Towards Community-Driven NLP: Measuring Geographic Performance Disparities of Offensive Language Classifiers

Anonymous ACL submission

Abstract

Text classifiers are applied at scale in the form of one-size-fits-all solutions. Nevertheless, many studies show that many classifiers are biased regarding different languages and dialects. Both language style and content change depending on the location that it is posted. For example, states that border Mexico may be more likely to discuss issues regarding immigration from Latin America. However, several questions remain, such as "Do changes in the style and content of text across geographic regions impact model performance?". We introduce a novel dataset called GeoOLID with more than 13 thousand examples across 15 geographically and demographically diverse cites to address this question. Furthermore, we perform a comprehensive analysis of geographical content and stylistic differences and their interaction in causing performance disparities of Offensive Language Detection models. Overall, we find that current models do not generalize across. Likewise, we show that understanding broad dialects (e.g., African American English) is not the only predictive factor of model performance when applied to cities with large minority populations. Hence, community-specific evaluation is vital for real-world applications. **Warning: This paper contains offensive language.**

1 Introduction

Text classification, especially when applied to social network data or online blogs, has been applied to wide array of tasks 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 cyber-bullying (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 ran 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 geographical-caused differences in language and content that may exist in the context of offensive language detection.

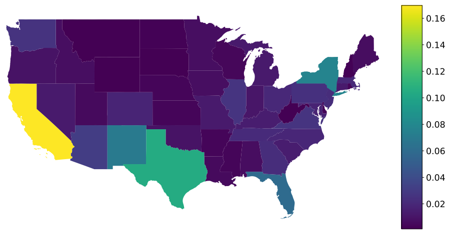

Figure 1: Proportion of border-related tweets in the GeoCOVID dataset (Qazi et al., 2020) for each state.

Several lines of research 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; Hovy et al., 2020; Cheke et al., 2020; Gaman et al., 2020). For instance, Hovy and Purschke (2018) show that doc2vec embedding frameworks which can be used to can help with the task of geolocation prediction. Hovy et al. (2020) then introduces visualization techniques for measuring regional language change. Kellert and Matlis (2021) shows that these differences exist at the city level as well. Our data consists of geo-tagged COVID-related tweets. In Figure 1, we measure the proportion of border-related tweets in each state to show case how topical content can be distributed geographically, finding most of the border-related tweets are in states near Mexico (e.g., Texas, Arizona, New Mexico, and California). Overall, much of the prior work has focused on better incorporating or identifying regional aspects of language data to improve performance in machine translation (Östling and Tiedemann, 2017) or geolocation prediction (Hovy and Purschke, 2018).

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)—can have a substantial impact on model performance. For instance, Gerz et al. (2018) show that fine-grained typological features must be incorporated into language modeling architectures for a single model to adequately perform across a wide array of languages. More specifcally, the absence of typological features is predictive of substantial performance disparities across languages. 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. 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. These findings show that models deployed at scale can adversely impact minority groups.

Overall, while there has been a substantial amount of research understanding how to identify and use regional (geographical) features and identify performance disparities across languages and dialects, to the best of our knowledge, there has been no prior work understanding geographical performance disparities across regional dialects. Do regional language differences, whether content or language style, impact offensive language model performance? While prior work has shown that dialect can impact, and dialect is correlated with social demographics of regional areas (Blodgett et al., 2016), . For instance, AAE is not spoken the same in every region of the United States (US). There are well-known sub-dialects such as Rural and Urban AAE. But, more importantly, certain features of AAE only appear within specific regions of the US (Jones, 2015). Does the interaction between AAE features and content spoken in one region adversely impact model performance more than other regions? There has been extensive research, understanding geographical health disparities, which are thought to be due to limited physical access to health care, but also to differences in demography, attitudes, lifestyle factors, and cultural practices in regional and rural settings (Eberhardt and Pamuk, 2004; Smith et al., 2008; Dixon and Welch, 2000). Prior work has shown that performance disparities can potentially increase health disparities for minority communities (Lwowski and Rios, 2021). Hence, depending on the applications in which text classifiers are applied (e.g., offensive language), geographical algorithmic disparities can further harm (e.g., the mental health) these regions.

To better understand the implications of geographical performance disparities, we make three major contributions: **(1.)** We introduce a new dataset called GeoOLID with more than 13 thousand tweets across 15 geographically and demographically diverse cities in the United States. **(2.)** We produce a comprehensive dataset analysis, analysing both the content and stylistic variations in each city. **(3.)** Finally, we perform a comprehensive analysis of performance disparities across a wide array of popular text classification models in each city, producing novel insights and discuss important future avenues of research.

2 Language Variation

Langauge variation is an important area of research for the NLP community. For example, understanding how different languages vary (e.g., Finnish vs Korean) typologically has been shown to be important to reduce performance disparities of language models (Gerz et al., 2018). While there has been some disagreement about whether morphology matters, recent work by (Park et al., 2021) has shown that incorporating information that can model morphological differences is important in improving model performance. Overall, much of the prior work has focused on either developing methods to identify language features within text or use various language features to improve model performance. For instance, VarDial Evaluation has been an annual competition to identify various dialects of different languages (e.g., German and Romanian) as well as geolocations (Gaman et al., 2020). For instance, 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 et al. (2021) show that combining modern NLP architectures like BERT with a double regression model can also provide success in determining the latitude and logitude points of the location of the text. The results of this shared task highlights the fact that semantic and lexical differences exist when locations of the text change. 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.

Overall, the major gap in prior work looking and language variation is that there has not been any studies evaluating the impact regional language variation has on the performance of downstream tasks. In this paper, we introduce a novel dataset called GeoCOVID. However, before addressing the research gap in understanding performance disparities, it is important to measure language variation across each city within the dataset. If cities do not vary with regards to content and langauge style, then we should not expect models to perform different within each city. Hence, we test the following hypothesis:

H1a. Text in the GeoOLID dataset is distributed differently (based on content and style) depending on the location it was posted.

H1b. Text is representative of the sociodemographic makeup of the area it was posted.

By expanding the analysis of prior work looking at dialectal variation (Abdul-Mageed et al., 2020; Lulu and Elnagar, 2018; Blodgett et al., 2016), we are able show that the results generalize to our newly collected dataset.

3 Performance Disparities

Performance disparities across languages and dialects recently have seen received much 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 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 (Font and Costa-jussà, 2019).

Overall, the major research gap in prior work is in the lack of fine-grained regional understanding of performance disparities. Many groups that are studied are broad, 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. Hence, if an Offensive language detection model performs poorly on one set of AAE patterns, it can impact one region much more than others. Unfortunately, it is neither possible to measure model performance for every minority sub-population nor all potential syntactic pattern given the ever evolving nature of language. Furthermore, it is not possible to understand how a model will perform on every "common" topic discussed within the US given the large variation in discussions (e.g., Texas may speak more about the Dallas Cowboys, while Ohio focuses on the Bengals for the topic of Football). Hence, we believe that community-driven analysis is a better future avenue to understand the real-world impact of NLP models. Instead of trying to understand all potential sub-populations and style variations to evaluate them all, we propose measuring performance on small communities instead. Overall, to begin addressing these gaps in understanding, we make the following hypothesis:

H2a. Because data is distributed differently in different geographic regions, model performance is not the same in each location.

H2b. Errors made by the models are caused by geographic-specific content and language style.

H3. Model choice depends on the community it will be deployed.

These hypotheses will provide a starting point towards what we term "community-driven NLP". By showing that model performance can vary location-to-location, we hope to bring awareness of adverse harms that the broad application of NLP can cause without carefully understanding the communities in which the models are deployed.

	Non Offensive	Offensive	Total	MDE
OLID	9,460	4,640	14,100	.014
		GeoOLID Dataset		
	Non Offensive	**Offensive**	**Total**	**MDE**
Unlabeled Data	—	—	34,724	—
All Cities (labeled)	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.

4 Data

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. The OLID datasets was split into 5 random 70/10/20 training, validation, and testing splits, respectively. The GeoOLID dataset is only used for testing.

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).

GeoOLID. In addition to the OLID dataset, we introduce a new offensive language dataset using tweets collected since the start of the COVID-19 pandemic. Qazi et al. (2020) and Lamsal (2021) originally collected more than 524 million multilingual tweets across 218 countries and 47,000 cities between the dates of February 1, 2020 and May 1, 2020. Given the large amount of politically divisive discourse, racist remarks, and social impact of COVID-19, GeoCOVID provides a unique testbed to understand geographic model variation.

Filtering: To filter down the 524 million tweets into a manageable set for this study, we first selected English Language tweets only. English Language tweets are then filtered by city, only keeping geocoded tweets with a coordinate point tied to a city of origination. We then subset all cities, removing any tweet not posted from a small group of 15 manually chosen cities that differ geographically and demographically.

With the goal of identifying offensive language, we wanted to guarantee there was a mixture of normal and offensive tweets present in each city. Our last filter included a keyword filter using the badword lexicon (von Ahn, 2009), hatebase lexicon (Davidson et al., 2017), offensive 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 that we deem relevant for potential discrimination against a race (chinese, china, asia, asian, wuhan). Along with the aforementioned filters we appended on a sub sample of tweet data for each city and drop any replicated tweets. The final counts of each city can be found in Table 1.

Cities: To measure the performance difference across varying geolocations, we decided on 15 cities based on multiple facets, data availability, annotation time, geographical diversity, and demographic diversity. When selecting the 15 cities we strategically selected locations in the United States that different dialects could be present, as well as the topic distribution. For example, 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 and Chicago. In addition we selected cities where we knew residents could use very distinct accents and phonics like New York and New Orleans. By selecting the 15 cities in Table 1, we created a diverse dataset with multiple ethnicities, language styles, and topical differences.

Annotation: In order to provide accurate labels for this study, samples of 500 tweets were assigned to 3 different graduate students to be labeled as offensive language using the logic provided by Zampieri et al. (2019). A total of 20 students were recruited

	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. Accuracy, Macro Precision, Macro Recall, and Macro F1.

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. The definition of an Offensive tweet was provided 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. First, the graduate students annotated the tweets, providing us with 3 separate independent labels of each tweet. We then calculated the agreement between annotators, resulting in a Fleiss Kappa of 0.47. This agreement score was not sufficient enough for us to feel comfortable running experiments on. Second, we (the authors) of the paper manually—and independently—adjudicated the annotation of each user, 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. Third, one of the authors went through the tweets once again correcting any final disagreements among the authors 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 and the final agreement score increased to 0.83. Finally, we 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.

Unlabeled GeoOLID Data. We also make use of unlabeled GeoCOVID data to addresses Hypotheses H1a and H1b. The basic stats of this dataset are available in Table 1 in the row titled "Unlabeled Data". The data is the same as the labeled GeoCOVID, except it was not labeled because of resource constraints.

5 Experiments

In order to address and test whether the hypothesis are supported, we ran multiple experiments and analyzed performance across the 15 cities in the GeoOLID dataset. In the next two subsections, we restate each hypothesis, provide the evidence to support it, and finally provide a discussion around the results, summarizing major implications.

5.1 Understanding Data Variation

In this section, we address hypotheses H1a and H1b. More specifically, we analyze how different the stylistic and content differs across each of the 15 cities in the GeoOLID dataset. Moreover, we look at the correlation between language style and the demographic makeup of each city.

Methods. To address these hypotheses we make use of two distinct methods. First, to answer H1a, we train a geolocation prediction model. Given a tweet, the goal of the geolocation 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 unlabeled GeoOLID dataset. The goal is to ensure that common content information is used, while avoid highly location-specific terminology. 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. The paper shows that the African-American proportions correlate with AAE language features. Finally, we train a Random Forest classifier on the unlabeled 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.

Second, to answer H1b, we we use the dialect information from the dialect inference tool from Blodgett et al. (2016) and correlate it with the demographic information of each city. 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) topic proportion for each city using the tool from Blodgett et al. (2016). Finally, we calculate the Pearson Correlation Coefficient (PCC) between the proportion of AA and H/L residents in each city.

H1a: Text in GeoOLID is distributed differently depending on the location it was posted. The results for the experiments addressing H1a are reported in Table 2. Using content and style features, we were able to predict the location of a tweet 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 point to the fact that there are unique stylistic and content-related differences in each location which is important for supporting Hypothesis H2 about variations in model performance across different locations.

H1b: Text made in certain geographic regions is representative of the sociodemographic makeup of the area. In addition to the classification of a tweets location, we present a strong correlation between the sociodemographic makeup of each city and the dialect style of the tweets within the dataset. If a cities population of residents has a higher percentage of African Americans, then prior work has shown that there is an increase in the number of occurrences of AAE-related language patterns (Blodgett et al., 2016). Specifically, using the average AAE (AAS) and Hispanic (HLS) scores from the Blodgett et al. (2016) tool over all tweets in each city, we correlate it with the proportion of AA and

	AAS	HLS	Tot.	AA	H/L
Bal	.168	.193	585,708	338,478	45,927
Chi	.147	.204	2,450,143	801,195	819,518
Col	.146	.201	905,748	259,483	70,179
Det	.196	.214	639,111	496,534	51,269
ElP	.158	.227	678,815	25,077	551,513
Hou	.161	.205	2,304,580	520,389	1,013,423
Ind	.151	.194	887,642	248,067	116,221
LA	.144	.204	3,898,747	336,096	1,829,991
Mem	.209	.220	633,104	389,779	62,167
Mia	.140	.175	442,241	57,254	310,472
NO	.182	.197	383,997	208,273	31,017
NY	.126	.182	8,804,190	1,943,645	2,490,350
Phi	.157	.204	887,642	248,067	116,221
Pho	.144	.208	1,608,139	125,260	661,574
SA	.175	.222	1,434,625	102,816	916,010
AA PCC	.565 (*p* value: .028)				
H/L PCC	.167 (*p* value: .55)				

Table 3: Pearson Correlation Coefficient (PCC) between the AAS and HLS. The abbreviations for the 15 cities in the GeoOLID dataset are 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).

Hispanic population within each city based on the 2020 US census data. This correlation can be seen in Table 3. Overall we find that there is significant correlation .565 (p value: 0.028) between the two variables. We find that cities like Baltimore, New Orleans and Detroit are more likely to have more AAE tweets 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 appeared was tagged in Phoenix, AZ.

5.2 Data Variation and Model Performance

In this subsection, we explore the central hypotheses of this paper looking at performance disparities between various locations within the US.

Methods. We train six different machine learning algorithms: Naive Bayes (NB), Linear Support Vector Machine (Linear SVM), Long Short Term Memory (LSTM), Bidirectional LSTM (BiLSTM), Convolutional Nueral 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.Each model is trained independently on

	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
Uniform	.484	.533	.500	.510	.465	.504	.479	.477	.487	.484	.495	.509	.503	.503	.513	.496
Prior	.695	.675	.672	.599	.554	.681	.662	.689	.605	.771	.651	.730	.651	.618	.596	.657
NB	.820	.765	.779	.764	.695	.781	.773	.794	.769	.863	.787	.797	.782	.710	.743	.775
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	.739
AVG	.816 (0)	.792 (1)	.782 (1)	.795 (1)	.734 (8)	.772 (1)	.793 (1)	.806 (1)	.787 (1)	.853 (0)	.812 (2)	.822 (0)	.815 (0)	.744 (6)	.768 (2)	

Table 4: Accuracy. In the bottom row, we mark the number of other cities that have a score greater than or equal to the MDE for that city given its score as a baseline.

	PCC	
	AA	**Hispanic**
NB	.186	-.216
Linear SVM	.142	-.272
LSTM	.279	-.362
CNN	.283	-.398
BiLSTM	.290	-.358
BERT	-.061	.056
AVG	.187	-.258
AAE vs SAE Results		
SAE Accuracy	.831	(3392)
AAE Accuracy	.854	(5789)

Table 5: PCC between AA and H/L population proportions of each city and Accuracy. This table also reports the SAE vs AAE Accuracy on the GeoOLID dataset—the total number in each group is in parenthesis.

each of the five different OLID training splits. The performance metrics are then averaged across the five different seeds as a way of measuring the robustness of the model and guaranteeing a high accuracy is not a coincidence when predicting on the same validation set. 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 for a complete listing of the evaluated embeddings). We also perform a comprehensive manual error analysis for H2b to better understand model performance differences beyond aggregate quantitative metrics.

H2a: Because data is distributed differently in different geographic regions, model performance is not the same in each location. In Table 4, we report the OLID model accuracy for each city. Overall, we find substantial variation in model accuracy across the 15 cities. The Naive Bayes (NB) classifier ranges from .695 to .863, resulting in around a 17% difference in accuracy between El Paso 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. Note that there are even larger differences in F1 score, please find the results in the Appendix.

The question remains, if the text in each city is correlated with demographic information, then why do we need location-specific performance analysis? The issue is that while demographic analysis provides broad insights, location-specific language is substantially more varied. Thus, unfortunately, demographic factors alone are not predictive of model performance for a given city. In Table 5, we use the Blodgett et al. (2016) tool to identify AAE and SAE (White-aligned) tweets in our GeoOLID dataset across all cities. When we calculate the accuracy across these two aggregate groups, we find similar conclusions to prior work (Sap et al., 2019) suggesting that offensive language models are biased towards AAE text. However, when we correlate (using PCC) model performance (Accuracy) with the proportion of Black or African American residents using US Census data, we find that the model is positively correlated (though not significant) with higher accuracy, which is contrary to the general demographic findings. We also correlate the performance of each model with Latino or Hispanic populations finding negative correlations. After manual analysis, we find that the models suffer for common topics in these areas (e.g., border-related topics). In summary, the major finding of this paper is listed below:

Major Finding: Broad dialectal analysis of model performance alone is not predictive of model performance for a specific community.

H2b: Errors made by the models are caused by geographic-specific content and language style. We perform a comprehensive manual analysis on the False Negatives made by the best model on the OLID dataset. The results are summarized in Fig-

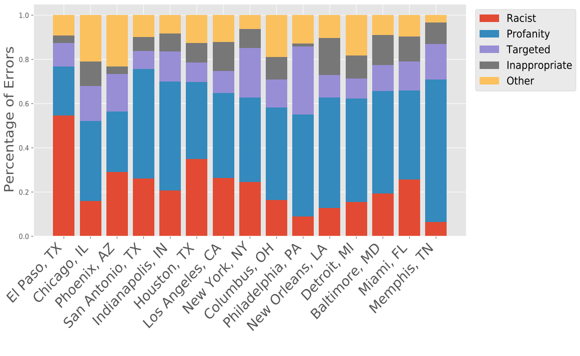

Figure 2: Category Percentages of False Negative Predictions per City.

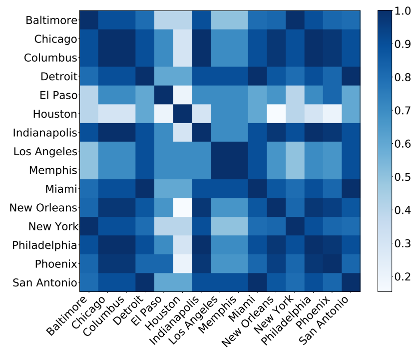

Figure 3: Model accuracy correlation between each pair of cities in GeoCOVID.

ure 2. Specifically, we categorized false negatives into four major categories: Racist, contained profanity, was a target attack on an individual, or was inappropriate (sexual references, insensitive jokes). 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). 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. Given the increase in border-related topics, this type of error is highly location-specific. Furthermore, non-location-specific errors include compound curse words and morphological variants of curse words that were a major cause for false negatives in multiple regions. 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 fucked, shits, damnit, fucks, phucking, effing, hoes, mothafucka, biatches.

5.3 Geographic Similarities

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

Methods. To answer this question, we analyze the performance of the models trained and described in Section 5.2. Specifically, we compare the PCC between the Accuracy of each applied for every pair of cities. Intuitively, if the rank of each model for New York based on Accuracy is the same as the rank of each model applied to Phoenix, then the correlation would be one. The more differences in rankings the lower the performance. In this experiment, we rank every model along with the variants of models (i.e., each model trained with different word embeddings listed in the Appendix are treated as independent models).

H3: The best model for each location is not the same. The results of the correlation analysis are shown in Figure 3. 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 than other cities with the exception of a few (e.g., Los Angeles). However, on further inspection, general architecture performance seems to be relatively similar across cities, e.g., the LSTM model is the best on OLID dataset and for most cities. Much of the variation comes from hyperparameter choice , or more specifically pretrained embedding choice(with more than 10% in Accuracy between the best and worst embeddings). This suggests that choosing the best hyperparameters based on a small subset of data is not optimal for each community. An interesting future research question would be 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?

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 community-driven NLP, where the impact and performance of NLP models are analyzed for specific communities, or even micro communities within a city. Moreover, finding communities that models perform poorly on can also provide unique testbeds as "hard test cases" similar to recent work on adversarial examples (Zhang et al., 2019).

References

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, and Lyle Ungar. 2020. Toward micro-dialect identification in diaglossic and code-switched environments. *arXiv preprint arXiv:2010.04900*.

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.

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.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. *arXiv preprint arXiv:2010.06595*.

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.

Jane Dixon and Nicky Welch. 2000. Researching the rural–metropolitan health differential using the 'social determinants of health'. *Australian Journal of Rural Health*, 8(5):254–260.

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.

Mark S Eberhardt and Elsie R Pamuk. 2004. The importance of place of residence: examining health in rural and nonrural areas. *American journal of public health*, 94(10):1682–1686.

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.

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.

Mihaela Gaman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer, et al. 2020. A report on the vardial evaluation campaign 2020. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*. International Committee on Computational Linguistics.

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.

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.

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.

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.

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

Leena Lulu and Ashraf Elnagar. 2018. Automatic arabic dialect classification using deep learning models. *Procedia computer science*, 142:262–269.

Brandon Lwowski and Peyman Najafirad. 2020. Covid-19 surveillance through twitter using self-supervised learning and few shot learning.

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.

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.

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.

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.

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.

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.

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 Conference of the Association for Computational Linguistics*, pages 1668–1678.

Yves Scherrer, Nikola Ljubešić, et al. 2021. Social media variety geolocation with geobert. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*. The 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.

Karly B Smith, John S Humphreys, and Murray GA Wilson. 2008. Addressing the health disadvantage of rural populations: how does epidemiological evidence inform rural health policies and research? *Australian journal of rural health*, 16(2):56–66.

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.

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.

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.

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.

A Appendix

A.1 Word Embeddings

In Table 6, 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 300. 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

Naive Bayes:
- alpha : 1.0
- fit prior: False

Linear SVM:
- penalty: l2
- C: 1.0

CNN:
- max words: 10000
- max sequence length: 125
- drop: 0.2
- 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: 0.2
- batch size: 128
- epochs: 30
- num filters: 512
- hidden layers: 1
- early stopping: 5 iterations

BiLSTM:
- max words: 10000
- max sequence length: 125
- drop: 0.2
- 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 : 0.2
- max length : 128
- epochs : 50
- batch size : 64
- fine tuned : after 5 epochs
- early stopping : 5 iterations

A.3 OLID Results

We report the OLID results for each model (NB, Linear SVM, CNN, LSTM, BiLSTM, and BERT) in Table 8. 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 (NB and 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.

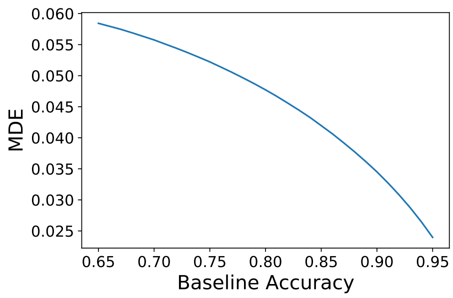

Figure 4: MDE given different baseline accuracy assumptions and a power of 80%.

Next, in Table 7 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 4, 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 F1 Scores per City

In Table 9, we reproduce Table 4 from the main component of our paper, but for F1 scores instead of Accuracy. Note that power analysis is possible for F1 score (Card et al., 2020), but many more assumptions are required. Based on our preliminary analysis, we found that significant differences can range between 2% and 5% depending on the assumptions. Overall, we we find that the CNN results in the best performance on average. Likewise, we find the best results on text in Baltimore, Detroit, and Philadelphia. The worst results are found in Houston, Phoenix, and New York.

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 6: List of word embeddings we use in our experiments.

Word Embedding	F1	Accuracy
BiLSTM		
FASTTEXT_en_300	0.580	0.760
GLOVE_twitter_27B_100d	0.627	0.785
GLOVE_twitter_27B_50d	0.5834	0.764
GLOVE_wiki_42B_300d	0.645	0.793
GLOVE_wiki_6B_100d	0.600	0.771
GLOVE_wiki_6B_200d	0.605	0.778
GLOVE_wiki_6B_300d	0.631	0.783
GLOVE_wiki_6B_50d	0.586	0.768
GLOVE_wiki_840B_300d	0.631	0.787
W2V_GoogleNews	0.616	0.781
W2V_PMC	0.488	0.730
W2V_PubMed_PMC	0.514	0.738
W2V_PubMed	0.402	0.704
LSTM		
FASTTEXT_en_300	0.524	0.749
GLOVE_twitter_27B_100d	0.618	0.782
GLOVE_twitter_27B_50d	0.591	0.770
GLOVE_wiki_42B_300d	0.619	0.790
GLOVE_wiki_6B_100d	0.607	0.774
GLOVE_wiki_6B_200d	0.616	0.781
GLOVE_wiki_6B_300d	0.609	0.782
GLOVE_wiki_6B_50d	0.577	0.762
GLOVE_wiki_840B_300d	0.624	0.788
W2V_GoogleNews	0.602	0.779
W2V_PMC	0.456	0.720
W2V_PubMed_PMC	0.495	0.730
W2V_PubMed	0.348	0.701
CNN		
FASTTEXT_en_300	0.611	0.778
GLOVE_twitter_27B_100d	0.657	0.792
GLOVE_twitter_27B_50d	0.635	0.788
GLOVE_wiki_42B_300d	0.642	0.793
GLOVE_wiki_6B_100d	0.621	0.779
GLOVE_wiki_6B_200d	0.621	0.786
GLOVE_wiki_6B_300d	0.621	0.785
GLOVE_wiki_6B_50d	0.612	0.775
GLOVE_wiki_840B_300d	0.648	0.794
W2V_GoogleNews	0.638	0.789
W2V_PMC	0.520	0.738
W2V_PubMed_PMC	0.541	0.743
W2V_PubMed	0.461	0.718

Table 7: Word Embedding Performance for Deep Learning Models

	Prec.	Rec.	F1	Acc.
Random Baselines				
Stratified	.324	.348	.336	.553
Uniform	.321	.505	.392	.493
Prior	.000	.000	.000	.676
Machine Learning Models				
NB	.722	.250	.371	.720
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 8: OLID Results

	Bal	Chi	Col	Det	ElP	Hou	Ind	LA	Mem	Mia	NO	NY	Phi	Pho	SA	AVG
NB	.607	.481	.538	.624	.562	.512	.521	.528	.614	.650	.588	.420	.570	.438	.576	.548
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	.653	.604	.633	.702	.644	.576	.616	.642	.689	.659	.673	.587	.675	.593	.657	

Table 9: F1 score for each city.