INSTITUTE FOR SOCIAL RESEARCH
UNIVERSITY OF MICHIGAN

CJARS Criminal Justice Administrative Records System

Hierarchical Approaches to Text-based Offense Classification

Jay Choi[1] David Kilmer[2] Michael Mueller-Smith[1] Sema Taheri[2]

[1]University of Michigan [2]Measures for Justice

Fall 2022

Motivation

CJARS Criminal Justice Administrative Records System

Offense classification plays critical roles for society, research, and public administration:

1. Criminal background checks (employment, public benefit eligibility, security credentialing, firearm purchases)
2. Evaluating sentencing disparities, testing behavioral theories, etc
3. Public funding, resource allocation, election outcomes

Yet, no comprehensive standard currently exists to cover the universe of illicit activity nor a mechanism to map free-entry descriptions into systematic codes, leading to:

- More discretion with little oversight
- Inconsistent definitions and mappings
- Irreproducible research

Project summary

CJARS Criminal Justice Administrative Records System

In this paper, we do the following:

1. Introduce the Uniform Crime Classification Standard (UCCS)
   - Schema satisfying 4 design principles laid out by the National Academy of Sciences
2. Develop the Text-based Offense Classification (TOC) tool
   - Hierarchical machine learning model trained on over 300k hand coded training records
3. Experiment with model variation in TOC to assess determinant of prediction success
   - Help identify potential lessons for other classification applications

Four design principles for offense classification

CJARS Criminal Justice Administrative Records System

National Academy of Sciences (2016) laid out four goals of offense schema:

1. Encompass new and emerging crime types
2. Fully realized classification for statistical purposes
3. Attribute-based classification
4. Enable comparisons between jurisdictions across time

	Summary Reporting System (SRS)	**National Incident-Based Reporting System (NIBRS)**	**National Corrections Reporting Program (NCRP)**	**Uniform Crime Classification Standard (UCCS)**
Principle 1:	×	✓	×	✓
Principle 2:	×	×	×	✓
Principle 3:	✓	✓	✓	✓
Principle 4:	×	×	×	✓

Uniform Crime Classification Standard

CJARS Criminal Justice Administrative Records System

Each UCCS code has four digits:

- First digit → Broad code
- Digits 2 and 3 → Offense code
- Fourth digit → Offense modifier

Example UCCS codes:

UCCS Code/Description		Broad Code/Description		Offense Code/Description		Offense Modifier Code/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

Text-based Offense Classification tool

CJARS Criminal Justice Administrative Records System

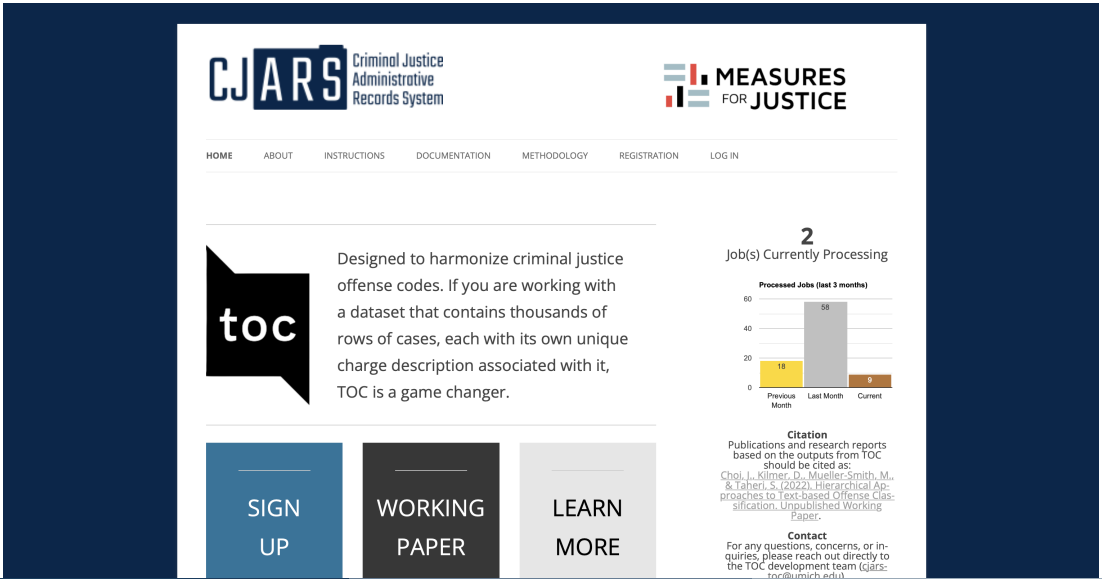
TOC tool maps free-entry offense description fields to UCCS codes

Built through a partnership between Measures for Justice (MFJ) and CJARS

- MFJ hand coded significant volume of offense descriptions, creating training data to develop TOC
- 313,209 hand-coded offense descriptions from 24 states

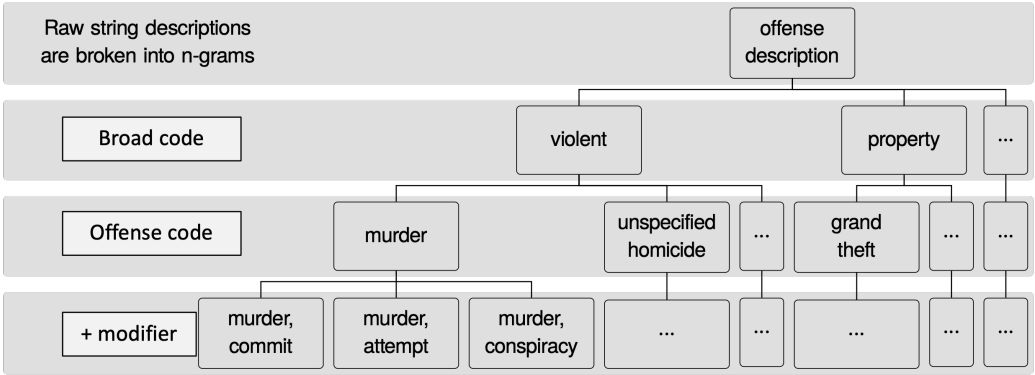
Freely available to general public online: `https://cjars-toc.isr.umich.edu`

Text-based Offense Classification tool



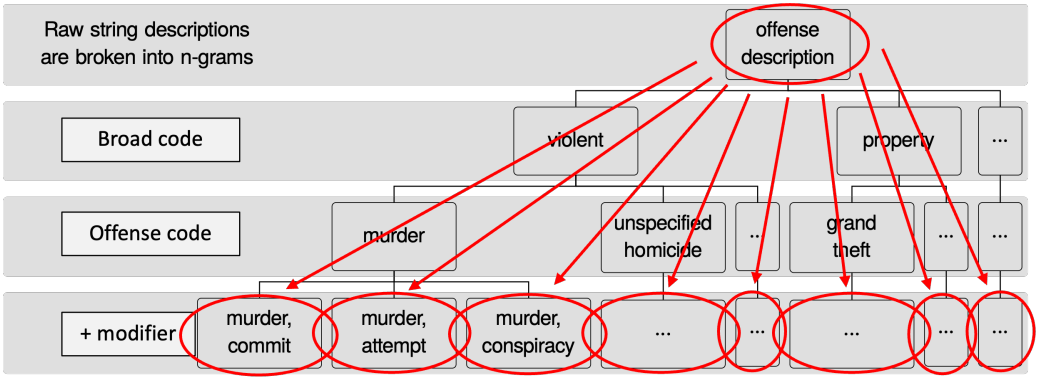
Prediction models

CJARS Criminal Justice Administrative Records System



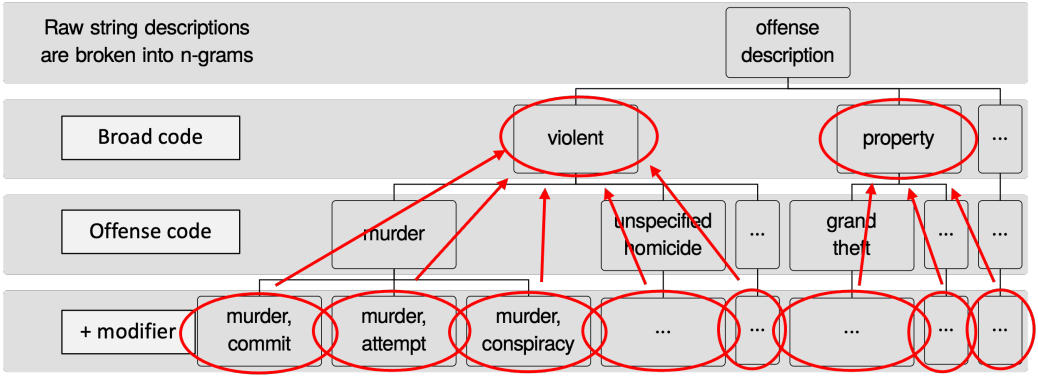
Prediction models: flat

CJARS Criminal Justice Administrative Records System



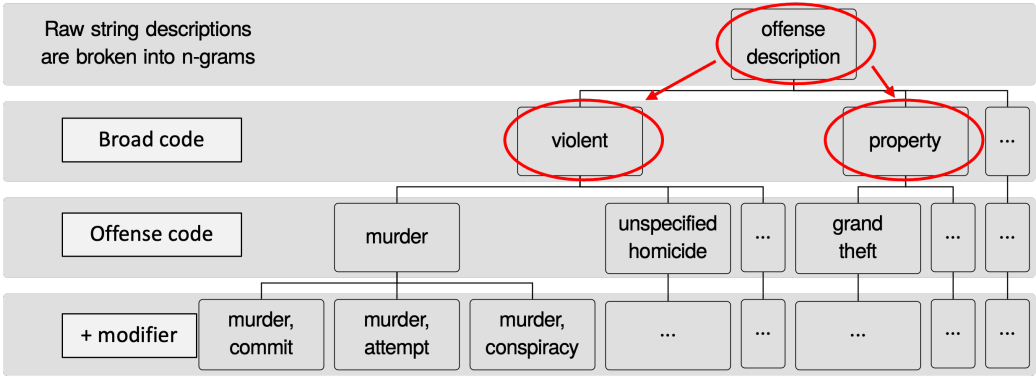
Prediction models: flat

CJARS Criminal Justice Administrative Records System

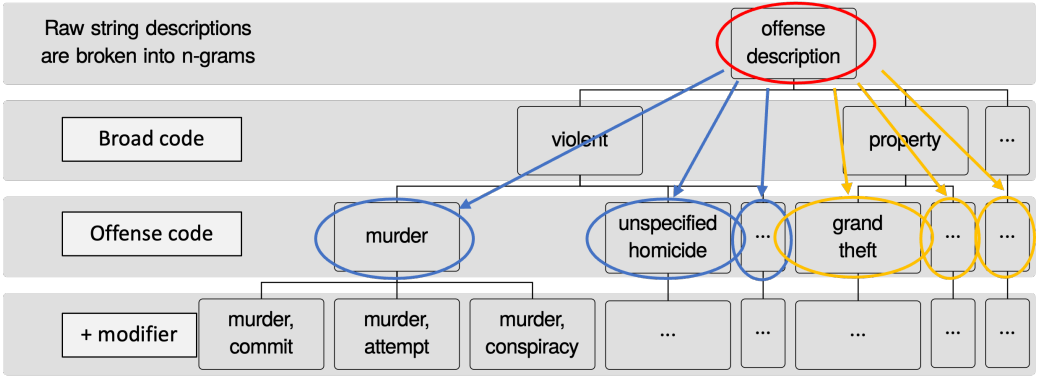


Prediction models: hierarchical

CJARS Criminal Justice Administrative Records System



Prediction models: hierarchical



Model details

CJARS Criminal Justice Administrative Records System

- 75% - 25% random split of hand-coded data for training vs testing purposes
  - Avoids known problems arising from data leakage
- Modest preprocessing: remove articles, punctuation, capitalization, word normalization
- Tokenization: 4-grams
- Feature selection: Term Frequency-Inverse Document Frequency
- Optimization algorithm: neural network
  - Multi-layer perceptron model with 1 hidden layer and 100 neurons

Feature extraction via n-grams

CJARS Criminal Justice Administrative Records System

(a) Contiguous sequence of 4 characters

theft from person > 65 value >300k<10k

t**heft** from person > 65 value >300k<10k

th**eft** from person > 65 value >300k<10k

...

theft from person > 65 value >30**0k<1**0k

theft from person > 65 value >3**00k<10**k

theft from person > 65 value >30**0k<10k**

(b) Extracted features

'thef', 'heft', 'eft ', 'ft f', 't fr', ' fro', 'from', 'rom ', 'om p', 'm pe', ' per', 'pers', 'erso', 'rson', 'son>', 'on>6', 'n>65', '>65 ', '65 v', '5 va', ' val', 'valu', 'alue', 'lue>', 'ue>3', 'e>30', '>300', '300<', '00<1', '0<10', '<10k', ...

Defining success

CJARS Criminal Justice Administrative Records System

We evaluate three related standard performance statistics:

- Precision: % of positive predictions that are correctly labelled
- Recall: % of labels that are correctly made
- F1 Score: harmonic mean of precision and recall

Model performance is assessed at both the broad crime type and full UCCS code levels

Performance Statistics

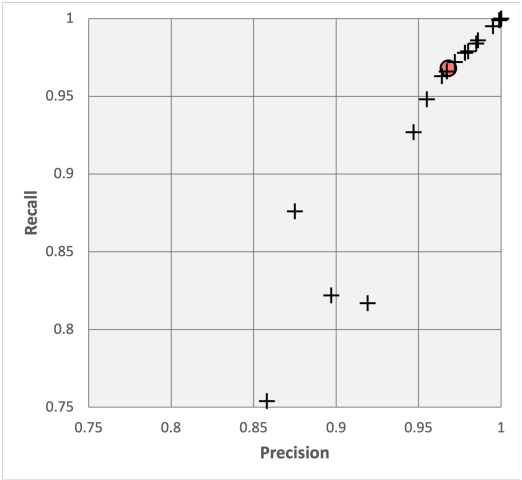
CJARS Criminal Justice Administrative Records System

	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

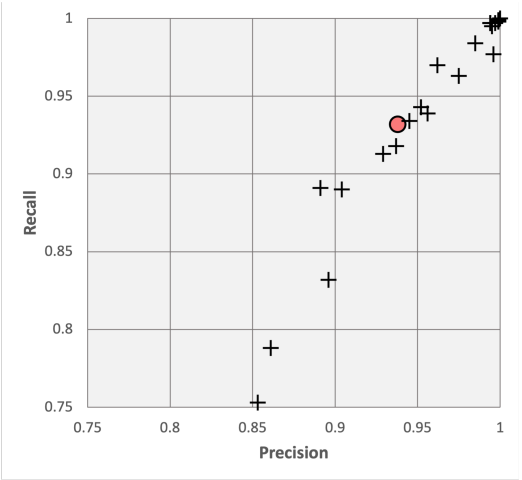
Out-of-state performance

CJARS Criminal Justice Administrative Records System

Broad crime type



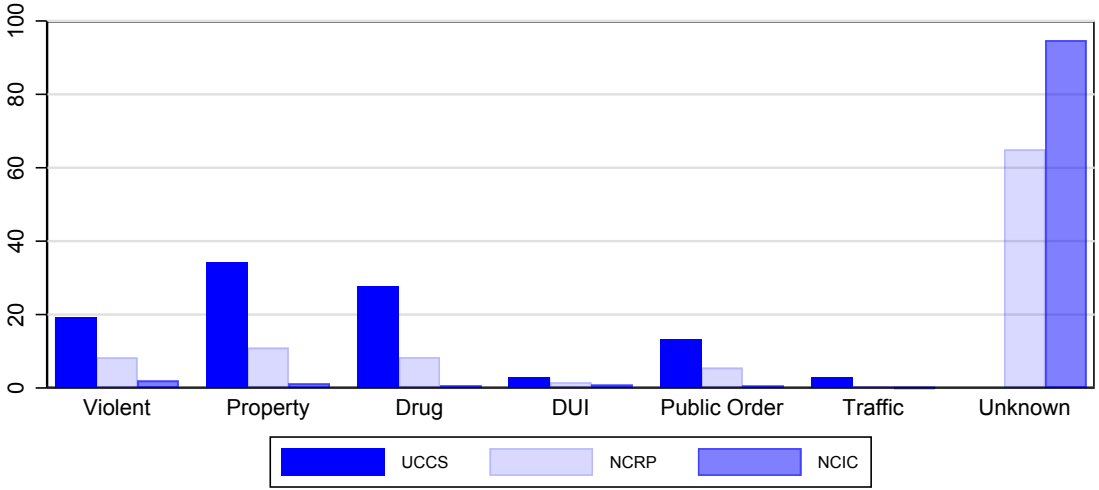
Full crime type



Comparison with existing resources

CJARS Criminal Justice Administrative Records System

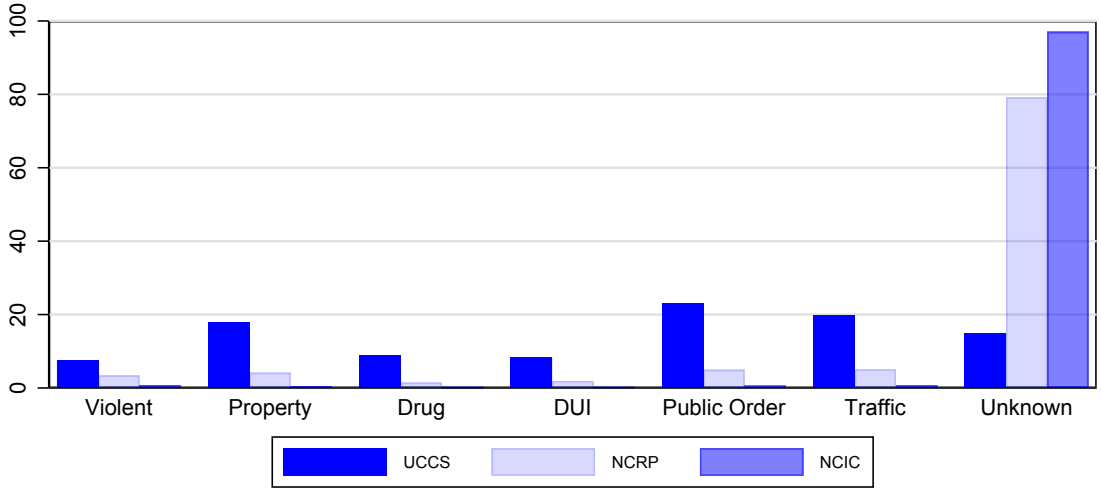
Distribution of offense types for felony charges in CJARS



Comparison with existing resources

CJARS Criminal Justice Administrative Records System

Distribution of offense types for misdemeanor charges in CJARS



Evaluating optimal model parameterization

CJARS Criminal Justice Administrative Records System

- Variation in size of training data: 5,000 – 200,000 training observations link
- Feature unit (n-grams, bag-of-words) and selection mechanism (Count Vectorizer vs. TF-IDF) link
- Total number of features: 100 – 10,000 selected features link
- Machine learning algorithm: Random Forest vs. Neural Network link

Next steps:

CJARS Criminal Justice Administrative Records System

1. Expand geographic coverage of training data to non-covered jurisdictions
2. Crowdsource prediction errors to improve corpus of training records via TOC portal
3. Incorporate feasibility to process cited statute numbers
   - Requires comprehensive database mapping statutes to offense descriptions
4. Develop mechanism to evaluate updates to schema

Summary and conclusion

CJARS Criminal Justice Administrative Records System

This project introduces the UCCS schema and the TOC tool to map free entry offense descriptions to standardized codes

Goal is to raise the bar for research and statistical reporting:

- Reduce researcher discretion as well as data wrangling burden
- Improve common definitions and comparability across jurisdictions
- Expand coverage of historically understudied (but large!) portion of the justice system: misdemeanor caseload
- Lower research barriers to increase diverse perspectives on justice system

Join our growing user base!

CJARS Criminal Justice Administrative Records System

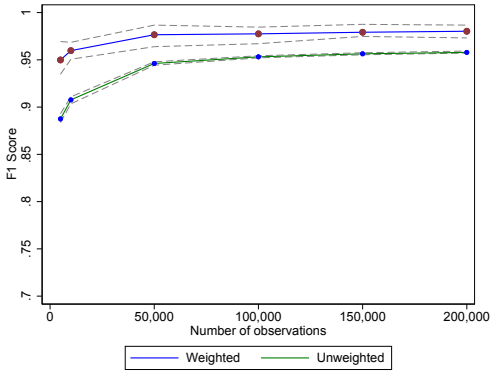
CJARS Criminal Justice Administrative Records System

Appendix Slides

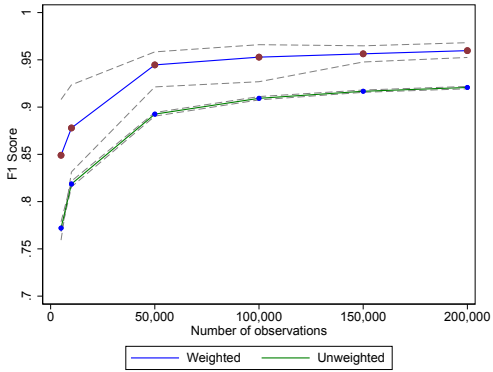
Variation in size of training data

CJARS Criminal Justice Administrative Records System

(a) Broad Crime Type



(b) Full UCCS Code

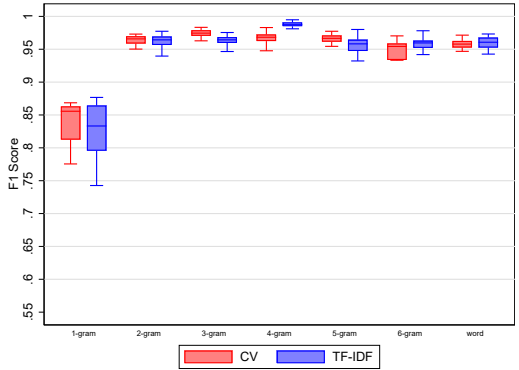


back

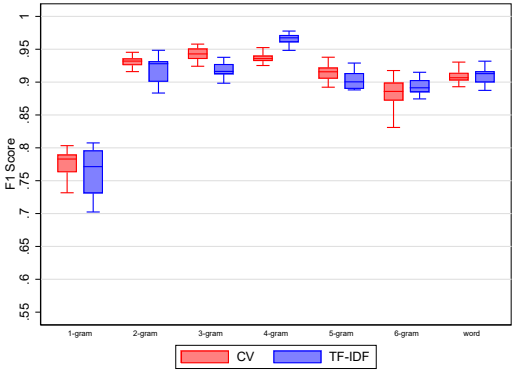
Feature unit selection mechanism

CJARS Criminal Justice Administrative Records System

(a) Broad Crime Type



(b) Full UCCS Code

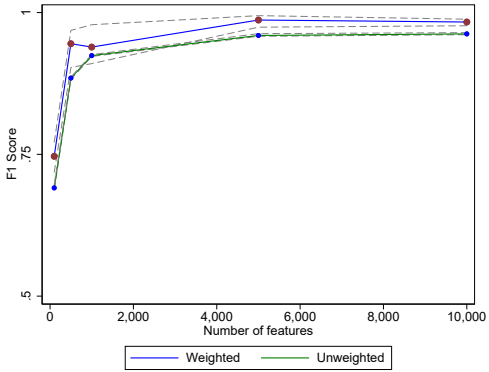


back

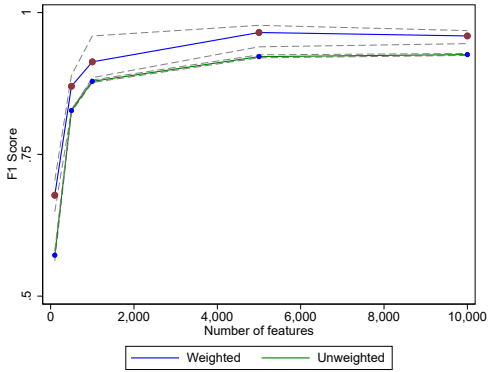
Total number of features

CJARS Criminal Justice Administrative Records System

(a) Broad Crime Type



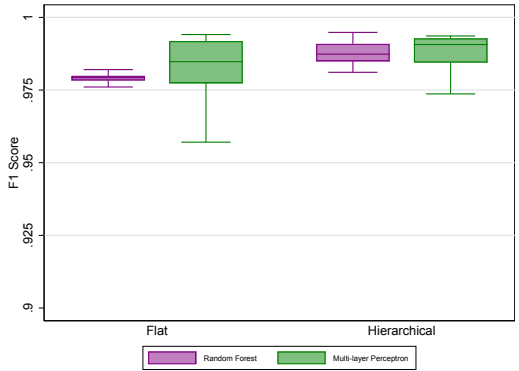
(b) Full UCCS Code



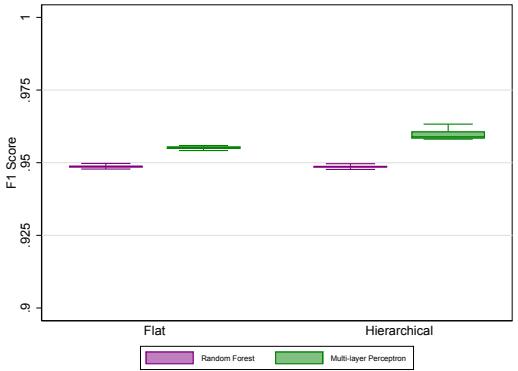
back

Machine learning algorithm

(a) Broad Crime Type



(b) Full UCCS Code



back