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

Xuhan Zhao, Yiwei Wang, Omar Abu-Rub

Introduction

- Crime rate is a critical issue that affects the public wellbeing. Researches have shown that social media platforms such as Twitter can be a practical tool for crime surveillance and monitoring [4].
- We hypothesize that a positive correlation exists between tweets and city crime rates, suggesting that social media can predict criminal activities.
- Crime prediction enables proactive measures to prevent crime by identifying high-risk areas and encouraging the implementation of prevention programs.
- Reducing crime on a national scale can lead to economic development, as high crime rates can deter investment and economic opportunities. Predicting crime helps governmental agencies create a better environment for economic growth.

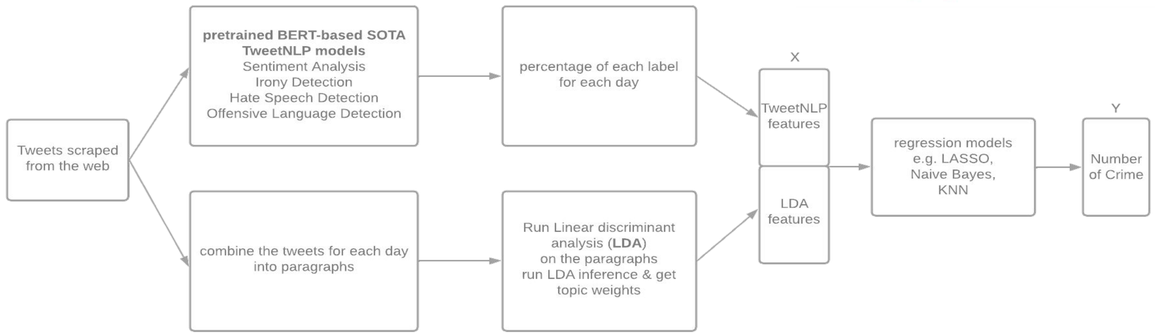
Data

- In this project, we collected Twitter data and crime rate data in Los Angeles and Boston for the year of 2022 with the goal to leverage natural language processing (NLP) and data analysis techniques to understand the association.
- Los Angeles crime data from 2020 to present [2]
  - 175.4MB in csv with 232330 records in 2022
- Boston crime data from 2015 to present [1]
  - 118.7 MB in csv with 70044 records in 2022
- Twitter Data
  - 4.6MB for Los Angeles with 36500 tweets
  - 5.28MB for Boston with 36500 tweets
  - Used snscrape package in Python to retrieve data in certain locations during our desired time span (01-01-2022 to 12-31-2022) [5].
  - Collected 100 tweets per day in the city of Los Angeles and Boston with the specific longitudinal and latitudinal coordinate location.

Twitter Logo [3]

Approach (algorithm, models, analysis)

- Latent Dirchlet Allocation (LDA): generates lists of topics with tweets' weights on the topics. Ex. *0.019*"work" + 0.015*"take" + 0.014*"stress" + 0.014*"need" + 0.014*"time" + 0.013*"much"*
- Lasso Regression: regularized linear regression that includes a penalty.
- Gaussian Naive Bayes Regression: a type of Naive Bayes algorithm with a Gaussian likelihood of features.
- K-Nearest Neighbors (KNN): supervised ML algorithm.



Experiment and Result

Evaluation: mean squared error (MSE) and accuracy when an error of ϵ is allowed, which means the predictions is considered to be true if our prediction falls in the range of [ground truth - ϵ * ground truth, ground truth + ϵ * ground truth] and false otherwise (ϵ = 0.05, 0.1, 0.2).

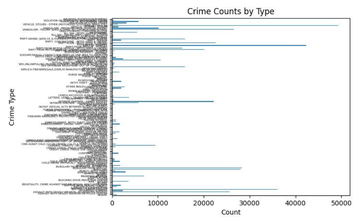
Los Angeles Dataset	Accuracy with ϵ = 0.05	Accuracy with ϵ = 0.1	Accuracy with ϵ = 0.2	MSE
Lasso Regression	0.54794	0.68493	0.87671	32567.11986
Gaussian Naive Bayes	0.52055	0.78082	0.87671	34938.24657
K-Nearest Neighbors	0.41096	0.68493	0.86301	34196.61432
Boston Dataset	Accuracy with ϵ = 0.05	Accuracy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

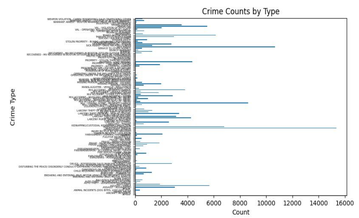
The top 5 crimes in Boston in 2022 with the most count are:

1.	INVESTIGATE PERSON	15354.0
2.	SICK ASSIST	10680.0
3.	M/V - LEAVING SCENE - PROPERTY DAMAGE	8622.0
4.	INVESTIGATE PROPERTY	6818.0
5.	TOWED MOTOR VEHICLE	6154.0

The top 5 crimes in Los Angeles in 2022 with the most count are:

1.	VEHICLE - STOLEN	49316.0
2.	THEFT OF IDENTITY	42290.0
3.	BATTERY - SIMPLE ASSAULT	36052.0
4.	BURGLARY FROM VEHICLE	28370.0
5.	BURGLARY	27978.0





Conclusion

- Found evidence that support positive correlation between tweets and city crime rates.
- Social media can provide valuable insights to law enforcement agencies and local government which enables them to take proactive measure to prevent crime.
- Gathering more data is definitely conducive to the performance of the model - we observe significant performance increase (0.1 increase for accuracy within 5%) when we move from 80% training data to 100% training data.
- To achieve better results,
  - Gather more randomized tweets
  - Gather tweets targeting at certain topics
  - Scrape tweets on more days
  - more powerful features (tweet embeddings, continuous representation of the topics)

References

1. Boston Police Department. 2023. Crime Incident Reports. Retrieved from https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system
2. Los Angeles Police Department. 2020. Crime Data from 2020 to Present. Retrieved from https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8
3. Wikipedia contributors. (2023). Twitter. *Wikipedia*. https://en.wikipedia.org/wiki/Twitter#/media/File:Logo_of_Twitter,_Inc..svg
4. 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/10.1177/2056305119834585arXiv:https://doi.org/10.1177/2056305119834585
5. snscrape: A social networking service scraper in Python. Retrieved from https://github.com/JustAnotherArchivist/snscrape