Leverage NLP and Data Analysis on Public Tweet Sentiment and Crime Data

Yiwei Wang
Georgia Institute of Technology
Atlanta, Georgia, USA
ywang3607@gatech.edu

Xuhan Zhao
Georgia Institute of Technology
Atlanta, Georgia, USA
xzhao395@gatech.edu

Omar Abu-Rub
Georgia Institute of Technology
Atlanta, Georgia, USA
oaburub3@gatech.edu

KEYWORDS

datasets, crime rate, machine learning, tweet text tagging

ACM Reference Format:
Yiwei Wang, Xuhan Zhao, and Omar Abu-Rub. 2023. Leverage NLP and Data Analysis on Public Tweet Sentiment and Crime Data. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 PROJECT WEBPAGE

Link to project webpage:
https://github.com/sherryxzhao/CSE8803IUC/wiki

2 INTRODUCTION

Crime rate is a critical issue that affects the public well-being. Researches have shown that social media platforms such as Twitter can be a practical tool for crime surveillance and monitoring. With the growing impact of social media, it is plausible that a correlation exists between crime rate and public sentiment. In this project, we collected Twitter data and crime data for the year of 2022 in Los Angeles and Boston with the goal to leverage natural language processing (NLP) and data analysis techniques to understand the connection. We hypothesized that there is a positive correlation between the frequency of negative sentiment and crime-related tweets and city crime rates which would provide evidence to support that social media data can be used to understand and predict criminal activities.

3 PROBLEM DEFINITION

Crime is a pervasive social problem that affects communities all around the planet. The impacts of crime ranges from an individual to community up to a national level. Amongst the detrimental impacts of crime are physical harm, psychological trauma, economical impact, social inequality, political instability, nation's international reputations, and the reduction of quality of life amongst other impacts. Traditional approaches to reduce crime include reactive measures such as policing. Research trends indicate that there is a growing interest in implementing proactive measures to predict and prevent crimes before occurring. Research on crime rate prediction entails the use of statistical models, machine learning algorithms, geospatial analysis, and social media analysis. This project will use such models on data collected from Twitter, to supplement traditional crime prediction methods. This project will explore the potential of using Twitter data for crime rate prediction, including its strengths and limitations, as well as its implications for crime prevention and control.

The ability to accurately predict the crime rate of a particular region plays a significant role in improving public safety, resource allocation, crime reduction, economic development, and social justice. Crime prediction helps law enforcement agencies and local governments take proactive measures to prevent crime, encouraging the implementation of crime prevention programs and identifying high-risk areas that require more attention and resources. Allocation of resources could involve additional funding to law enforcement or investing in community programs. Amongst the most important impacts of reducing crimes on a national scale is economic development as high crime rates can deter investment and economic development in a region. Furthermore this has a rather negative cascading impact on the reduction of economic opportunities for a nation's residents which may lead to increased crime rates. Therefore, being able to predict crime allows governmental agencies to create a more conducive environment for economic growth.

Twitter is a social media platform that allows users to share messages with their followers, called "tweets", composed of up to 280 characters at a time. Twitter is commonly used for a variety of purposes, including sharing news and current events, engaging in political discussions, sharing information, connecting with experts, promoting businesses or products, and connecting with others who share similar interests. It is therefore undeniable that Twitter and other large scale social media platforms are a great source of large and diverse data that can be used for various implementations. Furthermore, with the help of Twitter API, one is able to access real time publicly accessible data at a very low cost to be used for different statistical and machine learning applications. With respect to latency of the data collection it can depend on the type and amount of data that the user would like to collect, however, a major advantage is that the data collection process can be automated. With that said a major risk with this implementation are concerns around the privacy of users and data protection. Most users of social media platforms may not be aware that their data is being used to proactively monitored suspicious behavior of which they may not have given their consent. In addition, this approach is limited as only a subset of the population use Twitter. With that being said, the Twitter based crime prediction should be used as an additional resource to the already existing traditional reactive approach and not the primary resource to reduce crime.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

More formally, we will model the relationship between Twitter textual data and crime rate as a supervised learning problem. We will extract features from textual data on a city basis (to be more specific, sentiment, irony tone, hate speech, offensive language, ongoing topics) as training features **X**. Our goal is to infer the crime rate in each city **Y** from the mentioned features. Since the crime rate is continuous, we will employ regression models and minimize the mean squared error to achieve our goals.

4 SURVEY OF RELATED WORK

4.1 Sentiment analysis of Twitter Data: A survey of techniques

The paper provides a thorough study of the different methods applied to sentiment analysis of Twitter data including lexicon-based, machine learning-based, deep learning-based, and hybrid methods [1]. It begins by defining sentiment analysis and outlining its significance for comprehending public opinion and customer input. Additionally, it talks about the difficulties with mood analysis, including data scarcity, irony, and sarcasm. The paper examines the various pre-processing steps, including tokenization, stemming, and stop-word removal, that are usually carried out on Twitter data prior to sentiment analysis. It also discusses various feature extraction methods and sentiment analysis assessment metrics. Overall, the paper provides a comprehensive overview of the state-of-the-art techniques and challenges in sentiment analysis of Twitter data.

4.2 The relationship between social media data and crime rates in the United States

The study looks into the connection between social media data and criminal statistics in the US. This paper examines data from a variety of sources, including police department records, FBI Uniform Crime Reports, and social media analytics [13]. With other potential influences on crime rates, such as socioeconomic status, demographics, and policing levels, the paper uses regression analysis methods to investigate the relationship between social media chatter about crime and official crime rates. The findings demonstrate a positive relationship between official crime statistics and social media crime posts. Additionally, there is a greater correlation between social media activity and property crimes than violent crimes. For application, the paper suggests that policymakers should think about incorporating social media data into their crime prevention strategies because it finds that social media data can be a useful source of information for predicting and understanding crime rates. Overall, the paper offers insight into the connection between social media chatter and crime rates in the United States and gives evidence for the usefulness of social media data in crime analysis overall.

4.3 Twitter Sentiment for Analyzing different types of crimes

This paper examines the use of Twitter sentiment analysis to measure public opinion on different types of crimes [9]. It looks at the importance of public opinion on the issue of crime, and then explores various ways in which Twitter can be used to measure public sentiment on various types of crimes, such as violent crime, property crime, and white-collar crime. The research analyzes the tweets using both lexicon-based and machine learning-based sentiment analysis techniques. The findings demonstrate that sentiment analysis methods based on lexicons and machine learning can both be successful in examining Twitter sentiment related to various kinds of crimes. The study also reveals that various types of crimes have different emotional tones in tweets with tweets related to white-collar crimes have a more neutral tone. Overall, the study shows how Twitter sentiment analysis can be used to analyze crime data and emphasizes the significance of taking emotional responses into account when analyzing how the general public views crime.

4.4 Mining Twitter data for crime trend prediction

This paper mines information from Twitter data, extracting trending terms, sentiment of Tweets, and temporal topics, to predict the crime trend [2]. One highlight of this paper is that it utilizes Latent Dirichlet Allocation (LDA) techniques to infer latent topics from Bag-of-Words representation of the document (instead of having a fixed vocabulary as topics). Aside from the information from textual data, the authors also use features like unemployment rate, number of tweets, day of week, and events. Their results using temporal topics as features are over 20% better than the baseline. However, the biggest weakness of this paper is that it turns the increasing/decreasing of the crime rate into a binary classification task - given that crime rate going up by 3% (statistically significant) is quite different from crime rate increasing by 0.1% (probably just because of randomness) - which makes the result less meaningful in applications.

4.5 Predicting crime using Twitter and kernel density estimation

This paper presents an important aspect when working with Twitter data: the linguistics on Twitter is quite different from the "relatively clean domain of general written English" which the NLP models are trained on [7]. Therefore, the authors employ the LDA techniques and implement a topic modeling specific to tweets, using a Twitter-specific tokenizer and part-of-speech tagger. At the same time, this paper provides a comparison between the traditional methods (hot-maps) and later social media approaches using social media. However, the author split certain areas into finer grids in order to perform binary classification on every single grid and use kernel density estimation to approximate the probability of crime - both tasks are computationally expensive and it's unlikely that the method will scale up well (to predict crime in more general cases like counties, cities, states).

4.6 Crime rate detection using social media of different crime locations and Twitter part-of-speech tagger with Brown clustering

This paper utilized sentiment analysis skills to detect the crime rate in different cities without explicitly measuring the topics [12]. One highlight of this paper is that the authors utilize Brown clustering to help with the data cleaning: Brown clustering will cluster different spellings with similar semantic meaning together, which facilitate the prediction by the sentiment analysis model.

4.7 Mining Social Behavioral Biometrics in Twitter

This paper studies the behavior of people in online social networks, a field called social behavioral biometrics [11]. The article focuses on analyzing behavioral patterns of Twitter users through their tweets, followers, and engagement with other users. The paper mentions and references how public social media data can be used to develop machine learning models for different applications and even used in accurately identifying individuals. This paper covers a highly relevant topic and demonstrates with a detailed methodology how Twitter can be used as a valuable resource but fails to mention security implications that may be caused by malicious people intending to identify and target certain groups of people. Furthermore, it should be noted that the accuracy of social behavioral biometric data can be affected by a variety of factors such as changes in user behavior, data sampling biases, and the use of automated tools.

4.8 TweetNLP: Cutting-Edge Natural Language Processing for Social Media

This paper presents an intergrated platform (Python library) for natural language modeling specifically for social media as a toolkit to perform NLP tasks such as sentiment analysis, hate speech detection, offensive language identification [4]. In addition, it provides vector representations (i.e. embeddings) for words and tweets so that users are able to apply the model for their own applications, such as similarity analysis or tweet retrieval. Although the paper doesn't propose any new architecture or approach to the problem, it integrates State-Of-The-Art social media NLP techniques into a single Python package. This paper surveys SOTA Models and popular tasks in social media NLP, which functions as an excellent reference for NLP professionals. At the same time, the intergrated platform provide convenience for layman since the models are already trained or fine-tuned and the interface is much more user friendly compared to researchers' codebase.

4.9 Twitter Topic Classification

This paper indicates that Latent Dirichlet allocation (LDA) and its variations like Twitter-LDA, SKLDA, has flaws when we need to assign multiple topics to a document [3]. This paper collected, annotated a Tweet Topic classification dataset and use the constructed dataset to evaluate the topic classification performance of multiple Language Models such as BERT. The authors presented TweetTopic, which is the first large-scale dataset for tweet topic classification, and formalize the classification as a natural language processing task. However, although the large language models have excellent transfer learning performance, they are not applicable to our tasks given that even fine-tuning large language model will require large amount of training data and we lack annotated crime-specific datasets.

5 PROPOSED METHODOLOGY/APPROACH

5.1 Algorithms/Techniques/Models to Use

In section 4, we covered some literature on using social media as an indicator for the crime rate. However, most of them are using social media textual information as auxiliary features in addition to traditional crime-related features like unemployment rate. However, based on our research, there are two weakness of using these features:

- It takes time to gather and report the statistics (i.e. the current statistics are often outdated) so it's challenging to do any real time analysis or prediction
- There are more or less historical bias in the statistics. As a result, training models with biased data will make the prediction even more biased

Therefore, we decided to only utilize the information scraped from Twitter since tweets can be scraped (even streamed) instantaneously and there's much less historical bias in the tweets. We will use SOTA Twitter language models mentioned in 4.8 to extract features from tweets, and then utilize the extracted features to predict crime rates.

For the textual data collection, we choose the snscrape Python package for the data collection process to scrape the tweets needed from the web. Snscrape enables filtering based on time and geological information (if the users choose to include such info in their tweets). We first process the collected dataset using TweetNLP to get the sentiment analysis, hate speech detection, and offensive language detection for each tweet collected, and then we will compute the percentage of each label on a daily basis (e.g. the percentage of tweets with negative sentiment on 2022/01/01). After that, we will use Latent Dirichlet Allocation from the Gensim [10] package to generate popular topics (and the weight for each topic) from the scraped dataset. To compute the topics for a single day, we concatenate all tweets collected on that day into paragraphs and train the LDA model using the list of paragraphs from dates in the training set and run inference on both training set and test set. We concatenate the features inferred by BERT-based pre-trained models and the features inferred by LDA model, and use the concatenated features as the training features in our study. Data from separate locations are treated as separate modeling task. For the training process, we start with the most basic models in Sklearn: LASSO regression, Gaussian Naive Bayes regression model, and K-Nearest Neighbors regression model.

5.2 Evaluation/Testing

The models are evaluated and optimized using a least square error metric as a loss function which is a commonly used evaluation metric for regression problems. It measures the sum of the squared differences between the predicted and actual values. The goal of training the model is to minimize the loss function through optimization algorithms like gradient descent. We use 80% data as training data and the other 20% for testing. The models and the associated hyper-paramters are evaluated using cross-validation. In the cross-validation process, we divide the data into training and validation sets and perform 5-fold cross validation using different partitions. After we fix our choice of hyper-parameter, we will evaluate the performance of model on the test set. The success of the method is measured by calculating the mean squared error (MSE) which measures the average squared difference between the target and the actual values. The smaller the MSE, then the better the method's performance.

For the sake of interpretability, we also evaluate the models using accuracy of prediction allowing different levels of errors. Specifically, we compute the accuracy when an error level of ϵ is allowed, which means the predictions is considered to be true if our prediction falls in the range of $[\mu - \epsilon * \mu, \mu + \epsilon * \mu]$ and false otherwise with the ground truth label being μ. We are using $\epsilon = 0.05, 0.1, 0.2$ for our study.

5.3 Datasets

To train our model, we will use the crime data to compute the crime rate (labels) on a city basis for the city of Los Angeles and Boston in the year of 2022. Our main training features will be features extracted from Twitter textual data. Given the fact that Twitter API only allows tweets up to a week, we choose snscrape package in Python to retrieve data in certain locations during our desired time span. Details of the datasets we use can be found below.

- Los Angeles crime data from 2020 to present [6] (175.4 MB, csv)
- Boston crime data from 2015 to present [5] (118.7 MB, csv)
- Snscrape to Twitter data [8] (4.6 MB, 5.28 MB, csv)

5.4 Data Collection

The crime rate datasets for the city of Los Angeles and Boston are downloaded directly from various government official open data websites in CSV formats. To collect the data, we searched for crime rate dataset online, downloaded the relevant datasets, and inspected them to ensure they contain the relevant data fields. All datasets are preprocessed in Python using Jupyter Notebook. We removed duplicate records, missing values, and collected all relevant information of the incidents of crime, specifically the type of crime and its latitude and longitude using Python pandas package.

The social media data are collected through the Python Snscrape package, which allows us to scrape tweets from web and filter based on time, location, and other features on Tweet fields. The public Twitter API is not used as it limits access to public tweets published within the last 7 days. We collected the first 100 tweets in each time frame per location. Since we are doing a city-based analysis, it means we collect the first 100 tweets (only from tweets posted with location / tweets posted by a user who included location information in profile) in each city every day corresponding to the crime dataset. We use the following filters in scraping process:

- location: within 10 km of the city
- language: English
- start date and end date
- minimum number of fav or retweet

We use the Python enumerate function to iterate through results returned by TwitterSearchScraper. We extract the 'rawContent' field from each collected tweet, append the entries into a list, and use the Pandas Python package to convert the collected tweets into csv file. The details of implementation can be found in ur GitHub repo [1] under /data_collection.

[1] https://github.com/sherryxzhao/CSE8803IUC

5.5 Ethical Concerns

Studying social media data partially involves human subjects except that we will only study on the data that are publicly available. Therefore, all data should be anonymized and people shouldn't be able to recover a specific user from the published results. At the same time, crime-related topics are always sensitive given that the study results could be harmful for certain areas or certain neighborhood in reputation or economics. Therefore, we decided to study the correlation between social media data and crime rate in different cities (instead of neighborhood level) although the crime datasets come with low-level geographic information.

There exists a limitation in the data collection process that lies in the fact that only around 1-2 percent of existing tweets are geotagged. Snscrape tries to overlook this limitation through analyzing the bios of the users. The reason for that is many users add their geographical location in the bio of their twitter account. Thus, the assumption made in the data collection process is that for some of the collected tweets it is assumed that the tweets are uploaded from the same location indicated in the user's profile. Although not a weak assumption, this assumption can be manipulated by malicious users who may seek to maliciously manipulate the data collection process to potentially cause detrimental social discourse.

6 EXPERIMENTS/RESULTS

Our experiments on models are designed to answer the following questions:

- Which feature correlate the most with the crime rate?
- What is the overall sentiment of public tweets in Los Angeles and Boston in 2022?
- Is there a correlation between the sentiment of public tweets and the crime rates in Los Angeles and Boston?

Details of your experiments, observations, findings. Make sure you also interpret and explain your observations.

After cleaning and preprocessing the crime data for each city for the year of 2022, we calculate the frequency of each crime category reported by the police by counting the number of incidents in each category. Then we run the frequency distribution on calculated data and plot them in the form of histograms for visualization, shown in figure 1 and figure 2. We perform data analysis on the crime data and the top 5 crimes in Los Angeles in 2022 with the most count are VEHICLE - STOLEN with 49316 incidents, THEFT OF IDENTITY with 42290 incidents, BATTERY - SIMPLE ASSAULT with 36052 incidents, BURGLARY FROM VEHICLE with 28370 incidents, and BURGLARY with 27978 incidents and the top 5 crimes in Boston in 2022 with the most count are INVESTIGATE PERSON with 15354 incidents, SICK ASSIST with 10680 incidents, M/V - LEAVING SCENE - PROPERTY DAMAGE with 8622 incidents, INVESTIGATE PROPERTY with 6818 incidents, and TOWED MOTOR VEHICLE with 6154 incidents.

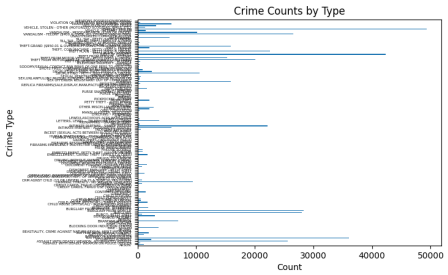

Figure 1: Histogram of crime in Los Angeles in 2022.

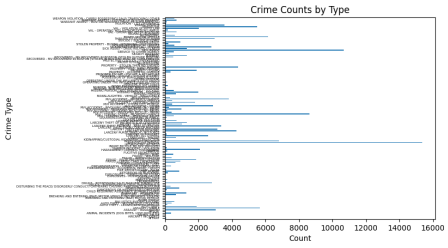

Figure 2: Histogram of crime in Boston in 2022.

To extract features from the textual data, we run inference with corresponding BERT-based classification models, a summary of statistics in Los Angeles is shown in Figure 3 and statistics in Boston is shown in Figure 4.

	positive	negative	neutral	non_irony	irony	non-hate	hate	non-offensive	offensive
count	292.000000	292.000000	292.000000	292.000000	292.000000	292.000000	292.000000	292.000000	292.000000
mean	0.433767	0.214795	0.351438	0.803253	0.196747	0.992637	0.007363	0.895548	0.104452
std	0.058235	0.051114	0.051239	0.041535	0.041535	0.008826	0.008826	0.034122	0.034122
min	0.260000	0.090000	0.210000	0.690000	0.100000	0.960000	0.000000	0.780000	0.020000
25%	0.390000	0.180000	0.317500	0.780000	0.170000	0.990000	0.000000	0.880000	0.080000
50%	0.440000	0.210000	0.350000	0.800000	0.200000	0.990000	0.010000	0.900000	0.100000
75%	0.470000	0.240000	0.390000	0.830000	0.220000	1.000000	0.010000	0.920000	0.120000
max	0.680000	0.470000	0.490000	0.900000	0.310000	1.000000	0.040000	0.980000	0.220000

Figure 3: Statistics of features from TweetNLP in Los Angeles.

	positive	negative	neutral	non_irony	irony	non-hate	hate	non-offensive	offensive
count	292.000000	292.000000	292.000000	292.000000	292.000000	292.000000	292.000000	292.000000	292.000000
mean	0.465218	0.193870	0.340912	0.786885	0.213115	0.995788	0.004212	0.926290	0.073710
std	0.071723	0.053068	0.057501	0.043948	0.043948	0.006237	0.006237	0.032175	0.032175
min	0.270000	0.060000	0.160000	0.630000	0.100000	0.970000	0.000000	0.790000	0.010000
25%	0.420000	0.160000	0.300000	0.760000	0.180000	0.990000	0.000000	0.910000	0.050000
50%	0.460000	0.190000	0.340000	0.790000	0.210000	1.000000	0.000000	0.930000	0.070000
75%	0.510000	0.230000	0.380000	0.820000	0.240000	1.000000	0.010000	0.950000	0.090000
max	0.660000	0.410000	0.510000	0.900000	0.370000	1.000000	0.030000	0.990000	0.210000

Figure 4: Statistics of features from TweetNLP in Boston.

The distributions for sentiment in LA and sentiment in Boston are relatively similar. There's slightly higher percentage of positive sentiment in Boston and slightly higher percentage of irony in Boston. However, the distribution for the LA labels (frequencies of crime) is significantly different from that of Boston, with a mean of 1725.2466 as shown in Figure 5 and a mean of 382.8356 for Boston labels as shown in Figure 6. Therefore, we train separate models (using the same architecture) since we are more interested in understanding the correlation between features and the labels, instead of predicting the right scale based on 1-bit encoding of the city.

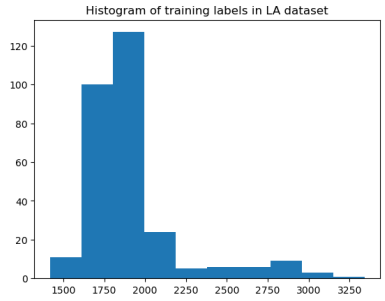

Figure 5: Histogram of training labels in Los Angeles dataset.

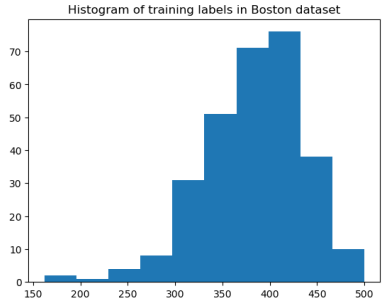

Figure 6: Histogram of training labels in Boston dataset.

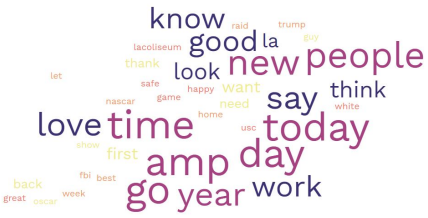

Figure 7: Word Cloud generated by LDA model.

We first run a simple Linear Regression model to see if there are any correlation between the features and the labels. A visualization of the coefficients is shown in Figure 8 and in Figure 9. For the TweetNLP features, we can observe that the percentage of hate speech is positively correlated with the crime rates and the percentage of non-hate speech is negatively correlated with the crime rates. Some features, like neutral and irony, doesn't seem to related to the labels much, and some coefficients are counter intuitive: for example, the percentage of offensive language has a negative correlation with the labels and the percentage of non-offensive language has a positive correlation with the labels.

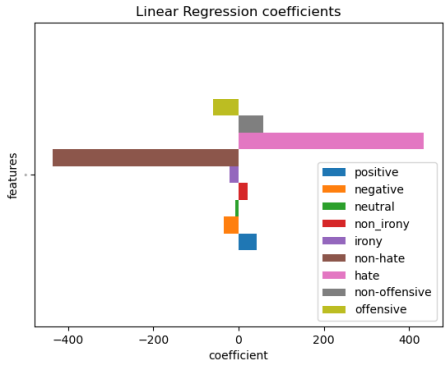

Figure 8: Coefficients for TweetNLP features from Linear Regression model.

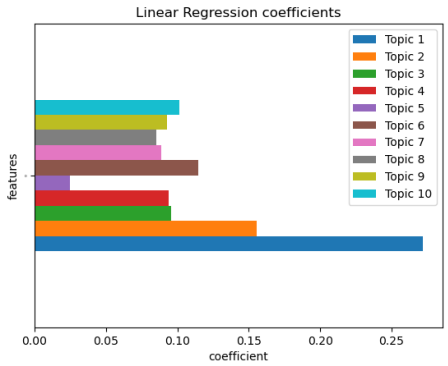

Figure 9: Coefficients for LDA topics from Linear Regression model.

We run experiment using datasets with Lasso Regression model, Gaussian Naive Bayes Regression model, and K-Nearest Neighbors classification model. The results are in table 1 and table 2.

Los Angeles Dataset	Accuracy with $\epsilon = 0.05$	Accuracy with $\epsilon = 0.1$	Accuracy with $\epsilon = 0.2$	MSE
Lasso Regression	0.54794	0.68493	0.87671	32567.11986
Gaussian Naive Bayes	0.52055	0.78082	0.87671	34938.24657
K-Nearest Neighbors	0.41096	0.68493	0.86301	34196.61432

Table 1: Table of results from three models on Los Angeles dataset.

Boston Dataset	Accuracy with $\epsilon = 0.05$	Accuracy with $\epsilon = 0.1$	Accuracy with $\epsilon = 0.2$	MSE
Lasso Regression	0.23287	0.58904	0.90410	2340.16571
Gaussian Naive Bayes	0.43836	0.71233	0.91781	38243.83562
K-Nearest Neighbors	0.19178	0.56164	0.89041	2597.03342

Table 2: Table of results from three models on Boston dataset.

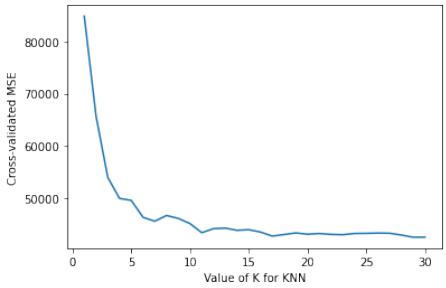

Figure 10: Hyper-paramter choice for KNN model.

7 CONCLUSION AND DISCUSSION

Although the results from the experiments don't entirely match our expectation and is far from, our study discovers the relationship between social media textual data and the crime rate. Moreover, future work can be conducted to further explore this relationship by improving the quality of data and increasing the amount of data used for the study.

gathering more data is definitely conducive to the performance of the model: we observe significant performance increase (abount 0.1 increase for accuracy within 5%) when we move from 80% training data during 5-fold cross validation to utilizing all of our training data, i.e. the performance of simple models (LASSO regression, Gaussian Naive Bayes, and KNN) has not saturated yet. When more data is gathered, we expect to see even better results by simply training on more data or moving to more complex models like neural networks. In addition, the model is able to utilize more powerful features like tweet embeddings and continuous representation of the topics when we have enough training data. Furthermore, the parameters in pre-trained model can also be fine-tuned specifically for our downstream regression task.

However, more data gathered brings computational overhead: time needed for BERT-based model inference is nontrivial. The runtime grows linearly in the number of days, number of tweets per day, and the number of features from pre-trained models. In addition, since the authors was using Google Colab to scrape from Twitter, the file writing and reading speed is also not negligible. It's recommended to scrape the data on local machine and then take advantage of GPUs to run the inference.

To achieve better results, there are some approaches to potentially improve the quality of scraped Twitter data

- Gather more randomized tweets instead of gathering the first 100 tweets on each day based on filter
- Gather tweets targeting at certain topics like crime-related, political, or other topics might be related to crime rate
- Scrape tweets on more days to have more training data

Moreover, auxiliary pre-trained NLP models could be employed to provide better features other than sentiment or offensive language, for example, classifier on whether a tweet is related to crime incident or on the category of crime will boost the performance of the crime rate prediction. Additionally, LDA topic extraction could be replaced by transformer-based model classifier fine-tuned on labelled tweets -> crime topic datasets.

REFERENCES

[1] Vishal A. and S.S. Sonawane. 2016. Sentiment Analysis of Twitter Data: A Survey of Techniques. *International Journal of Computer Applications* 139, 11 (apr 2016), 5–15. https://doi.org/10.5120/ijca2016908625

[2] Somayyeh Aghababaei and Masoud Makrehchi. 2018. Mining twitter data for crime trend prediction. *Intelligent Data Analysis* 22, 1 (2018), 117–141. https://doi.org/10.3233/ida-163183

[3] Dimosthenis Antypas, Asahi Ushio, Jose Camacho-Collados, Leonardo Neves, Vítor Silva, and Francesco Barbieri. 2022. Twitter topic classification. *arXiv.org* (Sep 2022). https://arxiv.org/abs/2209.09824

[4] Jose Camacho-Collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa Anke, Fangyu Liu, Eugenio Martínez-Cámara, and et al. 2022. TWEETNLP: Cutting-edge natural language processing for social media. *ACL Anthology* (2022). https://aclanthology.org/2022.emnlp-demos.5/

[5] Boston Police Department. 2023. *Crime Incident Reports.* Retrieved March 27, 2023 from https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system

[6] Los Angeles Police Department. 2020. *Crime Data from 2020 to Present.* Retrieved Febuary 22, 2023 from https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8

[7] Matthew S. Gerber. 2014. Predicting crime using Twitter and kernel density estimation. *Decision Support Systems* 61 (2014), 115–125. https://doi.org/10.1016/j.dss.2014.02.003

[8] JustAnotherArchivist. 2018. *snscrape: A social networking service scraper in Python.* Retrieved Febuary 23, 2023 from https://github.com/JustAnotherArchivist/snscrape

[9] Boppuru Rudra Prathap and K. Ramesha. 2018. Twitter Sentiment for Analysing Different Types of Crimes. In *2018 International Conference on Communication, Computing and Internet of Things (IC3IoT).* 483–488. https://doi.org/10.1109/IC3IoT.2018.8668140

[10] Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic* 3, 2 (2011).

[11] Madeena Sultana, Padma Polash Paul, and Marina Gavrilova. 2014. Mining Social Behavioral Biometrics in Twitter. *2014 International Conference on Cyberworlds* (2014). https://doi.org/10.1109/cw.2014.47

[12] Thanh Vo, Rohit Sharma, Raghvendra Kumar, Le Hoang Son, Binh Thai Pham, Dieu Tien Bui, Ishaani Priyadarshini, Manash Sarkar, and Tuong Le. 2020. Crime rate detection using social media of different crime locations and Twitter part-of-speech tagger with Brown clustering. *Journal of Intelligent Fuzzy Systems* 38, 4 (2020), 4287–4299. https://doi.org/10.3233/jifs-190870

[13] Yan Wang, Wenchao Yu, Sam Liu, and Sean D. Young. 2019. The Relationship Between Social Media Data and Crime Rates in the United States. *Social Media + Society* 5, 1 (2019), 2056305119834585. https://doi.org/10.1177/2056305119834585 arXiv:https://doi.org/10.1177/2056305119834585