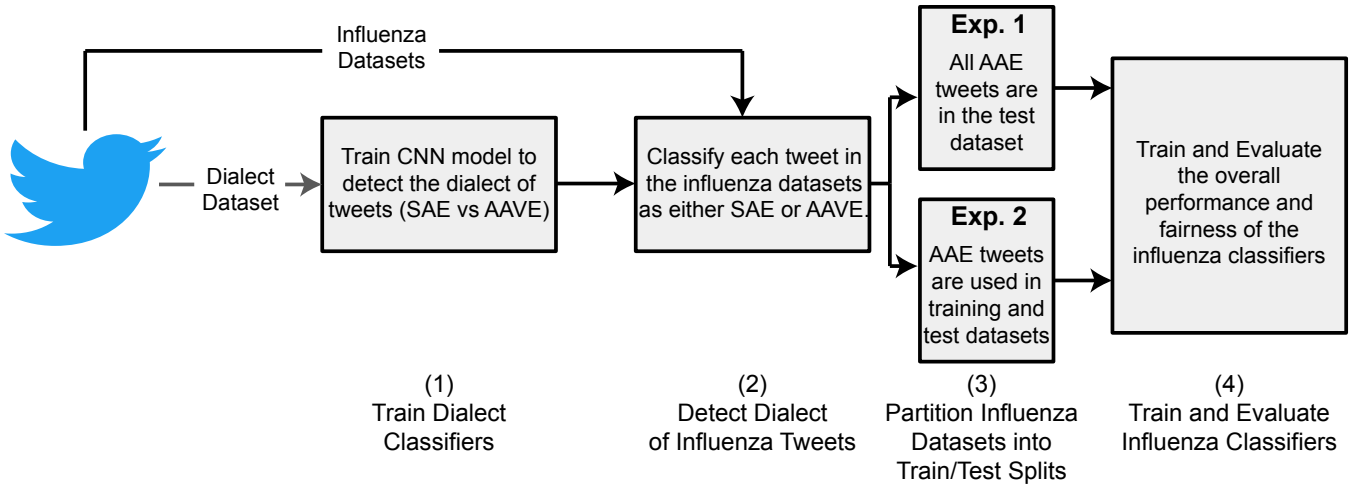


Figure 1: Overview of our data analysis pipeline. In summary, our pipeline has four major components: (1) training a dialect classifier to detect Standard American English (SAE) and African American Vernacular English (AAE), (2) training multiple machine learning models on influenza datasets, (3) partitioning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

our influenza datasets, all of the AAE tweets are reserved for testing. Why do we perform experiments without AAE tweets in the training data? While the datasets chosen for our analsyis do contain a small number of AAE tweets, in general, it is likely that many dialects will not appear in a training dataset (e.g., Chicano English or AAE variants such as Urban or Rural AAE). Therefore, these experiments will provide insight into how these models will perform for low-resource dialects that do not appear in the training dataset. Second, in Experiment 2, AAE tweets are used in the testing and training datasets. In this experiment, we also sub-sample different numbers of AAE tweets in the training data to measure the impact of "more AAE data" on the fairness metrics. **(4.)** Finally, we train and evaluate various models (i.e., neural networks and linear models) on multiple influenza datasets to understand the biases in them and its relationship with their overall performance. In the subsections below, 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 three 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. Furthermore, as previosuly mentioned, we perform experiments with and without AAE tweets in the training dataset. We describe each dataset in detail below:

Class	Total	SAE	AAE	% SAE
Related	2436	2334	102	95.81
Not Related	1900	1830	70	96.32
Awareness	1294	1242	52	95.98
Infection	1359	1303	56	95.88
Self	1392	1338	54	96.12
Other	664	638	28	96.08

(a) FluTrack Dataset Summary

Class	Total	SAE	AAE	% SAE
Vaccine Related	9517	9258	259	97.28
Not Related	483	466	17	96.48
Intent	3148	3027	121	96.16
No Intent	6365	6228	137	97.85
Received	3097	2981	116	96.25
Not Received	743	708	35	95.29

(b) FluVacc Dataset Summary

Table 1: Breakdown of Total examples in each influenza-related dataset, split into groups of Standard American English and African American Vernacular English.

Blodgett et al. (2016) [33] (Dialect). Dialect consists of more than 59 million tweets. Each tweet was annotated with various lingustic styles (e.g., SAE and AAE). It is important to note that the annotations were generated automatically, i.e., style was not manually annotated. Following the work by Elazar and Goldberg (2018) [41] and Rios (2020) [17], 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 Convolutional Neural Network (CNN) [42] 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.

Lamb et al. [5] (FluTrack). FluTrack consists of 11,990 tweets collected from years 2009 through 2012.[†] Each tweet is annotated with up to three labels (this is a multi-label classification task, not multi-class) [‡]: Related vs. Not Related, Awareness vs. Infection, and Self vs. Other. The first class (Related vs. Not Related) categorizes each tweet based on whether it discusses an influenza-related topic or not. If a tweet if related to influenza, then is 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 it is 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.

[†]Because the dataset was released using Tweet IDs, only a subset of the dataset was available for our study, i.e., some tweets and accounts were deleted since the original study.

[‡]It is important to not 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.

Huang et al. (2017) [4] (FluVacc). 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 [43]. The FluVacc dataset is from Huang et al. (2017) [4] and contains ten thousand annotated tweets. Each tweet is categorized with up to three 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". "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 also have 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 [42] to predict the dialect of individual tweets using the Dialect dataset. The dialect dataset is split into 80% for training/validation and 20% for testing. Following Rios (2020) [17], we use the CNN architecture from Kim (2014) [42]. The CNN model is trained with 900 filters that spans 3,4, and 5 words. The final CNN has an F1 of 0.87, with a Precision of 0.91 and Recall of 0.84. 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.

Influenza Classification Models

We compare three models on each of the influenza datasets in Step 4 (Figure 1): Linear Support Vector Machine (SVM), CNN, and Bi-directional Long Short-Term Memory Networks (BiLSTM). Furthermore, for both neural network models, we analyze the use of different pretrained word embeddings. We briefly describe each model below:

Linear SVM. In biomedical research using social media, linear models have been shown to outperform neural networks for some tasks (e.g., identifying adverse drug reactions) [44]. We trained a Linear SVM using term frequency inverse document frequency-weighting (TF-IDF) of unigrams and bigrams (i.e., single words, "vaccine", and pairs of words like "flu vaccine" are used as features) and L2 regularization. TF-IDF is a statistic measure that weights how important words are in a corpus. Furthermore, we searched for the best C value from the set

[§]Similar to the FluTrack dataset, this is a multi-label 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.

$\{0.0001, 0.001, 0.01, 0.1, 1, 10\}$ using a validation dataset. The SVM is implemented using the LinearSVC classifier in scikit-learn [45].

CNN. The CNN architecture has shown success in text classification across many biomedical tasks [46–48] For the CNN model implemented in this paper, we use the architecture from Kim (2014) [42]. Essentially, the CNN can discover patterns and identify semantics found in different sized n-grams for the purpose of classification. For each span of words a max-over-time pooling operation is applied to the outputs and concatenated together. The concatenated layer was fed forward through a fully connected layer. Specifically, for each task, the Kim (2014) CNN models were trained with 512 filters for each span width of three, four, and five words. Because of the cost of training the model, hyperparameters were chosen manually following some of the best practices described in Zang and Wallace (2015) [49]. In general we found the ngram ranges of three, four and five words (similar to the Kim (2015) [42]) to perform the best with 512 filters. From our limited tests, further increasing the filters did not improve the CNN's performance. The model was trained with the Adam optimizer [50] for 30 epochs. The best epoch were chosen based on a held-out validation dataset. The model was implemented using the Keras Python package [51].

BiLSTM. We trained a bi-directional LSTM model (BiLSTM), which has been shown to perform well across a wide variety of biomedical NLP tasks [47, 52]. Unlike the CNNs, BiLSTM models are recurrent networks that are able to capture dependencies between words. Long Short Term Memory 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. At each time step (i.e., at each word), the bidirectional layer provides two outputs, one hidden state for the forward pass and one hidden state for the backward pass. The hidden states for each time step are concatenated. Next, a max-over-time pooling operation is applied to all the hidden states, then passed through a fully-connected output layer. The BiLSTM model is trained with a hidden state size of 512 for each direction. The model was trained with the Adam optimizer [50] for 30 epochs. The best epoch was chosen based on a held-out validation dataset. For the BiLSTM , we tried a few other parameter configurations, by 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 model was implemented using the Keras Python package [51].

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 [42]. In this paper, 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 [53,54]. Specifically, we test the pretrained Twitter-specific embeddings GLOVE 27B embeddings [¶] with dimensions ranging from 50 to 200, GLOVE 6B embeddings [‖] trained on Wikipedia 2014 and Gigaword 5 with 300 dimensions, and Word2Vec Skip-Gram-based embeddings trained on Google News [**] with 300 dimensions.

Evaluation

We evaluate the three influenza classifiers using both overall performance (i.e., 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 most fair? Is fairness and overall performance related, i.e., is the most accurate classifier the most fair?

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

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

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

$$\text{FPED} = \sum_{t \in T} |FPR - FPR_t| \qquad \text{and} \qquad \text{FNED} = \sum_{t \in T} |FNR - FNR_t|,$$

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

[¶] http://nlp.stanford.edu/data/glove.twitter.27B.zip
[‖] http://nlp.stanford.edu/data/glove.6B.zip
[**] https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit

AAE	Ratio	SAE
iont/ion	141.405	I don't
finna	141.405	going to
nun	84.843	nothing
dey	56.562	they
nigga	35.351	guy[††]
tryna	23.567	trying to

(a)

Construction	Example	Ratio
O-be/b-V	I be vomiting bruh	8.783
done/dne-V	I done had a headache all day	8.783
gone/gne/gon-V	Then she gon be sick Af	6.920

(b)

Table 2: These tables summarize measurements of common AAE phonological variants of words and syntactic constructions as compared to SAE in the FluVacc dataset. Table 2a shows the top six AAE phonological variant words by ratio $r^{pho}_{AAE}(w)$. Table 2b shows the AAE syntactic patterns and the ratios of their occurrences in the AA- vs. SAE-aligned corpora.

RESULTS

For evaluation, following prior work in methodological testing procedures of machine learning in the biomedical context [55], we performed Monte Carlo Cross-Validation testing—sometimes referred to as repeated sub-sampling. Specifically, the dataset was split into ten unique training, validation, and test splits. 80% of the data was used for training and validation. 20% was used for testing. 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 ten times, i.e., each model was trained on each split ten times using different random seeds. This procedure results in a total of 100 instances trained of each model. The results reported in this paper are the average across both the data splits and multiple runs. The AAE tweets are split differently for Experiment 1 and Experiment 2—as shown in Figure 1. Specifically, for Experiment 1, no AAE tweets are used in the training dataset. For Experiment 2, 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 prior biomedical studies [55] using the Wilcoxon Signed Rank Test.

Dialect Detection Evaluation

Before we discuss the classification and fairness results on the influenza datasets, it is important to sanity check the performance of the dialect detection model. Unfortunately, it is non-trivial to manually annotate text as belonging to AAE or SAE. For instance, if a tweet contains a single phonological varient of a word associated with AAE (e.g., finna), is the tweet AAE? Should it contain at least two (or three) phonological variants? Rather than the phonological variants, does the text need to contain more complex syntactic constructions (e.g., habitual *be*)? Similar issues arise

[††]Nigga is more complex than this Table suggests. A more complete discussion of this topic can be found in Jones and Hall (2015) [56].

when trying to count the number of AAE speakers in general. Because of the complex nature of manually annotating AAE text—specifically, at the small scale of a single tweet—we measure well-known properties of AAE in both the AAE and SAE detected tweets based on our dialect classifier. Thus, following the work of Blodgett et al. (2016) [33], we calculate basic statistics of the AAE- and SAE-predicted text in the FluVacc dataset. The statistics are calculated on the entire FluVacc dataset (e.g., training, validation, and testing). Overall, we calculate the differences in the usage of well-known phonological word variants (e.g., sumn, finna, iont) and syntactic constructions (e.g., habitual *be*).

Phonology is the study of patterns of speech sounds in casual conversations. Orthography is the study of the style of writing, such as how words are spelled, hyphenated, and capitalized. Eisenstein (2013) [57] show that there is correlation between speech patterns and orthographic variation. While the correlation is not perfect, we are able to measure well-known phonetic patterns (which are associated with AAE) on social media by looking for their occurrences in text. There have been multiple studies of AAE phonology on Twitter [33,57–59]. Many phonological AAE features can be found in text, such as a distortion of the sound of the letter R (e.g. *brother* → *brotha*), deletion of initial *g* and *d* (e.g., *iont* → I don't), and expression of the sound of *th* as *d* (e.g., *they* → *dey*).

Following a similar framework as Blodgett et al. (2016) [33], we analyze the occurrence of *31* phonological AAE word variants (see the Supplementary Material for a complete listing). Let w represent a word. We measure the ratio

$$r^{pho}_{AAE}(w) = \frac{p(w|c = AAE)}{p(w|c = SAE)}$$

where $p(w|c = AAE)$ is the probability of w given the AAE category. Similarly, $p(w|c = SAE)$ is the probability of w given the SAE category. $p(w|c)$ is defined as

$$p(w|c) = \frac{\text{count}(w, c) + 1}{\text{count}(c) + |V|}$$

where V is the set all words in the dataset, count(w, c) is the count of w in class c, and count(c) is the count of all words in class c. The top six [‡‡] AAE phonological variant words are shown in Table 2a. Overall, we find that the variants are more likely in our AAE-predicted text. As a robustness check, we also calculated the scores for three stopwords (which we would expect to occur equally in both SAE and AAE text): *the*, *for*, and *a*. The stopwords had ratios $r^{pho}_{AAE}(w)$ of $r^{pho}_{AAE}(the) = 0.772$, $r^{pho}_{AAE}(for) = 0.634$, and $r^{pho}_{AAE}(a) = 0.735$, which are much closer to one (i.e., having an equal chance of occurring in both AAE and SAE text) than the results in Table 2a.

We also explore well-known AAE syntactic patterns. Specifically, we measure three patterns [35]: habitual *be*,

[‡‡]Given our dataset is small, these are the only six well-known AAE phonetic variants in the SAE *or* AAE tagged data.

	Related vs. Unrelated			Awareness vs. Infection			Self vs. Other		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Linear SVM	.766	.823	.793	.821	.816	.818	.766	.823	.793
CNN GloVe 300	.809***	**.850****	.827***	.903***	**.906*****	**.905*****	.809***	.847**	.827***
CNN Twitter GloVe 50	.813***	.832	.822***	.850***	.848***	.849***	.813***	.832	.823***
CNN Twitter GloVe 100	**.816*****	**.850****	**.832*****	**.919*****	.881***	.900***	**.816*****	**.850****	**.832*****
CNN Twitter GloVe 200	.800***	.822	.811***	.866***	.882***	.874***	.800***	.822	.811***
CNN Word2Vec 300	.796***	.839*	.817***	.902***	.903***	.903***	.796***	.839*	.817***
LSTM GloVe 300	.771	.836	.802**	.857***	.771	.812	.771*	.836	.802**
LSTM Twitter GloVe 50	.759	.845*	.799*	.748	.760	.754	.759	.845*	.799*
LSTM Twitter GloVe 100	.795***	.794	.794	.821	.752	.785	.795***	.794	.794
LSTM Twitter GloVe 200	.767	.837	.800*	.876***	.737	.800	.767	.837	.800*
LSTM Word2Vec 300	.788***	.829	.808***	.833*	.819	.826	.788***	.829	.808***

Table 3: The mean precision (P), recall (R) and F1 scores for the three labels in the FluTrack dataset: "Related vs Unrelated, "Awareness vs. Infection", and "Self vs. Other". A p-value (resulting from the Wilcoxon signed rank test) between 0.05 and 0.01 is indicated by *, a p-value between 0.01 and 0.001 is indicated by **, and a p-value that is less than or equal to 0.001 is indicated by ***.

future *gone*, and completive *done*. Habitual *be* is the use of the uninflected *be* to marks extended/on-going actions (e.g., He *be* puking). Future *gone* is used to mark something that will happen later (e.g., He *gone* be sick). The completive *done* is used to emphasize a recently completed action. In the context of misclassifying influenza-related tweets, it is feasible to think constructions such as the future *gone* may be incorrectly marking something that happened in the past.

Again, following the work of Blodgett et al. (2016) [33], we use simplified patterns to quantify their occurrences in the Influenza dataset. Specifically, we use the Twitter-specific part-of-speech tokenizer (Twokenizer) [60] to annotate each FluVacc tweet. The patterns we use are listed in Table 2b. The letter O represents a pronoun, while the letter V represents a verb. So, the pattern "O-be/b-V" (a pattern for the habitual *be*) represents a pronoun followed by the word be, which should be followed by a verb. Similar to phonetic word variants, we define a ratio to measure their occurrences in AAE and SAE text. Let s be a syntactic pattern, then we define the ratio as

$$r^{syn}_{AAE}(s) = \frac{p(s|c = AAE)}{p(s|c = SAE)}$$

where $p(s|c = AAE)$ is the probability of pattern s occurring given class *AAE*. The results of our experiment can be found in Table 2. Overall, we find the appearance of the patterns to be much more likely in the AAE annotated tweets compared to the SAE tweets. This result suggests that our dialect predictions our reliable, because they are more likely to contain well-known AAE phonetic and syntactic patterns.

	Related vs. Unrelated		Awareness vs. Infection		Self vs. Other	
	FPED	**FNED**	**FPED**	**FNED**	**FPED**	**FNED**
Linear SVM	.017	.020	.090	.028	.079	.027
CNN GloVe 300	.066*	.005	.169***	.095**	.332**	.116***
CNN Twitter GloVe 50	.136**	**.026**	.085	.165***	.241***	.039
CNN Twitter GloVe 100	**.152****	.015	.126**	**.229*****	.206**	.047*
CNN Twitter GloVe 200	.114**	.025	.145***	.113***	**.347*****	.031
CNN Word2Vec 300	.116**	.019	**.233*****	.092**	.222**	**.051** *
LSTM GloVe 300	.028	.008	.132**	.141***	.012	.008
LSTM Twitter GloVe 50	.098**	.022	.136***	.174***	.039	.014
LSTM Twitter GloVe 100	.081**	.025	.006	.142***	.047	.014
LSTM Twitter GloVe 200	.069**	.019	.052**	.101**	.036	.034
LSTM Word2Vec 300	.102**	.007	.156***	.153***	.052	.018

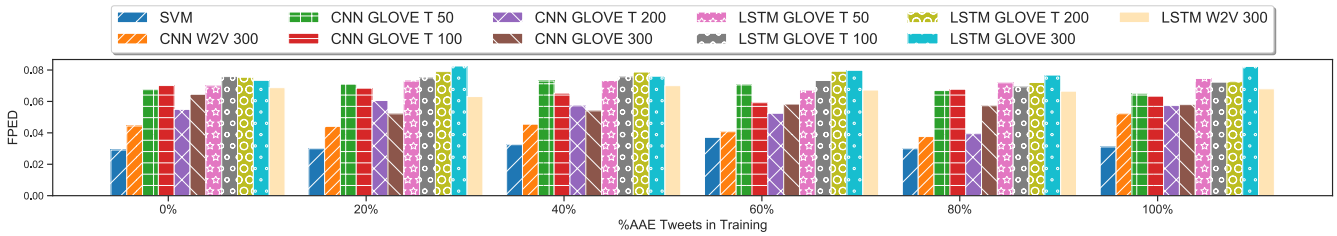
Table 4: The FPED and FNED fairness results for the three labels in the FluTrack dataset: "Related vs Unrelated, "Awareness vs. Infection", and "Self vs. Other". A p-value (resulting from the Wilcoxon signed rank test) between 0.05 and 0.01 is indicated by *, a p-value between 0.01 and 0.001 is indicated by **, and a p-value that is less than or equal to 0.001 is indicated by ***.

FluTrack Experiments

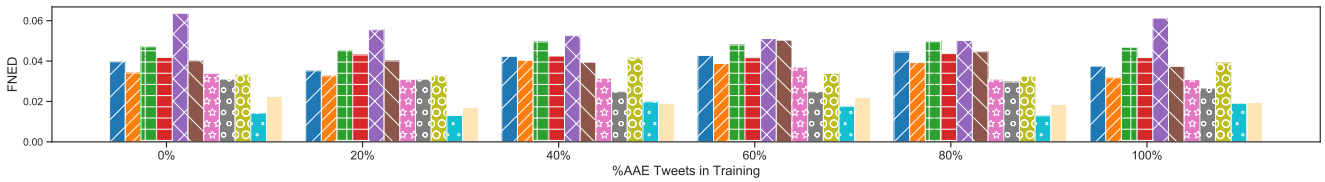
The overall performance results on the FluTrack [5] dataset are presented in Table 3. For Tables 3 and 4, the p-value obtained by the Wilcoxon signed rank test are also provided. Both neural network-based models (i.e., the CNN and LSTM) outperformed the baseline Linear SVM. When comparing the CNN to the LSTM, the CNN outperformed the LSTM 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 LSTM model. Similarly, the best Awareness CNN model outperforms the best LSTM model by nearly 0.08 (8%). With regard to the best pre-trained 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 LSTM, Word2Vec 300 generally performed the best in terms of F1.

The results of Experiment 1 (See Figure 1) for the FluTrack dataset, where the FPED and FNED fairness metrics for label are reported, can be found in Table 4. Overall, we find that the best overall performing model (the CNN) also produces the most unfair predictions. The Linear SVM, which performs similar to the LSTM, generally produces the fairer predictions than the best neural networks—with the exception of the FPED result for the Self label. However, even for FPED Self, the Linear SVM still performs similar to the most unfair LSTM model, with a difference less than 0.03 (3%).

With regard to word embeddings results in Table 4, there does not seem to be a single embedding type that



(a) The FluTrack's FPED results for the Related vs. Unrelated class.



(b) The FluTrack's FNED results for the Related vs. Unrelated class.

Figure 2: FluTrack's experimental results using AAE tweets in both the training and test datasets. The FPED and FNED scores are plotted using different percentages of AAE tweets in the training dataset.

produces the most unfair predictions. This result is in contrast with the overall results where GloVe 100 generally produced the best CNN. Thus, for the FluTrack dataset, **we find that model choice has a larger impact on fairness than pre-trained embedding choice**.

In Figure 2, we report the results of Experiment 2 (See Figure 1) on the FluTrack dataset. The results are reported for the FluTrack class with the most number of AAE examples, "Related vs Unrelated". 50% of AAE "Related vs Unrelated" examples are used for training and the other 50% are used for testing, so the results in Figure 2 and Table 4 are not directly comparable. 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%. There are two major observations. First, 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 (W2V) 300 embeddings, has similar FPED scores using 0% of the AAE examples as it does using 100%. We find similar results with for the FNED scores (e.g., LSTM W2V 300). Second, the results are similar to Experiment 1. For example, SVM has the smallest FPED score in Table 4 for the "Related vs. Unrelated" class. The SVM also results in the lowest FPED score, even if we add AAE examples to the training data.

Fluvacc Results

The overall performance results on the FluVacc [4] dataset are presented in Table 5. The results on FluVacc are similar to the findings on FluTrack. Specifically, we find that the CNN outperforms both the Linear SVM and LSTM models across the precision, recall, and F1 metrics for each label. Specifically, the best CNN model for Intent

	Related vs. Unrelated			Received vs. Not Received			Intent vs. No Intent		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Linear SVM	.987	.994	.991	.886	.939	.911	.829	.828	.828
CNN GloVe 300	**.993*****	.999***	**.996*****	.922**	**.961****	.944**	**.932*****	.876***	.903***
CNN Twitter GloVe 50	**.993*****	.999***	**.996*****	.917***	.942	.920*	.900***	**.904*****	.902***
CNN Twitter GloVe 100	.991***	.999***	.995***	.926***	.946	.936**	.931***	.893***	**.912*****
CNN Twitter GloVe 200	.991***	**1.00*****	.995***	**.945*****	.951*	**.948****	.923***	**.904*****	.902***
CNN Word2Vec 300	.992***	.999***	**.996*****	.922***	.949	.935**	.908***	.876***	.892***
LSTM GloVe 300	.987	.998***	.992***	.874	.936	.904	.833	.784	.808
LSTM Twitter GloVe 50	.987	.996**	.991	.828	.951	.885	.822	.750	.784
LSTM Twitter GloVe 100	.985	.997***	.991	.882	.892	.887	.770	.874***	.818
LSTM Twitter GloVe 200	.991***	.998***	.994***	.902*	.894	.898	.798	.865***	.830
LSTM Word2Vec 300	.987	.998***	.993***	.853	.920	.885	.837	.819	.828

Table 5: The mean precision (P), recall (R) and F1 scores for the three labels in the FluVacc dataset: "Related vs Unrelated, "Received vs. Not Received", and "Intent vs. No Intent". A p-value (resulting from the Wilcoxon signed rank test) between 0.05 and 0.01 is indicated by *, a p-value between 0.01 and 0.001 is indicated by **, and a p-value that is less than or equal to 0.001 is indicated by ***.

detection is 0.912, a nearly 10% absolute improvement over the Linear SVM (0.828) and the best LSTM model (0.828). The best CNN model for the Received label also outperformed the other methods by a large margin, e.g., by more than a 4% absolute improvement over the next best LSTM model. Moreover, unlike the FluTrack results, the Linear SVM model generally performs equivalent or better than the LSTM. For instance, the Linear SVM's F1 score for the received label is 0.01 (1%) better than the best performing LSTM 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 (e.g., "vaccine"). With regard to the overall results in Table 5, we find that the best pretrained word embeddings vary model-to-model. For instance, the best embeddings for the CNN are generally GloVe 100 and GloVe 300, while the best LSTM embeddings are GloVE 300 and Word2Vec 300.

FluVacc's fairness metrics are presented in Table 6. We find that the most unfair classifier varies between the neural network methods. The Linear SVM model generally makes fairer predictions than the most accurate CNN. For example, the CNN Twitter GloVe 200 has an Received F1 of 0.948 and FPED and FNED scores of 0.108 and 0.073, respectively. Yet, the Linear SVM only has FPED and FNED scores of 0.045 and 0.025 for the Received label with an F1 of 0.911. Finally, we find that the most unfair word embeddings vary, not just across models, but also within each model. As an example, depending on the metric (FPED or FNED), the most unfair embeddings for the CNN model are Twitter GloVe 50, GloVe 300, Twitter GloVe 200, and Twitter GloVe 100. Overall, based on our findings on FluVacc, we find that models trained to detect vaccine-related information on social media are biased.

	Related vs. Unrelated		Received vs. Not Received		Intent vs. No Intent	
	FPED	**FNED**	**FPED**	**FNED**	**FPED**	**FNED**
Linear SVM	.267	.002	.057	.025	.116	.020
CNN GloVe 300	.399***	.001	**.184****	.041*	.125*	.109***
CNN Twitter GloVe 50	.400***	.001	.046	.027	.185***	.104***
CNN Twitter GloVe 100	.356***	.001	.084*	.053**	.132**	**.120*****
CNN Twitter GloVe 200	.349***	.000	.118**	.073***	.159***	.108***
CNN Word2Vec 300	.385***	**.003***	.061	.041***	.119	.069***
LSTM GloVe 300	.247	.002	.127**	.041**	.128**	.028
LSTM Twitter GloVe 50	.319***	.001	.066	.029	.149***	.032
LSTM Twitter GloVe 100	.341***	**.003***	.077	**.076*****	**.245*****	.007
LSTM Twitter GloVe 200	.341***	.002	.063	.064***	.205***	.037*
LSTM Word2Vec 300	**.401*****	.038	.051	.051***	.110	.061***

Table 6: The FPED and FNED fairness results for the three labels in the FluVacc dataset: "Related vs Unrelated, "Received vs. Not Received", and "Intent vs. No Intent". A p-value (resulting from the Wilcoxon signed rank test) between 0.05 and 0.01 is indicated by *, a p-value between 0.01 and 0.001 is indicated by **, and a p-value that is less than or equal to 0.001 is indicated by ***.

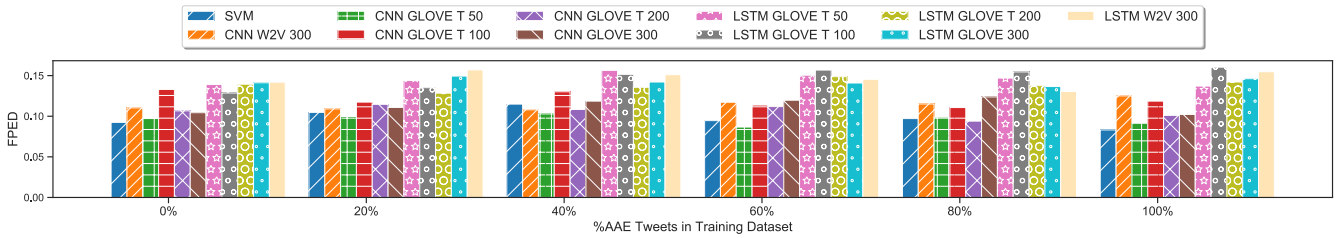
In Figure 3, we report the results of Experiment 2 (i.e., we use AAE tweets in the training set) for the FluVacc dataset. Again, we used the class with the largest number of evenly distributed AAE examples, the Intent vs No Intent class. For a majority of the models for the task of Intent classification, 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 affect 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. But, this is realistic, because, to the best of our knowledge, 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, we also observe similar patterns as found in Experiment 1's results reported in Table 6. Specifically, for the FPED scores, the SVM generally results in the smallest score.

Qualitative Error Analysis

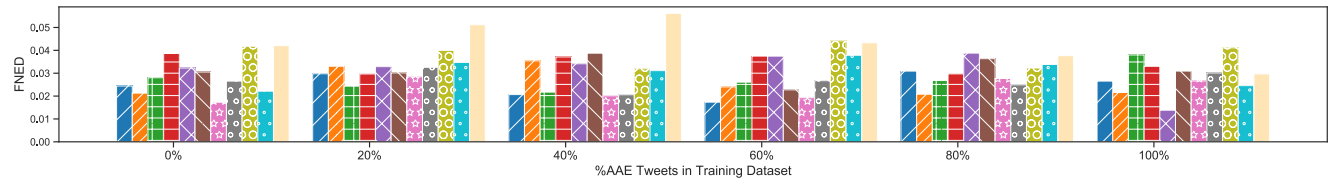
In this section, we provide a couple of AAE examples that resulted in incorrect predictions by the classifiers. We want to provide some insight into what aspects of AAE are potentially causing the problems.

We found many examples where the models had trouble classifying AAE text when they contain phonological variants of words. For instance, in the example from the FluVacc dataset

The examples have been slightly altered to preserve the privacy of users in the dataset.



(a) The FluVacc's FPED results for the Intent class.



(b) The FluVacc's FNED results for the Intent class.

Figure 3: FluVacc's experimental results using AAE tweets in both the training and test datasets. The FPED and FNED scores are plotted using different percentages of AAE tweets in the training 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 [59]—which means "I don't". The likely cause of the errors is the limited number of AAE tweets. However, since 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 [61]. 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*t y'all kiss my a*s already forced me to get the flu shot bullsh*t*"

was correctly classified by the SVM classifier as "Received". But, most of the neural network methods incorrectly classified it as "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 regularize the neural networks to reduce overfitting (e.g., with dropout or L2 regularization). Because it looks like the CNN and LSTM may rely on surface-level information, rather than real natural language understanding (NLU), it may be beneficial to explore novel methods of training neural networks by augmenting the data using adversarial learning [62].

DISCUSSION

Overall, the major finding of this paper is that machine learning methods for influenza-related tasks using social media data are biased. We do not simply detect bias, but we 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 which is not consistent with the general population in which the data is collected.

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. More importantly, we find that Linear SVMs *generally*, but not always, result in fairer predictions then the best neural network methods. 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, Linear SVMs provide a strong baseline while offering interpretability and fair results (as compared to the best neural network methods).

How can one model be more unfair than another? The interaction between data, features, and machine learning models is complex. As a toy example, let's assume we are using the Naive Bayes (NB) classifier with bag-of-words features. Furthermore, let's assume the task is to classify whether text is saying a disease is infectious (i.e., there are two classes infectious and noninfectious). Moreover, assume we have two groups in which we want to measure bias: Group 1 and Group 2. Group 1 always uses the word "noninfectious" (e.g., The new disease is noninfectious). Group 2 always uses "not infectious" (e.g., The new disease is not infectious). In this scenario, it is likely that the model will heavily correlate infectious with the infectious class, and noninfectious with the noninfectious class. Because the NB classifier assumes independence among the features (an inductive bias), it is unlikely the word "not" will highly correlate with the noninfectious class. Therefore, the classifier could be unfair towards Group 2. While this scenario is unlikely to be so extreme, it is likely to subtly appear. Moreover, many similar issues may compound resulting in more unfair predictions. For instance, other issues include models that are more prone to overfitting, which may result in unfair models because they will not generalize to novel word variants and syntactic patterns in low-resource dialects.

Finally, 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 policy makers, 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 a virus (i.e., the model has a large FNED score), 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 three 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], the classifier does a good job at identifying common AAE syntactic and phonetic constructions. Second, the number of AAE tweets is small. The effect caused by the small set of AAE tweets can be seen in the "Related vs. Unrelated" results on the FluVacc dataset (Table 6). With only 17 AAE unrelated tweets (see Table 1), the magnitude of the FPED and FNED scores are inflated. However, there is still evidence of bias in other classes with substantially more AAE data (e.g., Intent vs. No Intent which has more than 100 AAE tweets in each class). Third, we focus on dialect, which is not directly related to race or ethnicity. Because race and ethnicity is difficult to detect automatically, we believe it is best to perform controlled experiments where users are asked how they identify, rather than grouping them automatically. This approach (of asking rather than predicting) is also suggested for studies about gender [63].

CONCLUSION

In this paper, we used two influenza-related social media datasets to understand the potential biases in machine learning models trained on them. The major finding of this paper is that the resulting models are biased. Therefore, practitioners should be aware of the potential harms related to biased methods. We also establish that ngram-based Linear SVMs still provide a strong baseline while generally being fairer then the best neural network methods. As future work, it is important to expand this study to other tasks, machine learning models (e.g., BERT [24]), and demographic factors. Given the generalizability of the framework presented in this paper, it can easily be applied to other datasets. Beyond measuring bias, we believe it is also important to explore methods to reduce the bias of state-of-the-art machine learning approaches in biomedical NLP domains, which has already been explored in other application areas (e.g., abusive language) [15].

ACKNOWLEDGEMENTS

We would like to thank the reviewers for their insightful comments and help improving this paper.

FUNDING

This material is based upon work supported by the National Science Foundation under 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.

COMPETING INTERESTS

None

References

[1] Luca Ferretti, Chris Wymant, Michelle Kendall, Lele Zhao, Anel Nurtay, Lucie Abeler-Dörner, Michael Parker, David Bonsall, and Christophe Fraser. Quantifying sars-cov-2 transmission suggests epidemic control with digital contact tracing. *Science*, 2020.

[2] Iniobong Ekong, Emeka Chukwu, and Martha Chukwu. 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 and uHealth*, 8(4):e19139, 2020.

[3] Marcel Salathé, Clark C Freifeld, Sumiko R Mekaru, Anna F Tomasulo, and John S Brownstein. Influenza a (h7n9) and the importance of digital epidemiology. *The New England journal of medicine*, 369(5):401, 2013.

[4] Xiaolei Huang, Michael C Smith, Michael J Paul, Dmytro Ryzhkov, Sandra C Quinn, David A Broniatowski, and Mark Dredze. Examining patterns of influenza vaccination in social media. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[5] Alex Lamb, Michael J Paul, and Mark Dredze. 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*, pages 789–795, 2013.

[6] Courtney Corley, Armin R Mikler, Karan P Singh, and Diane J Cook. Monitoring influenza trends through mining social media. In *BIOCOMP*, pages 340–346, 2009.

[7] Courtney Corley, Diane Cook, Armin Mikler, and Karan Singh. Text and structural data mining of influenza mentions in web and social media. *International journal of environmental research and public health*, 7(2):596–615, 2010.

[8] Mauricio Santillana, André T Nguyen, Mark Dredze, Michael J Paul, Elaine O Nsoesie, and John S Brownstein. Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS computational biology*, 11(10):e1004513, 2015.

[9] Naheed Ahmed, Sandra C Quinn, Gregory R Hancock, Vicki S Freimuth, and Amelia Jamison. Social media use and influenza vaccine uptake among white and african american adults. *Vaccine*, 36(49):7556–7561, 2018.

[10] Kevin Fiscella. Commentary—anatomy of racial disparity in influenza vaccination. *Health services research*, 40(2):539, 2005.

[11] William K Bleser, Patricia Y Miranda, and Muriel Jean-Jacques. Racial/ethnic disparities in influenza vaccination of chronically-ill us adults: The mediating role of perceived discrimination in healthcare. *Medical care*, 54(6):570, 2016.

[12] Stephanie C Tse, Laura C Wyatt, Chau Trinh-Shevrin, and Simona C Kwon. Racial/ethnic differences in influenza and pneumococcal vaccination rates among older adults in new york city and los angeles and orange counties. *Preventing chronic disease*, 15:E159–E159, 2018.

[13] Kevin Fiscella, Richard Dressler, Sean Meldrum, and Kathleen Holt. Impact of influenza vaccination disparities on elderly mortality in the united states. *Preventive medicine*, 45(1):83–87, 2007.

[14] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Conference on AI, Ethics, and Society*, 2018.

[15] Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, 2018.

[16] Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *The World Wide Web Conference*, pages 49–59, 2019.

[17] Anthony Rios. FuzzE: Fuzzy fairness evaluation of offensive language classifiers on african-american english. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2020.

[18] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016.

[19] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning gender-neutral word embeddings. In *Proc. of EMNLP*, pages 4847–4853, 2018.

[20] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 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)*, pages 629–634, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[21] Anthony Rios, Reenam Joshi, and Hejin Shin. Quantifying 60 years of gender bias in biomedical research with word embeddings. In *Proceedings of the 2020 BioNLP Workshop, ACL*, 2020.

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

[23] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, 2019.

[25] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Quantifying social biases in contextual word representations. In *1st ACL Workshop on Gender Bias for Natural Language Processing*, 2019.

[26] Joel Escudé Font and Marta R Costa-jussà. Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154, 2019.

[27] Joel Escudé Font. Determining bias in machine translation with deep learning techniques. Master's thesis, Universitat Politècnica de Catalunya, 2019.

[28] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. IEEE, 2018.

[29] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. On the applicability of ml fairness notions. *arXiv preprint arXiv:2006.16745*, 2020.

[30] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy, August 2019. Association for Computational Linguistics.

[31] Maya Sen and Omar Wasow. Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science*, 19, 2016.

[32] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 1668–1678, 2019.

[33] Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 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*, pages 1119–1130, 2016.

[34] Su Lin Blodgett and Brendan O'Connor. Racial disparity in natural language processing: A case study of social media african-american english. *arXiv preprint arXiv:1707.00061*, 2017.

[35] Lisa J Green. *African American English: a linguistic introduction*. Cambridge University Press, 2002.

[36] Sarah Florini. Tweets, tweeps, and signifyin' communication and cultural performance on "black twitter". *Television & New Media*, 15(3):223–237, 2014.

[37] Jacob Eisenstein. Identifying regional dialects in on-line social media. *The Handbook of Dialectology*, pages 368–383, 2017.

[38] Nasser Zalmout and Nizar Habash. 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*, pages 1775–1786, 2019.

[39] Carl Pedersen. The obama dilemma: Confronting race in the twenty-first century. *Comparative American Studies An International Journal*, 10(2-3):128–141, 2012.

[40] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[41] Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. In *Proc. of EMNLP*, pages 11–21, 2018.

[42] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics.

[43] Ramez Kouzy, Joseph Abi Jaoude, Afif Kraitem, Molly B El Alam, Basil Karam, Elio Adib, Jabra Zarka, Cindy Traboulsi, Elie W Akl, and Khalil Baddour. Coronavirus goes viral: quantifying the covid-19 misinformation epidemic on twitter. *Cureus*, 12(3), 2020.

[44] Abeed Sarker, Maksim Belousov, Jasper Friedrichs, Kai Hakala, Svetlana Kiritchenko, Farrokh Mehryary, Sifei Han, Tung Tran, Anthony Rios, Ramakanth Kavuluru, 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. *Journal of the American Medical Informatics Association*, 25(10):1274–1283, 2018.

[45] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[46] Anthony Rios and Ramakanth Kavuluru. 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*, pages 258–267, 2015.

[47] Yifan Peng, Anthony Rios, Ramakanth Kavuluru, and Zhiyong Lu. Extracting chemical–protein relations with ensembles of svm and deep learning models. *Database*, 2018, 2018.

[48] Yifan Peng and Zhiyong Lu. Deep learning for extracting protein-protein interactions from biomedical literature. In *BioNLP 2017*, pages 29–38, 2017.

[49] Ye Zhang and Byron Wallace. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*, 2015.

[50] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[51] François Chollet et al. Keras. https://keras.io, 2015.

[52] Ramakanth Kavuluru, Anthony Rios, and Tung Tran. Extracting drug-drug interactions with word and character-level recurrent neural networks. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 5–12. IEEE, 2017.

[53] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[54] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[55] Pooya Mobadersany, Safoora Yousefi, Mohamed Amgad, David A Gutman, Jill S Barnholtz-Sloan, José E Velázquez Vega, Daniel J Brat, and Lee AD Cooper. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, 115(13):E2970–E2979, 2018.

[56] Taylor W Jones and Christopher S Hall. Semantic bleaching and the emergence of new pronouns in aave. In *LSA Annual Meeting Extended Abstracts*, volume 6, pages 10–1, 2015.

[57] Jacob Eisenstein. Phonological factors in social media writing. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 11–19, 2013.

[58] Anna Jørgensen, Dirk Hovy, and Anders Søgaard. Challenges of studying and processing dialects in social media. In *Proceedings of the workshop on noisy user-generated text*, pages 9–18, 2015.

[59] Taylor Jones. Toward a description of african american vernacular english dialect regions using "black twitter". *American Speech*, 90(4):403–440, 2015.

[60] Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 380–390, 2013.

[61] Shubhanshu Mishra, Sijun He, and Luca Belli. Assessing demographic bias in named entity recognition. *arXiv preprint arXiv:2008.03415*, 2020.

[62] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online, July 2020. Association for Computational Linguistics.

[63] Morgan Klaus Scheuerman, Jacob M. Paul, and Jed R. Brubaker. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019.