The Risk of Racial Bias while Tracking Virus-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*: 3988*

The Risk of Racial Bias while Tracking Virus-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. From the seasonal influenza virus to the development of COVID-19, 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 Vernacular 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 more fair predictions.

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

BACKGROUND AND SIGNIFICANCE

From seasonal outbreaks of the influenza virus to the advent of COVID-19, 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.

Current evidence suggests that there is a disproportionate incidence of disease and death among underrepresented minority groups. In the context of COVID-19, Garg et al. (2020) [10] show that among 580 patients hospitalized with lab-confirmed COVID-19, 45% of individuals for whom race or ethnicity data was available were white. This is in contrast to the surrounding community where 55% were white. Even worse, based on COVID-19 death data in New York City [11], Black/African American persons experience a death rate of 92.3 deaths per 100,000 population and Hispanic/Latino persons have a rate of 74.3. The rates are significantly higher than both white (45.2) and Asian (34.5) persons.

Health disparities are also present in Influenza cases. For example, there are significant racial disparities in influenza vaccinations [12, 13]. Tse et al. (2018) [14] report a nearly 10% difference in the influenza vaccination rate between non-Hispanic Black/African American adults over 50 than 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 [15].

In this paper, we look at the potential impact machine learning-based tools can have on health disparities—in the context 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 virus 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, virus-related datasets? Bias and fairness are abundant 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 [16–19] 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 [20–23]. For example, Bolukbasi et al. (2016) show that the word embedding for "man" is similar to "doctor", while "woman" is similar to "nurse" [20]. Garg et al. (2018) developed a technique to study 100 years of gender and racial bias using word embeddings [24]. Kurita et al. (2019) expanded on prior work [25] to generalize bias measurement metrics for word embedding to contextual word embeddings (e.g., BERT [26]) [27]. Machine translation systems have also been shown to exhibit biases [28, 29]. 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.

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, at the same time, have a high false negative rate for communities using low resource dialects like African American Vernacular English (AAVE). 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 [30]. Overall, if policy-related decisions are made using unfair models, then this can impact the governing bodies decision on where to intervene to stop the spread of a virus, as well as expanding potential economic harms. Therefore, it is important to understand how machine learning models will perform on the underrepresented populations we intend to apply them.

Finally, we summarize our three major contributions as follows: First, we study the performance differences between SAE and AAVE 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

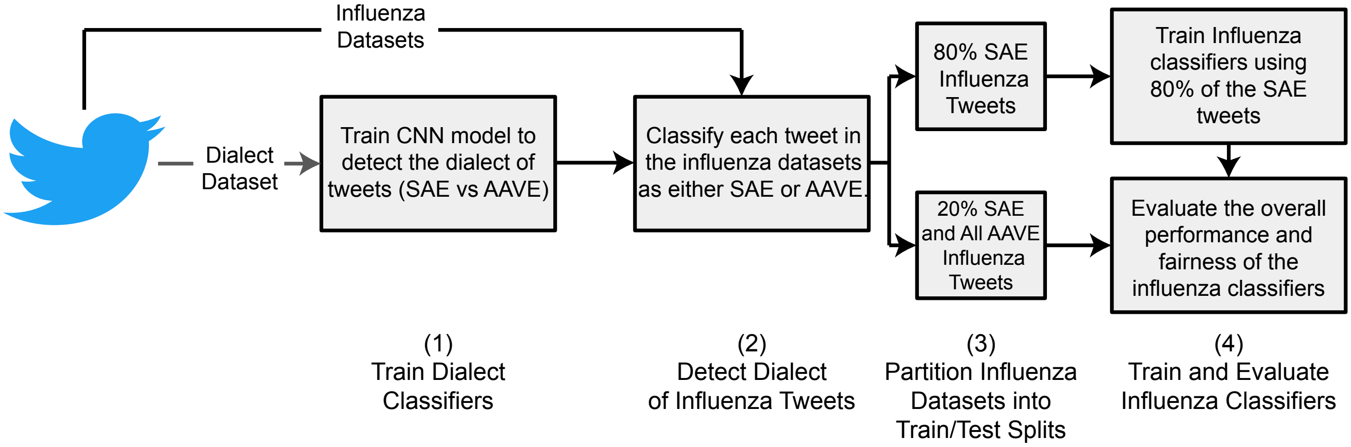
Influenza Datasets
Dialect Dataset
Train CNN model to detect the dialect of tweets (SAE vs AAVE)
Classify each tweet in the influenza datasets as either SAE or AAVE.
80% SAE Influenza Tweets
20% SAE and All AAVE Influenza Tweets
Train Influenza classifiers using 80% of the SAE tweets
Evaluate the overall performance and fairness of the influenza classifiers
(1) Train Dialect Classifiers
(2) Detect Dialect of Influenza Tweets
(3) Partition Influenza Datasets into Train/Test Splits
(4) Train and Evaluate Influenza Classifiers

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 (AAVE), (2) training multiple machine learning models on influenza datasets, (3) partitioning the influenza datasets to test fairness, and (4) the trained models are analyzed.

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 papers 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 AAVE). **(2.)** Next, the model is used to classify the dialect of each tweet in various influenza-related datasets. **(3.)** The influenza-related datasets are split into a 80% training/development split and 20% test split. Because of the limited number of AAVE tweets in our influenza datasets, all of the AAVE tweets are reserved for testing. **(4.)** Finally, we train and evaluate various models (i.e., neural networks and linear models) on multiple influenza datasets to understand the biases in the models and its relationship with the 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 and 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 [31], FluTrack [5], and FluVacc [4]. Dialect is used for quantitative evaluation of fairness by classifying tweets as SAE or AAVE. 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, AAVE tweets appear infrequently throughout every dataset used in our experiments. Because of the imbalance between SAE and AAVE, all AAVE tweets are only used

Class	Total	SAE	AAVE
Related	2436	2334	102
Not Related	1900	1830	70
Awareness	1294	1242	52
Infection	1359	1303	56
Self	1392	1338	54
Other	664	638	28

(a) FluTrack Dataset Summary

Class	Total	SAE	AAVE
Vaccine Related	9517	9258	259
Not Related	483	466	17
Intent	3148	3027	121
No Intent	6365	6228	137
Received	3097	2981	116
Not Received	743	708	35

(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.

in the testing dataset. We describe each dataset in detail below:

Blodgett et al. (2016) [31] (Dialect). Dialect consists of more than 59 million tweets. Each tweet was annotated with various lingustic styles (e.g., SAE and AAVE). 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) [32] and Rios (2020) [19], we limit our study to all tweets annotated with AAVE and SAE with a confidence of at least 80%. This resulted in 1.6 million AAVE 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) [33] to detect the dialect of each tweet. The CNN model is used in Step 2 or our data anlaysis process, as shown in Figure 1.

Lamb et al. [5] (FluTrack). FluTrack consists of 11,990 tweets collected from years 2009 through 2012.[*] This dataset can be used for many disease surveillance tasks [34], such as estimating influenza infection rates. Intuitively, standard methodologies to estimate infection rates can take weeks to generate. Thus, social media can potentially be used for quick and accurate estimates. Each tweet is annotated with up to three labels (this is a multi-label classification task, not mulit-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). Intuitively, many tweets may simply raise awareness, instead of discuss an infection. Meaning, tweets that discuss beliefs related to influenza infections or preventative influenza measures are not useful for disease surveillance. Furthermore, if a tweet is influenza-related, it is also labeled as Self or Other depending on whether it is about the user (Self) or it is about another person (Other).

[*] 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.

Huang et al. (2017) [4] (FluVac). Social media is not only useful for traditional disease surveillance tasks (e.g., infection rate estimation). 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 [35]. The FluVacc dataset is from Huang et al. (2017) [4] and contains ten thousand annotated tweets. Each tweet is categorized with 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 but receiving a vaccine and intent to receive it again.

Dialect Detection with Convolutional Neural Networks

As shown in Step 2 of Figure 1, we train CNN model [33] to predict the dialect of individual tweets using the Dialect dataset. Following Rios (2020) [19], we use the CNN architecture from Kim (2014) [33]. The CNN model is trained with 900 filters that spans 3,4, and 5 words. The final CNN has an F1 of 0.87. 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 Long Short-Term Memory Networks (LSTM). 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) [36]. We trained a Linear SVM using term frequency inverse document frequency-weighting (TF-IDF) of unigrams and bigrams and L2 regularization. We limited the number of features to the 500,000 most frequent ngrams. 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.

CNN. For the CNN model, we use the architecture from Kim (2014) [33]. For each task, the CNN models were trained with 512 filters for each span width of 3, 4, and 5 words. A max-over-time pooling operation was applied to the output of each span of words and concatenated together. The concatenated layer was fed forward through a fully connected layer. The model was trained with the Adam optimizer [37] for 30 epochs. The best epoch were chosen based on a held-out validation dataset.

LSTM. Instead of a standard LSTM model, we trained a bidirectional LSTM model, which has been shown to perform well across a wide variety of biomedical NLP tasks [38, 39]. At each time step 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 fully connected output layer in order to accurately make the prediction. The Bi-LSTM model is trained with a hidden state size of 512 for each direction. The model was trained with the Adam optimizer [37] for 30 epochs. The best epoch was chosen based on a held-out validation dataset.

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 [33]. In this paper, we also explore the overall performance of the CNN and Bi-LSTM models trained with different pretrained embeddings. We evaluate several variations of GLOVE and Word2Vec [40,41]. 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 AAVE [42]. 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 AAVE test examples independently.

[†] 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

	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 2: 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”

The FPR and FNR scores for each group are combined using the False Positive Equality Difference (FPED) and False Negative Equality Difference (FNED) [16]. 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 $T = \{\text{AAVE}, \text{SAE}\}$. FPR and FNR represent the overall false positive and false negative rates, respectively. FPR_t and FNR_t represent the group-specific (i.e., AAVE or SAE) false positive and false negative rates. 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.

RESULTS

For evaluation, we performed bootstrap testing. 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. The results reported in this section are the average across both the data splits and multiple runs.

FluTrack Experiments

The overall performance results on the FluTrack [5] dataset are presented in Table 2. 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 FPED and FNED fairness metrics for each FluTrack label are reported in Table 3. **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 more fair 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 3, there does not seem to be a single embedding type that 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**.

Finally, based on our findings, we find that models trained to detect infections on social media are biased. If these models were potentially used for virus surveillance, users of diverse English dialects would be adversely impacted, potentially increasing health disparities that already exist.

FluVacc Results

The overall performance results on the FluVacc [4] dataset are presented in Table 4. The results on FluVacc are similar to the findings on FluTrack. Specifically, we find that the CNN outperforms both the Linear SVC and LSTM 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 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 recieved label is 0.01 (1%) better than the best performing LSTM model. For the Related label, while the

	Related vs. Unrelated		Awareness vs. Infection		Self vs. Other	
	FPED	**FNED**	**FPED**	**FNED**	**FPED**	**FNED**
Linear SVM	.017	.020	.090	.028	.079	.002
CNN GloVe 300	.072	.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	.002
LSTM Twitter GloVe 50	.098	.022	.136	.174	.010	.006
LSTM Twitter GloVe 100	.081	.025	.006	.142	.043	.008
LSTM Twitter GloVe 200	.069	.019	.052	.101	.000	.031
LSTM Word2Vec 300	.102	.007	.156	.153	.052	.012

Table 3: The FPED and FNED fairness results for the three labels in the FluTrack dataset: “Related vs Unrelated, “Awareness vs. Infection”, and “Self vs. Other”.

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 4, 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 5. We find that the most unfair classifier varies between the neural network methods. The Linear SVM model generally makes more fair 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 0.015 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.

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

	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	.882	.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 4: 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".

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 results in more fair 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).

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.

	Related vs. Unrelated		**Received vs. Not Received**		**Intent vs. No Intent**	
	FPED	**FNED**	**FPED**	**FNED**	**FPED**	**FNED**
Linear SVM	.267	.002	.045	.015	.116	.009
CNN GloVe 300	.399	.001	**.180**	.041	.125	.109
CNN Twitter GloVe 50	.400	.001	.035	.027	.185	.104
CNN Twitter GloVe 100	.356	.001	.068	.053	.132	**.120**
CNN Twitter GloVe 200	.349	.000	.108	.073	.159	.108
CNN Word2Vec 300	.385	**.003**	.049	.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	.058	.064	.205	.037
LSTM Word2Vec	**.401**	.002	.001	.051	.110	.061

Table 5: 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”.

Limitations to this study

There are three limitations to this study. First, we rely on a “SAE vs. AAVE” 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 [19], the classifier does a good job at identifying common AAVE syntactic and phonetic constructions. Second, the number of AAVE tweets is small. The effect caused by the small set of AAVE tweets can be seen in the “Related vs. Unrelated” results on the FluVacc dataset (Table 5). With only 17 AAVE 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 AAVE data (e.g., Intent vs. No Intent which has more than 100 AAVE 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 [43].

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 more fair 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 [26]), and demographic factors. Given the generalizability of the framework presented in this paper, it an 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) [17].

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] Shikha Garg. Hospitalization rates and characteristics of patients hospitalized with laboratory-confirmed coronavirus disease 2019—covid-net, 14 states, march 1–30, 2020. *MMWR. Morbidity and Mortality Weekly Report*, 69, 2020.

[11] NYC Health. Covid-19 deaths, 2020. data retrieved from, https://www1.nyc.gov/assets/doh/downloads/pdf/imm/covid-19-deaths-race-ethnicity-04162020-1.pdf.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

[27] 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.

[28] 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.

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

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

[31] 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.

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

[33] 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.

[34] Nigel Collier, Ai Kawazoe, Lihua Jin, Mika Shigematsu, Dinh Dien, Roberto A Barrero, Koichi Takeuchi, and Asanee Kawtrakul. A multilingual ontology for infectious disease surveillance: rationale, design and challenges. *Language resources and evaluation*, 40(3-4):405, 2006.

[35] 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.

[36] 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.

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

[38] 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.

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

[40] 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.

[41] 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.

[42] 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.

[43] 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.