Project Proposal

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. Project Proposal. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 SURVEY OF RELATED WORK

1.1 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.

1.2 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.

1.3 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 [10]. 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.

1.4 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, weather, 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.

1.5 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

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

clean domain of general written English" which the NLP models are trained on [6]. 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).

1.6 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.

1.7 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.

2 PROPOSED METHODOLOGY/APPROACH

2.1 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.

More formally, we will model the relationship between Twitter textual data, information like weather, unemployment rate, etc, and the crime rate as a supervised learning problem. We will extract features from textual data on a county basis (e.g. sentiment, ongoing topics) and combine the features from tweets with features from other datasets (e.g unemployment rate, population) as training features $\mathbf{X}$. Our goal is to infer the crime rate in each county $\mathbf{Y}$ from the mentioned features. Since the crime rate is continuous, we will employ regression models to achieve our goals.

2.2 Algorithms/Techniques/Models to Use

For the data collection, we plan to use 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 will first process the collected dataset using Twitter-specific Part-of-Speech Tagger and Brown Clustering to "translate" the tweets into something interpretable for standard pretrained NLP models. After that, we will use Latent Dirichlet Allocation from the Sklearn package to generate popular topics (and the probability of each topic) from the scraped dataset. We will also use Sentiment Analysis results from the HuggingFace package as one of our features.

For the training and testing process, we will start with the most basic models in Sklearn: linear models like logistic regression, ridge regression, and LASSO. Depending on the performance of these basic models, we may add complexity to our model, moving to neural networks for better performance, or trying KNN, Naive Bayes models for better interpretation.

2.3 Evaluation/Testing

The method will be evaluated 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. The method will be tested using cross-validation which is used to assess performance on an independent dataset. We will use 80% data as training data and the other 20% for testing. In the cross-validation process, we will divide the data into training and validation sets and perform k-fold cross validation using different partitions. The success of the method will be 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.

2.4 Datasets

To train our model, we will use the crime data and the population data from the census to compute the crime rate (labels) on a county basis (data are mapped to county by the zip code attribute). 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 will use snscrape package in Python to retrieve data in certain locations on our desired time span. We will also use features that are found to be useful in previous papers like the Unemployment Data and weather data as auxiliary sources. If we want to include more features in our model, we can find more data specific to Los Angeles in LOS ANGELES OPEN DATA. Details of the datasets we use can be found below.

- Los Angeles Crime data [5] (170.9 MB, csv)
- Snscrape to Twitter data [7] (txt)
- Los Angeles Unemployment data [8] (csv)
- Population data from census [3] (14 KB, xlsx)
- Weather data [4]
- LOS ANGELES OPEN DATA [9]

2.5 Accomplishment/Goal

By the end of the semester the team expects to evaluate different algorithms/models and complete the analysis on the LA city dataset. Ideally, the team expects to complete a write up that is of quality to be published in a scientific conference. In the case that the students reach an impedance in achieving the desired outcome, the team expects to gain relevant experience with data collection using API's, data processing/feature extraction, and machine learning result validation using multiple models/algorithms.

3 EXPECTED TIMELINE AND TEAM MEMBER CONTRIBUTIONS

We are a team of three "horizontal" students. Yiwei Wang is a first-year Masters Computer Science student with concentration in Machine Learning. Xuhan Zhao is also a first-year Masters Computer Science student specializing in Machine Learning. Omar Abu-rub is a third-year Electrical Engineering PhD student with research interests in Machine Learning implementation in power systems.

As a team, we agree to divide the work equally among all three of us and contribute equally to the tasks assigned to us. We plan to collaborate and communicate effectively to ensure that all tasks are completed in a timely manner.

A breakdown of the project involves the following steps.

3.1 Data collection from twitter (1 week, til March 3rd)

This involves gathering tweet data using snscrape package and collecting public datasets such as surveys and online public records.

3.2 Data preprocessing (1 week, til march 10th)

The collected data needs to be preprocessed and cleaned to remove any errors, invalid, or missing values. This involves filtering, aggregating, and transforming the data into a format suitable for training and analysis.

3.3 Build model and train model (2 weeks, til March 24th)

This involves applying various statistical and machine learning models to analyze the data and build models that can make predictions and identify trends.

3.4 Results analysis/visualization (2 weeks, til April 7th)

The results of the models will be visualized using various visualization tools like charts and graphs to present the data in a way that is easy to understand.

3.5 Evaluation of methods (2 weeks, til April 21st)

The performance of each model will be evaluated using least squares method as a loss function, cross-validation, and mean squared error. Future improvement or directions may be proposed.

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] US Census Bureau. 2023. *County Population Totals: 2020-2021.* Retrieved January 25, 2023 from https://www.census.gov/data/tables/time-series/demo/popest/2020s-counties-total.html

[4] The Weather Channel. 2023. *National and Local Weather Radar, Daily Forecast, Hurricane and information from The Weather Channel and weather.com.* Retrieved Febuary 23, 2023 from http://www.weather.com

[5] 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

[6] 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

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

[8] US Bureau of Labor Statistics. 2022. *Local Area Unemployment Statistics.* Retrieved December 22, 2022 from https://www.bls.gov/lau/

[9] City of Los Angeles. 2023. *DataLA: Information, Insights, and Analysis from the City of Angels: Los Angeles - Open Data Portal.* Retrieved Febuary 23, 2023 from https://data.lacity.org/

[10] 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

[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