Final Report: Crime Prediction Model using Sentimental Analysis and Historical Crime Data

~ Varun Singh, Yash Mathur, Bhavya Khilrani, Siddhartha Dheer

Introduction

Public safety is a growing concern in today's society, with crime prevention playing a crucial role in enhancing it. Traditionally, crime prevention has relied on geographic methods such as hotspot mapping and analyzing past crime rates in specific locations. However, these approaches often reinforce data bias from over-policed areas and fail to account for fluctuations in social dynamics over time. In response, we propose a model that combines crime statistics with social media sentiment to improve the accuracy of early crime detection and prevention. By integrating historical crime data with real-time public sentiment, this approach offers greater adaptability to emerging behavioral patterns, enhancing the model's predictive capabilities.

Problem Definition

Traditional crime prediction models suffer from Bias through over – policing, lacks real time data adaptation, as well as neglect non-crime events. Our solution aims to address this through integration of sentiment data from social media.

Literature Review

Existing crime prediction models are more or less dependent on historical crime incidence data and spatial models like kernel density estimation (KDE) or heat maps. Such systems are typical for police departments where technologies such as PredPol, elaborating future crime anticipating settings rely on the analysis of prior patterns [5], [10]. As seen above, these traditional models do work but have huge drawbacks. One concern is that the database that existing models rely on contains a certain prejudice. This means that high policing in particular neighborhoods has a feedback cycle whereby the police end up recording more incidents in crime hotspots merely because of increased patrols, perpetuating biased policing recommendations [1], [3], [8], and [11] all highlight how biased policing practices and faulty data in predictive systems can reinforce discriminatory outcomes through harmful feedback loops that disproportionately impact minority communities, raising critical concerns over data integrity, machine learning model biases, and their real-world implications.

In addition, these models have no forms of dynamic data integration, such as sentiment obtained from Twitter or other social media. Opinions expressed in tweets can become early warning signals of social unrest in societies [2], but as of now, this information is hidden. For instance, while social unrest is going on, higher negative words on sources such as Twitter may predict actual incipient crime, allowing the police to act early [6]. Traditional models depend on historical data, and therefore they are not able to react to changes in public opinion or economic status as quickly and accurately as the system used here.

Another formal constraint is that current models neglect non-crime incidents and shifting of mass attitudes. Such crimes have cycles that are hard to model successfully with the old theories and paradigms; changes in the economic climate, political changes, changing demographics, and communal dissatisfaction are some factors that bring about changes in crime cycles [4], [8], [11]. All three papers serve a similar purpose of efficient policing through new approaches and methodologies. While [11] approach this using Machine Learning techniques by studying diverse datasets and challenges in crime prediction, [8] integrate temporal dynamics within a Spatio-temporal kernel density estimation framework to improve predictive crime hotspot mapping. Meanwhile, [4] focuses more on data-driven forecasting with due attention to geographical scale and spatial units. Collectively, their work attempts to enhance the accuracy to predict crime prevention through superior modeling: either Machine Learning techniques, Spatio-temporal analyses, or forecasting techniques. Some systems attempt to integrate socio-economic dimensionality and metrics, albeit ineffectively, and even fewer are capable of meeting sentiment-based dimensions that indicate the dynamics within communities [7].

Proposed Methods

Intuition

The hybrid model here attempts to provide a transformative approach towards crime prediction by combining real-time sentiment analysis from social media into historical data on crime occurrences. This would be a dynamic and adaptive system to capture conjunctions of patterns and social unrest, meeting the limitations of traditional crime prediction works. Instead of using static historical data and spatial patterns, we used community sentiment as a leading indicator of crime trends. Such ability would allow for the early detection of hot spots and social unrest, diminishing predictability bias imposed by over-policed areas while allowing for interventions in good time.

Detailed Description of your approaches

This model was developed with Playwright and other libraries to automate and streamline the extraction of social media data. Playwright was used to programmatically interact with Twitter's web interface, allowing for the retrieval of tweets that contained keywords related to crime and associated with specific geolocation data. This approach was viable for circumventing issues presented in the API as well as being extensible. PRAW was used to access the posts and comments on Reddit, from where crime-terms discussions could be identified. These tools facilitated scalable data collection and ensured compatibility with preprocessing workflows. We collected over 10,000 data points from Twitter and approximately 20GB of Reddit data, filtered them using crime-related keywords and processed them for analysis. Subsequently, we drew historical crime data from publicly available repositories, including Kaggle and state-level datasets, with efforts made to align this data with the regions analyzed in the sentiment analysis.

The preprocessing phase for social media data encompassed tokenization, stop-word removal, and lemmatization. To provide contextual insight, keywords related to crime were tagged with a 'crime_type' feature, and geolocation data was standardized at the state level using a geolocation dictionary. Similarly, to ensure consistency between data from different states, the data for all states were normalized with select features and then stratified and merged for further analysis. Hereafter, historical crime data underwent cleaning and normalization to ensure consistency with social media data formats, particularly in the categorization of crime types. These steps established a uniform dataset, ready for analysis. The following libraries were employed to support the data collection and processing pipeline: asyncio, which enabled asynchronous programming for concurrent task execution; nest_asyncio, which allowed nested event loops for seamless integration in interactive environments like Jupyter notebooks; and hashlib, which generated unique hashes for tweets to prevent duplicate entries during scraping. Together, these tools facilitated efficient, scalable, and reliable data extraction.

Sentiment analysis was performed on the gathered data using established natural language processing libraries, including NLTK and TextBlob, which assigned polarity scores ranging from -1 to 1 to classify posts as negative, neutral, or positive. We developed a custom scoring mechanism to enhance the analysis, assigning severity scores to each data point based on crime type and sentiment polarity. This was achieved through a weighted summation approach, where weights reflected the severity of crimes and the intensity of sentiments. This methodology allowed for the creation of a nuanced scoring system capable of contextualizing and comparing crime severity across different states.

For example, crimes that pose a significant threat to human life, such as murder, sexual assault, rape, homicide, and kidnapping, were assigned scores between 8 and 10 due to their high severity. Comparatively, crimes like armed robbery, burglary, and vandalism were scored in the range of 5 to 8. Non-violent crimes such as shoplifting, identity theft, and fraud received lower scores between 3 and 5. It is important to note that extreme and highly sensitive crimes, such as terrorism and mass shootings, were intentionally excluded from the dataset to maintain ethical considerations and avoid amplifying distressing

events. Similarly, minor crimes, such as petty theft or littering, were also excluded to focus on more impactful patterns in public sentiment and crime severity. This weighted scoring ensures that the model accurately reflects the seriousness of crimes being discussed while providing a comparative scale for analysis.

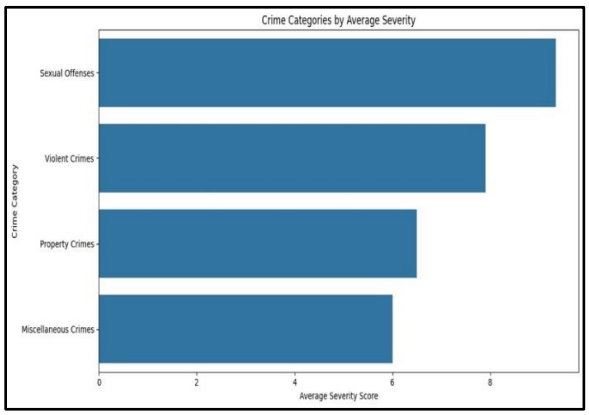


fig.(a)

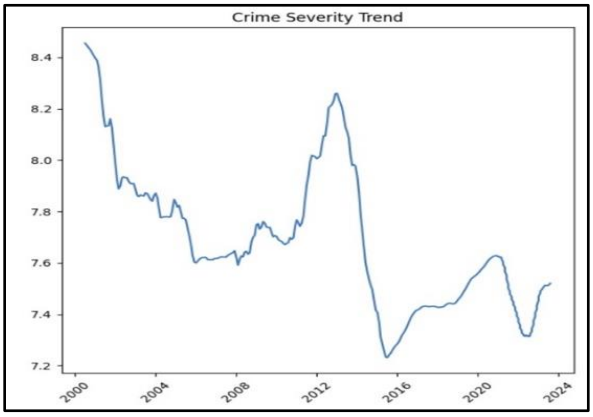
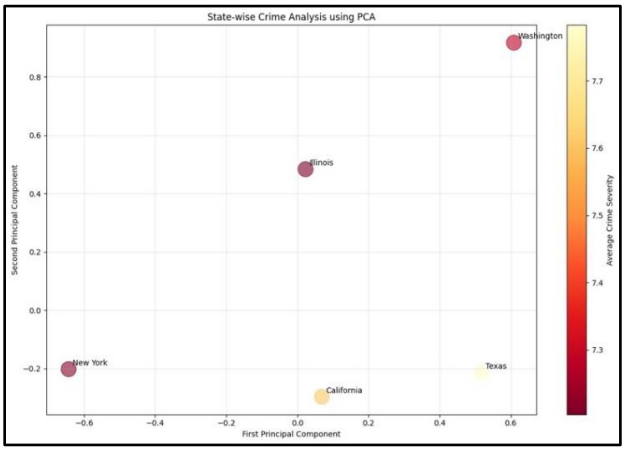


fig.(b)

Fig(a) illustrates the average severity scores assigned to various categories of crime. The most severe ratings were assigned for sexual offenses, followed by violent and property crimes that were considered to be moderate in their severity level. Miscellaneous crimes, such as minor offenses, scored the least on average. Thus, this classification is able to show the model differentiating between the impacts of crime, permitting a more nuanced severity appraisal.

Subsequently, **Fig(b)** presents how the average severity scores of crimes change over time from the years 2000 to 2024. As depicted in the figure, the average crime severity score fluctuates and reaches its peak value near 2012 before drastically decreasing starting from post-2014. The years with smaller peaks and slight declines after the year 2020 indicate the trend of crime severity as well as its impact on public safety. This study shows how the model can adapt to changing crime severity trends over time so that timely interventions can take place.



The integration of sentiment analysis with historical crime data culminated in the computation of safety scores for each state. Classification models such as PCA and Random Forest were employed to predict crime trends, incorporating temporal features to ensure adaptability to recent conditions. This hybrid approach enabled the model to dynamically adjust to changes in sentiment and historical patterns, thereby improving its predictive accuracy. The proposed method introduces several innovative elements. By incorporating live sentiment data, the model achieves real-time adaptability, addressing the static nature of

traditional approaches. The integration of community sentiment mitigates biases inherent in historical data, offering a balanced perspective. Additionally, the use of temporal dynamics ensures that predictions reflect current conditions, enhancing the relevance and accuracy of the model. Furthermore, the reliance on social media sentiment fosters a direct understanding of public concerns, strengthening community-police relations and contributing to proactive crime prevention strategies.

Experimental testbed

The test bed consists of data collection, preprocessing workflows, a sentiment scoring system, principal component analysis, and accuracy evaluation methods. Data from Twitter and Reddit were collected via tools and libraries such as PRAW and Playwright exemplifying community inclusion based on discussions around crime. Data preprocessing comprises cleaning, geolocation tagging, appropriate tagging of crime type by leveraging libraries such as pandas to manipulate the dataset by extracting location data and crime type from data points collected from twitter, reddit and historical crime descriptions. This step was followed up by allocating weights to crime type which was used to calculate severity scores. Sentiment Analysis is performed using NLP libraries such as NLTK, TextBlob. Further, we performed principal component analysis over historical data to classify state-wise crime severity levels. Our final model was evaluated against crime rates assigned to each state in consideration by Wikipedia and other credible sources.

Detailed description of the experiment

Our approach, being a hybrid of sentimental analysis and classification techniques, results in crime severity scores. We implemented sentimental scoring model on social media data from twitter and reddit into "positive," "neutral," or "negative", while assigning scores based on the sentimental polarity. These scores were further updated to quantify crime intensity by incorporating weights assigned to each crime. The accuracy of our sentimental analysis model was determined by manually labelling the dataset and calculating the accuracy score, which was 87% suggesting that the model captures community sentiment reliably. This accuracy score acts as a significant indicator of sentimental polarity, providing real-time indication of unrest. Thus, adding dynamic responsiveness to the model. The model initially sets a base for the assimilation of real-time data into the crime prediction system, marking a qualitative jump from traditional models restricted to only historical data.

Next, we performed classification modelling such as principal component analysis and random forest on historical crime data. We performed data cleaning by merging multiple data files into one, dropping irrelevant columns, and removing NA values. We then manipulated data to extract crime type from description of reports into primary and secondary crime types, for instance, primary category of ‘Violent Crime’ comprises secondary categories such as ‘Murder’, ‘Homicide’, ‘Assault’, ‘Battery’ and ‘Robbery’. Following data cleaning and preprocessing, we assigned a severity rating to each crime subcategory based on the rationale that a crime which pose significant threat to human life have been allocated more weight when compared to other kinds of crime. After assigning severity scores we analyzed crime trends over time and across five states. We then performed principal component analysis and random forest to identify relevant features and model severity score respectively. This element creates a really strong backbone in assessing safety at the state level, enabling a traditional yet robust form of analysis of criminal trends.

Lastly, we integrated the results from sentimental analysis with results from historical data modelling to develop a cumulative severity score by averaging out scores from both the models and assigning safety rankings to each of the five states. The combination of both the models resulted in crime severity rankings for five states, where higher scores represent sever crimes. We validated the results from our analysis by comparing the ranking of states based on scores generated by the model against states rankings given by Wikipedia, USA News and Numbeo based on their proprietary crime indexes. The combination of sentiment analysis and historical data classification offers multiple enhancements when compared to traditional models. This hybrid model provides enhancements by identifying patterns in social media posts indicative of any underlying unrest or concerns in the communities. This approach provides prompt warnings

regarding possible crime trends and hotspots, thereby improving model's predictive competencies. This integration of real-time community data and historical data showcase basis a more dynamic crime prediction system.

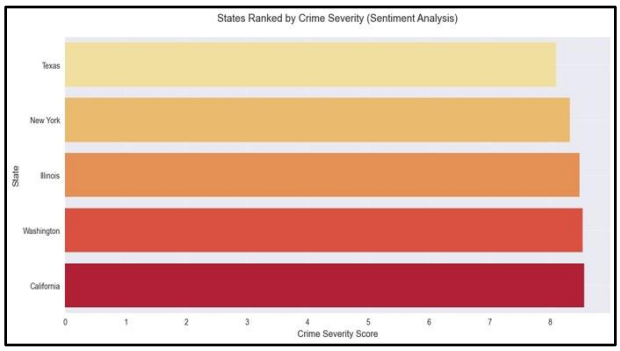
Conclusions and Discussion

This project successfully demonstrates a hybrid approach to crime prediction by integrating real-time social media sentiment analysis with historical crime data. The primary objective was to address the limitations of traditional models, such as bias in over-policed areas, reliance on static historical data, and lack of adaptability to societal changes. By combining these two data streams, our model offers a dynamic, adaptable, and proactive framework for predicting crime trends and identifying potential crime hotspots.

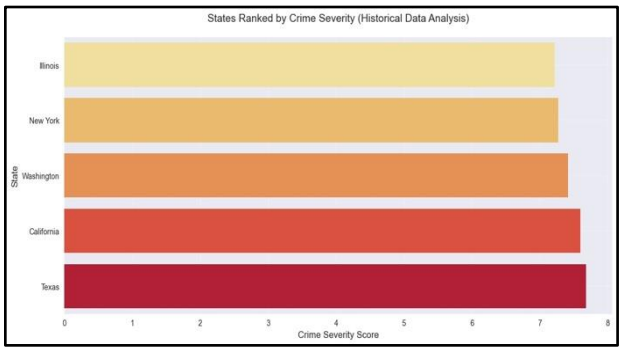
Key Results and Impacts

It was successful, estimating that through filtering 10,000+ tweets and pulling down 20-Gb of data from Reddit and Twitter with crime-related keyword terms, it acquires applicability. With the polarity scores from NLTK and TextBlob, one could gauge whether the mention had a positive or a negative stance. We can further classify this by providing severity score: murder, sexual assault, and homicide scored high on severity, 8-10, while fraud and shoplifting were rated much less, 3-5.

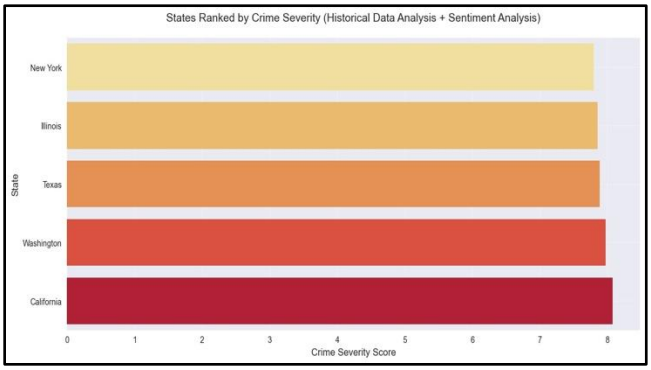
Accuracy tests were performed to validate the model's robustness and thereby confirm the reliability of the system in identifying community concerns and trends; achieved accuracy rate of 87% for sentiment analysis. Principal component analysis (PCA) and Random Forest Techniques were also integrated to analyze predictive nature of crime severity across regions. It generated cumulative safety scores that ranked states comparatively and presented actionable insights that can inform stakeholders.



(Sentiment Analysis)



(Historical Analysis)



(Sentiment + Historical)

Project Results	Countries	Severity Score
Scores From Sentiment Analysis	California	8.5610
	Illinois	8.4848
	New York	8.3250
	Texas	8.0943
	Washington	8.5353
Scores From Historical Data	California	7.5944
	Illinois	7.2171
	New York	7.2669
	Texas	7.6804
	Washington	7.4166
Averaged Score from both layers	California	8.0777
	Illinois	7.8509
	New York	7.7960
	Texas	7.8873
	Washington	7.9759

Limitations

We faced several limitations throughout the course of the semester for instance, the state level, historical crime data was unavailable, and most sources provided city specific data. For example, datasets for Dallas and Austin were available, but not for the entire state of Texas. The number of city level datasets to process was both time consuming and computationally intensive. Alongside that, many cities don't publish data online, nor do they encrypt public information for security reasons, making it available only to those having the 'proper' clearance. Furthermore, the data from social media platforms such as Twitter and Reddit are inherently noisy and require great preprocessing in order to filter out the irrelevant data. Often restricted by free-tier APIs to the number of rows that could be extracted and unable to perform any analysis in real time. Despite scraping tools like Playwright and PRAW making the process a little less restrictive, it was still resource intensive and time consuming. Lastly, the problem with computational resources made it difficult to process the geolocation data. However, we currently use a geo-dictionary to standardize city names using state level tags, but could not process more detailed information, like street names or neighborhood level locations. With further computational poses, this limitation can be mitigated to allow a finer analysis.

Implications and Future Extensions

Despite these challenges, we see significant opportunity for improvement and expansion from this project. Live tracking of emerging crime trends enables integration of real time monitoring dashboards for live interactive tracking of such trends with actionable, up to date insights for stakeholders. Advanced models like deep learning will provide the model the ability to make more accurate and scalable prediction, be able to process larger datasets and produce better results. Here, analysis can incorporate socio-economic and demographic data, which will deepen the understanding of what drives crime, so that root causes of the problem can be addressed and public safety improved. Additionally, based on computing resources successfully leveraged through cloud computing, high computational limitations will be overcome, enabling efficient processing of largescale datasets and better model performance

Contributions

All Team members have contributed equally to the project

Team members:

1. Varun Singh
2. Yash Mathur
3. Bhavya Khilrani
4. Siddhartha Dheer

References:

1. Richardson, R., Schultz, J. M., & Crawford, K. (2019). Dirty data, bad predic ons: How civil rights viola ons impact police data, predic ve policing systems, and jus ce.

2. Gerber, M. S. (2014). Predic ng crime using Twi er and kernel density es ma on.

3. Bran ngham, P. J., Valasik, M., & Mohler, G. O. (2018). Does Predic ve Policing Lead to Biased Arrests? Results From a Randomized Controlled Trial. Gorr, W. L., & Harries, R. (2003). Introduc on to crime forecas ng.

4. Mohler, G. O., Short, M. B., & Malinowski, S. (2011). Self-exci ng point process modeling of crime. Journal of the American Sta s cal Associa on.

5. Wang, X., Gerber, M. S., & Brown, D. E. (2012). Crime incident predic on using Tweets. IEEE Interna onal Conference on Data Mining.

6. Vo, T., Sharma, R., & Kumar, R. (2020). Crime rate detec on using social media of different crime loca ons and Twi er part-of-speech tagger with Brown clustering. Journal of Intelligent Systems.

7. Hu, Y., Wang, F., & Guin, C. (2018). A spa o-temporal kernel density es ma on framework for predic ve crime hotspot mapping and evalua on.

8. Aldossari, B. S., Alqahtani, F. M., & Alshahrani, N. S. (2020). A compara ve study of decision tree and Naive Bayes machine learning model for crime category predic on in Chicago.

9. D Das, M Nayak (2020). Crime pa ern detec on using data mining techniques.

10. Karabo Jenga, Gorkem Kar, Cagatay Catal (2023). “Machine Learning in Crime Predic on”

11. Nubani, L., Fierke-Gmazel, H., & ... (2023). Community engagement in crime reduc on strategies: A tale of three ci es.