Fusión Semántico-Espacio-Temporal de Datos de Redes Sociales Y Gubernamentales para el Análisis de Crimen

Leon F. Davis Coropuna*
National University of San Agustin
Arequipa, Peru
ldavis@unsa.edu.pe

Ana M. Cuadros Valdivia
National University of San Agustin
Arequipa, Peru
acuadrosv@unsa.edu.pe

ABSTRACT

Test

ACM Reference Format:
Leon F. Davis Coropuna and Ana M. Cuadros Valdivia. 2025. Fusión Semántico-Espacio-Temporal de Datos de Redes Sociales Y Gubernamentales para el Análisis de Crimen. In . ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introducción

El aumento de la criminalidad urbana y la creciente complejidad de los fenómenos delictivos han convertido el análisis de la seguridad pública en una tarea relevante para las sociedades actuales [8]. Estos eventos inciden de forma directa en la calidad de vida de las personas y generan implicancias sociales y económicas importantes [11]. En respuesta a ello, el análisis de datos provenientes tanto de redes sociales como de fuentes oficiales ha empezado a utilizarse como una alternativa para comprender mejor estas dinámicas.

Sin embargo, el análisis de datos criminales enfrenta dos desafíos principales. El primero es la fragmentación y heterogeneidad de las fuentes de información disponibles [2]. El segundo es la dificultad de integrar dimensiones clave —como la semántica, la espacial y la temporal— dentro de un mismo marco analítico [7]. Mientras que las fuentes oficiales suelen generar datos estructurados pero con cobertura limitada, las redes sociales proporcionan información más rica en contenido semántico, aunque también más desordenada y propensa al ruido [15]. Como resultado, muchos enfoques existentes abordan estas dimensiones de forma separada o secuencial, lo que limita su potencial para generar análisis integrales [4].

Desde el campo de la inteligencia urbana, se ha demostrado que la integración de información espacial, temporal y semántica resulta útil para estudiar dinámicas sociales complejas [10]. A pesar de ello, profesionales vinculados a la seguridad pública tienden a mostrar cautela ante el uso de modelos computacionales avanzados, especialmente por la falta de transparencia que caracteriza a muchos de estos enfoques y la ausencia de mecanismos claros para validar sus resultados desde el conocimiento experto [12]. Esta tensión no es exclusiva del análisis criminal: también se ha observado en áreas como la calidad del aire, donde existe una preferencia por modelos tradicionales más interpretables, aunque menos precisos que los basados en aprendizaje automático [1].

Con el objetivo de comprender mejor las necesidades del sector, se llevó a cabo un estudio preliminar con especialistas en análisis criminal y seguridad urbana. A partir de este estudio, se identificó que los profesionales requieren herramientas capaces de integrar múltiples fuentes de datos heterogéneos, correlacionar eventos delictivos en el espacio y en el tiempo, extraer patrones significativos a partir de contenido no estructurado y, finalmente, visualizar estos resultados de forma clara y comprensible.

En base a estos hallazgos, se propone una metodología que combina técnicas de fusión de datos heterogéneos [1] con modelos de procesamiento de lenguaje natural adaptados al dominio. Este enfoque permite abordar de manera simultánea las dimensiones semántica, espacial y temporal en el análisis de la criminalidad urbana. Además, incorpora visualizaciones interactivas que permiten a los especialistas explorar los datos y patrones de manera comprensible.

Este trabajo aporta tres contribuciones principales. En primer lugar, una metodología para integrar datos criminales heterogéneos provenientes de distintas fuentes. En segundo lugar, un marco analítico que combina información semántica, espacial y temporal en un solo sistema. Finalmente, un conjunto de herramientas visuales diseñadas para facilitar la interpretación y exploración de los resultados por parte de usuarios expertos en seguridad pública.

2 Trabajos Relacionados

La visualización espacio-temporal del crimen basada en datos abiertos es un campo de creciente interés, especialmente en contextos urbanos de países en desarrollo. Diversas investigaciones previas han abordado componentes esenciales del análisis criminal, con distintos enfoques y objetivos.

2.1 Análisis Criminal Tradicional y Visualización

El análisis criminal tradicional ha recurrido históricamente a enfoques estadísticos y espaciales para comprender la distribución y evolución delictiva en entornos urbanos. Diversos estudios han identificado patrones climáticos y estacionales asociados a ciertos tipos de crimen, como agresiones y robos, destacando la influencia de variables como la temperatura o la cantidad de luz solar [9, 3]. Estas correlaciones han permitido anticipar picos delictivos y orientar estrategias de prevención basadas en predicción estacional.

En el ámbito espacial, la identificación de hotspots —zonas de alta concentración delictiva— ha sido fundamental para la planificación policial y la asignación de recursos. Técnicas como la detección por

*Both authors contributed equally to this research.

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

KDE (Kernel Density Estimation), agrupamiento espacio-temporal y modelos predictivos han sido aplicadas con éxito en plataformas como CriPAV [5], que permite delimitar áreas de riesgo mediante análisis georreferenciado.

Por otro lado, la visualización de datos se ha convertido en una herramienta crucial para interpretar los resultados del análisis criminal. Herramientas como CrimAnalizer [6] y ST-UEx [13] permiten explorar visualmente la evolución del crimen en tiempo y espacio, incorporando dashboards interactivos, mapas de calor y filtros temporales. Estas plataformas no solo ayudan a expertos en criminología, sino también a autoridades locales y ciudadanía interesada en entender dinámicas delictivas específicas en sus entornos inmediatos.

2.2 Extracción de Información desde Noticias: CAMELON y CRIMSON

CAMELON [12] presenta un sistema de procesamiento de noticias digitales para la identificación automática de eventos delictivos, utilizando modelos de PLN como XLM-RoBERTa y WangchanBERTa para la clasificación multietiqueta. Incorpora técnicas de reconocimiento de entidades para extraer roles clave (víctima, perpetrador, arma, entre otros), almacena los datos en una base georreferenciada y ofrece una interfaz visual para el análisis espacio-temporal. En la tarea de clasificación de tipo de crimen, los modelos evaluados fueron diversos, incluyendo variantes de transformers, y se concluyó que el modelo XLM-RoBERTa fue el más eficiente. En términos de desempeño, se reportó una F1-score promedio por clase superior al 86% para las categorías principales, validando su uso práctico para el análisis automatizado de noticias delictivas en idiomas con recursos limitados como el tailandés. Además, el sistema también logró resultados destacados en tareas de extracción de metadatos del crimen, como roles y ubicaciones, sin embargo aqui F1-score baja bastante hasta un promedio de 51%.

CRIMSON [14] clasifica crímenes y accidentes en noticias tailandesas usando modelos como Naive Bayes, SVM, XGBoost y multilingües como mBERT, WangchanBERTa y XLM-RoBERTa (Base y Large). XLM-RoBERTa Large logró el mejor rendimiento (F1 macro = 0.86, AUC-ROC = 0.916), superando a los demás, salvo en la clase "gambling", donde XLM-RoBERTa Base fue superior. XLM-RoBERTa Large obtuvo F1 > 0.90 en murder (0.917), sexual abuse (0.904) y drug (0.916). La peor clase fue battery/assault (F1 = 0.753), posiblemente por ambigüedad lingüística. Aun así, mantuvo buen rendimiento general (F1 entre 0.75 y 0.92). SVM, con TF-IDF, fue competitivo (F1 = 0.83), apenas 3.6% menos que XLMR-Large, lo que indica que modelos clásicos aún son útiles en tailandés. mBERT tuvo bajo desempeño (F1 = 0.66), especialmente en "gambling" (F1 = 0.008), por escasez de ejemplos (2.91%). Además, CRIMSON analizó la correlación entre noticias clasificadas y datos reales (2016–2020). Hubo correlaciones altas en S1, como en battery/assault (r = 0.981) y moderadas en rape y accidentes (r = 0.57, 0.62). Murder y theft/burglary tuvieron correlaciones más bajas, quizás por sesgos mediáticos o en reportes policiales.

Los trabajos revisados abordan el análisis del crimen desde perspectivas complementarias: la visualización espacio-temporal con apoyo de datos estructurados por un lado, y la extracción automatizada desde noticias digitales por otro. Cada uno contribuye con herramientas y metodologías valiosas para entender la dinámica criminal, ya sea mediante mapas interactivos, análisis predictivo, o integración de fuentes periodísticas con técnicas de procesamiento del lenguaje natural.

3 PROPUESTA

La presente propuesta se estructura en torno a un flujo de trabajo integral para la visualización y análisis de delitos en Chicago, combinando fuentes oficiales y datos de percepción ciudadana. El proceso completo se articula en cuatro etapas: recopilación de datos, preprocesamiento de los datos, extracción de características, modelo de vinculación y visualización de resultados. Cada etapa cumple un rol fundamental en la transformación de datos crudos en información procesable y visualmente interpretable. En la Figura 1, se presenta un diagrama del pipeline general propuesto, el cual resume las fases principales del sistema y sus interrelaciones.

3.1 Preprocesamiento de Datos

El preprocesamiento de los tweets se realiza mediante dos procesos fundamentales. Para la clasificación del tipo de crimen, se emplea un modelo de lenguaje avanzado como XML-R o BERT, específicamente entrenado para identificar y categorizar eventos delictivos según la taxonomía oficial. Este modelo analiza el contenido textual de cada tweet, asignándolo a categorías predefinidas como robo, asalto u homicidio. Paralelamente, se ejecuta un sistema de reconocimiento de entidades nombradas (NER) que extrae metadatos cruciales, incluyendo ubicaciones geográficas (convertidas a coordenadas de latitud/longitud mediante geocodificación) y marcas temporales precisas (fecha y hora del evento reportado). Este proceso garantiza que los datos no estructurados de redes sociales adquieran un formato compatible con los registros oficiales.

3.2 Extracción de Características

La etapa de extracción de características transforma los datos preprocesados en representaciones numéricas adecuadas para el análisis computacional. Para las características semánticas, se generan embeddings de texto utilizando modelos contextuales como BERT, que capturan el significado lingüístico y las nuances del contenido de cada tweet. Estas representaciones vectoriales preservan las relaciones semánticas entre diferentes reportes. Simultáneamente, se calculan características espacio-temporales que incluyen no solo las coordenadas geográficas y timestamps, sino también derivados como día de la semana, hora pico, y proximidad a puntos de interés. Estas mismas transformaciones se aplican a los datos oficiales para asegurar consistencia dimensional.

3.3 Modelo de Vinculación

El modelo de vinculación emplea arquitecturas avanzadas de machine learning para establecer correspondencias entre tweets y registros oficiales. Se proponen dos enfoques complementarios: XGBoost, que ofrece alto rendimiento en datos tabulares y permite interpretabilidad de características, y transformers con mecanismos de atención cruzada, especializados para modelar interacciones complejas entre texto y datos espacio-temporales. La entrada del

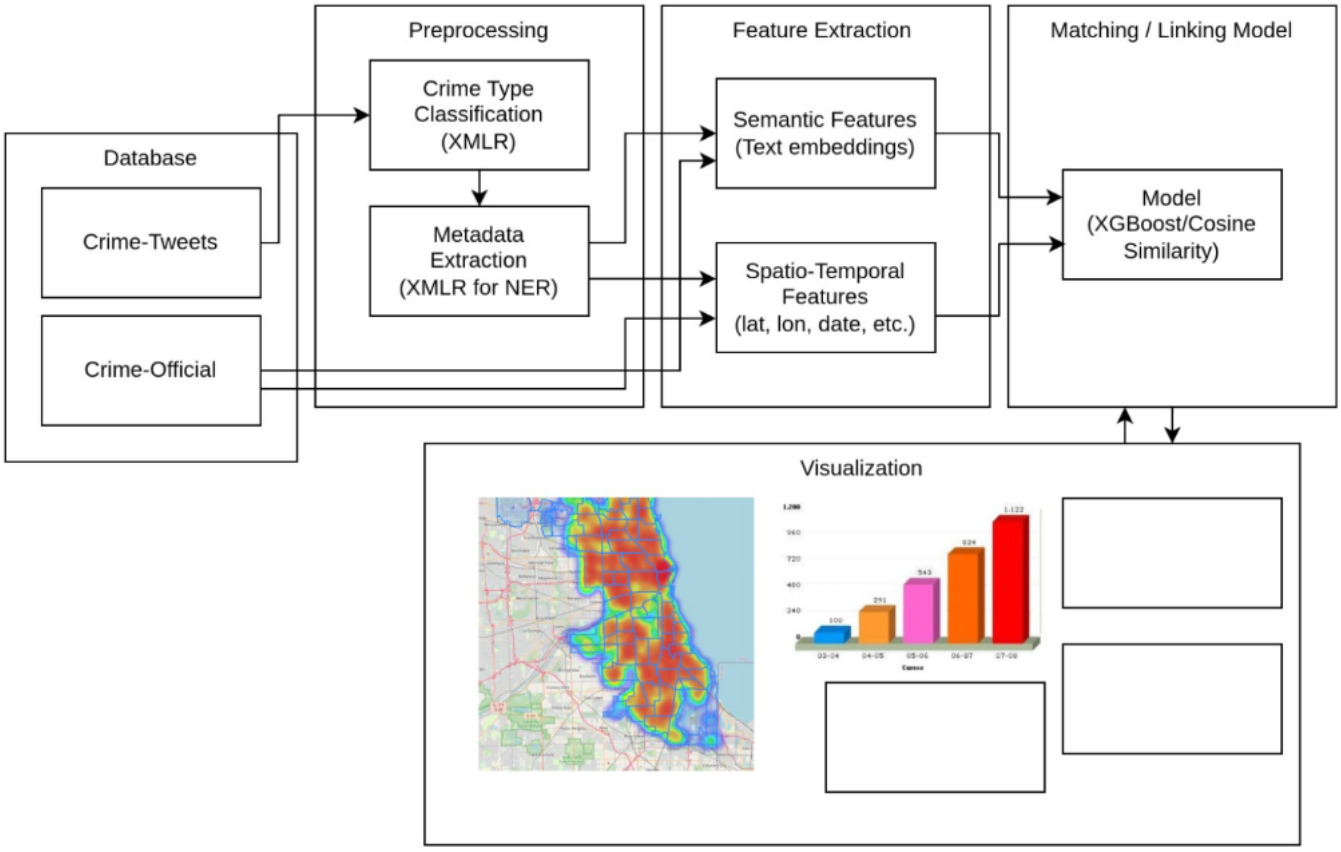

Figure 1: Pipeline general del método propuesto

modelo consiste en la concatenación de los embeddings semánticos y espacio-temporales de ambos tipos de fuentes. Como salida, el modelo genera puntuaciones de probabilidad que indican la verosimilitud de que un tweet y un registro oficial representen el mismo evento delictivo, considerando múltiples dimensiones de similitud.

3.4 Visualización de Resultados

La capa de visualización integra los resultados del modelo en un dashboard interactivo que combina múltiples representaciones gráficas. Mapas de calor superpuestos permiten comparar las distribuciones espaciales de reportes ciudadanos versus datos oficiales, mientras que líneas temporales sincronizadas revelan patrones y discrepancias cronológicas. Gráficos de dispersión multidimensionales facilitan el análisis de correlaciones entre diferentes variables. La interfaz incorpora herramientas para filtrar por tipo de crimen, período temporal y área geográfica, permitiendo a los usuarios explorar hipótesis específicas sobre la relación entre percepción pública y estadísticas institucionales.

References

[1] Sadaf Ahmed, Monica Gentili, Daniel Sierra-Sosa, and Adel S. Elmaghraby. 2022. Multi-layer data integration technique for combining heterogeneous crime data. *Information Processing Management* 59, 3 (2022), 102879. doi:10.1016/j.ipm.2022.102879

[2] Abdulaziz Almaslukh, Ahmed Almaalwy, Naif Allheeib, Abdulrahman Alajaji, Mohammed Almukaynizi, and Yazeed Alabdulkarim. 2024. Top-k sentiment analysis over spatio-temporal data. *PeerJ Computer Science* 10 (Sept. 2024), e2297. doi:10.7717/peerj-cs.2297

[3] Ysabel A. Castle and John M. Kovacs. 2021. Identifying seasonal spatial patterns of crime in a small northern city. *Crime Science* 10, 1 (oct 2021), 25. doi:10.1186/s40163-021-00161-w

[4] Muhammad G. Almatar, Huda S. Alazmi, Liuqing Li, and Edward A. Fox. 2020. Applying GIS and Text Mining Methods to Twitter Data to Explore the Spatiotemporal Patterns of Topics of Interest in Kuwait. *ISPRS International Journal of Geo-Information* 9, 12 (2020). doi:10.3390/ijgi9120702

[5] Germain García-Zanabria, Marcos M. Raimundo, Jorge Poco, Marcelo Batista Nery, Cláudio T. Silva, Sergio Adorno, and Luis Gustavo Nonato. 2022. CriPAV: Street-Level Crime Patterns Analysis and Visualization. *IEEE Transactions on Visualization and Computer Graphics* 28, 12 (2022), 4000–4015. doi:10.1109/TVCG.2021.3111146

[6] Germain Garcıa, Jaqueline Silveira, Jorge Poco, Afonso Paiva, Marcelo Batista Nery, Claudio T. Silva, Sérgio Adorno, and

Luis Gustavo Nonato. 2021. CrimAnalyzer: Understanding Crime Patterns in São Paulo. *IEEE Transactions on Visualization and Computer Graphics* 27, 4 (2021), 2313–2328. doi:10.1109/TVCG.2019.2947515

[7] D. Hanny and B. Resch. 2024. Multimodal Geo-Information Extraction from Social Media for Supporting Decision-Making in Disaster Management. *AGILE: GIScience Series* 5 (2024), 28. doi:10.5194/agile-giss-5-28-2024

[8] Mark Kibanov. 2019. Social Network Mining for Analysis of Social Phenomena. https://api.semanticscholar.org/CorpusID:208096044

[9] Shannon J. Linning, Martin A. Andresen, and Paul J. Brantingham. 2017. Crime Seasonality: Examining the Temporal Fluctuations of Property Crime in Cities With Varying Climates. *International Journal of Offender Therapy and Comparative Criminology* 61, 16 (2017), 1866–1891. doi:10.1177/0306624X16632259 arXiv:https://doi.org/10.1177/0306624X16632259 PMID: 26987973.

[10] Pablo Martí, Leticia Serrano-Estrada, and Almudena Nolasco-Cirugeda. 2019. Social Media data: Challenges, opportunities and limitations in urban studies. *Computers, Environment and Urban Systems* 74 (2019), 161–174. doi:10.1016/j.compenvurbsys.2018.11.001

[11] Mohd Suhairi Md Suhaimin, Mohd Hanafi Ahmad Hijazi, Ervin Gubin Moung, Puteri Nor Ellyza Nohuddin, Stephanie Chua, and Frans Coenen. 2023. Social media sentiment analysis and opinion mining in public security: Taxonomy, trend analysis, issues and future directions. *Journal of King Saud University - Computer and Information Sciences* 35, 9 (2023), 101776. doi:10.1016/j.jksuci.2023.101776

[12] Siripen Pongpaichet, Boonyapat Sukosit, Chitchaya Duangtanawat, Jiramed Jamjongdamrongkit, Chancheep Mahacharoensuk, Kantapong Matangkarat, Pattadon Singhajan, Thanapon Noraset, and Suppawong Tuarob. 2024. CAMELON: A System for Crime Metadata Extraction and Spatiotemporal Visualization From Online News Articles. *IEEE Access* 12 (2024), 22778–22802. doi:10.1109/ACCESS.2024.3363879

[13] Tiago Paulino Santos, João Matheus Siqueira Souza, Thales Vieira, and Luis Gustavo Nonato. 2024. Space-Time Urban Explorer: A Visual Tool for Exploring Spatiotemporal Crime and Patrolling Data. In *2024 37th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. 1–6. doi:10.1109/SIBGRAPI62404.2024.10716319

[14] Suppawong Tuarob, Phonarnun Tatiyamaneekul, Siripen Pongpaichet, Tanisa Tawichsri, and Thanapon Noraset. 2025. Beyond administrative reports: a deep learning framework for classifying and monitoring crime and accidents leveraging large-scale online news. *Neural Computing and Applications* 37, 10 (apr 2025), 7183–7205. doi:10.1007/s00521-024-10833-8

[15] Meghashyam Vivek and Boppuru Rudra Prathap. 2023. Spatiotemporal Crime Analysis and Forecasting on Twitter Data Using Machine Learning Algorithms. *SN Computer Science* 4, 4 (2023), 383. doi:10.1007/s42979-023-01816-y Accessed: 2023/05/06.