Measuring Geographic Performance Disparities of Offensive Language Classifiers

Brandon Lwowski[1], **Paul Rad,**[1,2] **and Anthony Rios**[1]
[1]Department of Information Systems and Cyber Security
[2]Department of Computer Science
University of Texas at San Antonio
{brandon.lwowski, peyman.najafirad, anthony.rios}@utsa

Abstract

Text classifiers are applied at scale in the form of one-size-fits-all solutions. Nevertheless, many studies show that classifiers are biased regarding different languages and dialects. When measuring and discovering these biases, some gaps present themselves and should be addressed. First, "Does language, dialect, and topical content vary across geographical regions?" and secondly "If there are differences across the regions, do they impact model performance?". We introduce a novel dataset called GeoOLID with more than 14 thousand examples across 15 geographically and demographically diverse cities to address these questions. We perform a comprehensive analysis of geographical-related content and their impact on performance disparities of offensive language detection models. Overall, we find that current models do not generalize across locations. Likewise, we show that while offensive language models produce false positives on African American English, model performance is not correlated with each city's minority population proportions. **Warning: This paper contains offensive language.**

1 Introduction

Many tasks revolving around text classification of social network data have been introduced including, but not limited to tracking viruses (Lamb et al., 2013; Corley et al., 2009, 2010; Santillana et al., 2015; Ahmed et al., 2018; Lwowski and Najafirad, 2020), providing help for (natural) disasters (Neubig et al., 2011; Castillo, 2016; Reuter and Kaufhold, 2018), detecting misinformation (Oshikawa et al., 2020), and identifying cyberbullying (Xu et al., 2012). Overall, text classifiers have been shown to be "accurate" across a wide range of applications. As deep learning models and packages have made substantial progress for the field of natural language processing (NLP), NLP models have become more accessible to the general public. Hence, models are being deployed in a production environment and run at scale at a growing pace. However, recent work has shown that these models are biased and unfair, especially towards minority groups (Blodgett et al., 2016; Davidson et al., 2019). In this paper, we expand on prior work by analyzing how model performance can fluctuate due do geographically-caused differences in language and topical content that exists in the context of offensive language detection.

Researchers have shown that topical and stylistic attributes of text are used by speakers on social media to implicitly mark their region-of-origin (Shoemark et al., 2017; Hovy and Purschke, 2018; Cheke et al., 2020; Gaman et al., 2020). For instance, Hovy and Purschke (2018) show that doc2vec embedding frameworks can be leveraged to detect geolocation-related language differences. Hovy et al. (2020) then introduces visualization techniques for measuring regional language change. Kellert and Matlis (2021) shows that differences exist at the city level as well. Hence, prior work has generally focused on incorporating or identifying regional aspects of language data to improve performance in machine translation (Östling and Tiedemann, 2017) or geolocation prediction and clustering (Hovy and Purschke, 2018).

For particular downstream tasks, recent work in understanding performance disparities has found differences across various languages (Gerz et al., 2018) (e.g., Finish vs. Korean) and dialects (Davidson et al., 2017; Sap et al., 2019)—such as African American English (AAE). Likewise, Davidson et al. (2019) and Sap et al. (2019) show that abusive and hate speech-related language classifiers are biased against AAE-like text and machine learning models can learn these biases when certain populations are not being represented, making the data unbalanced. These results have been shown to extend into other text classifications tasks, for example, Lwowski and Rios (2021) show that influenza detection models are also biased against AAE-like text. Similarly, Hovy and Søgaard (2015) find that part-of-speech

Proceedings of the 29th International Conference on Computational Linguistics, pages 6600–6616
October 12–17, 2022.

tagging performance correlates with age.

While there has been a substantial amount of research understanding, identifying, and measuring performance disparities across languages and dialects, to the best of our knowledge, there has been no prior work on measuring the performance of NLP classifiers across different geographic regions. Specifically, prior work has not measured how geographical variations in language and topical content—or stance towards certain topics—impacts the performance of offensive language classifiers. Complex interactions between topical content and style can impact model performance.

Even in the context of AAE-related studies (Sap et al., 2019), AAE is not spoken the same across different regions of the United States. There have been multiple studies in diversity, equity, and inclusion arguing against treating African Americans as a monolithic group of people (Tadjiogueu, 2014; Erving and Smith, 2021). Moreover, certain features of AAE only appear within specific regions of the US (Jones, 2015). Likewise, geographic factors have been known to impact social behaviors such as voting turnout (Zingher and Moore, 2019) and general health disparities (Thomas et al., 2014). Hence, these geographic factors can impact both *how* people write and *what* people write about on social media. Hence, to start addressing these issues, this paper proposes an initial study looking at how individual offensive language model's performance can vary geographically for the task of detecting offensive language due to the stylistic and topical differences in language.

Overall, to better understand the implications of geographical performance disparities offensive language models, we make three contributions:

(1.) To the best of our knowledge, we perform the first analysis of geographical performance variation of offensive language classification models, producing novel insights and a discussion of important avenues of future research.

(2.) We introduce a novel labeled offensive language dataset called GeoOLID [1] with more than 14 thousand tweets across 15 geographically and demographically diverse cities in the United States.

(3.) We produce a comprehensive manual error analysis, grounding some performance disparities to stance and topics.

[1] https://github.com/AnthonyMRios/Geographic-Performance-Disparities

2 Language Variation

To the best of our knowledge, the impact geographical variation in language style and topical content has not yet been studied in the context of offensive language detection to the best of our knowledge. Language variation is an important area of research for the NLP community. While there has been disagreement about whether morphology matters, Park et al. (2021) has shown that incorporating information that can model morphological differences is important in improving model performance. Prior work has generally focused on either developing methods to identify language features within text or use various language features to improve model performance. Early work by Bamman et al. (2014) showed that embeddings can capture geographically situated language, while Doyle (2014) explored ways to quantify regional differences against a background distribution. Recently, VarDial has hosted an annual competition to identify various dialects of different languages (e.g., German and Romanian) as well as geolocations (Gaman et al., 2020).

Cheke et al. (2020) use topic distributions to show that different topics can provide signal to determine where the text originated from. For the same shared task, Scherrer and Ljubešić (2021) show that combining modern NLP architectures like BERT with a double regression model can also provide success in determining the latitude and longitude points of the location for the given text. The results of this shared task highlights the fact that topical and lexical differences exist based on the location a tweet was written. Other work around regional variation of language (Hovy and Purschke, 2018; Hovy et al., 2020; Kellert and Matlis, 2021) further prove that these differences in dialect and lexical patterns are significant across geographies.

3 Performance Disparities

Performance disparities across languages and dialects recently have received attention in NLP. For example, recent research shows that performance drops in text classification models across different sub-populations such as gender, race, and minority dialects (Dixon et al., 2018; Park et al., 2018; Badjatiya et al., 2019; Rios, 2020; Lwowski and Rios, 2021; Mozafari et al., 2020). Sap et al. (2019) measure the bias of offensive language detection models on AAE. Likewise, Park et al. (2018) measure gender bias of abusive language detection models

	Non Offensive	Offensive	Total	MDE
OLID	9,460	4,640	14,100	.014
	Filtered/Unfiltered GeoOLID Dataset			
	Non Offensive	**Offensive**	**Total**	**MDE**
Unfiltered GeoOLID	—	—	5,013,474	—
GeoOLID	9,259	4,831	14,090	
City Name	**Non Offensive**	**Offensive**	**Total**	**MDE**
Baltimore, MD	630	277	907	.054
Chicago, IL	676	326	1002	.052
Columbus, OH	616	301	917	.054
Detroit, MI	549	367	916	.053
El Paso, TX	502	404	906	.055
Houston, TX	635	297	932	.054
Indianapolis, IN	600	307	907	.055
Los Angeles, CA	660	298	958	.053
Memphis, TN	564	368	932	.054
Miami, FL	726	216	942	.054
New Orleans, LA	607	325	932	.054
New York, NY	717	265	982	.053
Philadelphia, PA	629	337	966	.054
Phoenix, AZ	577	355	932	.054
San Antonio, TX	572	387	959	.053

Table 1: Dataset Statistics.

and evaluate various methods such as word embedding debiasing and data augmentation to improve biased methods. Davidson et al. (2019) shows that there is racial and ethnic bias when identifying hate speech online and show that tweets in the black-aligned corpus are more likely to get assigned as hate speech. Overall, performance disparities have been observed across a wide array of NLP tasks such as detecting virus-related text (Lwowski and Rios, 2021), coreference resolution (Zhao et al., 2018), named entity recognition (Mehrabi et al., 2020), and machine translation (Escudé Font and Costa-jussà, 2019).

Overall, the major gap in prior work investigating language variation is that there has not been any studies evaluating the impact regional language has on the performance of downstream tasks, particularly offensive language detection. Hence, we measure performance disparities across geographical regions for the task of detecting offensive language. Furthermore, many groups that are studied are "monolithic", such as male vs. female (using an unrealistic assumption of binary gender (Rios et al., 2020)), or AAE which is not universally spoken in the same way within different cities in the US. For example, Jones (2015) show that many well-known AAE patterns (e.g., sholl, an nonstandard spelling of "sure") do not appear uniformly across the US. Likewise, the discussion topics can also change regionally. Hence, if an offensive language detection model performs poorly on one set of AAE patterns or topics that only appear in a particular region, it can impact that location much more than

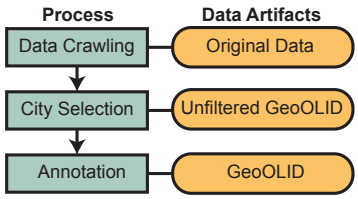

Figure 1: Data collection/annotation. Data collection steps are green and produced datasets are in orange.

others. Hence, we believe that fine-grain regional analysis is a better future avenue to understand the real-world impact of NLP models.

4 Data Collection and Annotation

In this section, we describe the two major datasets used in our experiments: the Offensive Language Identification Dataset (OLID) (Zampieri et al., 2019) and our newly constructed Geographical Diverse Offensive Language Identification Dataset (GeoOLID). A complete summary of the datasets can be found in Table 1. Furthermore, we provide a summary of the data collection and annotation pipeline in Figure 1. Intuitively, we have three main steps: Data Crawling, City Selection, and Data Annotation. We save the data after each step to be used throughout parts of our analysis. We describe the OLID dataset and each step below.

OLID. The OLID dataset introduced by Zampieri et al. (2019) contains 14,100 tweets labeled to identify different levels of offensiveness including, but not limited to, Not Offensive, Offensive, Targeted Offense, and Not Targeted Offense. Furthermore, Targeted Offenses are sub-categorized as targeting an individual, group, or other. For this study, we use the first level: Not Offensive (9,460 Total) and Offensive (4,640 Total).

Step 1: Data Crawling (Original Data). In addition to the OLID dataset, we introduce a new offensive language dataset using tweets collected since the start of the Covid-19 pandemic. The data set was crawled by Qazi et al. (2020) and Lamsal (2021), collecting more than 524 million multilingual tweets across 218 countries and 47,000 cities between the dates of February 1, 2020 and May 1, 2020. The data collection started on February 1, 2020 using trending hashtags such as #covid19, #coronavirus, #covid_19. See Qazi et al. (2020) for complete details. Given the large amount of politically divisive discourse, racist remarks, and social impact of Covid-19, the collection provides a unique testbed to understand geographic model

variation. Particularly, where researchers are exploring analyzing geospatial patterns of Covid-related content on Twitter (Stephens, 2020). If the models perform differently across locations, the it is difficult to interpret the results. We refer to this complete Covid-19 data as "Original Data".

Step 2: City Selection (Unfiltered GeoOLID). To measure the performance difference across varying locations, we decided on 15 cities based on multiple facets, data availability, geographic diversity, and demographic diversity. In deciding which cities to use for our study we first selected cities from different parts of the United States (North, South, East, West). Next we wanted cities that varied in size and were also demographically different. In table 3, the total populations (reported in the thousands) of the selected cities varies, ranging from around 400,000 to almost 9 million.

We also wanted cities that varied demographically, particularly with regard to African American and Hispanic/Latino population proportions.[2] In Table 9, cities like Baltimore, Memphis, New Orleans, and Detroit were chosen due to the high proportion of African Americans populations while, Indianapolis and Columbus had high proportions of White Non-Hispanic residents. El Paso, San Antonio and Phoenix have a close proximity to the Mexico boarder and higher percentage of Latino and Hispanic residents, which is very different from Columbus, having a smaller number of African American and Hispanic residents. In addition, we selected cities where we knew residents could use very distinct accents and phonics like New York and New Orleans. Overall, by selecting the 15 cities in Table 1, we created a diverse dataset with multiple ethnicities, language styles, and topical differences. We refer to this unlabeled dataset as "Unfiltered GeoOLID". The basic stats of this dataset are available in Table 1 in the row titled "Unfiltered GeoOLID".

Step 3: GeoOLID. Similar to prior work, we need to sample a large number of offensive and non-offensive tweets from Unfiltered GeoOLID (Zampieri et al., 2019). Hence, we filter Unfiltered GeoOLID using the following lexicons and keyword filters: the badword lexicon (von Ahn, 2009), hatebase lexicon (Davidson et al., 2017), offensive-related phrases used for the original OLID dataset (Zampieri et al., 2019) ("you are", "she is", "he is", "conservatives", "liberals", "MAGA", and "antifa"), and additional Covid-specific phrases we found to be correlated with potential discrimination in the dataset ("chinese", "china", "asia", "asian", "wuhan"). Along with the aforementioned filters, we randomly sampled a subset of tweets for annotation. The final counts of each city can be found in Table 1. This dataset is referred to as "GeoOLID".

Annotation. Overall, we performed multiple rounds of annotation until a quality dataset was created. First, in order to provide accurate labels for this study, samples of tweets were assigned to three graduate students to be labeled as "offensive" or "not-offensive" using the base guidelines provided by Zampieri et al. (2019) for the the OLID dataset. A total of 20 students were recruited and given a stipend of $100 for their time and effort. Several meetings were set up before labeling started to answer questions and address implications. We use the Offensive definition provided by Zampieri et al. (2019) is defined as *tweets containing any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct. This includes insults, threats, and posts containing profane language or swear words.*

Following general annotation recommendations for NLP (Pustejovsky and Stubbs, 2012), the annotation process was completed in three stages to increase the reliability of the labels across geographic regions. First, before assigning tweets, we assured every tweet was assigned to three graduate students for annotation, providing us with three independent labels for each tweet. We then calculated the agreement between annotators, resulting in a Fleiss Kappa of .47, indicating moderate agreement.

Second, we (the authors) of the paper manually—and independently—adjudicated (i.e., re-annotated) the labels of each student, correcting miss-annotated tweets that were not agreed on by all three annotators. Common issues found during the process were labels of "Not Offensive" for tweets with ad-hoc mentions of the "Wuhan Virus" and offensive content found in the hashtag. Specifically, based on the work by Dubey (2020), we decided that mentions of "Wuhan Virus" and other related terms like "China Flu" and "Kung Flu" were deemed offensive as it fit into the category of an targeted offense, which can be veiled or direct. The second round of agreement scores increased to .83

[2]We choose these groups because they align with classes in the Blodgett et al. (2016) dialect classifier.

representing "almost perfect agreement," (Landis and Koch, 1977).

To further ensure annotation quality, the authors went through the tweets once again discussing and correcting any final disagreements among the second round adjudications, forming the final dataset described in Table 1. After collecting and adjudicating the responses, the total number of Offensive tweets were 4,831 compared to 9,259 Not Offensive. We also report Minimum Detectable Effect (MDE) (Card et al., 2020) for Accuracy in Table 1. Specifically, use the Binomial Power Test, which assumes that samples are unpaired, i.e., the new model and baseline evaluation samples are drawn from the same data distribution but are not necessarily the same samples. The MDE numbers assume an accuracy of .75, which results in a significant difference between two models being around .05. We plot more potential MDE scores for different baseline numbers in the Appendix, Figure 5.

5 Experiments

In order to address and test whether performance disparities exist across geographic regions for offensive language classifiers, we ran multiple experiments. We analyzed performance across the 15 cities in the GeoOLID dataset. In the following subsections, we provide the details of our experiments and provide evidence supporting that our GeoOLID dataset is representative of the Unfiltered GeoOLID dataset and that offensive language classifier performance can vary by geolocation. In the final subsection, we explore the performance and language similarities across different geolocations that have similar demographics.

5.1 Data Representation Evaluation

In this section, we aim to measure how well the GeoOLID dataset matches the Unfiltered GeoOLID data from each city. Specifically, we want to ensure that patterns found in the unfiltered data are still present within our annotated GeoOLID sample. If patterns in the GeoOLID dataset are not in Unfiltered GeoOLID, it is hard to argue that the errors are location-specific. They could simply be caused by our data filtering strategy.

Methods. To measure how representative our sample is, we train a location prediction model. Given a tweet, the goal of the model is to predict the city in which the text was posted. To train the model we use two sets of features: Content Features and Stylistics Features. The content features are made up of the top 5000 unigrams in the Unfiltered GeoOLID dataset. It is also important to note that all of the GeoOLID tweets are removed from the Unfiltered GeoOLID dataset before processing.

	F1	Acc.
Stratified	.059	.056
Uniform	.062	.062
Prior	.008	.068
BoW	.430	.380
POS	.410	.356
Dialect	.374	.366
POS + Dialect	.419	.357
BoW + Dialect	**.436**	**.381**
BoW + POS + Dialect	.431	.370

Table 2: Location prediction. The Accuracy, Macro Precision, Macro Recall, and Macro F1 reported are the results when trained on a sample of the Unfiltered GeoOLID and predicted on the labeled GeoOLID dataset.

We also explore two sets of style Features: Part-of-Speech and Dialect Features. Specifically, we use unigram, bigram, trigram POS features. Moreover, the dialect features are the probabilities returned from the dialect inference tool from Blodgett et al. (2016). Given a tweet, the tool outputs the proportion of African-American, Hispanic, Asian, and White topics.

Finally, we train a Random Forest classifier on the Unfiltered GeoOLID dataset and the results are reported using the labeled GeoOLID dataset as the test set. Hyperparameters are optimized using 10-fold cross-validation on the training data. Because of the large size of the Unfiltered GeoOLID dataset, we sample a random subset of 35k examples from the Unfiltered GeoOLID dataset to reduce the training cost. The goal is not to achieve the most accurate predictions, but to simply see if we can predict location much better than random. If a small completely random sample shows this, that is better then requiring all of the data. We also compare the results to three random sampling methods to measure the difference between random guessing and the trained model: Stratified, Uniform, and Prior. Stratified makes random predictions based on the distribution of the cities in the training data, Uniform predicts cities with equal proportions, and Prior always predicts the most frequent city.

Results. The results of the experiments are reported in Table 2. Using content and style features, we were able to predict the location of a tweet

	AAS	HLS	Tot.	AA	H/L
Baltimore	.168	.193	585	338 (57.7%)	45 (7.8%)
Chicago	.147	.204	2,450	801 (32.7%)	819 (33.4%)
Columbus	.146	.201	905	259 (28.6%)	70 (7.7%)
Detroit	.196	.214	639	496 (77.7%)	51 (8.0%)
El Paso	.158	.227	678	25 (3.7%)	551 (81.2%)
Houston	.161	.205	2,304	520 (22.6%)	1,013 (44.0%)
Indianapolis	.151	.194	887	248 (28.0%)	116 (13.1%)
Los Angeles	.144	.204	3,898	336 (8.6%)	1,829 (47.0%)
Memphis	.209	.220	633	389 (61.6%)	62 (9.8%)
Miami	.140	.175	442	57 (12.9%)	310 (7.0%)
New Orleans	.182	.197	383	208 (54.2%)	31 (8.0%)
New York City	.126	.182	8,804	1,943 (22.1%)	2,490 (28.3%)
Philadelphia	.157	.204	887	248 (27.9%)	116 (13.1%)
Phoenix	.144	.208	1,608	125 (7.8%)	661 (41.1%)
San Antonio	.175	.222	1,434	102 (7.2%)	916 (63.9%)
AA PCC	.565 (p value: .028)				
H/L PCC	.167 (p value: .55)				

Table 3: Pearson Correlation Coefficient (PCC) between the AAS and HLS and city populations. Populations reported in thousands and percentages are in parenthesis.

more than 38% of the time, an increase of almost 140% in accuracy than the best random baseline, suggesting that both content and style features are predictive of the location a tweet is made. Likewise, using the POS and dialect features alone, the model achieves an accuracy of more than 35%, substantially higher than the random baselines. Given that there are only four dialect features, this is indicative that the group information detected by the Blodgett et al. (2016) is informative. Similarly, the POS results are also high, indicating that there are unique combinations of POS patterns that appear in each location. Overall, the findings show that our subsample (GeoOLID) is representative of patterns found in the Unfiltered GeoOLID dataset.

Discussion. Blodgett et al. (2016) show that the assumption that cities with large African American populations will have more text classified as AAE. Hence, as a simple robustness test, we use the tool provided by Blodgett et al. (2016) to correlate it with the demographic information of each city in our labeled GeoOLID dataset. Specifically, using the 2020 US Census data, we calculate the proportion of "Black or African American alone" (AA) and "Hispanic or Latino" (H/L) residents for each city. We also calculate the average African-American (AAS) and Hispanic (HS) scores for each city using the tool from Blodgett et al. (2016). Finally, we calculate the Pearson Correlation Coefficient (PCC) AA and AAS (and H/L and HS). The goal is to show that the findings found in Blodgett et al. (2016) hold on our GeoOLID dataset, i.e., their tool's scores correlate with minority populations. If they do, this provides further evidence that our dataset is representative of each location.

This correlation can be seen in Table 3. Overall we find that there is significant correlation .565 (p value: .028) between the two variables. We find that cities like Baltimore, New Orleans and Detroit are more likely to have more AAE tweets (in the labeled GeoOLID dataset) then cities like Miami, Columbus, and New York. For the Hispanic group we also find a positive correlation but the finding is not significant. We also manually analyzed the dataset and found other features indicative of a relationship between demographics of the city and language use. For example, we found Spanish curse words appearing in text in cities with higher Hispanic populations in our dataset, e.g., "Nationwide shutdown! pinché Cabron" is an slightly modified tweet that was tagged in Phoenix, AZ.

5.2 Data Variation and Model Performance

Next, we measure how much offensive language detection performance can vary location-to-location. Given the same model is applied to every city, ideally, we would have similar performance universally. However, if we see large variation in performance metrics and if the errors are caused by patterns also represented in the Unfiltered GeoOLID dataset, this is indicative of geographic performance disparities.

Methods. We train five different machine learning algorithms: Linear Support Vector Machine (Linear SVM), Long Short Term Memory (LSTM), Bidirectional LSTM (BiLSTM), Convolutional Neural Networks (CNN), and a Bidirectional Encoder Representations from Transformers (BERT). Each model is trained to classify Offensive and Non Offensive tweets using the OLID dataset.One thing to note is for the BiLSTM, CNN and LSTM, we also measure the performance of the model across multiple word embeddings. Specifically, each deep learning model is trained using different variations of Glove, Google Word2Vec and Fasttext word embedding (See the Appendix, Table 7, for a complete listing of the evaluated embeddings).

For evaluation, We train multiple models on the OLID dataset using a 5-fold shuffle-split cross-validation procedure. Specifically, a model is trained on each training split of the OLID dataset, then it is applied to the GeoOLID dataset to calculate each city's model performance. A 10% portion of the OLID training split for each fold is used for

	Bal	Chi	Col	Det	ElP	Hou	Ind	LA	Mem	Mia	NO	NY	Phi	Pho	SA	AVG
Stratified	.279	.328	.306	.367	.431	.348	.364	.303	.363	.253	.322	.296	.376	.394	.410	.343
Uniform	.362	.443	.398	.458	.453	.381	.390	.368	.420	.302	.412	.342	.412	.444	.456	.403
Linear SVM	.661	.615	.650	.714	.658	.568	.611	.643	.702	.660	.708	.625	.680	.613	.680	.653
BiLSTM	.678	.623	.651	.725	.660	.591	.643	.642	.720	.666	.694	.624	.705	.637	.669	.662
CNN	**.720**	**.662**	**.684**	**.745**	**.688**	**.611**	**.674**	**.670**	**.745**	**.701**	**.736**	**.663**	**.743**	**.657**	**.703**	**.694**
LSTM	.653	.614	.633	.709	.638	.570	.624	.620	.701	.661	.680	.600	.686	.615	.650	.643
BERT	. 601	.629	.641	.684	.661	.602	.621	.642	.651	.614	.635	.593	.668	.607	.665	. 634
AVG	.663	.629	.652	.715	.661	.588	.635	.643	.704	.660	.691	.621	.696	.626	.673	

Table 4: F1 score for each model-city combination. The largest and smallest average values are shaded in blue and red, respectively. The model with the *largest* F1 (larger is better) is **bolded** for each city. Each city's shortname is as follows: Chicago (Chi), Detroit (Det), Baltimore (Bal), El Paso (ElP), Los Angeles (LA), Houston (Hou), Columbus(Col), Indianapolis (Ind), Miami (Mia), Memphis (Mem), New York City (NYC), New Orleans (NO), San Antonio (SA), Philadelphia (Phi), and Phoinex (Pho).

	Bal	Chi	Col	Det	ElP	Hou	Ind	LA	Mem	Mia	NO	NY	Phi	Pho	SA	AVG
Linear SVM	.187	.193	.218	.233	.239	.211	.167	.172	.247	.152	.194	.151	.191	.247	.220	.201
BiLSTM	.111	.100	.144	.149	.154	.142	.115	.119	.142	.085	.118	.104	.116	.154	.143	.126
CNN	.124	.106	.155	.168	.168	.159	.112	.137	.167	.091	.115	.104	.126	.174	.164	.138
LSTM	.111	.092	.133	.137	.146	.134	.102	.114	.134	.079	.105	.099	.108	.145	.135	.118
BERT	**.105**	**.069**	**.113**	**.087**	**.075**	**.075**	**.068**	**.078**	**.118**	**.059**	**.086**	**.070**	**.091**	**.115**	**.097**	**.086**
AVG	.128	.112	.153	.155	.156	.145	.113	.124	.162	.093	.124	.106	.126	.167	.152	

Table 5: False positive rate (FPR) for each model-city combination. The largest and smallest average values are shaded in blue and red, respectively. The model with the *smallest* (smaller is better) FPR is **bolded** for each city.

	AA	Hispanic
Spearman	-.005	-.103
PCC	.102	-.018
AAE vs SAE Results		
AAE FPR	.154	(3392)
SAE FPR	.092	(5789)

Table 6: Correlation (Spearman and PCC) between the FPR scores and AA and H/L population proportions of each city.

hyperparameter selection. This procedure has been used in prior work to ensure robust results in similar social media-related NLP studies (Yin et al., 2017; Elejalde et al., 2017; Samory et al., 2020).

Results. In Table 4,we report the F1 of the OLID model applied to the GeoOLID dataset. Overall, we find substantial variation in model accuracy across the 15 cities. The average F1 for Houston (averaged across the non-random baselines) ranges from .588 (Houston) to .715 (Detroit), nearly 13% percent absolute difference and 22% relative difference. The CNN model achieves the best performance on average. Likewise, we find that CNN's best results are for the cities of Baltimore, Detroit, Memphis, and Philadelphia. Conversely, the CNN's worst results are found in Houston, Phoenix, Chicago, and New York.

Table 5 reports the False Positive Rates (FPR) for each city. Again, we see large variation, ranging from .093 to .167, nearly an 80% relative difference). On the other hand, we find that the best performing model is consistent across all cities. Hence, if a model performs better in Houston, it is likely it will perform better in Detroit. However, just because the model is better, the performance can be very low when compared to another location. Hence, decision-makers must carefully evaluate models based on the people impacted by them and not rely on evaluation metrics calculated on non-representative data before using the model. Finally, we report Accuracy results in the Appendix, Table 10, which show improvement greater than chosen MDE thresholds.

Prior work by Sap et al. (2019) show that offensive language detection models generate more false positives for text written in AAE. We evaluate this on our GeoOLID dataset following a similar strategy as Sap et al. (2019). In Table 6, we use the Blodgett et al. (2016) tool to identify AAE (African American English) and SAE (Standard American English) tweets in our GeoOLID dataset across all cities. When we calculate the false positive rate (FPR) across these two aggregate groups, we find similar conclusions to prior work (Sap et al., 2019) suggesting that offensive language models generate

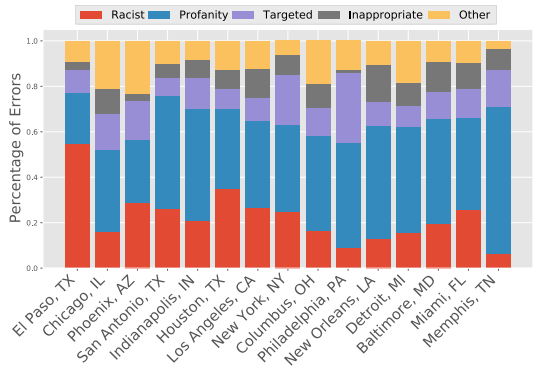

Figure 2: Manually coded false negatives per city.

more false positives on AAE text.

However, this does not mean that cities with larger African American populations will have larger FPRs. To test this using PCC and Spearman ρ, we correlate model performance (FPR) with the proportion of Black or African American and Hispanic/Latino residents (See Table 3) using US Census data. We find that there is weak to no correlation between FPR and minority population, which is surprising given AAE is correlated with population. There are two major reasons for this phenomena. First, AAE is not widely spoken. Even among African Americans, they do not always use an AAE dialect. Hence, if someone uses AAE sparingly, then the FPR on AAE will not accurately represent how the model will perform on text they write. Second, there are other factors that have a larger impact. In particular, this data is within the context of Covid-19. Hence, there are topics written with particular stances that the offensive langauge detection model is unable to handle, e.g., understanding the context of "wuhan virus".

Finally, we suggest that researchers should look at evaluating "contextual language", e.g., try to identify real people, ask them how they identify with regard to race and gender, then evaluate how models perform for them. This can provide insight into real bias issues and ground potential negative impact on real people. This idea fits with the narrative against treating certain groups as a monolith entities (Tadjiogueu, 2014; Erving and Smith, 2021).

Discussion. We perform a comprehensive manual analysis on the false negatives made by the best model on the OLID dataset. Specifically, we performed a qualitative open coding procedure to categorize the false negatives into commonly occurring groups. We allowed categories and meanings to emerge from posts in somewhat of an open coding fashion (Strauss and Corbin). We randomly sample up to 100 false negatives from each city, identifying the main categories. Next, a meeting was held where the main categories were discussed.

The final group of codes were identified as: Racist, Profanity, Targeted, Inappropriate, and Other. Racist was defined as a direct attack of mentioned of a race and/or ethnicity, Profanity as any sort of curse words, this could be in a hashtag or acronym. Targeted was defined as an attack on an individual, personal or group not associated to race or ethnicity, and finally Inappropriate is defined as any insensitive joke or sexual reference.

The results are summarized in Figure 2. A few important observations can be made from this graph. For instance, we find a large proportion of false negatives in the racist category in border cities, or cities in close proximity to Mexico (e.g., El Paso, Phoenix, San Antonio, and Houston). The reason the false negatives occur is based on the stance, topic, and way of Racist writing we found to be common in the border regions. For instance, we found many issues where the model did not detect language that refers to migrants being part of a "horde," meant to cause violence or destruction (this is common racist rhetoric at the time (Finley and Esposito, 2020)), as being offensive.

We counted the number of border-related topics using a small set of search terms (e.g., "border", "migrants", "immigrants", and "illegals") in the Original Data. We plot the results in Figure 3. We find that most of the border-related tweets are in states near Mexico (e.g., Texas, Arizona, New Mexico, and California). Hence, more false negatives caused by racist-categorized tweets about the border are more likely to be made in these cities, thus also increasing the likelihood of false negatives. Given the increase in border-related topics, this error is location-related.

Prior work has shown geographic differences in the use of swear words on social media (Carey, 2020; Grieve, 2015). We also found morphological variants of curse words in different cities that caused false negatives. For example, in New Orleans, Philadelphia, and Memphis there were many false negative tweets contain high percentages of Profanity due to multiple spellings of different swear words such as "phucking", "effing", "mothafucka", "biatches".

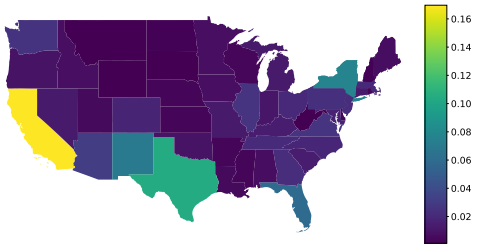

Figure 3: Proportion of border-related tweets in the "Original Data" for each state.

5.3 Geographic Similarities

In this subsection, we analyze the correlation between the best-performing models in each city.

Methods. We analyze the performance of the models trained and described in Section 5.2. Specifically, we compare the PCC between the Accuracy of each model applied to all cities. Intuitively, if the models for New York are sorted based on Accuracy, and the sorted order is the same as Phoenix, then the correlation would be one, showing a linear relationship. The more differences in sorted scores/models, the lower the performance (i.e., correlation). Overall, in this experiment, we rank every model along with the variants of models and compare every pair of cities rankings (i.e., each model trained with different word embeddings listed in the Appendix, Section A.2 are treated as independent models). Intuitively, if correlations are very high, this could indicate that you could choose the best hyperparameters for city A and they would be the best for city B. From the main results in Table 4 we saw that the best model was the same across all cities. However, this is with substantial hyperparameter optimization. Are the parameters the same for each city?

Results. The results of the correlation analysis are shown in Figure 4. Overall, similar to variations in model performance across cities, we find that the similarity in model performance correlations can vary substantially city-to-city. For instance, the best models for Houston are substantially different from other cities, except for a few (e.g., Los Angeles). However, on further inspection, general architecture performance seems to be relatively similar across cities, e.g., the CNN model is the best on the OLID dataset and for most cities. Much of the variation comes from hyperparameter choice or pretrained embedding choice (with more than 10% in Accuracy between the best and worst embeddings). The best embeddings can be substantially different city-to-city. This result suggests that choosing the best hyperparameters based on a small subset of data (e.g., from one city) is not optimal for each location, which can result in further performance disparities.

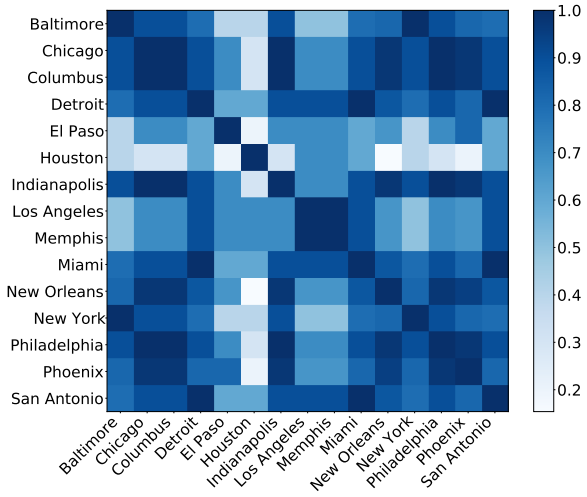

Figure 4: Model accuracy correlation between each pair of cities in GeoOLID.

Discussion. The results do provide us with a potential research avenue. An interesting question that could be explores is if we train a model with many hyperparameter options on a dataset, is it possible to predict which model to deploy in a given region? There has been some work in predicting model perform (Elsahar and Gallé, 2019). Hence, it would be interesting to expand that to predict the best hyperparameters.

6 Conclusion

We provide a comprehensive analysis of performance disparities of offensive language detection models. Furthermore, we introduce a novel dataset that provides more than 14 thousand examples for further analysis of geographical differences in model performance. The study points to the importance of geographically sensitive NLP, where the impact and performance of NLP models are analyzed for specific geographical regions, or even micro communities within a city. Moreover, finding regions where models perform poorly on can also provide unique testbeds as "hard test cases" similar to recent work on adversarial examples (Zhang et al., 2019).

Acknowledgements

This material is based upon work supported by the National Science Foundation (NSF) under Grant No. 1947697 and NSF award No. 2145357.

References

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

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

David Bamman, Chris Dyer, and Noah A. Smith. 2014. Distributed representations of geographically situated language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 828–834, Baltimore, Maryland. Association for Computational Linguistics.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. 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, Austin, Texas. Association for Computational Linguistics.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, Online. Association for Computational Linguistics.

Stan Carey. 2020. Mapping the united swears of america. Visited: 05/15/2022.

Carlos Castillo. 2016. *Big crisis data: social media in disasters and time-critical situations*. Cambridge University Press.

Nikhil Cheke, Joydeep Chandra, Sourav Kumar Dandapat, et al. 2020. Understanding the impact of geographical distance on online discussions. *IEEE Transactions on Computational Social Systems*, 7(4):858–872.

Courtney Corley, Diane Cook, Armin Mikler, and Karan Singh. 2010. 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.

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

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. 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. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

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

Gabriel Doyle. 2014. Mapping dialectal variation by querying social media. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 98–106, Gothenburg, Sweden. Association for Computational Linguistics.

Akash Dutt Dubey. 2020. The resurgence of cyber racism during the covid-19 pandemic and its aftereffects: analysis of sentiments and emotions in tweets. *JMIR Public Health and Surveillance*, 6(4):e19833.

Erick Elejalde, Leo Ferres, and Eelco Herder. 2017. The nature of real and perceived bias in chilean media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, pages 95–104.

Hady Elsahar and Matthias Gallé. 2019. To annotate or not? predicting performance drop under domain shift. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2163–2173, Hong Kong, China. Association for Computational Linguistics.

Christy L Erving and Monisola Vaughan Smith. 2021. Disrupting monolithic thinking about black women and their mental health: Does stress exposure explain intersectional ethnic, nativity, and socioeconomic differences? *Social Problems*.

Joel Escudé Font and Marta R. Costa-jussà. 2019. 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, Florence, Italy. Association for Computational Linguistics.

Laura Finley and Luigi Esposito. 2020. The immigrant as bogeyman: Examining donald trump and the right's anti-immigrant, anti-pc rhetoric. *Humanity & Society*, 44(2):178–197.

Mihaela Gaman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer, and Marcos Zampieri. 2020. A report on the VarDial evaluation campaign 2020. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–14, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. 2018. On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 316–327, Brussels, Belgium. Association for Computational Linguistics.

Jack Grieve. 2015. Research blog: Jack grieve's homepage swear word mapping. Visited: 05/15/2022.

Dirk Hovy and Christoph Purschke. 2018. Capturing regional variation with distributed place representations and geographic retrofitting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4383–4394, Brussels, Belgium. Association for Computational Linguistics.

Dirk Hovy, Afshin Rahimi, Timothy Baldwin, and Julian Brooke. 2020. Visualizing regional language variation across europe on twitter. *Handbook of the Changing World Language Map*, pages 3719–3742.

Dirk Hovy and Anders Søgaard. 2015. Tagging performance correlates with author age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 483–488, Beijing, China. Association for Computational Linguistics.

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

Olga Kellert and Nicholas H Matlis. 2021. Geolocation differences of language use in urban areas. *arXiv preprint arXiv:2108.00533*.

Alex Lamb, Michael J. Paul, and Mark Dredze. 2013. 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, Atlanta, Georgia. Association for Computational Linguistics.

Rabindra Lamsal. 2021. Design and analysis of a large-scale covid-19 tweets dataset. *Applied Intelligence*, 51(5):2790–2804.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Brandon Lwowski and Peyman Najafirad. 2020. COVID-19 surveillance through Twitter using self-supervised and few shot learning. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Brandon Lwowski and Anthony Rios. 2021. The risk of racial bias while tracking influenza-related content on social media using machine learning. *Journal of the American Medical Informatics Association*, 28(4):839–849.

Ninareh Mehrabi, Thamme Gowda, Fred Morstatter, Nanyun Peng, and Aram Galstyan. 2020. Man is to person as woman is to location: Measuring gender bias in named entity recognition. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, pages 231–232.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on bert model. *PloS one*, 15(8):e0237861.

Graham Neubig, Yuichiroh Matsubayashi, Masato Hagiwara, and Koji Murakami. 2011. Safety information mining — what can NLP do in a disaster—. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 965–973, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Ray Oshikawa, Jing Qian, and William Yang Wang. 2020. A survey on natural language processing for fake news detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6086–6093, Marseille, France. European Language Resources Association.

Robert Östling and Jörg Tiedemann. 2017. Continuous multilinguality with language vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 644–649, Valencia, Spain. Association for Computational Linguistics.

Hyunji Hayley Park, Katherine J. Zhang, Coleman Haley, Kenneth Steimel, Han Liu, and Lane Schwartz. 2021. Morphology matters: A multilingual language modeling analysis. *Transactions of the Association for Computational Linguistics*, 9:261–276.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. " O'Reilly Media, Inc.".

Umair Qazi, Muhammad Imran, and Ferda Ofli. 2020. Geocov19: A dataset of hundreds of millions of multilingual covid-19 tweets with location information. *ACM SIGSPATIAL Special*, 12(1):6–15.

Christian Reuter and Marc-André Kaufhold. 2018. Fifteen years of social media in emergencies: a retrospective review and future directions for crisis informatics. *Journal of contingencies and crisis management*, 26(1):41–57.

Anthony Rios. 2020. Fuzze: Fuzzy fairness evaluation of offensive language classifiers on african-american english. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 881–889.

Anthony Rios, Reenam Joshi, and Hejin Shin. 2020. Quantifying 60 years of gender bias in biomedical research with word embeddings. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 1–13, Online. Association for Computational Linguistics.

Anthony Rios and Brandon Lwowski. 2020. An empirical study of the downstream reliability of pre-trained word embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3371–3388, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Mattia Samory, Vartan Kesiz Abnousi, and Tanushree Mitra. 2020. Characterizing the social media news sphere through user co-sharing practices. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 602–613.

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

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

Yves Scherrer and Nikola Ljubešić. 2021. Social media variety geolocation with geoBERT. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 135–140, Kiyv, Ukraine. Association for Computational Linguistics.

Philippa Shoemark, James Kirby, and Sharon Goldwater. 2017. Topic and audience effects on distinctively Scottish vocabulary usage in Twitter data. In *Proceedings of the Workshop on Stylistic Variation*, pages 59–68, Copenhagen, Denmark. Association for Computational Linguistics.

Monica Stephens. 2020. A geospatial infodemic: Mapping twitter conspiracy theories of covid-19. *Dialogues in Human Geography*, 10(2):276–281.

A Strauss and J Corbin. Basics of qualitative research: Techniques and procedures for developing grounded theory sage publications, inc; 1998. translated by: Mohammadi, b. *Institute for Humanities and Cultural Studies.[In Persian]*.

Eteena J Tadjiogueu. 2014. Fifty shades of black: Challenging the monolithic treatment of black or african american candidates on law school admissions applications. *Wash. UJL & Pol'y*, 44:133.

Tami L Thomas, Ralph DiClemente, and Samuel Snell. 2014. Overcoming the triad of rural health disparities: How local culture, lack of economic opportunity, and geographic location instigate health disparities. *Health education journal*, 73(3):285–294.

Luis von Ahn. 2009. Offensive/profane word list. *Retrieved June*, 24:2018.

Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 656–666, Montréal, Canada. Association for Computational Linguistics.

Zhijun Yin, Bradley Malin, Jeremy Warner, Pei-Yun Hsueh, and Ching-Hua Chen. 2017. The power of the patient voice: learning indicators of treatment adherence from an online breast cancer forum. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 337–346.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. 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 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Huangzhao Zhang, Hao Zhou, Ning Miao, and Lei Li. 2019. Generating fluent adversarial examples for natural languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5569, Florence, Italy. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Joshua N Zingher and Eric M Moore. 2019. The power of place? testing the geographic determinants of african-american and white voter turnout. *Social Science Quarterly*, 100(4):1056–1071.

A Appendix

A.1 Word Embeddings

In Table 7, we link to the publicly available word embeddings we use in our experiments. We test three models: SkipGram, GLOVE, and FastText. We also explore different embeddings sizes, ranging for 25 dimensions to 30. Moreover, we explore embeddings trained on different corpora, ranging from biomedical text (PubMed) to social media data (Twitter). The best embeddings are chosen based on the OLID validation dataset for all reported results in the main manuscript.

A.2 Model Hyper-parameters

In this Section, we report the best hyperparmeters for each model. For the linear models we also report the best TF-IDF settings from the scikit-learn package.

TF-IDF:

- sublinear tf: True
- min df: 5
- norm: l2
- encoding: latin-1
- ngram range: (1,2)
- stop words: english

Linear SVM:

- penalty: l2
- C: 1.0

CNN:

- max words: 10000
- max sequence length: 125
- drop: .2
- batch size: 512
- epochs: 30
- filter sizes: 3,4,5
- num filters: 512
- early stopping: 5 iterations

LSTM:

- max words: 10000
- max sequence length: 125
- drop: .2
- batch size: 128
- epochs: 30
- num filters: 512
- hidden layers: 1
- early stopping: 5 iterations

BiLSTM:

- max words: 10000
- max sequence length: 125
- drop: .2
- batch size: 128
- epochs: 30
- num filters: 512
- hidden layers: 1
- early stopping: 5 iterations

BERT:

- tokenizer : bert-base-cased
- model : bert-base-cased
- dropout : .2
- max length : 128
- epochs : 50
- batch size : 64
- fine tuned : after 5 epochs
- early stopping : 5 iterations

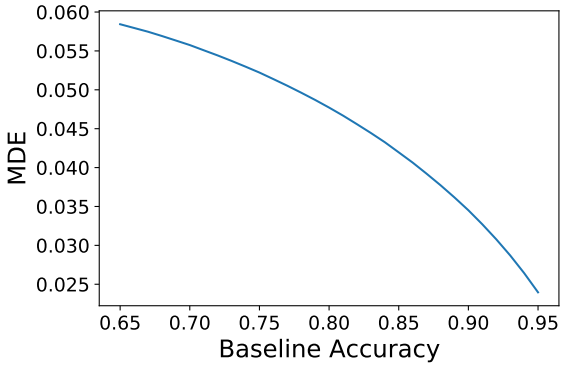

Figure 5: MDE given different baseline accuracy assumptions and a power of 80%.

A.3 OLID Results

We report the OLID results for each model (Linear SVM, CNN, LSTM, BiLSTM, and BERT) in Table 9. Interestingly, we find that the CNN model outperforms other methods, including the LSTM-based models and BERT. For instance, the CNN's F1 is more than 2% higher than the LSTM and BiLSTM models. Moreover, it is more than 6% higher than BERT. We also find that all methods outperform the traditional machine learning models (Linear SVM), with the CNN outperforming the Linear SVM by nearly 9% F1 and nearly 5% in Accuracy. The results support the results of the main paper with the CNN model generalizing better than other techniques.

Next, in Table 8 we report the performance of the CNN, LSTM, and BiLSTM models trained using different embeddings. Overall, we see variation across which embeddings result in teh best F1 score for each model, with wiki_42B_300d resulting in the highest F1 for the BiLSTM, wiki_840B_300d resulting in the best results for the LSTM, and GLOVE_twitter_27B_100d. This finding is similar to the results for H3 in the main paper, where embedding choice can vary city-to-city. We also find that it can vary model-to-model, which is also supported in Rios and Lwowski (2020).

A.4 Accuracy Power Analysis

In Figure 5, we report the MDE (Card et al., 2020) for Accuracy assuming different baseline scores and a power of 80%. For instance, if the baseline achieves an accuracy of .95, then we would need to see any improvement/difference of around .025 for it to be significant. Likewise, if the accuracy is around .65, then we need an improvement of nearly

.06 for it to be significant. Intuitively, the more accurate the results, the smaller the improvement can be for it be significant.

A.5 Accuracy Scores per City

In Table 10, we report the OLID model accuracy for each city. Overall, we find substantial variation in model accuracy across the 15 cities. The Linear SVM classifier ranges from .704 to .822, resulting in around a 12% difference in accuracy between Phoenix and Miami. Similar findings can be seen with the other models like CNN and BERT having a up to a 10% difference. Furthermore, given the MDE of around 5% for each city depending on the baseline score, we find that many of the differences are significant.

Model	Data Source	Dimension	Link
SkipGram	Google News	300	https://docs.google.com/file/d/0B7XkCwpI5KDYaDBDQm1tZGNDRHc/edit?usp=sharing
SkipGram	PubMed	200	http://evexdb.org/pmresources/vec-space-models/PubMed-w2v.bin
SkipGram	PubMed Central	200	http://evexdb.org/pmresources/vec-space-models/PMC-w2v.bin
SkipGram	PubMed and PubMed Central	200	http://evexdb.org/pmresources/vec-space-models/PubMed-and-PMC-w2v.bin
SkipGram	Wikipedia, PubMed, and PubMed Central	200	http://evexdb.org/pmresources/vec-space-models/wikipedia-pubmed-and-PMC-w2v.bin
GLOVE	Twitter	25	http://nlp.stanford.edu/data/glove.twitter.27B.zip
GLOVE	Twitter	50	http://nlp.stanford.edu/data/glove.twitter.27B.zip
GLOVE	Twitter	100	http://nlp.stanford.edu/data/glove.twitter.27B.zip
GLOVE	Twitter	200	http://nlp.stanford.edu/data/glove.twitter.27B.zip
GLOVE	Wikipedia 2014 and Gigaword 5	50	http://nlp.stanford.edu/data/glove.6B.zip
GLOVE	Wikipedia 2014 and Gigaword 5	100	http://nlp.stanford.edu/data/glove.6B.zip
GLOVE	Wikipedia 2014 and Gigaword 5	200	http://nlp.stanford.edu/data/glove.6B.zip
GLOVE	Wikipedia 2014 and Gigaword 5	300	http://nlp.stanford.edu/data/glove.6B.zip
GLOVE	Common Crawl V1	300	http://nlp.stanford.edu/data/glove.42B.300d.zip
GLOVE	Common Crawl V2	300	http://nlp.stanford.edu/data/glove.840B.300d.zip
FastText	Wikipedia 2017, UMBC webbase corpus, and statmt.org news dataset	300	https://dl.fbaipublicfiles.com/fasttext/vectors-english/wiki-news-300d-1M.vec.zip
FastText	Common Crawl	300	https://dl.fbaipublicfiles.com/fasttext/vectors-english/crawl-300d-2M.vec.zip

Table 7: List of word embeddings we use in our experiments.

Word Embedding	F1	Accuracy
BiLSTM		
FASTTEXT_en_300	.580	.760
GLOVE_twitter_27B_100d	.627	.785
GLOVE_twitter_27B_50d	.5834	.764
GLOVE_wiki_42B_300d	.645	.793
GLOVE_wiki_6B_100d	.600	.771
GLOVE_wiki_6B_200d	.605	.778
GLOVE_wiki_6B_300d	.631	.783
GLOVE_wiki_6B_50d	.586	.768
GLOVE_wiki_840B_300d	.631	.787
W2V_GoogleNews	.616	.781
W2V_PMC	.488	.730
W2V_PubMed_PMC	.514	.738
W2V_PubMed	.402	.704
LSTM		
FASTTEXT_en_300	.524	.749
GLOVE_twitter_27B_100d	.618	.782
GLOVE_twitter_27B_50d	.591	.770
GLOVE_wiki_42B_300d	.619	.790
GLOVE_wiki_6B_100d	.607	.774
GLOVE_wiki_6B_200d	.616	.781
GLOVE_wiki_6B_300d	.609	.782
GLOVE_wiki_6B_50d	.577	.762
GLOVE_wiki_840B_300d	.624	.788
W2V_GoogleNews	.602	.779
W2V_PMC	.456	.720
W2V_PubMed_PMC	.495	.730
W2V_PubMed	.348	.701
CNN		
FASTTEXT_en_300	.611	.778
GLOVE_twitter_27B_100d	.657	.792
GLOVE_twitter_27B_50d	.635	.788
GLOVE_wiki_42B_300d	.642	.793
GLOVE_wiki_6B_100d	.621	.779
GLOVE_wiki_6B_200d	.621	.786
GLOVE_wiki_6B_300d	.621	.785
GLOVE_wiki_6B_50d	.612	.775
GLOVE_wiki_840B_300d	.648	.794
W2V_GoogleNews	.638	.789
W2V_PMC	.520	.738
W2V_PubMed_PMC	.541	.743
W2V_PubMed	.461	.718

Table 8: Word Embedding Performance for Deep Learning Models

	Prec.	Rec.	F1	Acc.
Random Baselines				
Stratified	.324	.348	.336	.553
Uniform	.321	.505	.392	.493
Machine Learning Models				
Linear SVM	.643	.505	.566	.744
BiLSTM	.754	.551	.631	.783
CNN	.721	.603	.657	.792
LSTM	.768	.527	.624	.788
BERT	.652	.555	.592	.752

Table 9: OLID Results

	Bal	Chi	Col	Det	ElP	Hou	Ind	LA	Mem	Mia	NO	NY	Phi	Pho	SA	AVG
Stratified	.555	.555	.550	.536	.536	.570	.577	.567	.521	.592	.553	.588	.567	.544	.564	.558
Linear SVM	.779	.745	.751	.761	.694	.724	.748	.776	.752	.822	.787	.794	.771	.704	.740	.757
BiLSTM	.834	.809	**.799**	.803	.757	.774	.809	.824	**.818**	.861	.835	.842	.833	.768	.783	.809
CNN	**.843**	**.820**	.792	.823	.747	.773	**.819**	.805	.814	.851	**.842**	**.849**	.849	.760	**.788**	.811
LSTM	.832	.814	.790	**.834**	**.758**	.790	.817	**.829**	.810	**.873**	.837	.834	**.850**	**.772**	.783	**.815**
BERT	.786	.800	.788	.785	.755	**.791**	.790	.809	.761	.848	.785	.816	.803	.747	.771	.789
AVG	.815	.798	.784	.801	**.742**	.770	.797	.809	.791	**.851**	.817	.827	.821	.750	.773	.796

Table 10: Accuracy for each city in the GeoOLID dataset. In the bottom row (i.e., the average across all machine learning models), we mark the cities that have an average accuracy difference greater than or equal to the MDE compared to the city with the highest average accuracy.