Information Processing and Management 59 (2022) 102879

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Multi-layer data integration technique for combining heterogeneous crime data

Sadaf Ahmed[a], Monica Gentili[b,*], Daniel Sierra-Sosa[c], Adel S. Elmaghraby[a]

[a] Computer Science and Engineering, University of Louisville, 222 Eastern Parkway, Louisville, KY, 40208, United States
[b] Industrial Engineering, University of Louisville, 222 Eastern Parkway, Louisville, KY, 40208, United States
[c] Computer Science & Information Technology, Hood College, 401 Rosemont Ave, Frederick, MD, 21701, United States

ARTICLE INFO

Keywords:
Entity resolution
Data integration
Trafficking crimes
XGBoost
Binary classification
Natural language processing

ABSTRACT

Analysis of publicly available human and drug trafficking crime data faces the challenge of finding a comprehensive dataset that includes a sufficiently large number of crime incidents. Our proposed methodology attempts to address this challenge by using entity resolution techniques to merge multiple state-wide crime datasets and a county-wide incident report dataset to get a clearer picture of a category of criminal activity in a geographical area. This methodology combines incident reports, crime reports, and court records to close any gaps that may be present in a single data source. We apply this methodology to create a dataset that includes drug and human trafficking related crimes and incidents from three distinct sources (from Louisville Open Data Crime Reports, Federal Bureau of Investigation Kentucky Crime Incidents, and the Kentucky Online Offender Lookup website) to provide researchers data to study the link between drug and human trafficking related crimes. In a case study performed with the new merged dataset, an XGBoost classifier was able to label a 7-day sliding time window, within any given county, as containing a human trafficking related incident or not with a Matthews correlation coefficient of 0.86.

1. Introduction

Analysis of the relationship between different crime types is one area of data mining that may necessitate more than a single source of data to resolve. For example, separate datasets from different law enforcement agencies might allow for additional insights as each dataset may cover a different set of incidents (Robinson & Scogings, 2018; Zhou et al., 2021, 2020). Analysis of human trafficking in particular suffers from fragmented data, with information about networks, crimes, and victims being spread across multiple law enforcement organizations (Konrad et al., 2017). When considering datasets from different law enforcement agencies that cover overlapping areas and time frames, there is a possibility of separate datasets containing records that represent the same incident. Therefore, when analyzing these datasets, researchers should consider disambiguation and resolving missing information when dealing with multiple records covering the same incidents. For instance, information that is missing in one dataset, may be found in another dataset. Entity resolution (Getoor & Machanavajjhala, 2012) has been employed as a solution for individuating instances of potential crimes across data sources by Nagpal et al. (2017). In Nagpal et al. (2017), this methodology was applied to a proxy for crimes: online advertisements indicative of potential crime activity. Separating instances of crimes from criminal databases and crime reports databases could follow a similar framework of entity resolution as Nagpal et al.'s research.

[*] Corresponding author.
E-mail addresses: sfahme01@louisville.edu (S. Ahmed), monica.gentili@louisville.edu (M. Gentili), sierra-sosa@hood.edu (D. Sierra-Sosa), adel@louisville.edu (A.S. Elmaghraby).

https://doi.org/10.1016/j.ipm.2022.102879
Received 30 September 2021; Received in revised form 22 December 2021; Accepted 17 January 2022
Available online 14 February 2022

Using crime datasets requires different strategies for the varying types of data that describe illicit activity. Falade et al. (2019) provide a survey of crime prediction efforts wherein various machine learning methods have been applied to multiple types of datasets: criminal records, social media, news, and police reports. The authors note the different opportunities and challenges that each type of crime dataset presents, such as social media posts being highly unstructured and First Information Reports (FIRs) being unstructured but reliable. In Ku and Leroy (2011), an additional type of crime dataset, crime tips from the public, is examined with the intent to use entity resolution to find related crimes.

The United States Federal Bureau of Investigation (FBI) attempted to standardize and centralize crime reporting through its Uniform Crime Reporting (UCR) Program, now called the National Incident-Based Reporting System (NIBRS) (Farrell & Reichert, 2017; Maxfield, 1999; United States Department of Justice, Federal Bureau of Investigation, 2021). NIBRS provides incident-level data and has been used as a means for both understanding crime patterns in the United States, as well as a way to measure the effectiveness and progress of law enforcement initiatives transparently (Addington, 2019; Strom & Smith, 2017). Law enforcement agencies must meet technical reporting standards and voluntarily submit reports to the FBI to be included in the dataset. As of October 31, 2020, the FBI reported 48.9% of the United States population was covered by precincts that participate in the UCR program (United States Department of Justice, Federal Bureau of Investigation, 2021). This 51.1% coverage gap might be mitigated using data gathered from local law enforcement organizations. Local programs such as Louisville's Open Data Initiative (LOD) (Louisville Metro Government, 2021) and Kentucky Online Offender Lookup (KOOL) (Kentucky Department of Corrections, 2021) offer an opportunity to bridge the gap, as well as enrich the feature set available from NIBRS, should there be any overlap between the datasets. Merging these datasets then presents a series of challenges that need to be overcome to ensure the quality and integrity of the resulting merged dataset.

In this paper, we propose a strategy for entity resolution to create a database of unique crime instances related to specific categories of crime to address the data gaps described in Konrad et al. (2017). We apply a technique that allows combining multiple datasets containing different crime types and overlapping records to form a single, more complete dataset for deriving patterns between categories of drugs incidents and human trafficking incidents. The de-duplication is performed using a "merge and purge" strategy, introduced by Hernández and Stolfo (1995). The phrase "merge and purge" is used to describe the process of combining two duplicate records into a single, richer record before the two individual records are discarded. A "merge and purge" solution can also be found in Evans et al. (2019) and Nagpal et al. (2017). Data-centric frameworks have been proposed for analysis of drug abuse patterns in Sarker et al. (2019). Since quality of data can be a limiting factor in the performance of machine learning models, we aim to provide researchers with a data-centric framework to improve crime model performance by intelligently addressing data fragmentation typically encountered in crime studies. We also present a case study using the new dataset, wherein the drug trafficking and human trafficking incidents are used to train a binary classifier to classify 7 day windows and counties where a human trafficking incident is likely to occur.

1.1. Research objectives and motivation

The objective of our research was to attempt to remedy the data fragmentation problem (Konrad et al., 2017) common with human trafficking data analysis using crime records and police reports. We address an opportunity to create a more complete human trafficking and drug trafficking dataset using the publicly available data in the U.S. state of Kentucky. This research attempts to answer the following questions:

- Can crime data from multiple organizations within the same state be combined to form a richer dataset more representative of the crime activity in the state?
- What methods can be applied to identify and resolve duplicate crime events?
- Can this new dataset be used to explore relationships between crime types (i.e., drug trafficking and human trafficking related crimes)

Our research is motivated by the lack of insight into criminal networks that can be attributed to relevant details of crime incidents being spread across multiple agencies, or details of crime incidents being spread across multiple record types (crime records and police reports). The integrated data can be used to fill gaps in information and find missing links. This is a framework that can be used for heterogeneous data in general, but in this context targeting the domain of crime data. Once crime networks have been constructed using the integrated data, the work by Marciani et al. (2017) surveys state-of-the-art techniques for analysis of criminal networks. These existing techniques demonstrate the applications for which improved crime data would be immediately useful. In addition to assisting with providing a more complete picture of a certain type of crime, our proposed crime data integration framework can be used to explore relationships between crime types. Separate law enforcement agencies might focus on different types of crime, therefore, integrating the records held by separate agencies may uncover links between these crime types. The case study presented in Section 5 explores the relationship between drug crimes and human trafficking. This methodology and the analyses that could be performed on the integrated data can be used for additional decision support for actionable law enforcement policy or crime prevention. Bahulkar et al. (2018) discuss data-driven methods for disrupting transnational criminal organizations that limit crime organizations ability to operate effectively through strategic allocation of law enforcement.

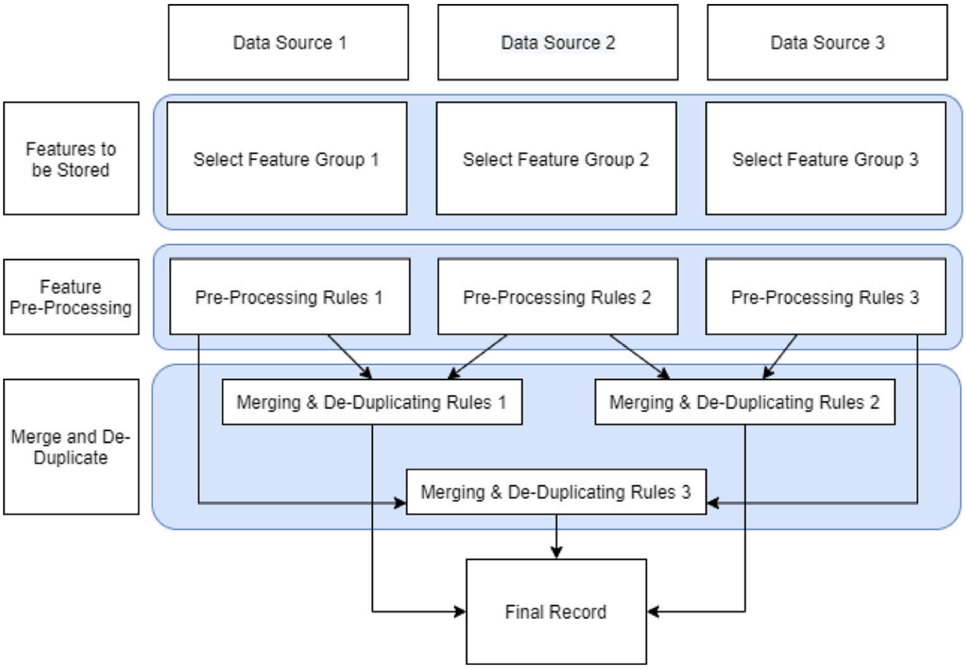

Fig. 1. Diagram of proposed multi-layer data integration technique.

1.2. Contributions

Our contribution is a proposed and demonstrated multi-layer data integration technique for creating a dataset that combines crime records from different types of law enforcement agencies (i.e., corrections department, local police, federal police) at multiple levels of agency (city-wide, state-wide, and national) for expanded analysis. The process for this combination and analysis is broadly diagrammed in Fig. 1, with each data source flowing through a tailored preprocessing layer before being merged and de-duplicated to create the final dataset. Additionally, we use natural language processing and fuzzy matching for finding duplicate crime records. Next, we apply the proposed methodology to derive a novel dataset of crimes related to human and drug trafficking in the U.S. state of Kentucky, which is now publicly available on uofllogistics.org. Finally, we analyze the merged dataset to explore the spatio-temporal relationship between drug and human trafficking related crimes with the assistance of a machine learning classifier.

2. Background and related works

Data fragmentation refers to when data representing a singular event or entity exists as broken pieces spread across multiple data sources. Aggregating the data fragments is the first step to resolving fragmentation. Laura and Me (2017) describes an active effort to collect drug data from primary sources by discovering illicit activity online using a semantic search engine. Developed under the Semantic Illegal Content Hunter (SICH) Project, this search engine endeavors to build a queryable dataset built on data found across the web as a means of gathering the necessary information to study crime and criminal networks.

Once the data is collected, fragmentation is an issue that can be addressed with entity resolution. Yang et al. (2018) explores social media solutions to the fragmentation problem by integrating crime and urban infrastructure data with geotagged social media posts. The authors sought to implement a crime hotspot prediction framework using the integrated datasets and compare the performance to the same framework using only the crime records dataset. The results show consistently higher performance in predicting crime hotspots using the integrated datasets across traditional machine learning methods (logistic regression, naive bayes, support vector machines, decision trees, and random forest) and deep learning methods (a neural network). Neural networks and random forest offered the best performance. While the addition of the social media and urban infrastructure data improved the performance in every case, crime report data was the most influential. We see an opportunity to improve upon this work by expanding the reach of the crime data, as Yang et al. (2018) only used a single official crime data source: New York Police Department crime records provided through the New York City open data portal (New York City, 2021). Analysis of crimes might be improved through utilizing crime data sourced from other agencies and levels of government.

Konrad et al. (2017) defines human trafficking as “the recruitment, transportation, or harboring of persons by means of force, coercion, abduction, or deception for the purpose of exploitation”. Broadly, human trafficking related crimes tend to promote prostitution and/or forced labor (Burke & Bruijn, 2017). For the purpose of this study, we focused on sex trafficking related crimes only, due to the lack of incidents related to labor trafficking. Crimes relevant to sex trafficking were identified with the assistance of a criminology expert as being represented by NIBRS codes 40A (Prostitution), 40B (Assisting or Promoting Prostitution), and 370 (Pornography/Obscene Material). Commercial sex and pornography related crimes will, therefore, be the focus of the research. It should be noted that not all acts of commercial sex involving adults are considered human trafficking (Polaris Project, 2021). However, for the purposes of this dataset, all sex crimes in the available data sources are included.

One method of studying human trafficking networks is through online advertisements. Multiple research efforts into human trafficking, such as Alvari et al. (2016), Boecking et al. (2019), Dubrawski et al. (2015) and Nagpal et al. (2017) have focused on online advertisements from classified advertisement websites like Backpage.com (Portnoff et al., 2017). Advertisement datasets offer their own challenges separate from court record datasets, as an advertisement may be human trafficking-related or may be an innocent listing. Alvari et al. (2016) created a manually labeled training dataset to identify suspected human trafficking ads. A semi-supervised learning approach was then applied and was able to achieve a precision over 90% on identifying human trafficking advertisements. Other work by Nagpal et al. (2017) tackled the entity resolution problem regarding a dataset of known human trafficking advertisements. A classifier was trained to identify if two trafficking ads were created by the same source. In Tong et al. (2017), a dataset of 10,000 known trafficking records is provided, called Trafficking-10k. Zhu et al. (2019) expands on this research. They use Trafficking-10k to study language structures present within human trafficking advertisements.

Another angle of human trafficking data is business advertisements and reviews (Bouche & Crotty, 2018; Diaz & Panangadan, 2020). These are differentiated from the personal and classified ads discussed earlier in that business advertisements and reviews are meant to represent brick-and-mortar establishments that might engage in commercial sex activity alongside legitimate business activity. Diaz and Panangadan (2020) integrates a dataset of known illegal businesses generated from an illicit business review website (Rubmaps) with a general-purpose review website (Yelp) to train a classifier to identify sex trafficking businesses based on reviews.

Analysis of human trafficking can be difficult as real data on trafficking instances is not available (Farrell & Reichert, 2017; Goodey, 2008; Kangaspunta, 2007). One option to the verified data problem is using crime and conviction reports as in Bales et al. (2020), Farrell et al. (2019) and Gholizadehy et al. (2020). Gholizadehy et al. (2020) propose a solution to analyze court records as a proxy for human trafficking. This sort of analysis is incomplete as only human trafficking incidents that were caught by law enforcement and prosecuted by the courts are represented, but was still shown to be useful in predicting the number of human trafficking crimes that were encountered by the court, as well as other types of crimes that were highly indicative of human trafficking. Additionally, the authors noted that any human trafficking records that are present in the dataset can be identified as human trafficking with a high confidence, as they were identified through due process in a court of law. The court records were used to train a random forest classifier and measure the correlation between the predicted number of trafficking cases and the actual number, achieving correlations above 0.8.

The complex relationship between drug crimes and human trafficking crimes has been documented and discussed in Roe-Sepowitz (2019), Shelley (2012) and Tripp and McMahon-Howard (2016). Not only is there an overlap in drug and human trafficking routes, but also various types of drugs intersect in different ways with the many forms of human trafficking. Drugs have an important role in recruiting new victims, maintaining them in an exploited state, and maximizing their exploitation (U.S. Department of Health and Human Services, 2018; US v. Pipkins, 2004). Meanwhile, stimulants are used to make individuals capable of working longer hours or more intensely (The Daily Star, 2011). As seen in this review, human trafficking crime networks have been studied individually and there is a lack of research in analyzing drug networks and human trafficking networks jointly, despite work by Bahulkar et al. (2018) showing that a better understanding of the interactions of crime networks can be used to more efficiently combat them. We try to fill this gap in by providing a dataset containing data that suits the needs of both drug and human trafficking research. The next section describes how a combination of these related works were expanded on to create a useful human trafficking and drug dataset of crime records and incident reports for the state of Kentucky, United States. New datasets and angles could provide additional insight into these complex relationships.

3. Dataset integration

Three crime data sources were collected and merged for this study. All three crime sources were either only reporting on the U.S. state of Kentucky (Kentucky Online Offender Lookup [KOOL] and Louisville Open Data [LOD]), or filtered to only contain results for the U.S. state of Kentucky (FBI). Each data source contains unique features, such as crime classifications, and unique challenges in collection and cleaning. If an activity happened to involve multiple crimes, for example a single suspect was arrested for a drug trafficking crime and a human trafficking crime at the same time, we assume this would be represented in the data using multiple records. Since the data sources did not have explicit classifications for human trafficking crimes, specific NIBRS codes and crime descriptions that were indicative of human trafficking were chosen with the oversight of a criminologist. This section describes the specifics of each crime data source.

3.1. Federal Bureau of Investigation National Incident-Based Reporting System

The United States Federal Bureau of Investigation (FBI) issues a variety of query-able crime related data on their website (Maxfield, 1999; United States Department of Justice, Federal Bureau of Investigation, 2021). This data is sourced from law enforcement agencies across the U.S. as part of their National Incident-Based Reporting System (NIBRS) and its standards. The goal of gathering, standardizing, and providing this information is to facilitate research into crime and law enforcement patterns (Addington, 2019).

Table 1
Features of the FBI dataset.

Feature name	Data type	Description
`incident_date`	datetime	The date the incident occurred
`county_name`	string	The county in which the crime occurred
`offense_name`	string	A description of the crime
`offense_category_name`	string	a broad category for the crime
`offense_code`	string	The NIBRS code of the offense
`suspect_using_name`	string	Any weapons or items that were used during the incident
`suspected_drug_name`	string	The type of drug (if any) that might have been present during the incident
`criminal_act_desc`	string	A standardized description of the crime
`prop_loss_desc`	string	A standardized description of any property lost during the incident (stolen/destroyed)

The information is provided as a collection of CSV files with instructions and code for importing into a SQL database. For the purposes of this research, we utilized the crime databases for the years 2017, 2018 and 2019, containing a total of 1,939,990 unique incidents. The database provides many tables and features. A selection of relevant features is provided in Table 1. The `offense_code` feature is the NIBRS code. This feature allows for filtering to extract only the crimes relevant to this study. The human trafficking codes are 40A (Prostitution), 40B (Assisting or Promoting Prostitution), and 370 (Pornography/Obscene Material). The drug incidents were found using codes 35A (Drug/Narcotic Violations) and 35B (Drug Equipment Violations).

The dataset does not provide a useful crime description field as can be found in Louisville Open Data's `UOR_DESC` feature and KOOL's Offense feature described in the following sections. Instead, the FBI's NIBRS dataset contains two separate fields; `criminal_act_desc` and `suspected_drug_name`, that can be concatenated to create a description of the crime. For example, an instance of a `criminal_act_desc` of "buying/receiving" and an instance of a `suspected_drug_name` of "heroin" can be concatenated to create the crime description "buying/receiving heroin".

3.2. Louisville Open Data

The Louisville Open Data Initiative (LOD) (Louisville Metro Government, 2021; Nguyen & Boundy, 2017) is a program from the city of Louisville, Kentucky, U.S.A. to increase the transparency of the city government and promote technological innovation. As part of LOD, a dataset of crime reports is made available online (Louisville Metro Government, 2021). The records contained within the LOD dataset represent any call for police service where a police incident report was generated. This does not necessarily mean a crime was committed, as an incident report can be generated before an investigation has taken place. The City of Louisville makes yearly sets of records available in comma-separated values (CSV) format, as well as a year-to-date report in CSV that is updated nightly. The LOD "Crime Reports" datasets for the years 2017 through 2019 contain 233,962 records. The features for this dataset are detailed in Table 4.

Once collected, LOD data was filtered for drug and human trafficking incidents through use of the `NIBRS_CODE` feature. The human and drug trafficking incidents were found using the NIBRS codes listed in Section 3.1. After filtering, the LOD "Crime Reports" dataset contains 38,421 records.

3.3. Kentucky Online Offender Lookup

The Kentucky Department of Corrections, as a service to the public, provides an online lookup of people currently in its custody called Kentucky Offender Online Lookup (KOOL) (Kentucky Department of Corrections, 2021). This web application offers users tools to search for sets of inmates based on features such as name, crime date, crime name, race, and gender. The data that KOOL searches contains only people who are currently under supervision of the state of Kentucky (or should be under supervision in the case of escape). This means that while a crime might have been committed on a given day, the person who committed the crime would only show up in the results if they were currently serving time. A person who has served their time and is no longer under supervision would no longer appear in the search results. Additionally, if a person is currently under supervision, the search results will include any past crimes for which that person served time. Practically, this means the data KOOL returns for a given set of search parameters, even searches that filter for crimes committed in previous years, can change daily as people are incarcerated and released.

The Kentucky Department of Corrections does not provide the underlying data that KOOL searches. Collecting this data required scraping the website that displayed the search results. This scraping was done using a Python script. The script queries KOOL for all the inmates with crimes committed during a given year, then for each result, follows the link to the inmate's profile and collects any relevant fields from the HTML tables contained within.

KOOL was queried for all crimes committed from the year 2017 through the year 2019. The script produced 129,479 crimes spread across 38 unique crime types. The features that were collected are described in Table 3. The collection process did not allow for a direct link between the crime description (a human-written description of the incident) and the crime type (e.g., "Dangerous Drugs" or "Assault"). In order to filter the KOOL data for drug crimes, a manual process of generating string matching rules for drugs was completed. A sample of the string matching rules are shown in Table 2. To filter the KOOL data for human trafficking crimes, a process of fuzzy string matching, explained further in Section 4.2, was used. If the crime description from the KOOL record in question matched *any* human trafficking crime description from Louisville Open Data above a threshold of 94, it was labeled human trafficking. This process resulted in each record being labeled with its appropriate crime category.

Table 2
Example of string matching rules for labeling KOOL crimes; if the crime description contains the string in column 1, it is considered a match.

KOOL string matching rule	Crime category
"POSS CONT"	Dangerous Drugs
"CULTIVATE IN MARIJUANA"	Dangerous Drugs
"RESISTING ARREST"	Obstructing the Police
"GIVING OFFICER FALSE"	Obstructing the Police
"ASSAULT"	Assault
"STRANGULATION"	Assault

Table 3
Features of the KOOL dataset.

Feature name	Data type	Description
PID	string	Person ID (PID) is a unique identifier for the offender
Offense	string	The crime that was committed
Offense_Type	string	Every crime category the offender has ever committed
Incident_Date	datetime	Date of crime incident
County	string	County offender is currently jailed in
Race	string	Offender's race
Gender	string	Offender's gender
Classification	string	The type of facility in which the offender is held
stint_start	datetime	Start of term
expected_end	datetime	Expected date of release
min_end	datetime	Earliest date of release
parole	datetime	Date of parole eligibility
height	string	Offender's height
risk	string	Danger posed by offender
spr_vsn_start	datetime	Probation start date
spr_vsn_end	datetime	Probation end date

Table 4
Features of the LOD dataset.

Feature name	Data type	Description
DATE_REPORTED	DateTime	The date the incident was reported to LMPD
DATE_OCCURED	DateTime	The date the incident actually occurred
UOR_DESC	String	Uniform offense reporting code for the criminal act committed
CRIME_TYPE	String	The crime type category
NIBRS_CODE	String	The code that follows the guidelines of the National Incident Based Reporting System.
ATT_COMP	String	Status indicating whether the incident was an attempted crime or a completed crime.
LMPD_DIVISION	String	The LMPD division in which the incident actually occurred
LMPD_BEAT	String	The LMPD beat in which the incident actually occurred
PREMISE_TYPE	String	The type of location in which the incident occurred (e.g., Restaurant)
BLOCK_ADDRESS	String	The location the incident occurred
CITY	String	The city associated to the incident block location
ZIP_CODE	Int	The zip code associated to the incident block location
ID	Int	Unique identifier for internal database

4. De-duplication strategy

In this section, the merging strategy for the three datasets will be elaborated. First, the data is cleaned and standardized. Next, all tables from the same data source are appended together to form three tables. Finally, the three tables are appended to each other after de-duplicating instances across all combinations of tables. Each de-duplication technique will be detailed in the following subsections. An example of a merged record can be seen in Appendix A. The number of records from each year, as well as the number of duplicates found, can be seen in Table 5. It should be noted that some features have not been included in the dataset for anonymization purposes. These features, however, were collected and used for the entity resolution tasks that follow in this section and are highlighted in Fig. 2. The list of features that were removed from the final dataset are as follows: Race, Gender, Person Identifier (PID), and Height.

This methodology broadly follows a similar data integration process used by Evans et al. (2019), wherein the data is collected from each source, the text is pre-processed as appropriate for the application, entity resolution is performed to find matching entities across multiple datasets, and relevant features are saved to a final dataset. While Evans et al. (2019) task of resolving Twitter data with cashtags could rely on Named Entity Recognition and standardized formatting for discovering entities, this crime data integration project requires techniques to handle descriptions of events that do not contain named entities. Additionally, semantic analysis of crime descriptions is required to determine if two unique descriptions refer to the same activity. We found that some

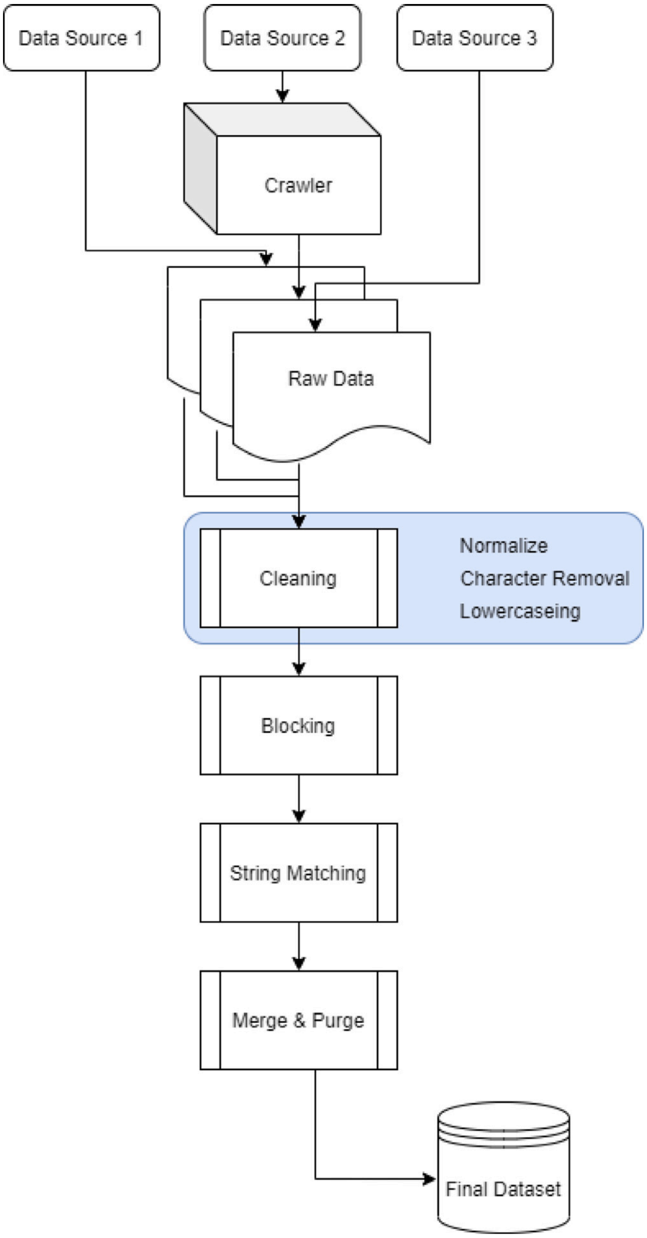

Fig. 2. Entity resolution flowchart.

of the entity resolution strategies used in Evans et al. (2019) were not applicable to our crime data. Our proposed solution looks to supplement the state-of-the-art data integration techniques with layers catered to the unique needs of entity resolution in crime data.

4.1. Cleaning

Text pre-processing followed methods highlighted in Baccouche et al. (2020) All dates were standardized to yyyy-mm-dd format. Lower-casing of text was performed across all features that were of string data type. All records without county data were imputed with the string “None” in place of the county name. `Premise Type` from LOD, which corresponds to `Location` in FBI, are categorical variables that were normalized using mapping. Additionally, a similar mapping process was applied to all Race and Gender variables across the tables. The normalization process also removed any special characters and extra white spaces. For example, “MALL/SHOPPING CENTER” is mapped to “shopping mall”.

Table 5
De-duplication results.

	Human trafficking			Drug trafficking		
Year	2017	2018	2019	2017	2018	2019
Before	1771	1831	1393	282,134	302,027	279,130
After	1696	1747	1352	276,158	296,169	275,554
Duplicates	75	84	41	5976	5858	3576

4.2. LOD and KOOL de-duplication for drug crimes

We start the de-duplication process by grouping together crime incidents in the KOOL and LOD tables if they shared both incident date and county into a subset. Next, the crime descriptions for all incidents in the subset were compared against each other. If two records matched on incident dates, county and crime descriptions, the data were all combined into one record. Crime descriptions were matched using the token_set_ratio method from Python's FuzzyWuzzy (Cohen, 2011; Rao et al., 2018) fuzzy string matching library. This method transforms the strings to be compared by reordering the words; words that are common to both strings are placed at the start of the string and the remaining words are placed at the end. The edit distance between the two strings is then calculated using Levenshtein distance (Zhang et al., 2017) with a threshold set to 0.94. The threshold was chosen after testing a range of threshold values from 0.5 to 0.99 on a random sample of annotated data. String matching alone was sufficient for this de-duplication due to the two sources sharing very similar standards for describing crimes, and major differences were in formatting and word order.

4.3. LOD and FBI de-duplication for drug crimes

All records present in the LOD and FBI datasets were first blocked together if incident date, county, premise type and NIBRS code values matched exactly. Crime descriptions in the blocked set were then compared using a two-layer method. The first layer uses the token_set_ratio method with a threshold set to 0.7. All pairs of records passing the first layer are passed to a Phrase2Vec embedding layer, trained on our crime description data, with a threshold of 0.5 for the cosine similarity metric. Phrase2Vec (Artetxe et al., 2018; Wu et al., 2020) is an adaptation of the skip-gram Word2Vec (Mikolov et al., 2013) text embedding at the phrase level. In Asghari et al. (2020), Word2Vec embeddings of short snippets of natural language, in this case tweets from Twitter streaming data, are used as a component of an entity resolution framework. A two-layer method was more suitable for this set of data because the embedding layer captures semantics that are missed with the string matching method alone.

4.4. KOOL and FBI de-duplication for drug crimes

Duplicates were removed between all drug incidents in the KOOL and FBI set, similarly to how they were removed between the LOD and FBI set. Records were first blocked together if they matched exactly on incident date, county, sex, and race. Finally, the crime descriptions for the blocked records were passed through the same two-layer method as described in Section 4.3.

4.5. De-duplication for human trafficking related crimes

Due to the relatively smaller size of the set of crimes related to human trafficking in the KOOL table, the de-duplication process for any records that occurred in the KOOL tables utilized a human-assisted approach. Additionally, crime descriptions for sex trafficking related crimes in the FBI tables were shorter and lacked sufficient detail to be used in the de-duplication process. With the exception of the aforementioned modifications, the de-duplication process for the subset of crimes that were related to human trafficking was identical to the process for drug trafficking related crimes detailed in the previous sections. Once duplicate records are found, the "merge and purge" strategy is executed. A detailed diagram of this process can be seen in Fig. 3.

4.6. Data curation and feature generation

The final set of features included in the dataset are a mix of features collected from the different data sources and generated features that might be useful for analysis. Examples of excluded features are source-specific case numbers and personally identifiable information such as age, sex, and race. The features collected from the data sources can be found in Table 6. Some granularity was lost during the merging process. Incidents are labeled by county in the final dataset, but are labeled by zip code, block address, and city in LOD. These features were omitted because they were only available for a subset of records from a single county. New features of drug classes were generated as detailed in Table 7 and included in the final dataset. All crime descriptions were run against every rule in the table. Some drug classes were combined to create larger sets (i.e., the rules for cocaine, amphetamines, and crack being used to generate a set of drugs labeled "stimulants".) The drug class for each drug record is one-hot encoded and appended to the dataset. For example, a cocaine possession incident would have attributes `is_cocaine`, `is_stimulant`, `is_all`, and `is_possessing` being 1 and all other such labels being 0. A total of twenty-one features were added as a result of the process. The final dataset can be accessed at http://uofllogistics.org/?page_id=2649.

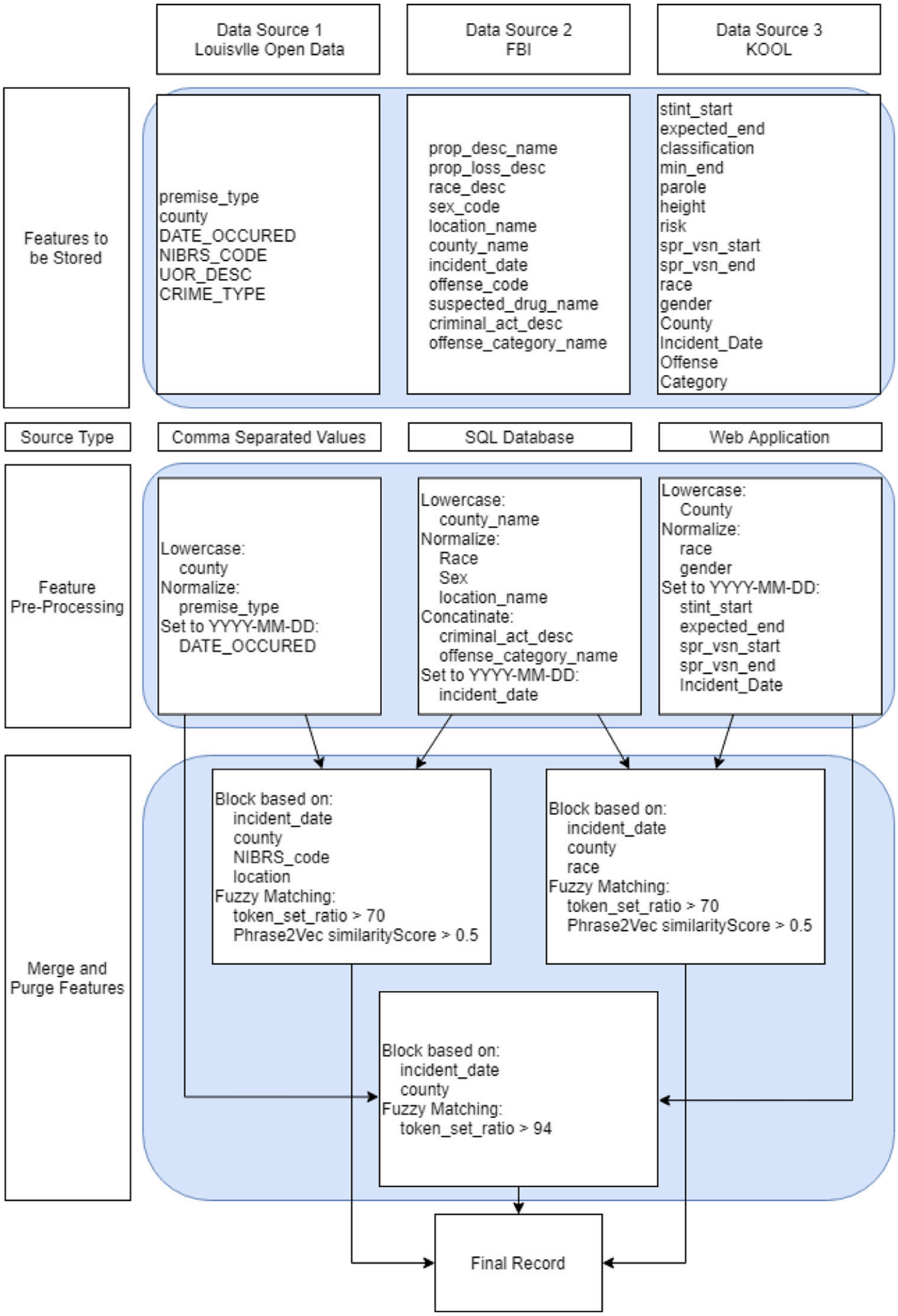

Fig. 3. Feature pipeline with record merging.
Source: Adapted from Evans et al. (2019).

5. Case study: A spatio-temporal analysis of drug and human trafficking incidents

We present a case study using the combined dataset for exploration of potential interactions between drug crimes and other crimes such as human trafficking. We hypothesize that we can use this new combined dataset to predict human trafficking incidents given specific patterns in drug crimes. Spatio-temporal analysis and prediction of crime is a continually evolving field that has

Table 6
Final non-generated feature set along with the source of each feature. The generated features in the final set can be found in Table 7.

Feature	Source
prop_desc_name	FBI
prop_loss_desc	FBI
stint_start	KOOL
expected_end	KOOL
classification	KOOL
min_end	KOOL
parole	KOOL
risk	KOOL
spr_vsn_start	KOOL
spr_vsn_end	KOOL
location	merged
county	all
incident_date	all
NIBRS_code	LOD + FBI
crime_desc	merged
crime_cats	FBI
merged_categories	merged
preprocessed_desc	KOOL + LOD

Table 7
Generated features and the rules for each feature according to drug type.

Drug feature	Rules	Notes
is_fentanyl	“fent” in crime_desc	Matches carfentanyl too
is_synthetic	“synthetic” in crime_desc or matches is_opioid	Lab-developed compounds
is_meth	“meth” or “amp” in crime_desc	
is_little_meth	“<2”, “<10 du”, “<10 d.u.”, or “<= 2” in crime_desc	Based on strings found in crime description feature
is_big_meth	“2 or >”, “>= 10 d.u.”, “>= 10 du”, or “>= 2” in crime_desc	Based on strings found in crime description feature
is_heroin	“hero” in crime_desc	
is_crack	“crack” in crime_desc	
is_cocaine	“cocaine” in crime_desc and not “crack” in crime_desc	“crack cocaine” often seen in description
is_trafficking	“traf”, “manufacturing”, “importing”, “selling”, “promoting”, or “distributing” in crime_desc	Signifies intent to sell the drug in question
is_posessing	“pos”, “using”, “consuming”, or “buying” in crime_desc	Signifies intent to use the drug in question
is_opiate	“opium”, “morphine”, “opiate”, or “codeine” in crime_desc	
is_opioid	Matches is_opioid or is_heroin	
is_stimulant	“stimulant” in crime_desc or matches is_crack, is_cocaine, or is_meth	
is_stim_poppy	matches is_stimulant, is_opioid, is_opiate, is_heroin	Grouping of stimulants and poppy-based drugs
is_hallucinogen	“lsd”, “hall”, “pcp” in crime_desc	
is_weed	“marij” in crime_desc	
is_big_weed	matches is_weed and “>5 lbs”, “gt 5 lbs”, “more than 5 lbs”, “5 plants or more”, “5 plants or gt”, or “5 plants or >”in crime_desc	Based on strings found in crime description feature
is_little_weed	matches is_weed and “less than 8 oz-”, “<8 oz”, “lt 5 plants”, “<8 oz”, “<5 plants”, “under 8 oz”	Based on strings found in crime description feature
is_meth_ops	matches is_meth, is_opioid, or is_opiate	Grouping of meth and poppy-based drugs
is_other	does not match any rule above	Any drug type not accounted for above
is_all	all records	

adapted to availability of new datasets and crime data types (Catlett et al., 2019; Hossain et al., 2020; Zhao & Tang, 2017). The features we determined were relevant for this analysis were each generated feature in Table 7 along with the spatio-temporal features county and date. The data to be analyzed for this case study was processed by generating 2 × 2 sliding contingency tables with a ninety-eight-day reference window followed by a 7 day analysis window as seen in Table 8, similar to temporal analysis work done by Boecking et al. (2019). Fisher’s Exact Test was then implemented to calculate the significance in the local (county level) and regional (state level) exceedances in the per-capita crime rate for the 7 day analysis window and the per-capita crime rate calculated using the 98 day reference window for each drug type in Table 7. Exceedance was determined using the p-value generated from Fisher’s Exact Test, shown in Eq. (1) where sr, cr, sa, and ca can be found in Table 8 and N is the sum of sr, cr, sa, and ca.

$$p = \frac{(sr+cr)!(sa+ca)!(sr+sa)!(cr+ca)!}{N!(sr!cr!sa!ca!)} \tag{1}$$

Table 8
Contingency table format for analyzing drug crime rate. All numbers are crimes per 1 million people.

	State-Wide	County
98 day reference window	*sr*	*cr*
7 day analysis window	*sa*	*ca*

The contingency tables were computed for all drug classes and all windows from April 9, 2017 (the first date to have 98 days of data preceding it to calculate the reference window crime rate) to December 24, 2019 (the last 7 days of data for which an analysis window can be computed). This resulted in roughly 110,000 analysis windows spread over 3 years and 120 counties. The analysis windows were annotated with a binary target variable that identified each record by whether or not a human trafficking incident occurred in the same county and time frame. The final generated dataset included the state-wide reference window crime rate, the state-wide analysis window crime rate, the county-wide reference window crime rate, the county-wide analysis window crime rate, and the p-value for each drug type as features, for a total of 105 features. XGBoost classifiers (Chen & Guestrin, 2016; Chen et al., 2018) were fit on the data to test our hypothesis that this new combined dataset can be used to predict human trafficking incidents by analyzing drug incidents.

Two separate models were trained: one for the entire state of Kentucky and one for just Jefferson County. Jefferson County is singled out because it has its own dataset, Louisville Open Data, that contains not only convicted crimes, but reported incidents that may not have resulted in a criminal charge. The contingency windows still used the data from the other counties to calculate the regional reference and analysis window, but the training was only performed on records for Jefferson County. A truncated example of the analyzed data can be seen in Appendix B.

5.1. Results and discussion

Model performance for all counties can be seen in Table 9, while performance for Jefferson County can be seen in Table 10. Principal Component Analysis (PCA) (Nobre & Neves, 2019) was performed as part of the hyperparameter tuning for dimensionality reduction to reduce processing time. PCA was chosen because the features of the dataset exhibited multicollinearity. The hyperparameters were tuned using SciKit Learn's RandomizedSearchCV function (Buitinck et al., 2013). The number of principal components used, as well as other hyperparameter values, can be found in Table 12. For this sort of analysis, avoiding false negatives is more important than avoiding false positives, as we would prefer to incorrectly label a window as having human trafficking than miss a human trafficking window. With this consideration, the confusion matrix for both models can be seen in Table 11. False negatives present a concern for the All Counties model, while false positives are a concern for the Jefferson County model. The feature importance of each model was judged by the number of correct classifications that would have been incorrect if not for that feature, also called the gain. The top ten features by gain can be seen in Figs. 4 and 5. The local reference window (the crime rate for a drug type across the preceding 98 days in that county) seemed to be a particularly useful indicator, making 6 of the top 10 features by gain. In addition to F-1 score and accuracy, Matthews correlation coefficient (MCC) was calculated as it can offer a better metric for evaluating the performance of machine learning models on unbalanced datasets such as this (Chicco & Jurman, 2020). MCC can be seen in Eq. (2), where *TP*, *TN*, *FP*, and *FN* are the number of true positive results, true negative results, false positive results, and false negative results respectively.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \tag{2}$$

The Jefferson County model did not perform as strongly as the state-wide model. This may be due to the fewer number of records for the Jefferson County model (118,542) compared to the All Counties model (758,621), or due to the nature of the additional Louisville Open Dataset records containing incidents that may not have resulted in a criminal conviction. Previous experiments that we tried did not use data differentiated by drug types, instead we trained a model on aggregate values of any drug related crimes. Higher classification performance was achieved once the model could analyze each drug individually. Additionally, the MCC score of the analysis performed on the integrated dataset (0.86), was higher than the highest performing individual dataset (0.8). Other models with different window lengths may work better in similar case studies using data from different states. What works for the state of Kentucky might not capture the relationships present in another area of the United States.

During the exploratory analysis, we tried correlating drug crimes and human crimes by month, first looking for relationships between the spikes in drug trafficking and human trafficking in the same month, then looking for relationships in succeeding months. However, this method did not provide a clear pattern. This may be because using the month in which a crime occurred as a blocking scheme is arbitrary. With this method, a crime that occurred on January 31st would be analyzed separately from a crime that occurred on February 1st even though they occurred only a day apart. The windowing method removed this restriction, providing better results.

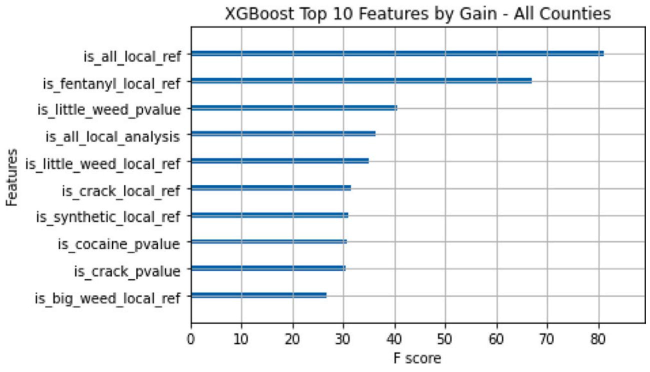

Fig. 4. Feature importance for all counties.

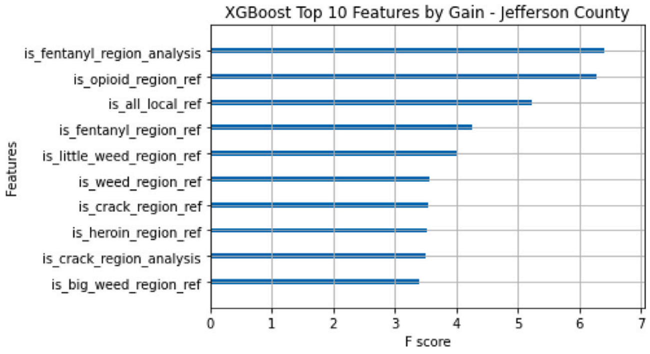

Fig. 5. Feature importance for Jefferson County.

Table 9
Performance of XGBoost Binary Classifier classifying human trafficking windows for all counties.

Class	Precision	Recall	F1-score
Human trafficking not in window (Class 0)	0.99	0.99	0.99
Human trafficking in window (Class 1)	0.84	0.92	0.88
Accuracy	0.98		
MCC	0.86		

Table 10
Performance of XGBoost Binary Classifier classifying human trafficking windows for Jefferson County.

Class	Precision	Recall	F1-score
Human trafficking not in window (Class 0)	0.95	0.59	0.73
Human trafficking in window (Class 1)	0.82	0.98	0.89
Accuracy	0.85		
MCC	0.66		

6. Conclusions and future work

Studies of crime have struggled to find a complete picture of criminal activity for an area or time frame, as crime data can be difficult to collect. A practical method of advancing crime research would thus be a data-centric approach rather than model-centric. Integrating data from multiple crime data sources, such as different law enforcement agencies across different levels of government, provides a clearer picture of crime in an area. This helps by adding crimes exclusive to a dataset and unique information about

Table 11
Confusion matrices for best performing models for all counties and Jefferson County only.

	All counties	
N = 23,760	Pred No	Pred Yes
Actual No	21,902	284
Actual Yes	126	1448

	Jefferson County	
N = 198	Pred No	Pred Yes
Actual No	41	28
Actual Yes	2	127

Table 12
Hyperparameters for best performing models for Jefferson County and all counties.

	Jefferson County	All counties
estimators	20	10
subsample	0.8	0.6
scale_pos_weight	1	14
pca_n_components	5	2
min_child_weight	1	1
max_depth	8	8
learning_rate	0.01	0.05
gamma	1.5	0.5
colsample_bytree	0.6	1.0

crimes that multiple datasets have in common. The methods we have presented have been demonstrated by building and sharing a human trafficking and drug trafficking dataset for further research.

We have presented a dataset and generalized dataset building framework to address the issue of finding sufficient human trafficking data to allow for machine learning solutions to analyzing human trafficking data. This solution combines crimes datasets from multiple sources for the state of Kentucky to allow researches to discover patterns and information that would not be detectable otherwise. We make this dataset available so that other can expand on the drug and human trafficking analysis in the future. Possible avenues of future work are leveraging other datasets made available from Louisville's Open Data initiative to expand on smart city crime analysis research done by Catlett et al. (2019) to the city of Louisville. Additionally, implementing the methodology proposed in this paper to other states that make crime data available to complement the NIBRS dataset could offer additional insights, especially in states that see inconsistent NIBRS participation. Future work that could improve the de-duplication strategy is exploring generative adversarial network (GAN) models for text features of crime records, as seen in Khorshidi et al. (2020). This could be used to improve the classification, entity resolution, and de-duplication process by using GAN models that have been trained on labeled data to classify new crime data to be integrated.

CRediT authorship contribution statement

Sadaf Ahmed: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Monica Gentili:** Conceptualization, Validation, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Daniel Sierra-Sosa:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Adel S. Elmaghraby:** Conceptualization, Methodology, Validation, Resources, Writing – review & editing, Supervision, Project administration, Funding acquistion.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This work has been supported by the Logistics and Distribution Institute (LoDI) at the University of Louisville, United States. This study was also partially funded by National Science Foundation, USA (EAGER: ISN:/1838306). This research was partially funded by a National Centers of Academic Excellence in Cybersecurity, USA grant (H98230-21-1-0154), which is part of the National Security Agency. The funding sources had no role in the design, analysis, write-up, or decision to submit for publication.

Appendix A. Record merging results

Table A.13
Before and after of a merge and purge operation on two duplicate records for a case of drug possession.

Feature	KOOL record	FBI record	Final record
prop_desc_name	nan	drug equipment	drug equipment
prop_loss_desc	nan	seized (to impound property that was not previously stolen)	seized (to impound property that was not previously stolen)
risk	low	nan	low
spr_vsn_start	10/9/2018	nan	10/9/2018
spr_vsn_end	10/8/2022	nan	10/8/2022
location	nan	convenience store	convenience store
county	barren	barren	barren
incident_date	2017-08–06	2017-08-06	2017-08-06
NIBRS_code	nan	35A	35A
crime_desc	nan	poss of marijuana	poss of marijuana
crime_cats	nan	dangerous drugs	dangerous drugs
merged_categories	drugs	drugs	drugs
preprocessed_desc	poss of marijuana	nan	poss of marijuana

Appendix B. Contingency window analysis

Table B.14
An example of analyzed data showing the contingency matrix values for is_weed for a sample of 7 day time windows starting at the date indicated date in the "day" column for the indicated county in the "county" column. The column "ht" indicates a human trafficking incident occurred in this county during this window. This example is truncated, the actual data has the contingency matrix values for each drug type as features.

county	day	is_weed loc ref	is_weed reg ref	is_weed loc a	is_weed reg a	is_weed pvalue	...	ht
garrard	4/9/2017	95928	126679	0	127378	0		0
wayne	4/9/2017	29798	126679	0	127378	0		0
jefferson	4/9/2017	140784	126679	162697	127378	2.61E−147		1
trigg	4/9/2017	99576	126679	39547	127378	0		0
leslie	4/9/2017	992	126679	0	127378	1.19E−299		0

References

Addington, L. A. (2019). NIBRS as the new normal: What fully incident-based crime data mean for researchers. In *Handbook on crime and deviance* (pp. 21–33). Springer.

Alvari, H., Shakarian, P., & Snyder, J. E. K. (2016). A non-parametric learning approach to identify online human trafficking. In *2016 IEEE conference on intelligence and security informatics* (pp. 133–138). http://dx.doi.org/10.1109/ISI.2016.7745456.

Artetxe, M., Labaka, G., & Agirre, E. (2018). Unsupervised statistical machine translation. In *Proceedings of the 2018 conference on empirical methods in natural language processing*. Brussels, Belgium: Association for Computational Linguistics.

Asghari, M., Sierra-Sosa, D., & Elmaghraby, A. S. (2020). A topic modeling framework for spatio-temporal information management. *Information Processing & Management*, *57*, Article 102340.

Baccouche, A., Ahmed, S., Sierra-Sosa, D., & Elmaghraby, A. (2020). Malicious text identification: Deep learning from public comments and emails. *Information*, *11*, 312.

Bahulkar, A., Baycik, N. O., Sharkey, T., Shen, Y., Szymanski, B., & Wallace, W. (2018). Integrative analytics for detecting and disrupting transnational interdependent criminal smuggling, money, and money-laundering networks. In *2018 IEEE international symposium on technologies for homeland security* (pp. 1–6). IEEE.

Bales, K., Murphy, L. T., & Silverman, B. W. (2020). How many trafficked people are there in Greater New Orleans? Lessons in measurement. *Journal of Human Trafficking*, *6*, 375–387.

Boecking, B., Miller, K., Kennedy, E., & Dubrawski, A. (2019). Quantifying the relationship between large public events and escort advertising behavior. *Journal of Human Trafficking*, *5*, 220–237. http://dx.doi.org/10.1080/23322705.2018.1458488.

Bouche, V., & Crotty, S. M. (2018). Estimating demand for illicit massage businesses in Houston, Texas. *Journal of Human Trafficking*, *4*, 279–297.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., & Varoquaux, G. (2013). API design for machine learning software: Experiences from the scikit-learn project. In *ECML PKDD workshop: Languages for data mining and machine learning* (pp. 108–122).

Burke, M. C., & Bruijn, B. (2017). Introduction to human trafficking: Definitions and prevalence. In *Human trafficking* (pp. 3–24). Routledge.

Catlett, C., Cesario, E., Talia, D., & Vinci, A. (2019). Spatio-temporal crime predictions in smart cities: A data-driven approach and experiments. *Pervasive and Mobile Computing*, *53*, 62–74.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *KDD '16, Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/2939672.2939785.

Chen, Z., Jiang, F., Cheng, Y., Gu, X., Liu, W., & Peng, J. (2018). XGBoost classifier for DDoS attack detection and analysis in SDN-based cloud. In *2018 IEEE international conference on big data and smart computing* (pp. 251–256). http://dx.doi.org/10.1109/BigComp.2018.00044.

Chicco, D., & Jurman, G. (2020). The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*, 1–13.

Cohen, A. (2011). URL https://chairnerd.seatgeek.com/fuzzywuzzy-fuzzy-string-matching-in-python/.

Diaz, M., & Panangadan, A. (2020). Natural language-based integration of online review datasets for identification of sex trafficking businesses. In *2020 IEEE 21st international conference on information reuse and integration for data science* (pp. 259–264). IEEE.

Dubrawski, A., Miller, K., Barnes, M., Boecking, B., & Kennedy, E. (2015). Leveraging publicly available data to discern patterns of human-trafficking activity. *Journal of Human Trafficking*, *1*, 65–85.

Evans, L., Owda, M., Crockett, K., & Vilas, A. F. (2019). A methodology for the resolution of cashtag collisions on Twitter–A natural language processing & data fusion approach. *Expert Systems with Applications*, *127*, 353–369.

Falade, A., Azeta, A., Oni, A., & Odun-ayo, I. (2019). Systematic literature review of crime prediction and data mining. *Review of Computer Engineering Studies*, *6*, 56–63.

Farrell, A., Dank, M., Kafafian, M., Lockwood, S., Pfeffer, R., Hughes, A., & Vincent, K. (2019). Capturing human trafficking victimization through crime reporting.

Farrell, A., & Reichert, J. (2017). Using US law-enforcement data: Promise and limits in measuring human trafficking. *Journal of Human Trafficking*, *3*, 39–60.

Getoor, L., & Machanavajjhala, A. (2012). Entity resolution: Theory, practice & open challenges. *Proceedings Of The VLDB Endowment*, *5*, 2018–2019.

Gholizadehy, S., Phillipsy, M., Hosseinabadi, M. T., Leon, D., & Rozier, J. (2020). Analysis of human trafficking in North Carolina based on criminal records: A framework to measure human trafficking trends. In *2020 IEEE international conference on big data* (pp. 1309–1315). IEEE.

Goodey, J. (2008). Human trafficking: Sketchy data and policy responses. *Criminology & Criminal Justice*, *8*, 421–442.

Hernández, M. A., & Stolfo, S. J. (1995). The merge/purge problem for large databases. *ACM Sigmod Record*, *24*, 127–138.

Hossain, S., Abtahee, A., Kashem, I., Hoque, M. M., & Sarker, I. H. (2020). Crime prediction using spatio-temporal data. In *International conference on computing science, communication and security* (pp. 277–289). Springer.

Kangaspunta, K. (2007). Collecting data on human trafficking: Availability, reliability and comparability of trafficking data. In *Measuring human trafficking* (pp. 27–36). Springer.

Kentucky Department of Corrections (2021). URL http://kool.corrections.ky.gov/.

Khorshidi, S., Mohler, G., & Carter, J. G. (2020). Assessing GAN-based approaches for generative modeling of crime text reports. In *2020 IEEE international conference on intelligence and security informatics* (pp. 1–6). IEEE.

Konrad, R. A., Trapp, A. C., Palmbach, T. M., & Blom, J. S. (2017). Overcoming human trafficking via operations research and analytics: Opportunities for methods, models, and applications. *European Journal of Operational Research*, *259*, 733–745.

Ku, C. -H., & Leroy, G. (2011). A crime reports analysis system to identify related crimes. *Journal of the American Society for Information Science and Technology*, *62*, 1533–1547.

Laura, L., & Me, G. (2017). Searching the web for illegal content: The anatomy of a semantic search engine. *Soft Computing*, *21*, 1245–1252.

Louisville Metro Government (2021). URL https://data.louisvilleky.gov/.

Marciani, G., Porretta, M., Nardelli, M., & Italiano, G. F. (2017). A data streaming approach to link mining in criminal networks. In *2017 5th International conference on future Internet of Things and cloud workshops* (pp. 138–143). IEEE.

Maxfield, M. G. (1999). The national incident-based reporting system: Research and policy applications. *Journal of Quantitative Criminology*, *15*, 119–149.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. arXiv preprint arXiv:1310.4546.

Nagpal, C., Miller, K., Boecking, B., & Dubrawski, A. (2017). An entity resolution approach to isolate instances of human trafficking online. In *Proceedings of emnlp '17 3rd workshop on noisy user-generated text* (pp. 77–84).

New York City (2021). NYC open data. URL https://opendata.cityofnewyork.us/.

Nguyen, M. T., & Boundy, E. (2017). Big data and smart (equitable) cities. In *Seeing cities through big data* (pp. 517–542). Springer.

Nobre, J., & Neves, R. F. (2019). Combining principal component analysis, discrete wavelet transform and XGBoost to trade in the financial markets. *Expert Systems with Applications*, *125*, 181–194.

Polaris Project (2021). Myths, facts, and statistics. URL https://polarisproject.org/myths-facts-and-statistics/.

Portnoff, R. S., Huang, D. Y., Doerfler, P., Afroz, S., & McCoy, D. (2017). Backpage and bitcoin: Uncovering human traffickers. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1595–1604).

Rao, G. A., Srinivas, G., VenkataRao, K., & Prasad Reddy, P. (2018). A partial ratio and ratio based fuzzy-wuzzy procedure for characteristic mining of mathematical formulas from documents. *ICTACT Journal on Soft Computing*, *8*, 1728–1732.

Robinson, D., & Scogings, C. (2018). The detection of criminal groups in real-world fused data: Using the graph-mining algorithm "GraphExtract". *Security Informatics*, *7*, 1–16.

Roe-Sepowitz, D. (2019). A six-year analysis of sex traffickers of minors: Exploring characteristics and sex trafficking patterns. *Journal of Human Behavior in the Social Environment*, *29*, 608–629.

Sarker, A., DeRoos, A., & Perrone, J. (2019). Mining social media for prescription medication abuse monitoring: A review and proposal for a data-centric framework. *Journal of The American Medical Informatics Association*, *27*, 315–329. http://dx.doi.org/10.1093/jamia/ocz162, arXiv:https://academic.oup.com/jamia/article-pdf/27/2/315/34152138/ocz162.pdf.

Shelley, L. (2012). The relationship of drug and human trafficking: A global perspective. *European Journal on Criminal Policy and Research*, *18*, 241–253.

Strom, K. J., & Smith, E. L. (2017). The future of crime data: The case for the national incident-based reporting system (NIBRS) as a primary data source for policy evaluation and crime analysis. *Criminology & Public Policy*, *16*, 1027–1048.

The Daily Star (2011). Sex workers forced to take harmful drug. URL https://www.thedailystar.net/news-detail-195013.

Tong, E., Zadeh, A., Jones, C., & Morency, L. -P. (2017). Combating human trafficking with multimodal deep models. In *Long Papers*: *Vol. 1*, *Proceedings of the 55th annual meeting of the association for computational linguistics* (pp. 1547–1556). Vancouver, Canada: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P17-1142, URL https://www.aclweb.org/anthology/P17-1142.

Tripp, T. M., & McMahon-Howard, J. (2016). Perception vs. reality: The relationship between organized crime and human trafficking in metropolitan Atlanta. *American Journal of Criminal Justice*, *41*, 732–764.

United States Department of Justice, Federal Bureau of Investigation (2021). National incident-based reporting system. URL https://www.fbi.gov/services/cjis/ucr/nibrs.

U. S. Department of Health and Human Services (2018). Fact sheet: Human trafficking. URL https://www.acf.hhs.gov/otip/fact-sheet/resource/fshumantrafficking.

US v. Pipkins (2004). *US v. Pipkins: vol. 378* (p. 1281). Court of Appeals, 11th Circuit.

Wu, Y., Zhao, S., & Li, W. (2020). Phrase2Vec: Phrase embedding based on parsing. *Information Sciences*, *517*, 100–127.

Yang, D., Heaney, T., Tonon, A., Wang, L., & Cudré-Mauroux, P. (2018). CrimeTelescope: Crime hotspot prediction based on urban and social media data fusion. *World Wide Web*, *21*, 1323–1347.

Zhang, S., Hu, Y., & Bian, G. (2017). Research on string similarity algorithm based on Levenshtein distance. In *2017 IEEE 2nd advanced information technology, electronic and automation control conference* (pp. 2247–2251). IEEE.

Zhao, X., & Tang, J. (2017). Modeling temporal-spatial correlations for crime prediction. In *Proceedings of the 2017 ACM on conference on information and knowledge management* (pp. 497–506).

Zhou, B., Chen, L., Zhao, S., Zhou, F., Li, S., & Pan, G. (2021). Spatio-temporal analysis of urban crime leveraging multisource crowdsensed data. *Personal and Ubiquitous Computing*, 1–14.

Zhou, B., Chen, L., Zhou, F., Li, S., Zhao, S., & Das, S. K. (2020). Escort: Fine-grained urban crime risk inference leveraging heterogeneous open data. *IEEE Systems Journal*.

Zhu, J., Li, L., & Jones, C. (2019). Identification and detection of human trafficking using language models. In *2019 European intelligence and security informatics conference* (pp. 24–31). IEEE.