Hierarchical Approaches to Text-based Offense Classification

Jay Choi
University of Michigan

David Kilmer
Measures for Justice

Michael Mueller-Smith*
University of Michigan

Sema A. Taheri
Measures for Justice

January 19, 2022

Abstract

Researchers and statisticians working with multi-jurisdictional crime data often must classify offense information into a common scheme. No comprehensive standard currently exists, nor a mapping tool to transform raw description information into offense type variables. This paper introduces a new offense classification schema, the Uniform Crime Classification Standard (UCCS), and the Text-based Offense Classification (TOC) tool to address these shortcomings. The UCCS schema draws from existing Department of Justice efforts, with the goal of better reflecting offense severity and improving offense type disambiguation. The TOC tool is a machine learning algorithm that uses a hierarchical, multilayer perceptron classification framework, built on 313,209 unique hand-coded offense descriptions from 24 states, to translate raw description information into UCCS codes. In a series of experiments, we test how variations in data processing and modeling approaches impact recall, precision, and F1 scores to assess their relative influence on the TOC tool's performance. The code scheme and classification tool are collaborations between Measures for Justice and the Criminal Justice Administrative Records System.

Keywords: criminal justice, offense description, text analysis, data science
JEL Codes: K42, C45, C55

*Corresponding author: mgms@umich.edu.

1 Introduction

Criminal justice systems across the world are tasked with responding to a wide range of activity deemed to be illegal and against public interest. These offenses range from driving while intoxicated to vehicular manslaughter, from possessing illegal narcotics to possessing stolen property, and from conspiracy to commit murder to capital murder offenses. Subtle distinctions in offense descriptions reflect deep differences in offense severity, potential motive, risk to public safety, and optimal responses by law enforcement. Accurately classifying offenses helps us to better understand the nature of crime, allow policy makers to evaluate the effectiveness of criminal justice policies, and provide the research community and the public a common measurement system to effectively analyze crime trends over time and across jurisdictions (Maxfield 1999; National Academies of Sciences, Engineering, and Medicine 2016; Strom and Smith 2017; Langton, Planty, and Lynch 2017). Organizing and classifying criminal activity is a core premise of a well-functioning system of criminal justice (Baumer, Velez, and Rosenfeld 2018; Wormeli 2018).

This paper introduces the text-based offense classification (TOC) tool to map unstructured offense description information to the Uniform Crime Classification Standard (UCCS) charge codes. Prior schema efforts have been inadequate due to their exclusive focus on felony-level offenses, lack of internal consistency, or lack of specificity on emerging crime types (e.g. possession of methamphetamine or possession of heroin versus possession of illegal drugs). The UCCS schema addresses a comprehensive set of violent, property, drug, traffic, and public order offenses at all levels of criminal severity, with modifiers to distinguish between completed, attempted, and conspired acts.

The production and maintenance of criminal justice data largely records offense information in the form of free entry text fields (National Academies of Sciences, Engineering, and Medicine 2016). In order to utilize the UCCS schema, one must identify how to map a specific text of an offense description to a corresponding UCCS code. While there are only 257 potential UCCS values, the set of potential offense descriptions is boundless. For instance, the Criminal Justice Administrative Records System (CJARS), which currently holds data on over 175 million criminal justice events from 25 states and has roughly 4 million unique offense descriptions (Finlay, Mueller-Smith, and Papp 2021). Several factors drive this high number of unique descriptions: (1) varying abbreviations used by local jurisdictions across

the country, (2) differences in cited state and municipal statute numbers, (3) typographical errors at data entry, and (4) varying degrees of detail contained within the field. As a consequence, even if two agencies or researchers agree on a common classification schema and have sufficient implementation resources, it is unlikely that they will arrive at consistent data classifications in practice due to the multitude of discretionary choices required to map the raw data to analyzable codes.

We leverage a combination of text classification and supervised machine learning methods to build a bridge between free entry offense descriptions and UCCS codes. This bridge takes the form of a hierarchical, multilayer perceptron model trained on 313,209 hand-coded observations vetted by trained professionals, which we refer to as the TOC tool.[1] In this framework, there are three levels of classification: (i) broad offense code, (ii) specific offense code, and (iii) offense modifier.[2] Our approach helps leverage meaningful common descriptors that otherwise might be ignored by the algorithm (e.g. *possession* of stolen property, *possession* of a illegal narcotics, and distribution of illegal narcotics). We find the TOC tool generates F1 scores, which combine precision and recall, in the range of 0.957-0.995 for broad offense type (e.g. violent, drug, property, etc), and 0.845-0.991 for full offense codes (e.g. Violent - Murder, Attempted). Importantly, out-of-state predictions yield similar performance levels (overall F1 score for broad type prediction: 0.983 $\rightarrow$ 0.968; overall F1 score for full UCCS offense code prediction: 0.963 $\rightarrow$ 0.935), suggesting the TOC tool will perform well when applied to the 26 states not currently covered in the training sample.

We explore a range of factors that contribute to our realized performance statistics, including: text preprocessing, feature construction and selection, hierarchical versus flat models, random forest versus neural network machine learning approaches, and the size of the training sample. Training sample size and feature construction have the largest relative impact on performance statistics, while variations in preprocessing, feature selection, and classification approaches yield more minor gains in performance in this context.

The UCCS schema and the TOC tool are intended to be used as an open-source system and thus provide administrative users, the research community, and the general public with a common classification system to ensure reproducible statistics for conducting comparative analysis. This system will especially

[1]These records come from the following states: Alabama, Arkansas, Arizona, California, Colorado, Connecticut, Florida, Illinois, Indiana, Kansas, Maryland, Michigan, Minnesota, Mississippi, North Carolina, North Dakota, Nebraska, New Jersey, Ohio, Oregon, Pennsylvania, Texas, Utah, and Wisconsin.

[2]Together these constitute what we refer to as a *full offense code*.

be useful for processing big data in which a large workforce would otherwise be needed to manually classify thousands of offense descriptions. The remainder of the paper is organized as follows: Section 2 reviews prior classification efforts and introduces the UCCS schema. Section 3 introduces the data sources used in producing and evaluating the TOC tool. Section 4 discusses the TOC tool, baseline performance statistics, and performance stability when applied to states not covered in the training data. Section 5 investigates the relative importance of a range of modeling decisions, including training data size, feature unit and selection method, number of selected features, and machine learning algorithm type and classification method. Section 6 provides discussion of the findings, while Section 7 concludes.

2 A new offense classification scheme

National Academies of Sciences, Engineering, and Medicine (2016) recently proposed four design principles for modernizing crime statistics in the United States:

Principle 1: "classification should not be limited to current crime statistics' traditional focus on violent or street crime, and should encompass new and emerging crime types"

Principle 2: "classification should satisfy all the properties of a fully realized classification for statistical purpose"

Principle 3: "classification should follow - to the greatest extent possible - an attribute-based approach, yet should also be a hybrid with a code- or definition-based approach due to the nature of the topic"

Principle 4: "classification should be designed in order to enable and promote comparisons between jurisdictions, between periods of time, and across state and national boundaries"

Although existing schemes (described below) satisfy many of the aforementioned design choices, no single schema meets all of the recommended criteria. As a result, we developed the new UCCS scheme, which is built from adopting key features of existing schemes while satisfying all four design principles recommended by National Academies of Sciences, Engineering, and Medicine (2016).

2.1 Uniform Crime Reporting Program

As one of the most prominent sources of crime statistics in the United States, the Federal Bureau of Investigation's Uniform Crime Reporting (UCR) Program has collected data from more than 18,000 participating agencies since its inception in 1930. Historically, the UCR Program collected aggregated monthly reports using the UCR's Summary Reporting System (SRS) which collected information for only ten "Part I offenses," which are described in Table A1. Meanwhile, "Part II offenses" are designated for less severe crimes that may not always be captured by the police. In effect, SRS collects information for "Part II offenses" only from recorded arrests.

However, these reports using the SRS were compiled using a hierarchy rule such that only the most severe crime for a given incident was reported. In an effort to improve the quality of data collection, the Bureau of Justice Statistics (BJS) and the FBI began a multiyear study in the early 1980s to re-evaluate the SRS for user needs, to identify potential improvements in the existing system, and to design a new data collection system which accounts for system changes (Poggio et al. 1985). The design recommendations from this study, such as the omission of hierarchy rule and the transition from summary-based reporting to "unit-record" reporting, provided the "Blueprint" for the modern UCR system, the National Incident-Based Reporting System (NIBRS).

Starting January 1, 2021, the FBI officially retired the SRS and adopted the NIBRS as the new standard criminal data collection system moving forward. Unlike the SRS's summary-based reporting system which uses the hierarchy rule, NIBRS uses an incident-based crime reporting system that provides information on 46 different classifications and 53 other contextual elements such as victim information (Federal Bureau of Investigation 2019). Although the NIBRS aims to provide more information in regards to the specific circumstances and context of a crime, one of the key challenges for researchers and public users is the complexity of the data collection system which has led to slow adoption rate by law enforcement agencies (Strom and Smith 2017; National Academies of Sciences, Engineering, and Medicine 2016; Maxfield 1999). As an additional challenge, users must have the technical knowledge to aggregate the incident-level data as well as have the necessary understanding of the NIBRS data infrastructure to design their own sub-classification of offense types for their analysis.

With the addition of the "hacking/computer trespassing" offense type to NIBRS in 2017 (US Department of Justice 2021), NIBRS has demonstrated a willingness to adapt to new crime types while maintaining mutually exclusive offense categories. At the same time, the omission of certain crime types makes it less ideal for classifying offenses over time due to inconsistent coverage of crime types[3]. However, NIBRS weakly satisfies the third design principle since it is largely mapped using statute information despite its attribute-based approach where letters A through Z are used to provide additional context. For instance, NIBRS code of 26 maps broadly to Fraud Offenses while the suffix is used to denote specific types of fraud such as impersonation (26C) and welfare fraud (26D). Finally, NIBRS cannot be used to compare crime data reported in SRS due to latter's hierarchy rule (US Department of Justice 2011; Federal Bureau of Investigation 2017). In effect, Principle 4 cannot be satisfied until every participating law enforcement agencies have transitioned from SRS to NIBRS.

2.2 National Corrections Reporting Program

While the FBI's UCR program aims to consolidate arrest data from participating U.S. law enforcement agencies, the BJS's National Corrections Reporting Program (NCRP) collects offender-level data on prison and supervision admissions and releases. Similar to NIBRS, the NCRP is also comprised of multiple data files but one distinct characteristic of this data collection system is that it also provides a standardized offense classification schema and crosswalks for each state (Bureau of Justice Statistics 2020b). In addition, NCRP offense classification schema also designates specific offense codes to indicate whether an offense was attempted (offense code ends with 1 or 6) or conspiracy (ends with 2 or 7) to provide additional context to the offense as well as convenience to researchers interested in inchoate crimes. Finally, the NCRP offense classification schema also provides broader classification categories to facilitate research on sub-classification of offenses (Table A3).

Although the publicly available offense crosswalks make NCRP convenient for analyzing crime data, there remain limitations. Foremost, NCRP offense crosswalks do not provide significant information for the

[3]In 2008, the Advisory Policy Board (APB) of the Criminal Justice Information Services (CJIS) approved the removal of "90I - Runaway" from NIBRS collection efforts. Per APB recommendations in 2018, "90A - Bad Checks," "90E - Drunkenness," and "90H - Peeping Tom" were also omitted, effective January 1, 2021 (US Department of Justice 2021).

multitude of misdemeanor and low level offense types since the data collection itself is focused on prison and post-confinement records. As a result, the available crosswalks are mostly sufficient for researchers working with felony level offense data while lacking for other, broader research projects.[4] As part of the Conversion of Criminal History Records into Research Databases (CCHRRD) project initiated in 2009 (Bureau of Justice Statistics 2009), the National Opinion Research Center (NORC) at the University of Chicago has been developing the Criminal History Record Assessment and Research Program (CHRARP), primarily to better conduct recidivism studies (Bureau of Justice Statistics 2015, 2020a). As one of the project goals, NORC has been working on an algorithm to classify string offense descriptions to their charge codes, similar to TOC (Bureau of Justice Statistics 2015, 2020a). However, the publicly available files for NORC's CHRARPS only contain the bare minimum code and lack necessary system components to classify new offense descriptions, thus limiting the pool of users who can leverage this classification tool for their research.

A remaining issue is that the NCRP offense codes may not always be consistent for a given offense description. For instance, the description "MANSLAUGHTER" is associated with "013 – Homicide," "015 – Voluntary Manslaughter," and "030 – Manslaughter" in the 2020 version (Table A4). In effect, users without all of the identifying variables such as the statute code of the offense description will have to rely on subjective deduplication for consistent offense classification in their data. As a result, the inconsistent charge codes for a given description and the redundant offense categories (e.g. "220 - Forgery/Fraud", "810 - Forgery (Federal)", and "820 - Fraud (Federal)") in NCRP scheme does not make it the ideal classification scheme.

2.3 Uniform Crime Classification Standard (UCCS)

In order to address shortcomings of prior schemas, we created the UCCS schema. It is grounded in the original offense type delineations developed for the National Corrections Reporting Program (NCRP) in the early 1990s, but with modifications including reordering offenses to reflect their seriousness (e.g. moving felony forced sexual assault into more serious categories), adding clarifications to the NCRP codes to ensure coding consistency (e.g. blood alcohol levels, conspiracy), reclassifying DUI to its own

[4]Even for those focused on felony level offenses, because the crosswalks are static, users working with offense descriptions that are not currently included in the NCRP offense crosswalks will have to classify the descriptions themselves.

offense type, reclassifying many of the “other” and “public order” offenses to more specific definitions, and adding new codes for previously omitted offenses including human trafficking, amphetamine drug offenses, opiate drug offenses, and other prescription drug offenses.

UCCS is operationalized as a four digit offense code, that is hierarchical in nature (see Table A5). The first digit represents the *broad crime type*, which can take on values: 1 – Violent, 2 – Property, 3 – Drug Offense, 4 – DUI, 5 – Public Order, or 6 – Criminal Traffic. For each broad crime type, offense category codes are generated by enumerating from 01 to 99 where 99 is reserved for "Other" category within the broad crime type (e.g. if broad crime type is 1, then 99 maps to "Other Violent Offense"). The final digit is reserved as an offense modifier. This can be used to delineate whether: 0 – an offense was committed, 1 – an offense was attempted, or 2 – an offense was conspired.

Since UCCS offense categories are generated by enumerating from 01 to 99, the scheme satisfies Principle 1 as there are unmapped category codes in each broad crime type for adding new offense categories (e.g. Violent categories go up to 27 for "Hit and Run with Bodily Injury, Conspiracy", and then 99 for "Other Violent Offense"). It satisfies Principle 2 given that broad crime types and offense categories are defined as mutually exclusive and exhaustive categories. Furthermore, the 4-digit UCCS codes preserve the hierarchical data taxonomy such that the last 3-digits are used to provide additional context to the offense. Finally, UCCS codes fulfill Principles 3 and 4 through using an attribute-based approached, distinct from statute numbers or leveraging other local features that might limit cross-jurisdiction comparisons, through being generated from text descriptions of offense types.

3 Offense description data

In this paper, we estimate a supervised machine learning model to classify text-based offense descriptions to our new UCCS schema. In order to generate this model, we pool two novel sources of hand-coded offense description information and caseload count data from Measures for Justice (MFJ)[5] and the

[5]Measures for Justice Institute is a non-partisan non-profit with a mission to make reliable criminal justice system data available at the county level to spur dialogue and reform.

Criminal Justice Administrative Records System (CJARS)[6]. The pooled set of data draws on multiple decades of electronic criminal justice records from across the United States.

Overall, we employ 313,209 hand-coded unique offense descriptions to create the TOC tool. Each individual description was categorized by a human reviewer, who has been trained on the charge coding schema, the ordering logic, and the nature of inchoate classification.[7] Additional oversight of the resulting classification for every description occurs through validation within the coding tool, and through audit by a senior staff member. Finally, once individual descriptions are added to the overall coded repository, a final round of auditing occurs to ensure classification consistency in the context in which they were originally provided, and across disparate data sources.

In addition, we utilize caseload counts per offense description to weigh observations according to their relative prevalence in the estimation procedure. Due to the free-entry nature of many of the text fields in the data that was collected, rare typos and obscure abbreviations represent a non-trivial share of the unique offense descriptions but a negligible number of cases overall. Together, the 313,209 unique offense descriptions represent 439,534,275 total criminal justice events that have occurred in the United States over recent decades. Figure 1 plots a histogram of the distribution of total cases per unique description, showing that a fair number of unique descriptions happen quite infrequently in the data.[8]

While the production version of the TOC tool[9] leverages the full set of data described above, to evaluate model performance and identify optimal parameterization we subset the data into mutually exclusive training and test datasets to avoid overfitting biases. To generate the training and test data, we employ a 75%–25% mutually exclusive split of the unweighted unique descriptors. Allocation to the training or testing data is randomly assigned. To ensure coverage of each UCCS value in both the training and test data, the random assignment process was stratified at the UCCS level.[10] In cases where there

[6]The Criminal Justice Administrative Records System (CJARS) is a partnership between the University of Michigan and the U.S. Census Bureau, creating an integrated data repository to track involvement in the U.S. justice system that is linkable with socio-economic data.

[7]This training is overseen by senior staff. Additional training occurs as needed when consistent errors are identified in classification audits.

[8]In practice, we set a maximum case level count to 100 to balance the focus of the estimated model between regularly occurring descriptions without typos (which are a large share of the data and are easier to classify without machine learning) and rare occurring descriptions with typos (which are a smaller share of the data and harder to classify without machine learning).

[9]This free to use tool can be found online at: `https://cjars.isr.umich.edu/toc-tool/`.

[10]Kang, Ryu, and Kwon (2004) provides evidence of improved performance from stratified sampling over non-stratified

were fewer than 4 descriptors coded to a given UCCS code, the 75%–25% ratio was suspended to ensure that each UCCS code appeared in both the training and testing data.[11]

4 The TOC tool

4.1 Parameterization of the TOC tool

As a text classification tool, TOC consists of 5 main components: (1) preprocessing, (2) tokenization, (3) feature selection, (4) classifier, and (5) classification framework.

In the *preprocessing* stage, raw text descriptions are cleaned to reduce noise from sources such as articles ("a/an", "the"), punctuation, capitalization, and grammatical tense (word normalization). This is a crucial step in text classification since the reduction in the overall size of input data can improve classification performance. For instance, Uysal and Gunal (2014) evaluate various combinations of preprocessing techniques using e-mail and news articles from English and Turkish sources, and found that lowercase conversion significantly improved classification performance. However, the authors also noted that the optimal combination of preprocessing techniques for improving performance is largely dependent on both the domain and language of the data. Similarly, Toman, Tesar, and Jezek (2006) analyze the effects of word normalization methods and stop-word removal on English and Czech data and found that only stop-word removal yielded significant improvement in performance while word normalization only resulted in slight improvement. For our production model, we applied lowercase conversion, stop-word removal, word normalization using Porter stemming algorithm, and a custom filter for keeping only alphanumeric characters and relational operators (">", "<", "=").

Tokenization then segments text into individual tokens, or features, so that the information can be represented as numeric variables. In general, there are two ways to generate tokens for text classification. On one hand, word-based tokenization generates tokens delimited by leading and trailing spaces. On the other hand, character-based tokenization uses contiguous sequence of N characters, or N-grams, to

sampling procedures.

[11] Random assignment between training and testing was still enforced in these cases.

generate individual features. In the TOC tool, data is tokenized using 4-grams as character-based approach tends to outperform the former when abbreviations and typographical errors are prevalent in the corpus (Stamatatos 2013; Koppel, Schler, and Argamon 2009).

In *feature selection*, each token or feature is scored using a selection metric. Then, a subset of the best N features based on their score is kept as inputs for the machine learning algorithm. The TOC tool uses Term Frequency-Inverse Document Frequency (TF-IDF) which scores each term through an inverse proportion of its frequency in a description to the percentage of descriptions the term appears in. For a given feature f in description d from document D, $TF-IDF(f,d,D)$ is then calculated as the product of the term frequency, $TF(f,d)$, and the inverse document frequency, $IDF(f,D)$. These terms are defined below,

$$\begin{aligned} TF(f,d) &= \frac{\text{Frequency of feature f in description d}}{\text{Total number of features in d}} \\ IDF(f,D) &= log\left(\frac{N}{DF(f)}\right) \\ TF-IDF(f,d,D) &= TF(f,d)*IDF(f,D) \\ &= TF(f,d)*log\left(\frac{N}{DF(f)}\right) \end{aligned}$$

where, $DF(f)$ is the number of descriptions in the document that contains the term f, and N is the total number of descriptions in the data.

The selected features are then used as inputs for *classification*. We use a hierarchical multi-layer perceptron (MLP) model to generate the mapping from selected features to predicted full offense code. MLP is a type of artificial neural network that consists of interconnected network of nodes, or neurons, in three different layers: an input layer, one or more hidden layers, and an output layer. In MLP, each node is associated with a weight which is adjusted during the training phase to minimize classification error, or the difference between the predicted class and true class. In our application, we utilize one hidden layer.

To leverage the taxonomic characteristic of UCCS schema, we induced class hierarchy to the TOC tool using local classifier per parent node method where one or more classifiers are trained at each level in the hierarchy. As a result, the TOC tool starts by training a MLP classifier at the parent level for classifying Broad Crime Type (1st digit of UCCS). Then for each Broad Crime Type, a new MLP classifier is trained to predict

the Offense Category (2nd and 3rd digits of UCCS) at the sub-parent level. As the last step, a single MLP classifier is trained across all records to predict the Offense Modifier (4th digit). In total, TOC tool consists of a group of 100 MLP classifiers (1 for Broad Crime Type, 6 for Offense Category, and 93 for Offense Modifier) that together provide a predicted UCCS classification. To summarize, Figure 2 provides an overview of the TOC production model's workflow, and Table 3 shows the performance results by Broad Crime Types.

4.2 Performance of the TOC tool

The overall out-of-sample performance of the TOC tool is presented in Table 3. We evaluate the performance of the model at predicting both the broad crime type as well as the full UCCS code using standard metrics in the literature: precision ($\frac{\text{True Positive}}{\text{True Positive + False Positive}}$), recall ($\frac{\text{True Positive}}{\text{True Positive + False Negative}}$) and F1 scores ($2 * \frac{\text{Precision * Recall}}{\text{Precision + Recall}}$). Overall, we observe high levels of performance, with all three metrics delivering performance statistics at 0.983 for the broad crime type level, and 0.963 for the full UCCS code level. The vast majority of offenses in the data are being accurately mapped to their true categorizations. While there is some drop-off from the broad to full code level, it is quite modest. It still remains though that utilizing TOC to make higher level offense type predictions (an application that will be sufficient for many researchers) will be more reliable than full UCCS code predictions.

Table 3 also shows performance statistics within broad crime type codes to assess whether overall performance is masking subtype heterogeneity. At the broad crime type level, we measure precision scores in the range of 0.927 – 0.999 and recall scores in the range of 0.938 – 0.994. The TOC tool does appear to be generating high quality predictions across a range of offense types. The Public Order broad crime type shows the lowest relative performance on recall among the set, which is not surprising; in contrast to the other broad crime types, Public Order encompasses a more diverse array of behavior (e.g. prostitution, bribery, or weapons offenses) with less commonly used words and phrases within the category.The Property broad crime type shows the lowest relative performance on precision, which likely reflects the fact that these types of offenses often include words like *possession* which do show up occasionally in other types of offenses.

At the full UCCS code level, the TOC tool yields F1 Scores in the range of 0.845 – 0.991. The largest declines in performance from broad to full code is observed within the Drug broad crime type. This is

perhaps unsurprising given the prevalence of similarly described but distinct offense types within this category (e.g. possession/use of heroin, distribution of heroin, and distribution of prescription drugs) which the algorithm has a difficult time differentiating between. Overall, however, performance statistics remain quite high, suggesting promising opportunities from the widespread adoption of the TOC tool.

An important concern that remains is how well the TOC tool will perform when applied to new states not contained in the training data. The TOC tool is built off of administrative data from 24 states in the U.S., and we observe that some offense descriptions are unique to specific states. This could be due to state-specific abbreviations or data entry practices, or the inclusion of state-specific statute numbers when describing the offense in the free entry text field. While we do not hold data from 26 remaining states, we can explore performance stability when excluding individual covered states from estimating the TOC tool, making predictions for those excluded states, and evaluate the resulting out-of-scope performance. The results of this exercise are presented in Table 4.

We find that there is only modest performance degradation when applying the TOC tool to new states not included in the training data. Comparing results from Tables 3 and 4, we observe a decline of precision, recall, and f1 scores consistently in the range of 1.5 percentage points (0.983 → 0.968) at the broad crime type level and 2.8 percentage points (0.963 → 0.935) at the full UCCS code level. While performance remains high in the out-of-state predictions, the relatively larger drop-off at the full code level is consistent with our motivation in creating the TOC tool in the first place, that local differences in offense descriptions are pervasive and require an algorithmic approach to classify offenses at scale. As more training data becomes available, it will be important to update the TOC tool to improve performance in jurisdictions currently uncovered.

5 Determining optimal model parameterization

In order to build the TOC tool, we conducted experiments on a number of permutations of our modeling choices, including: (1) training data size, (2) feature unit and selection method, (3) number of selected features, and (4) machine learning algorithm type and classification method. The goal in these exercises

is to both identify the parameterizations that yield the strongest out-of-sample performance measures, and to better understand which choices are more or less consequential. For each perturbation of the model, we ran 20 bootstrapped iterations, sampling the fixed 75% training data sample with replacement in each iteration. All other features of the model described in Section 4 remain unchanged.

Sample Size. We first explore the role of the size of the training data. We have the luxury of having hundreds of thousands of training observations available to build the TOC tool, yet if other researchers were interested in building their own classification tool for other topics (e.g. civil case filing types) or jurisdictions (non-U.S. criminal offenses), this would be important information for assessing how much should be invested in developing an original training dataset.

Figure 6 shows the results of this exercise. In the first column, F1 scores are shown for Broad Crime Type predictions, and the second column shows corresponding estimates for the full UCCS code prediction. In the solid blue line, we plot the average out-of-sample performance, with 95% bootstrapped confidence intervals in the dotted lines.

Performance monotonically improves with additional training sample observations, both in terms of better average performance and more consistent performance. The largest gains in average performance accrue from 1,000 through 50,000 training observations, yet some improvements persist beyond that point, especially with regard to decreasing performance variability across the bootstrapped iterations. Predicting the full UCCS code also benefits from substantial numbers of training observations, which is not surprising given the more challenging goal of predicting offense type at such a fine-grained level of detail compared to just the broad crime type.

Tokenization and feature selection. We examine several aspects of how tokenization and feature selection influence the overall performance of the TOC tool. In the first exercise, we eliminate the prepocessing stage of the TOC tool, leaving all other aspects of the model untouched. Table 5 shows the resulting performance statistics overall and by broad crime type. Including the preprocessing stage led to only marginal improvements in performance at the broad crime type level, although Public Order and Drug offenses show larger than average gains. Overall, F1 scores increase from 0.960 to 0.983 with the inclusion of the prepocessing;

Public Order and Drug-specific F1 scores improve by 0.066 and 0.042 respectively. . Predictions of full UCCS code do appear to benefit slightly more from preprocessing; F1 scores at this more detailed level increase from 0.929 to 0.963. And, here, we see substantial improvement in performance for Public Order and DUI offenses, with significant, but more modest, improvements for Violent and Drug offenses.

The second exercise in this theme explores variations in tokenization and feature extraction through varying the size of the N-grams (1 to 6 characters), introducing a bag-of-words option, as well as allowing feature extraction to be determined by either the TF-IDF or CountVectorizer (CV) algorithms.[12] Performance peaks at 4grams for both feature selection approaches (Figure 7), indicating value being generated from increasing specificity of the tokenization process that is constrained by a fixed number of features. The TF-IDF feature selection method modestly outperforms the CV feature selection method for almost all types of features, but the performance gain is usually quite small, likely reflecting the fact that offense descriptions are brief and do not contain repetitive extraneous terms, especially once preprocessed.

The final exercise holds tokenization (4gram) and feature extraction method (TF-IDF) constant, but varies the number of features extracted to use as inputs to estimate the MLP models (see Figure 8). We examine 100, 500, 1,000, 5,000, and 10,000 features. While more features can improve model performance, there is a tradeoff with computing efficiency, as processing time grows non-linearly with decreasing returns to scale in performance statistics, as well as increasing risk of overfitting the model.

Model performance improves significantly as features increase from 100 to 1,000 without a meaningful different in computing time. Additional gains accrue with 5,000 features, especially at the full UCCS code level of prediction, but with a corresponding fourfold increase in time to train the model. At 10,000 features, model performance declines and processing time is substantially longer, indicating that the

[12]CV selects features by using the most frequently occurring terms in the entire document. Given a list of offense descriptions, D, CV generates x by y sparse matrix, F, where x is the number of offense descriptions in the data, y is the total number of unique features found in D, and f_{xy} is the total number of times yth feature appear in xth description. By summing the columns, CV is then able to generate total frequency for yth feature in D.

$$F = \begin{bmatrix} f_{11} & f_{12} & \dots & f_{1y} \\ f_{21} & f_{22} & \dots & f_{2y} \\ \dots & \dots & \dots & \\ f_{x1} & f_{x2} & \dots & f_{xy} \end{bmatrix}$$

feature space has potentially been over-saturated.

Classification. Finally, we compare the role of classification along two dimensions. First, we evaluate relative performance of MLP and random forest classifiers,[13] and second, hierarchical versus flat modeling approaching.[14,15] Figure 9 documents how MLP models and hierarchical classification methods systematically outperform random forest models and flat classification methods. At the broad crime type level, performance across all considered approaches is relatively similar, yet stronger differences emerge at the full UCCS code level.

6 Discussion

Both the UCCS schema and the TOC tool are intended to evolve over time. New crime types (or differentiating important differences within existing pooled groups of offenses) will be incorporated at regular intervals into UCCS to ensure that the schema remains current and valuable. Receiving feedback on the schema will be critical for ensuring the categorization matures with the criminal justice system.

As the UCCS schema evolves, the TOC tool will necessarily need to be updated. In addition to adding new offense types, there are several additional features which may improve the TOC tool's performance and utility, which have not yet been incorporated.

First, the out-of-state exercises suggest that there can be fundamental differences between states in how illicit behavior is described. This raises the question of whether the TOC tool should incorporate geographic information on the location of the offense into the prediction model, or alternatively build state-specific tools that focus exclusively on predictions generated from within-jurisdiction training data. Given that not all jurisdictions in the U.S. are yet incorporated into the corpus of training data for the TOC tool, pursuing these options come with the trade-off of potentially decreasing the utility of the TOC

[13] In random forest models, an ensemble of individual decision trees is used to generate predictions for each tree. Then, a vote is performed across the predicted results and the model selects the final prediction value using majority vote rule.

[14] In flat classification, a single classifier is used to assign a class (Figure 4a). As a result, the flat classification method uses a single set of input features to directly predict the 4-digit UCCS codes.

[15] In the context of hierarchical classification, there are two additional methods: local classifier per level (Figure 4c) and local classifier per node (Figure 4d). However, these methods were ultimately excluded from our experiments since they are susceptible to hierarchical inconsistency (Figure 5).

tool when applying to non-covered jurisdictions or national level data.

The second feature would be to leverage the implicit information contained within cited statute numbers in offense descriptions. Statute numbers are entered into a number of observed offense descriptions as short-hand for more lengthy information on classes of criminal activity defined in statute, which supplement free entry offense description fields. The challenge in leveraging this information is twofold. First, statute numbers are cited irregularly in inconsistent formats, requiring the need to develop a technique to identify and interpret a statute number when it appears in an offense description. Second, there does not currently exist a comprehensive database that maps statute numbers to their offense descriptions, and so additional effort would be required to translate the statute numbers into a structure that the TOC tool could interpret.

7 Conclusion

In this paper, we introduce a new schema (UCCS) to organize and categorize criminal activity in the U.S., provide a robust machine learning algorithm (TOC tool) to implement the new schema in practice based on data fields that are systematically collected in most jurisdictions, and evaluate the relative importance of numerous features of this model. None of this would have been possible without the joint collaboration of two organizations pioneering major advances in data infrastructure and statistical reporting on the U.S. criminal justice system: Measures for Justice and the Criminal Justice Administrative Records System.

The UCCS schema and TOC tool lower barriers to working with cutting edge data from the U.S. criminal justice system. These initiatives promote inclusive research dialogues on pressing social policy issues through providing researchers without a background in data science with automated classification systems at their disposal. We also hope that these contributions will help encourage consistent and reproducible research in the field through encouraging researchers to utilize common definitions of offense types and minimizing the need for researcher discretion in wrangling administrative records for research purposes, a process that can be opaque and have minimal oversight.

8 Acknowledgements

We would like to thank Shawn Bushway, Keith Finlay, Magaret Levenstein, James Lynch, Jeffrey Morenoff, Amy O'Hara, Jordan Papp, Anne Piehl, JJ Prescott, Steven Raphael, William Sabol, and seminar participants at the Michigan Institute for Data Science, the University of Michigan Institute for Social Research, the Federal Committee on Statistical Methodology, the American Society of Criminology annual conference, and the 2021 SEARCH Symposium on Justice Information Technology, Policy and Research for their helpful comments. We would also like to acknowledge the significant work of the MFJ Research team members who made the training dataset possible: Alexandra Ackles, Mauricio Alvarez, Gipsy Escobar, Robert Hutchison, Hillary Livingston, Nathaniel LeMahieu, Trevariana Mason, Dominic Testino, Ian Thomas, and Raelynn Walker. Finally, we thank Alex Albright, CJARS, Measures for Justice, NYU Justice Lab, Recidiviz, and Vera Institute of Justice for participating in beta testing for the TOC tool.

9 Funding

This work was generously supported in part by the The John D. And Catherine T. MacArthur Foundation's Safety Justice Challenge, Arnold Ventures, and the National Science Foundation.

References

Baumer, E. P., M. B. Velez, and R. Rosenfeld. 2018. Bringing crime trends back into criminology: A critical assessment of the literature and a blueprint for future inquiry. *Annual Review of Criminology* 1:39–61.

Bureau of Justice Statistics. 2009. 2009 Conversion of Criminal History Records into Research Databases. https://bjs.ojp.gov/sites/g/files/xyckuh236/files/media/document/cchrrd09sol.pdf.

———. 2015. 2015 Criminal History Record Assessment and Research Program (CHRARP). https://www.bjs.gov/content/pub/pdf/chrarp15_sol.pdf.

———. 2020a. 2020 Criminal History Record Assessment and Research Program (CHRARP). https://www.bjs.gov/content/pub/pdf/chrarp2020_sol.pdf.

———. 2020b. National Corrections Reporting Program, [United States], 2000-2017. https://www.icpsr.umich.edu/web/NACJD/studies/37608.

Federal Bureau of Investigation. 2013. Summary Reporting System (SRS) User Manual. https://www.fbi.gov/file-repository/ucr/ucr-srs-user-manual-v1.pdf/view.

———. 2017. Uniform Crime Reporting Statistics: Their Proper Use. https://ucr.fbi.gov/ucr-statistics-their-proper-use.

———. 2019. NIBRS Offense Definitions. https://ucr.fbi.gov/nibrs/2019/resource-pages/nibrs_offense_definitions-2019.pdf.

Finlay, Keith, and Michael Mueller-Smith. 2020. Criminal Justice Administrative Records System (CJARS) [dataset]. Ann Arbor, MI: University of Michigan. https://cjars.isr.umich.edu.

Finlay, Keith, Michael Mueller-Smith, and Jordan Papp. 2021. *The Criminal Justice Administrative Records System: A Next-Generation Data Platform for Studying the U.S. Criminal Justice System.* Working paper.

Kang, Jaeho, Kwang Ryel Ryu, and Hyuk-Chul Kwon. 2004. Using Cluster-Based Sampling to Select Initial Training Set for Active Learning in Text Classification. In *Advances in Knowledge Discovery and Data Mining,* edited by Honghua Dai, Ramakrishnan Srikant, and Chengqi Zhang, 384–88. Berlin, Heidelberg: Springer Berlin Heidelberg.

Koppel, Moshe, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology* 60 (1): 9–26. https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20961.

Langton, L., M. Planty, and J. P. Lynch. 2017. Second major redesign of the National Crime Victimization Survey (NCVS). *Criminology & Public Policy* 16 (4): 1049–74.

Maxfield, M.G. 1999. The National Incident-Based Reporting System: Research and Policy Applications. *Journal of Quantitative Criminology* 15:119–49.

National Academies of Sciences, Engineering, and Medicine. 2016. *Modernizing Crime Statistics: Report 1: Defining and Classifying Crime*. Edited by Janet L. Lauritsen and Daniel L. Cork. Washington, DC: The National Academies Press. https://www.nap.edu/catalog/23492/modernizing-crime-statistics-report-1-defining-and-classifying-crime.

Poggio, Eugene, Stephen Kennedy, Jan Chaiken, and Kenneth Carlson. 1985. Blueprint for the future of the Uniform Crime Reporting Program: Final report of the UCR Study (Contract No. J-LEAA-011-82). (Washington, DC). https://www.ojp.gov/pdffiles1/bjs/98348.pdf.

Stamatatos, Efstathios. 2013. On the Robustness of Authorship Attribution Based on Character N-gram Features. *Journal of Law and Policy* 21 (2): 421–39. https://brooklynworks.brooklaw.edu/cgi/viewcontent.cgi?article=1048&context=jlp.

Strom, Kevin, and Erica Smith. 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 (4). https://www.ojp.gov/ncjrs/virtual-library/abstracts/future-crime-data-case-national-incident-based-reporting-system.

Toman, Michal, Roman Tesar, and Karel Jezek. 2006. Influence of Word Normalization on Text Classification. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.83.6363.

US Department of Justice. 2021. 2021.1 National Incident-Based Reporting System User Manual. https://www.fbi.gov/file-repository/ucr/ucr-2019-1-nibrs-user-manua-093020.pdf.

US Department of Justice, Federal Bureau of Investigation. 2011. Hate Crime Statistics, 2011. https://ucr.fbi.gov/hate-crime/2011/resources/variablesaffectingcrime_final.pdf.

Uysal, Alper Kursat, and Serkan Gunal. 2014. The impact of preprocessing on text classification. *Information Processing Management* 50 (1): 104–12. https://www.sciencedirect.com/science/article/pii/S0306457313000964.

Wormeli, P. 2018. Criminal justice statistics - An evolution. *Criminology & Public Policy* 17 (2): 483–96.

Tables

Table 1: Examples of data at "descriptor" level and "case" level

(a) Descriptor level data

State	Statute	Description	UCCS Code	Frequency
IN	35 43 4 2 A	FD THEFT	2070	32,503
IN	35 43 4 2 A	MA THEFT	2070	36,809

(b) Case level data

State	Statute	Description	UCCS Code
IN	35 43 4 2 A	FD THEFT	2070
IN	35 43 4 2 A	FD THEFT	2070
..	..	..	..
IN	35 43 4 2 A	MA THEFT	2070
IN	35 43 4 2 A	MA THEFT	2070
..	..	..	..

Notes: This table shows the difference between data represented at the descriptor level (a), and at the case level (b). At the descriptor level, each row represents a unique offense description and contains a variable for its frequency in our data while the latter is a duplicated version of the former. There are two reasons why TOC makes this distinction. First, it is possible to weight individual features by multiplying the weight, or count, of each description with the sparse matrix generated using Count Vectorizer. However, the same method is not applicable for TF-IDF since it is dependent on the corpus size, or the number of descriptions in the data. In other words, TF-IDF will set N, the total number of descriptions in the data, to 2 at the descriptor level while at the case level, N is set to 69,312 for better document weightage of a term. Second, Random Forest classifiers allow users to pass in class weights but Multi-layer Perceptron classifiers do not have support for different class weights which are important when working with imbalanced classes. As a result, training Multi-layer Perceptron at the case level helps with building a more balanced dataset by oversampling minority classes.

Table 2: Examples of input/output of preprocessing stage

Original Description	**Preprocessed Description**
893136A-DRUG3101 (FT) POSSESSION OF HERO	89313fdrug3100 ft possess hero
SAO1000 (MS) DRIVING UNDER THE INF	sao1000 ms drive under inf
Theft from Persons >65 VALUE >300<10k	theft from person >65 value >300<10k

Notes: This table shows how offense descriptions are transformed through preprocessing. The TOC tool applies lowercase conversion, stop-word removal, word normalization (e.g. “possession” to “possess”, “driving” to “drive”), and relational operator filter.

Table 3: Performance of the TOC tool, by broad crime type

	Broad Crime Type			Full UCCS Code		
	Precision	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
All Crime Types	0.983	0.983	0.983	0.963	0.963	0.963
Broad Crime Type Code:						
Violent	0.997	0.994	0.995	0.993	0.989	0.991
Property	0.927	0.990	0.957	0.884	0.944	0.913
Drug	0.999	0.960	0.979	0.862	0.828	0.845
DUI	0.987	0.986	0.986	0.942	0.941	0.941
Public Order	0.993	0.938	0.965	0.977	0.923	0.949
Criminal Traffic	0.987	0.991	0.989	0.986	0.991	0.988

Notes: This table shows the classification performance of the production TOC model at the parent class (Broad Crime Type) and at the child class (UCCS Code) weighted by the case count of each offense description. The model uses hierarchical classification method with Multi-layer Perceptron classifier trained at each parent node using 5,000 4-grams selected by TF-IDF from preprocessed descriptions.

Table 4: Performance of the TOC tool on out-of-state predictions

	Broad Crime Type			Full UCCS Code		
State	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
All Crime Types	0.968	0.968	0.968	0.938	0.932	0.935
State:						
Alabama	0.919	0.817	0.865	0.895	0.712	0.793
Arkansas	0.980	0.979	0.979	0.896	0.832	0.863
Arizona	0.978	0.978	0.978	0.975	0.963	0.969
California	0.985	0.984	0.984	0.985	0.984	0.984
Colorado	1.000	1.000	1.000	0.997	0.997	0.997
Connecticut	0.897	0.822	0.858	0.853	0.753	0.800
Florida	0.978	0.978	0.978	0.956	0.939	0.947
Illinois	0.995	0.995	0.995	0.995	0.995	0.995
Indiana	0.955	0.948	0.951	0.937	0.918	0.927
Kansas	1.000	1.000	1.000	0.994	0.997	0.995
Maryland	0.858	0.754	0.803	0.798	0.535	0.641
Michigan	0.999	0.999	0.999	0.929	0.913	0.921
Minnesota	0.999	0.999	0.999	0.999	0.999	0.999
Mississippi	1.000	1.000	1.000	0.994	0.997	0.995
North Carolina	0.972	0.972	0.972	0.945	0.934	0.939
North Dakota	0.967	0.966	0.966	0.952	0.943	0.947
Nebraska	0.999	0.999	0.999	0.999	0.998	0.998
New Jersey	1.000	1.000	1.000	0.998	0.682	0.810
Ohio	1.000	1.000	1.000	1.000	1.000	1.000
Oregon	0.986	0.986	0.986	0.962	0.970	0.966
Pennsylvania	0.964	0.963	0.963	0.891	0.891	0.891
Texas	1.000	1.000	1.000	0.996	0.977	0.986
Utah	0.947	0.927	0.937	0.904	0.890	0.897
Wisconsin	0.875	0.876	0.875	0.861	0.788	0.823

Notes: Summary statistics of out-of-state experiment weighted by case count. Each subset of the data by state contains unique offense descriptions that may not be mutually exclusive (e.g. "cruelty to animals" is in 21 of 24 states in the data). The state-specific data is treated as out-of-sample testing data while the remaining descriptions from other states are used for training the model.

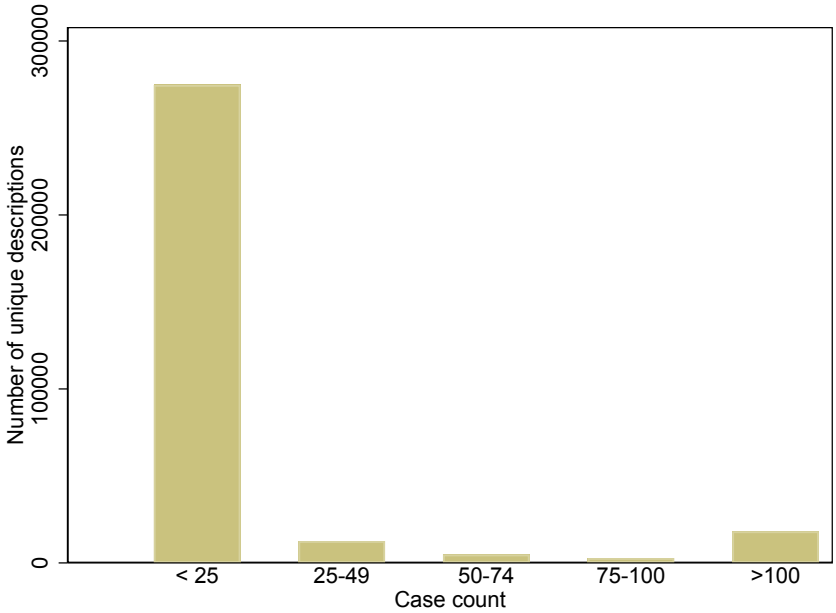
Table 5: Performance of TOC tool without the preprocessing stage

	Broad Crime Type			Full UCCS Code		
	Precision	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
All Crime Types	0.961	0.960	0.960	0.930	0.929	0.929
Broad Crime Type Code:						
Violent	0.980	0.951	0.965	0.945	0.917	0.931
Property	0.923	0.959	0.941	0.902	0.936	0.919
Drug	0.922	0.953	0.937	0.782	0.809	0.795
DUI	0.980	0.953	0.966	0.827	0.744	0.783
Public Order	0.882	0.886	0.899	0.827	0.839	0.833
Criminal Traffic	0.990	0.978	0.984	0.990	0.978	0.984

Notes: This table shows the classification performance without preprocessing at the parent class (Broad Crime Type) and at the child class (UCCS Code), weighted by case count. The remaining parameters are the same as that of the production model (hierarchical method with Multi-layer Perceptron using 5,000 4-grams selected by TF-IDF on raw descriptions).

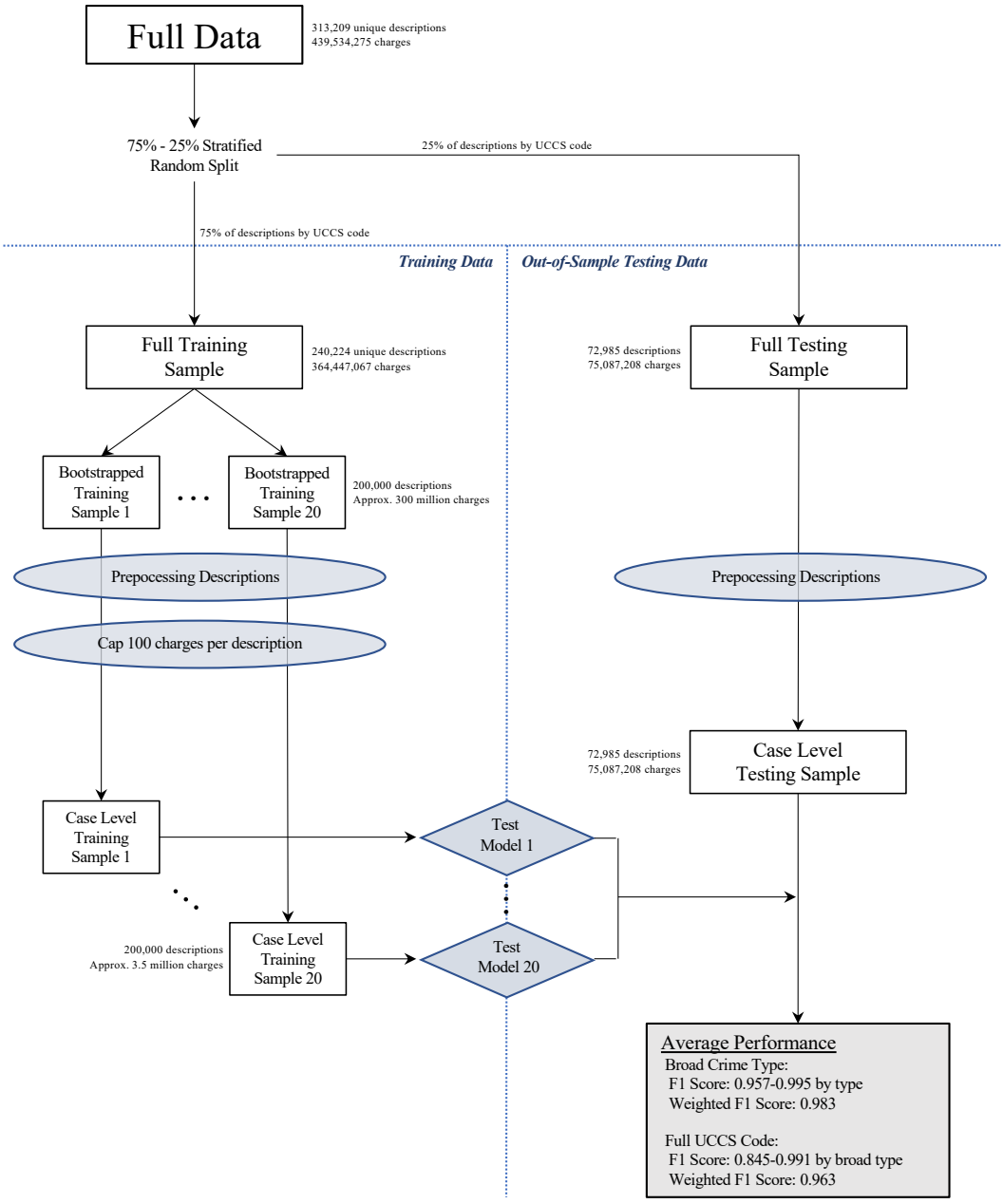
Figures

Figure 1: Distribution of total cases per unique description



Notes: This figure shows the distribution of unique descriptions by their case counts in the full data. 274,744 descriptions out of the 313,209 unique descriptions (87.7%) occur less than 25 times, and 18,202 descriptions occur more than 100 times in the data with a mean case count of 23,950.

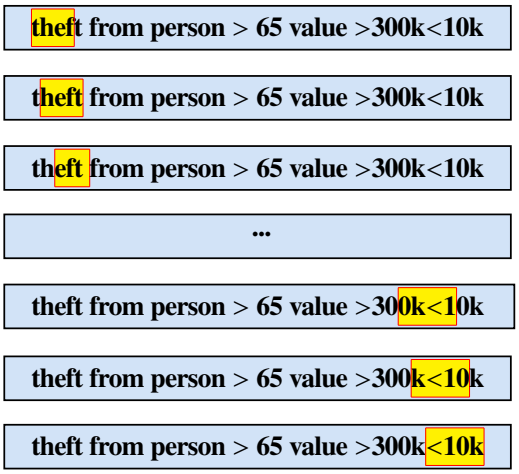
Figure 2: Sequence for Optimal Model Parameterization



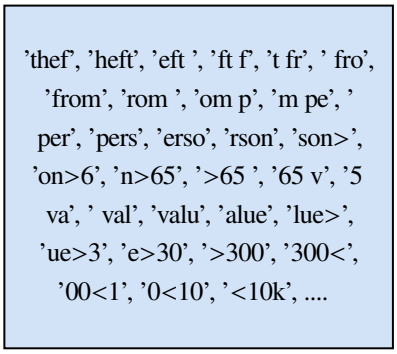
Notes: This figure shows the workflow for identifying the optimal model parameters for the TOC tool. The full data is stratified at the descriptor level using 75%-25% ratio by UCCS code to ensure mutually exclusive split of offense descriptions and coverage of each UCCS values. From 240,224 unique descriptions in the full training data, 200,000 descriptions are sampled with replacement for each iteration while 72,985 descriptions stay constant. Both bootstrapped training and testing data are preprocessed at the descriptor level to reduce the overall program run time. In the training phase, the maximum case count is set to 100 to generate training set with average of 3,500,000 charges while in the testing phase, the true case count is used. The bottom right box shows the range of F1 scores by Broad Crime Type and the weighted average using optimal model parameters at the Broad Crime Type and at the UCCS code.

Figure 3: Feature extraction

(a) Contiguous sequence of 4 characters

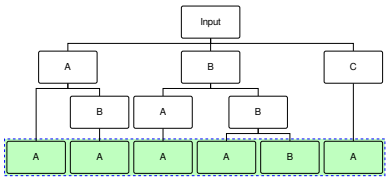


(b) Extracted features

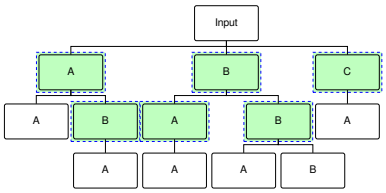


Notes: This figure shows a visual representation of tokenization using 4-grams for the offense description "theft from person > 65 value >200k<10k." Using this method, contiguous sequence of 4 characters are selected from the beginning to the end of the text as highlighted by yellow box in (Figure 3a). Each of the extracted tokens, or features, in (Figure 3b) are then scored using a vectorizer. For Count Vectorizer, each feature is assigned an integer to represent the number of times it appears in the entire data. In the case of TF-IDF, each feature is assigned a score from 0 to 1. The values derived from CV or TF-IDF are then used for selecting which features to keep for the model. CV will select the most frequently occurring features whereas TF-IDF will select features with scores closest to 1.

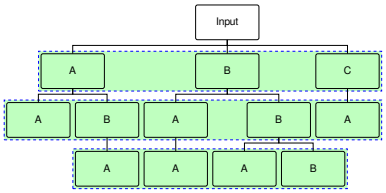
Figure 4: Hierarchical Classification Methods



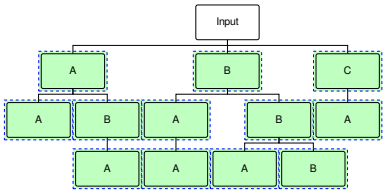
(a) Flat Classifier: One multiclass classifier for child nodes



(b) Local Classifier Per Parent Node: One multiclass classifier for each parent node



(c) Local Classifier Per Level: One multiclass classifier for each level

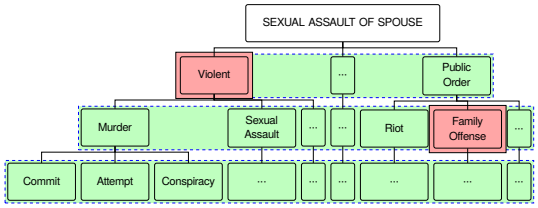


(d) Local Classifier Per Node: One binary classifier for each node

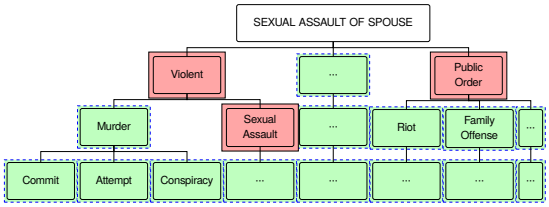
Notes: This figures shows the different types of hierarchical classification methods. The green boxes are used to denote both the predicted classes of a model, or a classifier, as well as the number of models required in total. In a flat classifier (Figure 4a), the text data is used to predict the child class (UCCS code) without any intermediate steps. Although this approach requires training only one multi-class classifier, or a machine learning algorithm designed to classify an input into one of three or more categories, it fails to capture information on class hierarchy. This issue is addressed by employing the local classifier per parent node method (Figure 4b) where a flat classifier is trained for each parent class in order to distinguish its child classes. In effect, the total number of classifiers involved in local classifier per parent node method is equal to the number of parent classes plus 1 for the initial classification. In local classifier per level method (Figure 4c), a flat classifier is trained at each level of the hierarchy. Although this method only requires 3 flat classifiers (one for each level), one of the drawbacks is that the predicted results can ignore the hierarchical taxonomy of the data due to level independence (Figure 5a). Finally, the local classifier per node method (Figure 4d) trains a binary classifier for each node. Similar to the previous approach, the local classifier per node method can ignore class hierarchy due to node independence (Figure 5b). For these reasons, this paper focuses only on flat classifier and local classifier per parent node methods.

Figure 5: Hierarchical Inconsistencies

(a) Vertical Inconsistency

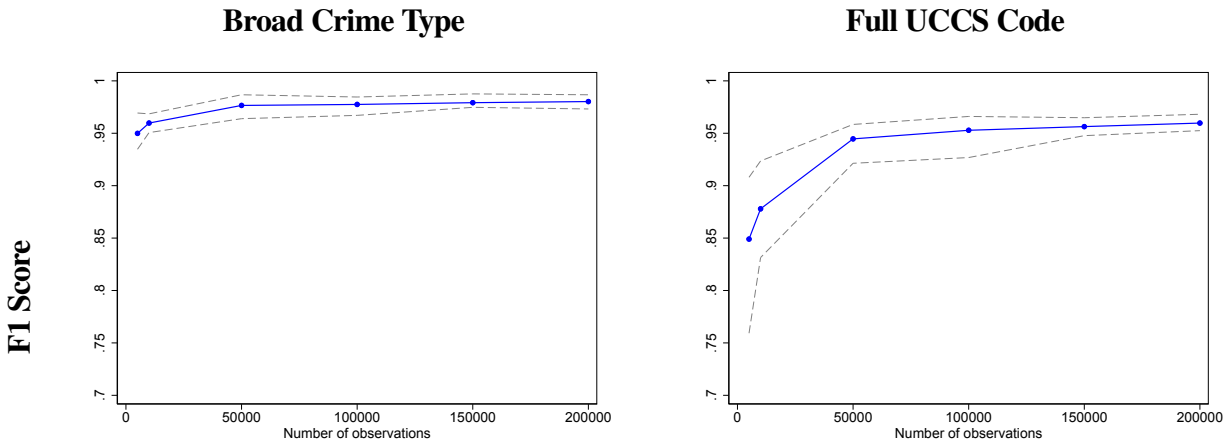


(b) Horizontal Inconsistency



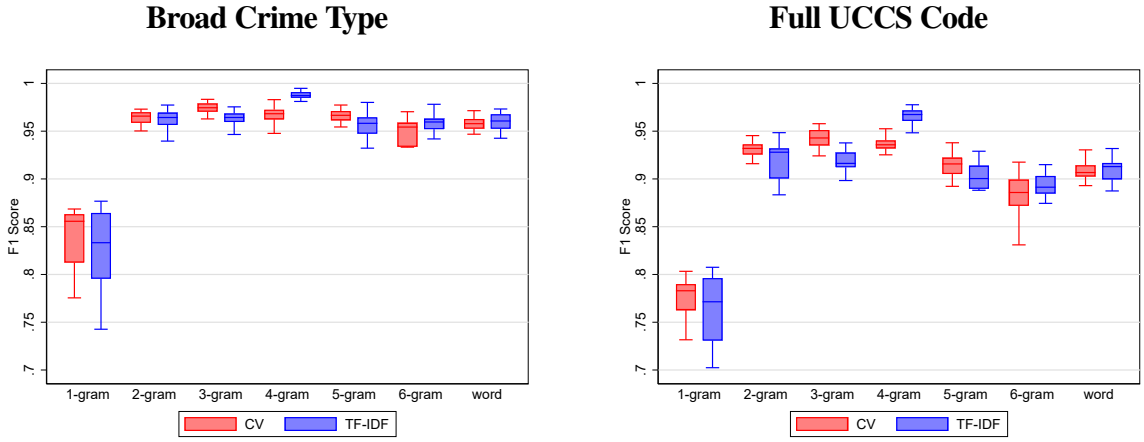
Notes: Consider the offense description "SEXUAL ASSAULT OF SPOUSE" as an example. In the case of (Figure 5a) local classifier per level method, the word "ASSAULT" could be a strong predictor for the model to classify the Broad Crime Type as Violent. However, because each level is independently classified, the word "SPOUSE" could influence the model to classify it as Family Offense which is a category under Public Order. In (Figure 5b) local classifier per node method, each node is predicted using a binary classifier. As a result, "SEXUAL ASSAULT OF SPOUSE" can be classified to multiple Broad Crime Types such as Violent and Public Order. Furthermore, local classifier per node method is also susceptible to vertical inconsistency due to independent classifiers at each node.

Figure 6: Relationship between size of training data and out-of-sample performance statistics



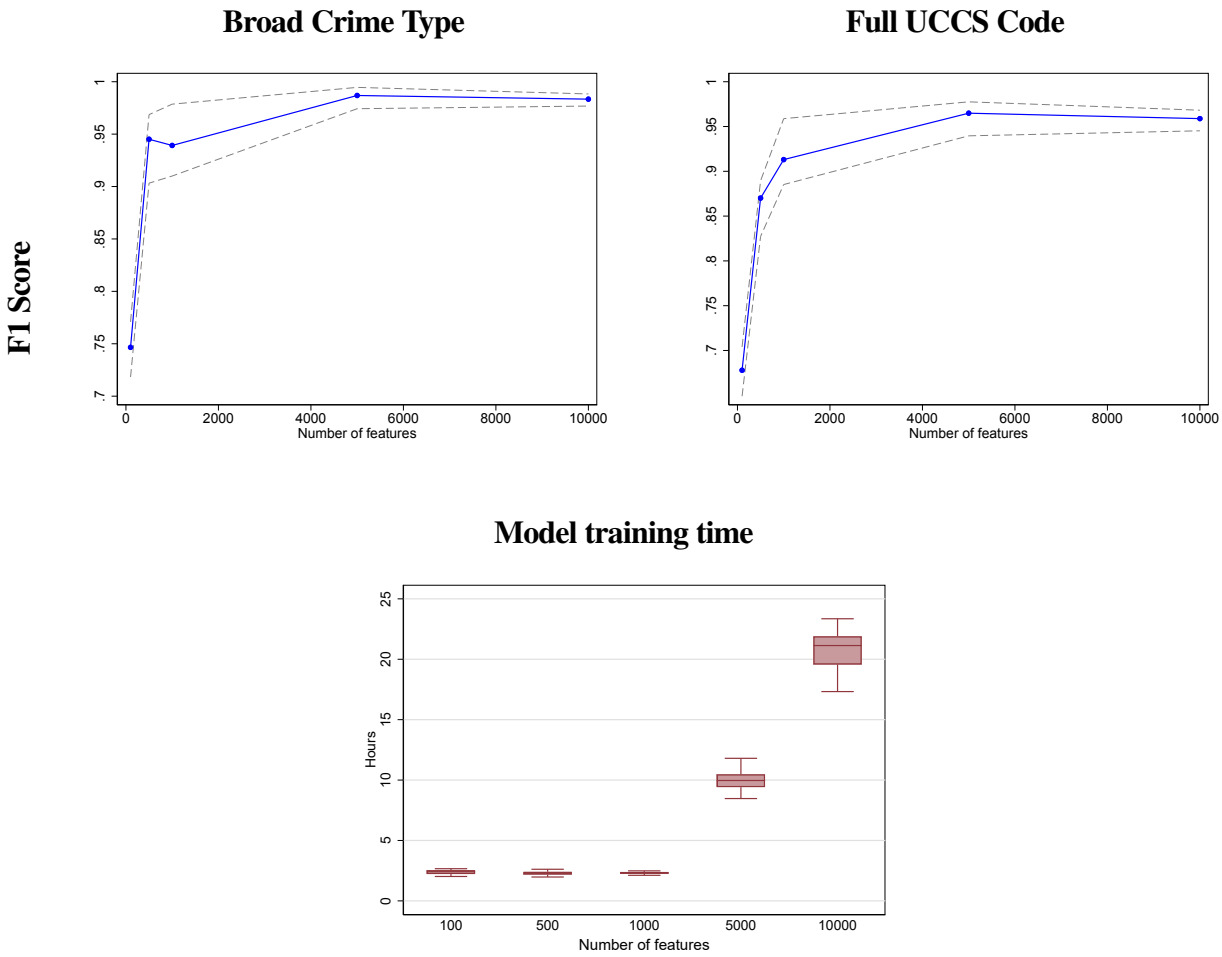
Notes: This figure shows the convergence of out-of-sample model performance as the size of the training sample is increased from 5,000 training observations to 200,000 training observations. 240,224 observations of the total 313,209 unique offense descriptions were selected at random for use in the training sample; the remaining 72,985 descriptions were used as out-of-sample testing data for this exercise. 20 bootstrapped hierarchical multi-layer perceptron models were estimated for each level of training data, with training observations selected at random (with replacement) from 240,224 unique descriptions. The left panel shows the change in average as well as 5th/95th percentile model performance at the Broad Crime Type as the number of training observations grow while the right panel shows the change in average as well as 5th/95th percentile model performance at the UCCS Code at different sample sizes.

Figure 7: Relationship between feature unit and out-of-sample performance statistics



Notes: Box plots show the comparison of out-of-sample model performance between Count Vectorizer (red) and TF-IDF (blue) using different units of features including word-level unigram (delimited by space). Each box plot incorporates the lower and upper adjacent values, 25th and 75th percentiles, and median performance. At the character-level, the number of contiguous characters is increased from 1 character (1-gram) to 6 characters (6-grams). 20 bootstrapped hierarchical multi-layer perceptron models were estimated for each feature unit, with 200,000 training observations selected at random (with replacement) from 240,224 unique descriptions. For each unit, both Count Vectorizer and TF-IDF selected a maximum of 5,000 features. The box plot on the left shows the F1 score of each feature selection method at the Broad Crime Type while the box plot on the right shows the F1 score of each method at the UCCS Code.

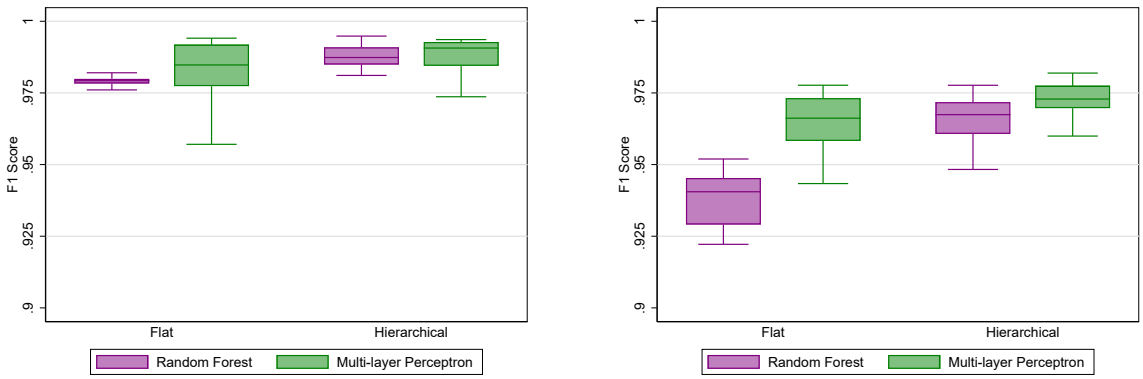
Figure 8: Relationship between number of features, out-of-sample performance statistics, and time to train model



Notes: This figure shows the convergence of out-of-sample model performance as the number of selected features is increased from 100 4-grams to 10,000 4-grams selected using TF-IDF. 20 bootstrapped hierarchical multi-layer perceptron models were estimated for each level of feature space, with 200,000 training observations selected at random (with replacement) from 240,224 unique descriptions. The left panel shows the change in average and 5th/95th percentile model performance at the Broad Crime Type as the number of selected features increase while the right panel shows the change in average and 5th/95th percentile model performance at the UCCS Code at different feature space. The bottom panel summarizes the distribution of training time for each feature space on the Criminal Justice Administrative Record System's (CJARS) server which has 256 GB of RAM and 12 virtual processors. Each box plot incorporates the lower and upper adjacent values, 25th and 75th percentiles, and median performance.

Figure 9: Out-of-sample performance across variations in classification technique

Broad Crime Type **Full UCCS Code**



Notes: Box plots show the comparison of out-of-sample performance between random forest and multi-layer perceptron using flat and hierarchical classification techniques. Each box plot incorporates the lower and upper adjacent values, 25^{th} and 75^{th} percentiles, and median performance. 20 bootstrapped models were estimated for each classification method, with 200,000 training observations selected at random (with replacement) from 240,224 unique descriptions. For each classification method, a maximum of 5,000 4-grams were selected using TF-IDF. The box plot on the left shows the F1 score of each method at the Broad Crime Type while the box plot on the right shows the F1 score of each classification method at the UCCS Code.

Appendix (for online publication)

Detailed information on UCR, NCRP, and UCCS Schemas

This sections provides detailed list of offense categories used in existing offense classification schemes. Table A1 shows list of offenses reported in the SRS schema and Table A2 shows offense categories used in the NIBRS schema from the UCR Program. Table A3 shows list of offenses in the NCRP schema while Table A4 provides examples of inconsistent offense code mappings for a given description. Table A5 provides a full list of offenses covered in the UCCS schema.

Table A1: SRS Part I & Part II Offense Categories

Part I Offenses	*Part II Offenses*	
Criminal Homicide	Forgery/Counterfeiting	Driving Under the Influence
Rape	Fraud	Liquor Laws
Robbery	Embezzlement	Drunkenness
Aggravated Assault	Stolen Property	Disorderly Conduct
Burglary	Vandalism	Vagrancy
Larceny/Theft	Weapons	All Other Offenses
Motor Vehicle Theft	Prostitution/Commercialized Vice	Suspicion
Arson	Sex Offenses (Except Rape, Prostitution, Commercialized Vice)	Curfew and Loitering Laws
Human Trafficking - Commercial Sex Acts	Drug Abuse Violations	Runaways
Human Trafficking - Involuntary Servitude	Offenses Against the Family and Children	Human Trafficking
	Gambling	Assault

Notes: "Part I offenses" are reserved for serious offense types while "Part II offenses" are used for lesser offenses that may not always be reported to the police. In the event that a person is charged with both Part I and Part II offenses (e.g. "Aggravated Assault" and "Vandalism"), only the former will be reported due to SRS' hierarchy rule. Source: Federal Bureau of Investigation (2013).

Table A2: NIBRS Offense Categories

NIBRS Code	NIBRS Description	NIBRS Category	NIBRS Offense Group	Crime Against Type
100	Kidnapping/Abduction	Kidnapping/Abduction	A	Person
120	Robbery	Robbery	A	Property
200	Arson	Arson	A	Property
210	Extortion/Blackmail	Extortion/Blackmail	A	Property
220	Burglary/Breaking & Entering	Burglary/Breaking & Entering	A	Property
240	Motor Vehicle Theft	Motor Vehicle Theft	A	Property
250	Counterfeiting/Forgery	Counterfeiting/Forgery	A	Property
270	Embezzlement	Embezzlement	A	Property
280	Stolen Property Offenses	Stolen Property Offenses	A	Property
290	Destruction/Damage/Vandalism of Property	Destruction/Damage/Vandalism of Property	A	Property
370	Pornography/Obscene Material	Pornography/Obscene Material	A	Society
500	Violation of No Contact/Protection Order	Violation of No Contact/Protection Order	A	Person
510	Bribery	Bribery	A	Property
520	Weapon Law Violations	Weapon Law Violations	A	Society
720	Animal Cruelty	Animal Cruelty	A	Property
09A	Murder & Non-negligent Manslaughter	Homicide Offenses	A	Person
09B	Negligent Manslaughter	Homicide Offenses	A	Person
09C	Justifiable Homicide	Homicide Offenses	A	Person/Not a Crime
11A	Forcible Rape	Sex Offenses	A	Person
11B	Forcible Sodomy	Sex Offenses	A	Person
11C	Sexual Assault With An Object	Sex Offenses	A	Person
11D	Forcible Fondling	Sex Offenses	A	Person
13A	Aggravated Assault	Assault Offenses	A	Person
13B	Simple Assault	Assault Offenses	A	Person
13C	Intimidation	Assault Offenses	A	Person
23A	Pocket-picking	Larceny/Theft Offenses	A	Property
23B	Purse-snatching	Larceny/Theft Offenses	A	Property
23C	Shoplifting	Larceny/Theft Offenses	A	Property
23D	Theft From Building	Larceny/Theft Offenses	A	Property
23E	Theft From Coin-Operated Machine or Device	Larceny/Theft Offenses	A	Property
23F	Theft From Motor Vehicle	Larceny/Theft Offenses	A	Property
23G	Theft of Motor Vehicle Parts or Accessories	Larceny/Theft Offenses	A	Property
23H	All Other Larceny	Larceny/Theft Offenses	A	Property
26A	False Pretenses/Swindle/Confidence Game	Fraud Offenses	A	Property
26B	Credit Card/Automated Teller Machine Fraud	Fraud Offenses	A	Property
26C	Impersonation	Fraud Offenses	A	Property
26D	Welfare Fraud	Fraud Offenses	A	Property
26E	Wire Fraud	Fraud Offenses	A	Property
26F	Identity Theft	Fraud Offenses	A	Property
26G	Hacking/Computer	Fraud Offenses	A	Property
35A	Drug/Narcotic Violations	Drug/Narcotic Offenses	A	Society
35B	Drug Equipment Violations	Drug/Narcotic Offenses	A	Society
36A	Incest	Sex Offenses, Consensual	A	Person
36B	Statutory Rape	Sex Offenses, Consensual	A	Person
39A	Betting/Wagering	Gambling Offenses	A	Society
39B	Operating/Promoting/Assisting Gambling	Gambling Offenses	A	Society
39C	Gambling Equipment Violations	Gambling Offenses	A	Society
39D	Sports Tampering	Gambling Offenses	A	Society
40A	Prostitution	Prostitution Offenses	A	Society
40B	Assisting or Promoting Prostitution	Prostitution Offenses	A	Society
40C	Purchasing Prostitution	Prostitution Offenses	A	Society

Table A2: NIBRS Offense Categories - Continued

NIBRS Code	NIBRS Description	NIBRS Category	NIBRS Offense Group	Crime Against Type
90A	Bad Checks	Bad Checks	B	Property
90B	Curfew/Loitering/Vagrancy Violations	Curfew/Loitering/Vagrancy Violations	B	Society
90C	Disorderly Conduct	Disorderly Conduct	B	Society
90D	Driving Under the Influence	Driving Under the Influence	B	Society
90E	Drunkenness	Drunkenness	B	Society
90F	Family Offenses, Nonviolent	Family Offenses, Nonviolent	B	Society
90G	Liquor Law Violations	Liquor Law Violations	B	Society
90H	Peeping Tom	Peeping Tom	B	Society
90I	Runaway	Runaway	B	Not a Crime
90J	Trespass of Real Property	Trespass of Real Property	B	Society
90Z	All Other Offenses	All Other Offenses	B	Person/Property/Society
64A	Human Trafficking, Commercial Sex Acts	Human Trafficking	A	Person
64B	Human Trafficking, Involuntary Servitude	Human Trafficking	A	Person

Notes: NIBRS users require additional crosswalks to analyze crime trends with more granularity. For example, identifying the specific type of drug involved in "Drug/Narcotic Violations" (35A) requires users to merge in a separate table that contains such information. Source: US Department of Justice (2021).

Table A3: NCRP Offense Categories

BJS Code	BJS Description	BJS Category	BJS Broad Category
010	Murder	Murder	Violent
011	Assault with Intent to Kill	Murder	Violent
012	Conspiracy to Commit Murder	Murder	Violent
013	Unspecified Homicide - Willful Kill	Unspecified Homicide	Violent
014	Unspecified Homicide, Attempted/Conspiracy	Unspecified Homicide	Violent
015	Voluntary/Nonnegligent Manslaughter	Non-negligent Manslaughter	Violent
016	Voluntary/Nonnegligent Manslaughter, Attempted/Conspiracy	Non-negligent Manslaughter	Violent
020	Manslaughter, Vehicular	Negligent Manslaughter	Violent
021	Manslaughter, Vehicular, Attempted	Negligent Manslaughter	Violent
022	Manslaughter, Vehicular, Conspiracy	Negligent Manslaughter	Violent
030	Involuntary Manslaughter	Negligent Manslaughter	Violent
031	Attempted Manslaughter	Negligent Manslaughter	Violent
032	Manslaughter, Non Vehicular, Conspiracy	Negligent Manslaughter	Violent
040	Kidnapping/Abduction	Kidnapping	Violent
041	Kidnapping/Abduction, Attempted	Kidnapping	Violent
042	Kidnapping/Abduction, Conspiracy	Kidnapping	Violent
050	Forcible Rape	Sexual Assault	Violent
051	Forcible Rape, Attempted	Sexual Assault	Violent
052	Forcible Rape, Conspiracy	Sexual Assault	Violent
060	Statutory Rape	Sexual Assault	Violent
061	Statutory Rape, Attempted	Sexual Assault	Violent
062	Statutory Rape, Conspiracy	Sexual Assault	Violent
070	Sexual Abuse	Sexual Assault	Violent
071	Sexual Assault, Attempted	Sexual Assault	Violent
072	Sexual Assault, Conspiracy	Sexual Assault	Violent
080	Lewd Act with a Child	Sexual Assault	Violent
081	Lewd Act with a Child, Attempted	Sexual Assault	Violent
082	Lewd Act with a Child, Conspiracy	Sexual Assault	Violent
090	Armed Robbery	Robbery	Violent
091	Armed Robbery, Attempted	Robbery	Violent
092	Armed Robbery, Conspiracy	Robbery	Violent
100	Unarmed Robbery	Robbery	Violent
101	Unarmed Robbery, Attempted	Robbery	Violent
102	Unarmed Robbery, Conspiracy	Robbery	Violent
110	Forcible Sodomy	Sexual Assault	Violent
111	Attempted Forcible Sodomy	Sexual Assault	Violent
112	Conspiracy to Commit Forcible Sodomy	Sexual Assault	Violent
120	Aggravated Assault	Assault	Violent
121	Aggravated Assault, Attempted	Assault	Violent
122	Aggravated Assault, Conspiracy	Assault	Violent
130	Simple Assault	Assault	Violent
131	Simple Assault, Attempted	Assault	Violent
132	Simple Assault, Conspiracy	Assault	Violent
140	Assault on a Public Safety Officer	Assault	Violent
141	Assault on a Public Safety Officer, Attempted	Assault	Violent
142	Assault on a Public Safety Officer, Conspiracy	Assault	Violent
150	Blackmail/Intimidation/Extort	Other Violent	Violent
151	Blackmail/Intimidation/Extort, Attempted	Other Violent	Violent
152	Blackmail/Intimidation/Extort, Conspiracy	Other Violent	Violent
160	Hit and Run with Bodily Injury	Other Violent	Violent
161	Hit and Run with Bodily Injury, Attempted	Other Violent	Violent
162	Hit and Run with Bodily Injury, Conspiracy	Other Violent	Violent
170	Child Abuse	Other Violent	Violent
171	Child Abuse, Attempted	Other Violent	Violent
172	Child Abuse, Conspiracy	Other Violent	Violent
180	Violent Offenses - Other	Other Violent	Violent

Table A3: NCRP Offense Categories - Continued

BJS Code	BJS Description	BJS Category	BJS Broad Category
190	Burglary	Burglary	Property
191	Burglary, Attempted	Burglary	Property
192	Burglary, Conspiracy	Burglary	Property
200	Arson	Arson	Property
201	Arson, Attempted	Arson	Property
202	Arson, Conspiracy	Arson	Property
210	Auto Theft	Motor Vehicle Theft	Property
211	Auto Theft, Attempted	Motor Vehicle Theft	Property
212	Auto Theft, Conspiracy	Motor Vehicle Theft	Property
220	Forgery/Fraud	Fraud	Property
221	Forgery/Fraud, Attempted	Fraud	Property
222	Forgery/Fraud, Conspiracy	Fraud	Property
230	Grand Larceny/Theft, $200+	Larceny	Property
231	Grand Larceny/Theft, $200+, Attempted	Larceny	Property
232	Grand Larceny/Theft, $200+, Conspiracy	Larceny	Property
240	Petty Larceny/Theft, Under $200	Larceny	Property
241	Petty Larceny/Theft, Under $200, Attempted	Larceny	Property
242	Petty Larceny/Theft, Under $200, Conspiracy	Larceny	Property
250	Larceny/Theft Value Unknown	Larceny	Property
251	Larceny/Theft Value Unknown, Attempted	Larceny	Property
252	Larceny/Theft Value Unknown, Conspiracy	Larceny	Property
260	Embezzlement	Other Property	Property
261	Embezzlement, Attempted	Other Property	Property
262	Embezzlement, Conspiracy	Other Property	Property
270	Receiving Stolen Property	Stolen Property	Property
271	Receiving Stolen Property, Attempted	Stolen Property	Property
272	Receiving Stolen Property, Conspiracy	Stolen Property	Property
280	Stolen Property Trafficking	Stolen Property	Property
281	Stolen Property Trafficking, Attempted	Stolen Property	Property
282	Stolen Property Trafficking, Conspiracy	Stolen Property	Property
290	Destruction of Property	Other Property	Property
291	Destruction of Property, Attempted	Other Property	Property
292	Destruction of Property, Conspiracy	Other Property	Property
300	Hit and Run Driving - Property Damage	Other Property	Property
310	Unauthorized Use of a Motor Vehicle	Other Property	Property
311	Unauthorized use of Vehicle, Attempted	Other Property	Property
312	Unauthorized use of Vehicle, Conspiracy	Other Property	Property
320	Trespass Against Property	Other Property	Property
321	Trespass Against Property, Attempted	Other Property	Property
322	Trespass Against Property, Conspiracy	Other Property	Property
330	Other Property Offenses, Other Types	Other Property	Property
331	Other Property Offenses, Attempted	Other Property	Property
332	Other Property Offenses, Conspiracy	Other Property	Property
333	Possession of Burglary Tools	Other Property	Property
334	Possession of Burglary Tools, Attempted	Other Property	Property
335	Possession of Burglary Tools, Conspiracy	Other Property	Property

Table A3: NCRP Offense Categories - Continued

BJS Code	BJS Description	BJS Category	BJS Broad Category
340	Drug Trafficking - Heroin	Drug Trafficking	Drug
341	Drug Trafficking - Heroin, Attempted	Drug Trafficking	Drug
342	Drug Trafficking - Heroin, Conspiracy	Drug Trafficking	Drug
345	Drug Trafficking - Cocaine/Crack	Drug Trafficking	Drug
346	Drug Trafficking - Cocaine/Crack, Attempted	Drug Trafficking	Drug
347	Drug Trafficking - Cocaine/Crack, Conspiracy	Drug Trafficking	Drug
350	Drug Trafficking - Other	Drug Trafficking	Drug
351	Drug Trafficking - Other, Attempted	Drug Trafficking	Drug
352	Drug Trafficking - Other, Conspiracy	Drug Trafficking	Drug
360	Drug Trafficking - Marijuana	Drug Trafficking	Drug
361	Drug Trafficking - Marijuana, Attempted	Drug Trafficking	Drug
362	Drug Trafficking - Marijuana, Conspiracy	Drug Trafficking	Drug
370	Drug Trafficking - Unspecified	Drug Trafficking	Drug
371	Drug Trafficking - Unspecified, Attempted	Drug Trafficking	Drug
372	Drug Trafficking - Unspecified, Conspiracy	Drug Trafficking	Drug
380	Drug Possession/Use - Heroin	Drug Possession/Use	Drug
381	Drug Possession/Use - Heroin, Attempted	Drug Possession/Use	Drug
382	Drug Possession/Use - Heroin, Conspiracy	Drug Possession/Use	Drug
385	Drug Possession/Use - Cocaine/Crack	Drug Possession/Use	Drug
386	Drug Possession/Use - Cocaine/Crack, Attempted	Drug Possession/Use	Drug
387	Drug Possession/Use - Cocaine/Crack, Conspiracy	Drug Possession/Use	Drug
390	Drug Possession/Use - Other	Drug Possession/Use	Drug
391	Drug Possession/Use - Other, Attempted	Drug Possession/Use	Drug
392	Drug Possession/Use - Other, Conspiracy	Drug Possession/Use	Drug
400	Drug Possession/Use - Marijuana	Drug Possession/Use	Drug
401	Drug Possession/Use - Marijuana, Attempted	Drug Possession/Use	Drug
402	Drug Possession/Use - Marijuana, Conspiracy	Drug Possession/Use	Drug
410	Drug Possession/Use - Unspecified	Drug Possession/Use	Drug
420	Drug Offense Unspecified - Heroin	Other Drug Offenses	Drug
425	Drug Offense Unspecified - Cocaine/Crack	Other Drug Offenses	Drug
430	Drug Offense Unspecified - Other	Other Drug Offenses	Drug
440	Drug Offense Unspecified - Marijuana	Other Drug Offenses	Drug
450	Drug Offense Unspecified - Unspecified (e.g. Utter Rx, Possess Drug Paraphernalia)	Other Drug Offenses	Drug
460	Escape from Custody	Other Public Order	Public Order
461	Escape from Custody, Attempted	Other Public Order	Public Order
462	Escape from Custody, Conspiracy (includes harboring)	Other Public Order	Public Order
470	Flight to Avoid Prosecution	Other Public Order	Public Order
471	Flight to Avoid Prosecution, Attempted	Other Public Order	Public Order
472	Flight to Avoid Prosecution, Conspiracy	Other Public Order	Public Order
480	Weapons Offense	Weapons Offense	Other
481	Weapons Offense, Attempted	Weapons Offense	Other
482	Weapons Offense, Conspiracy	Weapons Offense	Other
490	Parole Violation	Other Public Order	Public Order
500	Probation Violation	Other Public Order	Public Order
510	Riot	Other Public Order	Public Order
511	Riot, Attempting to Incite	Other Public Order	Public Order
512	Riot, Conspiracy to Incite	Other Public Order	Public Order
520	Habitual Offender	Other Public Order	Public Order
530	Contempt of Court/Violate Prot or Rest Order/Fail to Pay Fines	Other Public Order	Public Order
540	Other Court Offenses (e.g. Bond Jump, FTA, Intimidate Witness, Perjury, Tampering)	Other Public Order	Public Order
541	Other Court Offenses, Attempted	Other Public Order	Public Order
542	Other Court Offenses, Conspiracy	Other Public Order	Public Order
550	Minor Traffic Offenses	Other Public Order	Public Order
560	Driving While Intoxicated	Driving While Intoxicated	Other
565	Driving Under the Influence - Alcohol/Unspecified	Driving While Intoxicated	Other
570	Driving Under the Influence - Drugs	Driving While Intoxicated	Other
580	Family Offenses	Other Public Order	Public Order
590	Drunk/Vagrant/Disorderly Conduct	Other Public Order	Public Order
600	Offense Against Morals/Decency	Other Public Order	Public Order
601	Offense Against Morals/Decency, Attempted	Other Public Order	Public Order
602	Offense Against Morals/Decency, Conspiracy	Other Public Order	Public Order
610	Immigration Violation (e.g. Harboring, Smuggling, Illegal Entry)	Other Public Order	Public Order
620	Obstruction of Law Enforcement	Other Public Order	Public Order
621	Obstruction of Law Enforcement, Attempted	Other Public Order	Public Order
622	Obstruction of Law Enforcement, Conspiracy	Other Public Order	Public Order

Table A3: NCRP Offense Categories - Continued

BJS Code	BJS Description	BJS Category	BJS Broad Category
630	Invasion of Privacy	Other Public Order	Public Order
640	Commercialized Vice (e.g. Gambling, Prostitution)	Other Public Order	Public Order
650	Contributing to the Delinquency of a Minor	Other Public Order	Public Order
660	Liquor Law Violations Excluding Drunkenness and DWI	Other Public Order	Public Order
670	Public Order Offenses	Other Public Order	Public Order
671	Public Order Offenses, Attempted	Other Public Order	Public Order
672	Public Order Offenses, Conspiracy	Other Public Order	Public Order
673	Bribery excluding Public Officer	Other Public Order	Public Order
674	Bribery excluding Public Officer, Attempted	Other Public Order	Public Order
675	Bribery excluding Public Officer, Conspiracy	Other Public Order	Public Order
680	Juvenile Offenses	Other Public Order	Public Order
690	Felony Unspecified	Other Public Order	Public Order
691	Felony Unspecified, Attempted	Other Public Order	Public Order
692	Felony Unspecified, Conspiracy	Other Public Order	Public Order
700	Misdemeanor Unspecified	Other Public Order	Public Order
710	Other/Unknown Offense	Other	Other
800	Embezzlement	Federal Offense	Property
810	Fraud	Federal Offense	Property
820	Forgery	Federal Offense	Property
830	Counterfeiting	Federal Offense	Property
840	Regulatory Offense	Federal Offense	Property
850	Tax Law	Federal Offense	Property
860	Racketeering/Extortion	Federal Offense	Property
995	Illegal Entries	Other	
997	Not Known, Seeking State Clarification	Other	
998	Blanks	Blank	
999	Not Known	Other	

Notes: In the NCRP schema, "BJS Broad Category" is used to classify offenses by Violent, Drug, Property, Public Order, or Other offenses. Within these broader categories, "BJS Category" is then used to provide further sub-classification of offenses categories. Finally, "BJS Code" provides information on whether the offense involved committed or inchoate crime. Source: Bureau of Justice Statistics (2020b).

Table A4: Examples of Inconsistent Offense Classification

State	Offense Description	BJS Code
Alaska	MANSLAUGHTER	013
Alabama	MANSLAUGHTER	030
Arkansas	MANSLAUGHTER	015
Arizona	MANSLAUGHTER	013
California	VOLUNTARY MANSLAUGHTER	015
Kentucky	VOLUNTARY MANSLAUGHTER	710
Nevada	VOLUNTARY MANSLAUGHTER	010
Pennsylvania	VOLUNTARY MANSLAUGHTER	030
Tennessee	VOLUNTARY MANSLAUGHTER	120
Tennessee	VOLUNTARY MANSLAUGHTER	015
Virginia	VOLUNTARY MANSLAUGHTER	010
Virginia	VOLUNTARY MANSLAUGHTER	011

Notes: Inconsistent offense code mappings for a given descriptions pose challenge for machine learning algorithms. For instance, the most recent NCRP crosswalks 8 different offense codes for the offense description "VOLUNTARY MANSLAUGHTER." Furthermore, these inconsistencies exist for a state specific crosswalk. Source: Bureau of Justice Statistics (2020b).

Table A5: Uniform Crime Classification Standard Schema

UCCS Code	UCCS Description	Broad Code	Broad Description	Offense Code	Offense Description	Offense Modifier Code	Offense Modifier Description
1010	Murder	1	Violent	01	Murder	0	
1011	Attempted Murder	1	Violent	01	Murder	1	Attempt
1012	Conspiracy to Commit Murder	1	Violent	01	Murder	2	Conspiracy
1020	Unspecified Homicide	1	Violent	02	Unspecified homicide	0	
1021	Unspecified Homicide, Attempted	1	Violent	02	Unspecified homicide	1	Attempt
1022	Unspecified Homicide, Conspiracy	1	Violent	02	Unspecified homicide	2	Conspiracy
1030	Voluntary Manslaughter	1	Violent	03	Voluntary/nonnegligent manslaughter	0	
1031	Voluntary Manslaughter, Attempted	1	Violent	03	Voluntary/nonnegligent manslaughter	1	Attempt
1032	Voluntary Manslaughter,Conspiracy	1	Violent	03	Voluntary/nonnegligent manslaughter	2	Conspiracy
1040	Vehicular Manslaughter	1	Violent	04	Voluntary/nonnegligent manslaughter	0	
1041	Vehicular Manslaughter, Attempted	1	Violent	04	Voluntary/nonnegligent manslaughter	1	Attempt
1042	Vehicular Manslaughter, Conspiracy	1	Violent	04	Voluntary/nonnegligent manslaughter	2	Conspiracy
1050	Involuntary Manslaughter	1	Violent	05	Manslaughter - non-vehicular	0	
1051	Involuntary Manslaughter, Attempt	1	Violent	05	Manslaughter - non-vehicular	1	Attempt
1052	Involuntary Manslaughter, Conspiracy	1	Violent	05	Manslaughter - non-vehicular	2	Conspiracy
1060	Kidnapping	1	Violent	06	Kidnapping	0	
1061	Kidnapping, Attempted	1	Violent	06	Kidnapping	1	Attempt
1062	Kidnapping, Conspiracy	1	Violent	06	Kidnapping	2	Conspiracy
1070	Rape	1	Violent	07	Rape - force	0	
1071	Rape, Attempted	1	Violent	07	Rape - force	1	Attempt
1072	Rape, Conspiracy	1	Violent	07	Rape - force	2	Conspiracy
1080	Statutory Rape	1	Violent	08	Rape - statutory - no force	0	
1081	Statutory Rape, Attempted	1	Violent	08	Rape - statutory - no force	1	Attempt
1082	Statutory Rape, Conspiracy	1	Violent	08	Rape - statutory - no force	2	Conspiracy
1090	Child Molestation	1	Violent	09	Lewd act with children	0	
1091	Child Molestation, Attempted	1	Violent	09	Lewd act with children	1	Attempt
1092	Child Molestation, Conspiracy	1	Violent	09	Lewd act with children	2	Conspiracy
1100	Sexual Assault	1	Violent	10	Sexual assault - other	0	
1101	Sexual Assault, Attempted	1	Violent	10	Sexual assault - other	1	Attempt
1102	Sexual Assault, Conspiracy	1	Violent	10	Sexual assault - other	2	Conspiracy
1110	Human Trafficking, Sex - child	1	Violent	11	Human Trafficking	0	
1111	Human Trafficking, Sex - child, Attempted	1	Violent	11	Human Trafficking	1	Attempt
1112	Human Trafficking, Sex - child, Conspiracy	1	Violent	11	Human Trafficking	2	Conspiracy
1120	Human Trafficking, Sex - adult or no age specified	1	Violent	12	Human Trafficking	0	
1121	Human Trafficking, Sex - adult or no age specified, Attempted	1	Violent	12	Human Trafficking	1	Attempt
1122	Human Trafficking, Sex - adult or no age specified, Conspiracy	1	Violent	12	Human Trafficking	2	Conspiracy
1130	Human Trafficking, Labor - child	1	Violent	13	Human Trafficking	0	
1131	Human Trafficking, Labor - child, Attempted	1	Violent	13	Human Trafficking	1	Attempt
1132	Human Trafficking, Labor - child, Conspiracy	1	Violent	13	Human Trafficking	2	Conspiracy
1140	Human Trafficking, Labor - adult or no age specified	1	Violent	14	Human Trafficking	0	
1141	Human Trafficking, Labor - adult or no age specified, Attempted	1	Violent	14	Human Trafficking	1	Attempt
1142	Human Trafficking, Labor - adult or no age specified, Conspiracy	1	Violent	14	Human Trafficking	2	Conspiracy
1150	Human Trafficking, Unspecified - child	1	Violent	15	Human Trafficking	0	
1151	Human Trafficking, Unspecified - child, Attempted	1	Violent	15	Human Trafficking	1	Attempt
1152	Human Trafficking, Unspecified - child, Conspiracy	1	Violent	15	Human Trafficking	2	Conspiracy
1160	Human Trafficking, Unspecified - adult or no age specified	1	Violent	16	Human Trafficking	0	
1161	Human Trafficking, Unspecified - adult or no age specified, Attempted	1	Violent	16	Human Trafficking	1	Attempt
1162	Human Trafficking, Unspecified - adult or no age specified, Conspiracy	1	Violent	16	Human Trafficking	2	Conspiracy
1170	Human Trafficking	1	Violent	17	Human Trafficking	0	
1171	Human Trafficking, Attempted	1	Violent	17	Human Trafficking	1	Attempt
1172	Human Trafficking, Conspiracy	1	Violent	17	Human Trafficking	2	Conspiracy
1180	Armed Robbery	1	Violent	18	Armed robbery	0	
1181	Armed Robbery, Attempted	1	Violent	18	Armed robbery	1	Attempt
1182	Armed Robbery, Conspiracy	1	Violent	18	Armed robbery	2	Conspiracy
1190	Unarmed Robbery	1	Violent	19	Unarmed robbery	0	
1191	Unarmed Robbery, Attempted	1	Violent	19	Unarmed robbery	1	Attempt
1192	Unarmed Robbery, Conspiracy	1	Violent	19	Unarmed robbery	2	Conspiracy
1200	Aggravated Assault	1	Violent	20	Aggravated assault	0	
1201	Aggravated Assault, Attempted	1	Violent	20	Aggravated assault	1	Attempt
1202	Aggravated Assault, Conspiracy	1	Violent	20	Aggravated assault	2	Conspiracy
1210	Assault of an Officer	1	Violent	21	Assaulting public officer	0	
1211	Assault of an Officer, Attempted	1	Violent	21	Assaulting public officer	1	Attempt
1212	Assault of an Officer, Conspiracy	1	Violent	21	Assaulting public officer	2	Conspiracy
1220	Child Abuse	1	Violent	22	Child abuse	0	
1221	Child Abuse, Attempted	1	Violent	22	Child abuse	1	Attempt
1222	Child Abuse, Conspiracy	1	Violent	22	Child abuse	2	Conspiracy
1230	Simple Assault	1	Violent	23	Simple assault	0	
1231	Simple Assault, Attempted	1	Violent	23	Simple assault	1	Attempt
1232	Simple Assault, Conspiracy	1	Violent	23	Simple assault	2	Conspiracy
1240	Extortion/Threat	1	Violent	24	Blackmail/extortion/intimidation	0	
1241	Extortion/Threat, Attempted	1	Violent	24	Blackmail/extortion/intimidation	1	Attempt
1242	Extortion/Threat, Conspiracy	1	Violent	24	Blackmail/extortion/intimidation	2	Conspiracy
1250	Hit and Run with Bodily Injury	1	Violent	25	Hit and run driving - injury	0	
1251	Hit and Run with Bodily Injury, Attempted	1	Violent	25	Hit and run driving - injury	1	Attempt
1252	Hit and Run with Bodily Injury, Conspiracy	1	Violent	25	Hit and run driving - injury	2	Conspiracy
1990	Violent Offense, Other	1	Violent	99	Violent offenses - other	0	
1991	Violent Offense Other, Attempted	1	Violent	99	Violent offenses - other	1	Attempt
1992	Violent Offense Other, Conspiracy	1	Violent	99	Violent offenses - other	2	Conspiracy
2010	Burglary	2	Property	01	Burglary	0	
2011	Burglary, Attempted	2	Property	01	Burglary	1	Attempt
2012	Burglary, Conspiracy	2	Property	01	Burglary	2	Conspiracy

Table A5: Uniform Crime Classification Standard Schema - Continued

UCCS Code	UCCS Description	Broad Code	Broad Description	Offense Code	Offense Description	Offense Modifier Code	Offense Modifier Description
2020	Arson	2	Property	02	Arson	0	
2021	Arson, Attempted	2	Property	02	Arson	1	Attempt
2022	Arson, Conspiracy	2	Property	02	Arson	2	Conspiracy
2030	Auto Theft	2	Property	03	Auto theft	0	
2031	Auto Theft, Attempted	2	Property	03	Auto theft	1	Attempt
2032	Auto Theft, Conspiracy	2	Property	03	Auto theft	2	Conspiracy
2040	Forgery/Fraud	2	Property	04	Forgery/fraud	0	
2041	Forgery/Fraud, Attempted	2	Property	04	Forgery/fraud	1	Attempt
2042	Forgery/Fraud, Conspiracy	2	Property	04	Forgery/fraud	2	Conspiracy
2050	Grand Theft (>$500)	2	Property	05	Grand larceny - theft over $500	0	
2051	Grand Theft (>$500), Attempted	2	Property	05	Grand larceny - theft over $500	1	Attempt
2052	Grand Theft (>$500), Conspiracy	2	Property	05	Grand larceny - theft over $500	2	Conspiracy
2060	Petty Theft (=<$500)	2	Property	06	Petty larceny - theft equal or under $500	0	
2061	Petty Theft (=<$500), Attempted	2	Property	06	Petty larceny - theft equal or under $500	1	Attempt
2062	Petty Theft (=<$500), Conspiracy	2	Property	06	Petty larceny - theft equal or under $500	2	Conspiracy
2070	Theft, Value Unknown	2	Property	07	Larceny/theft - value unknown	0	
2071	Theft, Value Unknown, Attempted	2	Property	07	Larceny/theft - value unknown	1	Attempt
2072	Theft, Value Unknown, Conspiracy	2	Property	07	Larceny/theft - value unknown	2	Conspiracy
2080	Financial Crimes	2	Property	08	Financial Crimes	0	
2081	Financial Crimes Attempted	2	Property	08	Financial Crimes	1	Attempt
2082	Financial Crimes Conspiracy	2	Property	08	Financial Crimes	2	Conspiracy
2090	Sale of Stolen Property	2	Property	09	Stolen property - trafficking	0	
2091	Sale of Stolen Property, Attempted	2	Property	09	Stolen property - trafficking	1	Attempt
2092	Sale of Stolen Property, Conspiracy	2	Property	09	Stolen property - trafficking	2	Conspiracy
2100	Receiving Stolen Property	2	Property	10	Stolen property - receiving	0	
2101	Receiving Stolen Property, Attempted	2	Property	10	Stolen property - receiving	1	Attempt
2102	Receiving Stolen Property, Conspiracy	2	Property	10	Stolen property - receiving	2	Conspiracy
2110	Destruction of Property	2	Property	11	Destruction of property	0	
2111	Destruction of Property, Attempted	2	Property	11	Destruction of property	1	Attempt
2112	Destruction of Property, Conspiracy	2	Property	11	Destruction of property	2	Conspiracy
2120	Hit and Run Driving with Property Damage	2	Property	12	Hit and run driving - property damage	0	
2121	Hit and Run Driving, Attempted	2	Property	12	Hit and run driving - property damage	1	Attempt
2122	Hit and Run Driving, Conspiracy	2	Property	12	Hit and run driving - property damage	2	Conspiracy
2130	Unauthorized use of Vehicle	2	Property	13	Unauthorized use of vehicle	0	
2131	Unauthorized use of Vehicle, Attempted	2	Property	13	Unauthorized use of vehicle	1	Attempt
2132	Unauthorized use of Vehicle, Conspiracy	2	Property	13	Unauthorized use of vehicle	2	Conspiracy
2140	Criminal Trespass	2	Property	14	Trespassing	0	
2141	Criminal Trespass, Attempted	2	Property	14	Trespassing	1	Attempt
2142	Criminal Trespass, Conspiracy	2	Property	14	Trespassing	2	Conspiracy
2150	Possession of Property Crime Tools	2	Property	15	Property offenses - other	0	
2151	Possession of Property Crime Tools, Attempted	2	Property	15	Property offenses - other	1	Attempt
2152	Possession of Property Crime Tools, Conspiracy	2	Property	15	Property offenses - other	2	Conspiracy
2990	Other Property Offense	2	Property	99	Property offenses - other	0	
2991	Other Property Offense, Attempt	2	Property	99	Property offenses - other	1	Attempt
2992	Other Property Offense, Conspiracy	2	Property	99	Property offenses - other	2	Conspiracy
3010	Distribution Heroin	3	Drug	01	Distribution - heroin	0	
3011	Distribution, Heroin, Attempted	3	Drug	01	Distribution - heroin	1	Attempt
3012	Distribution, Heroin, Conspiracy	3	Drug	01	Distribution - heroin	2	Conspiracy
3020	Distribution of amphetamines	3	Drug	02	Distribution - amphetamines	0	
3021	Distribution of amphetamines, Attempted	3	Drug	02	Distribution - amphetamines	1	Attempt
3022	Distribution of amphetamines, Conspiracy	3	Drug	02	Distribution - amphetamines	2	Conspiracy
3030	Distribution Cocaine or Crack	3	Drug	03	Distribution - cocaine or crack	0	
3031	Distribution Cocaine or Crack, Attempted	3	Drug	03	Distribution - cocaine or crack	1	Attempt
3032	Distribution Cocaine or Crack, Conspiracy	3	Drug	03	Distribution - cocaine or crack	2	Conspiracy
3040	Distribution of opioids	3	Drug	04	Distribution of opioids	0	
3041	Distribution of opioids, Attempted	3	Drug	04	Distribution of opioids	1	Attempt
3042	Distribution of opioids, Conspiracy	3	Drug	04	Distribution of opioids	2	Conspiracy
3050	Distribution of prescription drugs	3	Drug	05	Distribution of prescription drugs	0	
3051	Distribution of prescription drugs, Attempted	3	Drug	05	Distribution of prescription drugs	1	Attempt
3052	Distribution of prescription drugs, Conspiracy	3	Drug	05	Distribution of prescription drugs	2	Conspiracy
3060	Distribution Other Controlled Substances	3	Drug	06	Distribution - other controlled substances	0	
3061	Distribution Other Controlled Substances, Attempted	3	Drug	06	Distribution - other controlled substances	1	Attempt
3062	Distribution Other Controlled Substances, Conspiracy	3	Drug	06	Distribution - other controlled substances	2	Conspiracy
3070	Distribution Marijuana	3	Drug	07	Distribution marijuana/hashish	0	
3071	Distribution Marijuana, Attempted	3	Drug	07	Distribution marijuana/hashish	1	Attempt
3072	Distribution Marijuana, Conspiracy	3	Drug	07	Distribution marijuana/hashish	2	Conspiracy
3080	Distribution, Drug Unspecified	3	Drug	08	Distribution - drug unspecified	0	
3081	Distribution, Drug Unspecified, Attempted	3	Drug	08	Distribution - drug unspecified	1	Attempt
3082	Distribution, Drug Unspecified, Conspiracy	3	Drug	08	Distribution - drug unspecified	2	Conspiracy
3090	Possession/Use of Heroin	3	Drug	09	Possession/use - heroin	0	
3091	Possession/Use of Heroin, Attempted	3	Drug	09	Possession/use - heroin	1	Attempt
3092	Possession/Use of Heroin, Conspiracy	3	Drug	09	Possession/use - heroin	2	Conspiracy
3100	Possession of amphetamines	3	Drug	10	Possession of amphetamines	0	
3101	Possession of amphetamines, Attempted	3	Drug	10	Possession of amphetamines	1	Attempt
3102	Possession of amphetamines, Conspiracy	3	Drug	10	Possession of amphetamines	2	Conspiracy

Table A5: Uniform Crime Classification Standard Schema - Continued

UCCS Code/Description		Broad Code/Description		Offense Code/Description		Offense Modifier Code/Description	
3110	Possession/Use of Cocaine or Crack	3	Drug	11	Possession/use - cocaine or crack	0	
3111	Possession/Use of Cocaine or Crack, Attempted	3	Drug	11	Possession/use - cocaine or crack	1	Attempt
3112	Possession/Use of Cocaine or Crack, Conspiracy	3	Drug	11	Possession/use - cocaine or crack	2	Conspiracy
3120	Possession of opioids	3	Drug	12	Possession of opioids	0	
3121	Possession of opioids, Attempted	3	Drug	12	Possession of opioids	1	Attempt
3122	Possession of opioids, Conspiracy	3	Drug	12	Possession of opioids	2	Conspiracy
3130	Possession of prescription drugs	3	Drug	13	Possession of prescription drugs	0	
3131	Possession of prescription drugs, Attempted	3	Drug	13	Possession of prescription drugs	1	Attempt
3132	Possession of prescription drugs, Conspiracy	3	Drug	13	Possession of prescription drugs	2	Conspiracy
3140	Possession/Use of Other Controlled Substance	3	Drug	14	Possession/use - other controlled substances	0	
3141	Possession/Use of Other Controlled Substance, Attempted	3	Drug	14	Possession/use - other controlled substances	1	Attempt
3142	Possession/Use of Other Controlled Substance, Conspiracy	3	Drug	14	Possession/use - other controlled substances	2	Conspiracy
3150	Possession/Use of Marijuana	3	Drug	15	Possession/use - marijuana/hashish	0	
3151	Possession/Use of Marijuana, Attempted	3	Drug	15	Possession/use - marijuana/hashish	1	Attempt
3152	Possession/Use of Marijuana, Conspiracy	3	Drug	15	Possession/use - marijuana/hashish	2	Conspiracy
3160	Possession/Use of Unspecified Drug	3	Drug	16	Possession/use - drug unspecified	0	
3161	Possession/Use, Drug Unspecified, Attempted	3	Drug	16	Possession/use - drug unspecified	1	Attempt
3162	Possession/Use, Drug Unspecified, Conspiracy	3	Drug	16	Possession/use - drug unspecified	2	Conspiracy
3170	Heroin Violation, Offense Unspecified	3	Drug	17	Heroin violation - offense unspecified	0	
3180	Amphetamines, Offense unspecified	3	Drug	18	Amphetamines - offense unspecified	0	
3190	Cocaine/Crack Violation, Offense Unspecified	3	Drug	19	Cocaine or crack violation offense unspecified	0	
3200	Prescription of opioid drugs, offense unspecified	3	Drug	20	Prescription of opioid drugs - offense unspecified	0	
3210	Prescription, offense unspecified	3	Drug	21	Prescription - offense unspecified	0	
3220	Other Controlled Substance Violation, Offense Unspecified	3	Drug	22	Controlled substance - offense unspecified	0	
3230	Marijuana Violation, Offense Unspecified	3	Drug	23	Marijuana/hashish violation - offense unspecified	0	
3240	Fraudulent Drug Offense	3	Drug	24	Other Drug Offense/Paraphernalia	0	
3241	Fraudulent Drug Offense, Attempted	3	Drug	24	Other Drug Offense/Paraphernalia	1	Attempt
3242	Fraudulent Drug Offense, Conspiracy	3	Drug	24	Other Drug Offense/Paraphernalia	2	Conspiracy
3250	Drug Paraphernalia	3	Drug	25	Other Drug Offense/Paraphernalia	0	
3251	Drug Paraphernalia, Attempted	3	Drug	25	Other Drug Offense/Paraphernalia	1	Attempt
3252	Drug Paraphernalia, Conspiracy	3	Drug	25	Other Drug Offense/Paraphernalia	2	Conspiracy
3990	Other Drug Offense	3	Drug	99	Other Drug Offense/Paraphernalia	0	
3991	Other Drug Offense, Attempt	3	Drug	99	Other Drug Offense/Paraphernalia	1	Attempt
3992	Other Drug Offense, Conspiracy	3	Drug	99	Other Drug Offense/Paraphernalia	2	Conspiracy
4010	Driving While Intoxicated	4	DUI Offense	01	Driving while intoxicated	0	
4011	Driving While Intoxicated, Attempted	4	DUI Offense	01	Driving while intoxicated	1	Attempt
4012	Driving While Intoxicated, Conspiracy	4	DUI Offense	01	Driving while intoxicated	2	Conspiracy
4020	Driving Under the Influence of Alcohol	4	DUI Offense	02	Driving Under the Influence	0	
4021	Driving Under the Influence of Alcohol, Attempted	4	DUI Offense	02	Driving Under the Influence	1	Attempt
4022	Driving Under the Influence of Alcohol, Conspiracy	4	DUI Offense	02	Driving Under the Influence	2	Conspiracy
4030	Driving Under the Influence of Drugs	4	DUI Offense	03	Driving under influence - drugs	0	
4031	Driving Under the Influence of Drugs, Attempted	4	DUI Offense	03	Driving under influence - drugs	1	Attempt
4032	Driving Under the Influence of Drugs, Conspiracy	4	DUI Offense	03	Driving under influence - drugs	2	Conspiracy
5010	Riot	5	Public Order	01	Rioting	0	
5011	Riot, Attempting to Incite	5	Public Order	01	Rioting	1	Attempt
5012	Riot, Conspiracy to Incite	5	Public Order	01	Rioting	2	Conspiracy
5020	Escape from Custody	5	Public Order	02	Escape from custody	0	
5021	Escape from Custody, Attempted	5	Public Order	02	Escape from custody	1	Attempt
5022	Escape from Custody, Conspiracy	5	Public Order	02	Escape from custody	2	Conspiracy
5030	Flight to Avoid Prosecution	5	Public Order	03	Flight to avoid prosecution	0	
5031	Flight to Avoid Prosecution, Attempted	5	Public Order	03	Flight to avoid prosecution	1	Attempt
5032	Flight to Avoid Prosecution, Conspiracy	5	Public Order	03	Flight to avoid prosecution	2	Conspiracy
5040	Weapons Offense	5	Public Order	04	Weapon offense	0	
5041	Weapons Offense, Attempted	5	Public Order	04	Weapon offense	1	Attempt
5042	Weapons Offense, Conspiracy	5	Public Order	04	Weapon offense	2	Conspiracy
5050	Habitual Offender	5	Public Order	05	Habitual offender	0	
5060	Parole Violation	5	Public Order	06	Parole violation	0	
5070	Probation Violation	5	Public Order	07	Probation violation	0	
5080	Contempt of Court/Violate Court Order	5	Public Order	08	Contempt of court	0	
5081	Contempt of Court/Violate Court Order, Attempted	5	Public Order	08	Contempt of court	1	Attempt
5082	Contempt of Court/Violate Court Order, Conspiracy	5	Public Order	08	Contempt of court	2	Conspiracy
5090	Other Court Offense	5	Public Order	09	Offenses against courts, legislatures and commissions	0	
5091	Other Court Offense, Attempted	5	Public Order	09	Offenses against courts, legislatures and commissions	1	Attempt
5092	Other Court Offense, Conspiracy	5	Public Order	09	Offenses against courts, legislatures and commissions	2	Conspiracy
5100	Family or Custody Related Offense	5	Public Order	10	Family related offenses	0	
5101	Family or Custody Related Offense, Attempted	5	Public Order	10	Family related offenses	1	Attempt
5102	Family or Custody Related Offense, Conspiracy	5	Public Order	10	Family related offenses	2	Conspiracy
5110	Offense Against Morals/Decency	5	Public Order	11	Morals/decency - offense	0	
5111	Offense Against Morals/Decency, Attempted	5	Public Order	11	Morals/decency - offense	1	Attempt
5112	Offense Against Morals/Decency, Conspiracy	5	Public Order	11	Morals/decency - offense	2	Conspiracy
5120	Immigration Violation	5	Public Order	12	Immigration violations	0	
5121	Immigration Violation, Attempted	5	Public Order	12	Immigration violations	1	Attempt
5122	Immigration Violation, Conspiracy	5	Public Order	12	Immigration violations	2	Conspiracy
5130	Obstruction/Resisting	5	Public Order	13	Obstruction - law enforcement	0	
5131	Obstruction/Resisting, Attempted	5	Public Order	13	Obstruction - law enforcement	1	Attempt
5132	Obstruction/Resisting, Conspiracy	5	Public Order	13	Obstruction - law enforcement	2	Conspiracy
5140	Invasion of Privacy	5	Public Order	14	Invasion of privacy	0	
5141	Invasion of Privacy, Attempted	5	Public Order	14	Invasion of privacy	1	Attempt
5142	Invasion of Privacy, Conspiracy	5	Public Order	14	Invasion of privacy	2	Conspiracy
5150	Commercialized Vice	5	Public Order	15	Commercialized vice	0	
5151	Commercialized Vice, Attempted	5	Public Order	15	Commercialized vice	1	Attempt
5152	Commercialized Vice, Conspiracy	5	Public Order	15	Commercialized vice	2	Conspiracy

Table A5: Uniform Crime Classification Standard Schema - Continued

UCCS Code	UCCS Description	Broad Code	Broad Description	Offense Code	Offense Description	Offense Modifier Code	Offense Modifier Description
5160	Contributing to the Delinquency of a Minor	5	Public Order	16	Contributing to delinquency of a minor	0	
5161	Contributing to the Delinquency of a Minor, Attempted	5	Public Order	16	Contributing to delinquency of a minor	1	Attempt
5162	Contributing to the Delinquency of a Minor, Conspiracy	5	Public Order	16	Contributing to delinquency of a minor	2	Conspiracy
5170	Disorderly Conduct Offense	5	Public Order	17	Drunkenness/Vagrancy/Disorderly Conduct	0	
5171	Disorderly Conduct Offense, Attempted	5	Public Order	17	Drunkenness/Vagrancy/Disorderly Conduct	1	Attempt
5172	Disorderly Conduct Offense, Conspiracy	5	Public Order	17	Drunkenness/Vagrancy/Disorderly Conduct	2	Conspiracy
5180	Liquor Law Violation	5	Public Order	18	Liquor law violations	0	
5181	Liquor Law Violation, Attempted	5	Public Order	18	Liquor law violations	1	Attempt
5182	Liquor Law Violation, Conspiracy	5	Public Order	18	Liquor law violations	2	Conspiracy
5190	Taxation Offense	5	Public Order	19	Taxation Offenses	0	
5191	Taxation Offense, Attempted	5	Public Order	19	Taxation Offenses	1	Attempt
5192	Taxation Offense, Conspiracy	5	Public Order	19	Taxation Offenses	2	Conspiracy
5200	Bribery/Conflict of Interest	5	Public Order	20	Bribery and conflict of interest	0	
5201	Bribery/Conflict of Interest, Attempt	5	Public Order	20	Bribery and conflict of interest	1	Attempt
5202	Bribery/Conflict of Interest, Conspiracy	5	Public Order	20	Bribery and conflict of interest	2	Conspiracy
5990	Public Order Offense, Other	5	Public Order	99	Public order offenses - other	0	
5991	Public Order Offense, Other, Attempted	5	Public Order	99	Public order offenses - other	1	Attempt
5992	Public Order Offense, Other, Conspiracy	5	Public Order	99	Public order offenses - other	2	Conspiracy
6010	Traffic Offense, Minor	6	Criminal traffic	01	Traffic offenses - minor	0	

Notes: The UCCS is operationalized as a four digit offense code that is hierarchical in nature. Each UCCS code is concatenation of Broad Crime Type Code (1st digit), Offense Code (2nd and 3rd digits), and Offense Modifier code (4th digit).