NSM 2022
NORDIC STATISTICAL MEETING

Using machine learning to classify theft offences to International Classification of Crime

Kimmo Haapakangas, Statistics Finland, kimmo.haapakangas@stat.fi

Abstract

The use of International Classification of Crime for Statistical Purposes (ICCS) is gradually increasing. However, mapping an existing national classification to correspond ICCS classes has its own challenges. National classifications are often based on national criminal law and its definitions, whereas ICCS is based on internationally agreed definitions. National laws might not always correspond to ICCS definitions.

Producing statistics according to ICCS would often require someone to read the definition of each reported offence or given sentence. Use of Natural Language Processing (NLP) and Machine Learning (ML) techniques might provide a solution to this problem.

Statistics Finland has gained access to prosecutor's text descriptions of offence for the years 2019 and 2020. There are around 24.000 descriptions for theft, aggravated theft, and petty theft in the data. 2.000 of these are manually read and classified into ICCS 0501 and 0502 subcategories. Based on these observations, a Random Forrest model is trained to classify these texts. Overall accuracy of the model is 78.9 per cent. The model predicts theft from a shop (ICCS 050231) and theft of personal property from a person (ICCS 050221) very well with sensitivity and specificity of over 93 per cent.

Due to heavy class imbalance, some infrequent classes, like theft of public property, have only 15 per cent sensitivity. However, specificity is over 95 per cent.

Once trained, machine learning models are useful tools for mapping ICCS, and these models can be reused easily. For example, the distribution of theft offences in 2020 in ICCS is similar to the 2019 distribution. As a drawback, training a NLP model can take some time and running it will require computational power, but in this case the investment is well worth it.

Keywords: International Classification of Crimes, machine learning, text classification

1. Introduction

Many countries have their own crime classification that is based on local law. The International Classification of Crime for Statistical Purposes was introduced to harmonize data collection and to ease international comparison. While some offences are easy to map to the ICCS there are some problematic ones.

Using machine learning and natural language processing might prove useful when classifying national offences to ICCS. This study focuses on theft offences in Finland, but this method is likely to be useful when classifying other offences as well.

1.1 International Classification of Crimes for Statistical Purposes

The International Classification of Crime for Statistical Purposes (ICCS) is a classification of criminal offences which is based on internationally agreed concepts, definitions and principles in order to enhance the consistency and international comparability of crime statistics and improve analytical capabilities at both the national and international levels *(International Classification of Crime for Statistical Purposes, 2015*).

The ICCS is a new classification. It was endorsed by the United Nations Statistical Commission at its 46th session in March 2015, and the Commission on Crime Prevention and Criminal Justice at its 24th session in May 2015 as an international statistical standard for data collection. United Nations Office on Drugs and Crime (UNODC) was confirmed as the custodian of the ICCS.

The ICCS provides a framework for the systematic production and comparison of statistical data across different criminal justice institutions and jurisdictions. This means that the ICCS is applicable to all forms of crime data, whatever the stage of the criminal justice process (police, prosecution, conviction, imprisonment) at which they are collected, as well as to data collected in crime victimization surveys.

At international level, the ICCS improves the comparability of crime data between countries. Standardized concepts and definitions allow for the systematic collection, analysis, and dissemination of data, and also respond to the demand for in-depth research and analysis of transnational crime.

Several criteria have been used to build the hierarchical structure of the ICCS, in the attempt to build categories that can respond to a variety of information needs. In particular, the following criteria have been used to form categories of the ICCS:

- policy area of the act/event (protection of property rights, protection of health, etc.)
- target of the act/event (e.g. person, object, natural environment, state, etc.)
- seriousness of the act/event (e.g. acts leading to death, acts causing harm, etc.)
- means by which the act/event is perpetrated (e.g. by violence, threat of violence, etc.).

Based on these criteria, 19 criminal offences can be grouped in homogenous categories, which are aggregated at four different hierarchical levels: Levels 1, 2, 3 and 4. There are 11 Level 1 categories designed to cover all acts or events that constitute a crime within the scope of the ICCS. Criminal offences at Levels 2, 3 and 4 can be summed to provide observations at more aggregated levels, while observations at higher levels can be subdivided into lower-level categories. The numerical coding of the categories is in accordance with their level in the classification: Level 1 categories are the broadest categories and have a two-digit code (e.g. 01); Level 2 categories have a four-digit code (e.g. 0101); Level 3 categories have a five-digit code (e.g. 01011); and Level 4 categories, the most detailed level, have a six-digit code (e.g. 010111).

This study will focus on category 5. Act against property only, and more specific subcategories 0501 Burglary and 0502 Theft.

Tabel 1 Level 1 ICCS categories

1 Acts leading to death or intending to cause death
2 Acts leading to harm or intending to cause harm to the person
3 Injurious acts of a sexual nature
4 Acts against property involving violence or threat against a person
5 Acts against property only
6 Acts involving controlled psychoactive substances or other drugs
7 Acts involving fraud, deception or corruption
8 Acts against public order, authority and provisions of the State
9 Acts against public safety and state security
10 Acts against the natural environment
11 Other criminal acts not elsewhere classified

1.2 Crime classification in Finland

The crime classification currently used in Finland and at Statistics Finland is based on sections and paragraphs of criminal code. Like theft, petty theft and attempted aggravated theft etc. This is the situation in many other countries also.

When the Finnish Parliament imposes/decree a new law, expert at Statistics Finland reads it at finlex.fi web site and decides whether a new code is required or an old code is to be abolished.

Some of the national offences, like assault, can be mapped to correspond the ICCS categories quite easily.

Unfortunately, burglary or shoplifting for example are not specifically described as punishable acts. Burglary is mentioned as subsection under aggravated theft, but classification codes include only to section not subsection. Theft is defined in law (769/1990) as follows.

Theft

1. A person who appropriates movable property from the possession of another shall be sentenced for theft to a fine or to imprisonment for at most one year and six months.

2. An attempt is punishable.

Aggravated theft

1. If in the theft
    1. the object of the appropriation is very valuable,...
    2. the offender breaks into an occupied residence...

2. An attempt is punishable. *(The Criminal Code of Finland, 2016)*

Currently Finland can provide very limited data on theft offences in ICCS classification. This is shown in table 2. National offences classified as theft are almost all mapped to highest ICCS theft level (0502). And there are no offences mapped to ICCS 0501 burglary.

Tabel 2 Persons sentenced by principal offence rule / found quilty for theft offences in ICCS classification 2017

ICCS offence	Number of persons sentenced by principal offence rule in 2017
0501 Burglary	-
05011 Burglary of business premises	-
05012 Burglary of private residential premises	-
050121 Burglary of permanent private residences	-
050122 Burglary of non-permanent private residences	-
05013 Burglary of public premises	-
05019 Other acts of burglary	-
0502 Theft	7574
05021 Theft of a motorized vehicle or parts thereof	343
050211 Theft of a motorized land vehicle	-
050212 Illegal use of a motorized land vehicle	87
050213 Theft of parts of a motorized land vehicle	-
050219 Other theft of a motorized vehicle or parts thereof	-
05022 Theft of personal property	-
050221 Theft of personal property from a person	-
050222 Theft of personal property from a vehicle	-
050229 Other theft of personal property	-
05023 Theft of business property	-
050231 Theft from a shop	-
050239 Other theft of business property	-
05024 Theft of public property	-
05025 Theft of livestock	-
05026 Theft of services	-
05029 Other acts of theft	17

Offences known to the authorities can be mapped to a more detailed level of ICCS, since police uses so-called “specificators” which allow to distinct theft from burglary to private residence etc.

1.3 Prosecutor’s text descriptions

For some years Statistics Finland has had access to prosecutor’s text description of the criminal act charged. These are not the conclusion of judgements/sentences, but rather shorter text descriptions of what has happened by the prosecutor’s point of view. Length of these text varies from one sentence to several dozens of sentences.

2. About the study

Since it is not possible to fill all required ICCS categories about the persons prosecuted and sentenced with originally available data, the idea about using text descriptions was introduced.

Changing national crime classification is not an easy process and changing criminal matters application (RITU), software used by the district courts to record issued decisions is not cost effective. The use of already available but not yet used data might be better solution.

2.1 Manual classification and pre-processing

Prosecutor’s text descriptions (later text data) include text in Finnish and Swedish. For this study, only Finnish texts for theft offences (theft, petty theft, aggravated theft, Chapter 28 sections 1-3) were chosen. There were only a few texts in Swedish. There were total of 23.600 texts.

Of these texts, a weighted random sample of 2.000 texts were chosen for manual classification. These texts were read and manually classified to ICCS categories 0501, 0502 and their sub-classes.

This sample included only few observations from ICCS 050130 Burglary of public premises and 050260 Theft of services for example. Number of these observations in these classes in sample data was increased by using some keyword searches from the whole data.

Final manually classified data had 2.100 observations. This was divided to train and test data with 75/25 proportions. The ICCS category 05025 Theft of livestock was not included in the analysis, since there were no observations in sample data, and it is not very relevant offence in Finland.

Below is an example text of one offence translated to English and original Finnish text. Names are replaced with letters.

“Person A has stolen property of plaintiff’s B and C by force entry to their dwelling from open window located in address X and taking an iPad tablet from there. The stolen property has been returned to the owners”.

”A on anastanut asianomistajien B ja C omaisuutta tunkeutumalla heidän osoitteessa X sijaitsevaan asuntoon avoimesta ikkunasta ja ottamalla sieltä iPad-tabletin. Anastettu omaisuus on palautettu asianomistajalle.”

Tabel 3 Number of observation in train and test sets

ICCS class	Train (n)	Test (n)
05011 Burglary of business premises	148	49
050121 Burglary of permanent private residences	123	40
050122 Burglary of non-permanent private residences	56	18
05013 Burglary of public premises	13	4
05019 Other acts of burglary	22	7
050211 Theft of a motorized land vehicle	22	7
050213 Theft of parts of a motorized land vehicle	22	7
050219 Other theft of a motorized vehicle or parts thereof	20	6
050221 Theft of personal property from a person	90	29
050222 Theft of personal property from a vehicle	72	24
050229 Other theft of personal property	235	78
050231 Theft from a shop	513	170
050239 Other theft of business property	200	66
05024 Theft of public property	22	7
05026 Theft of services	10	3
05029 Other acts of theft	15	5
Total	1583	520

2.2 Tools and data pre-processing

R-programming language and Rstudio (workstation) were chosen as main tools for text analysis. Text data was lemmatized, words were taken to their basic form, with Udpipe-package. Pronouns and conjunctions were removed *(Udpipe, 2022).*

Package TM was used to build text-corpus and document-term-matrix. Terms were also cleaned by removing punctuation, numbers and changing to lower case letters. Also, Finnish stop-words were removed *(TM package, 2022).*

Terms were weighted with term frequency-inverse document frequency function (tf-idf), so that more rare terms got bigger weight *(Tidytextmining, 2022)*.

Machine learning model was done with Caret-package and random forest algorithm. Model parameters were searched with cross-validation (5 times). Despite the weighted sampling, the training data had high class imbalance issues, so up-sampling/oversampling was used to correct that. Up-sampling produced better results than down-sampling/undersampling.

Also, linear support vector machine (SVM) and Naïve Bayes algorithms were tested, but the results were not as good as with random forest model. SVM gave poorer results especially with more rare classes.

3. Results

3.1 Model results

Final random forest model had total accuracy of 78.9 per cent, which is quite good for an NLP model, especially when there were 16 classes in ML problem. Some classes had over 90 per cent sensitivity/recall and specificity values.

- Sensitivity/recall describes how many predictive positive classes were predicted correctly. Sensitivity/recall is how certain we are that we are not missing any positives. This is good measure when we want to rather have some extra false positives than saving some false negatives *(https://towardsdatascience.com/decoding-the-confusion-matrix-bb4801decbb, 2022).*
- Specificity determines the proportion of actual negatives that are correctly identified. Specificity is chosen when we want to cover all true negatives. For example, driving while intoxicated we don’t want anyone drug-free going to jail.

- Precision tells how often a yes prediction was correct. It should be used when one wants to be confident about true positives. In this case we want to be sure that those classified as burglary to private residence are correct.

Classes with larger number of observations were better recognized. For example, shoplifting (ICCS 05231) had almost 98 per cent sensitivity/recall and precision of 97 per cent. On one hand, out of 170 observations in this class the model predicted 166 correctly, but on the other hand the model predicted 6 observations not in this class to belong to it (false positives).

Some offences and their descriptions are complex and can cover more than one offence. Like the following example in Finnish. Two persons have stolen property from business premises and from a van. They have also stolen private property from those premises. This has been manually classified as theft of business property since first victim is the company.

”A ja B ovat yhdessä anastaneet yritys C:n, (yksityishenkilöiden) D:n, E: ja F:n omaisuutta X sijaitsevista C:n tiloista ja pakettiautosta syytekohdassa 1 kuvatun teon yhteydessä. A ja B ovat anastaneet C:n omistamia autonavaimia, työkaluja, elektroniikkaa ja alkoholijuomia, D: omistamia vaatteita ja kassin, E:n omistamia vaatteita, makuupussin, alkoholijuomia ja kassin sekä E: omistamia kalastusvälineitä, vaatteita, elektroniikkaa, retkeilyvarusteita, auton varaosia ja autonavaimet.”

Tabel 4 Model metrics from Random Forest model

Class	Sensitivity/ Recall	Specificity	Precision	Balanced Accuracy
05011 Burglary of business premises	0.897	0.963	0.721	0.93
050121 Burglary of permanent private residences	0.85	0.991	0.894	0.92
050122 Burglary of non-permanent private residences	0.666	1	1	0.833
05013 Burglary of public premises	0.25	0.992	0.2	0.621
05019 Other acts of burglary	0.142	1	1	0.571
050211 Theft of a motorized land vehicle	0.857	0.998	0.857	0.927
050213 Theft of parts of a motorized land vehicle	0.428	1	1	0.714
050219 Other theft of a motorized vehicle or parts thereof	0.666	0.996	0.666	0.831
050221 Theft of personal property from a person	0.931	0.965	0.613	0.948
050222 Theft of personal property from a vehicle	0.708	0.971	0.548	0.84
050229 Other theft of personal property	0.743	0.923	0.63	0.833
050231 Theft from a shop	0.976	0.982	0.965	0.979
050239 Other theft of business property	0.484	0.982	0.8	0.733
05024 Theft of public property	0.142	0.996	0.333	0.569
05026 Theft of services	0.333	0.998	0.5	0.665
05029 Other acts of theft	0.6	1	1	0.8

However, some classes, like other acts of burglary (ICCS 050190) had only 14.2 per cent sensitivity/recall. A lot of positives were missed for this class. Nonetheless, it had 100 per cent precision, so all positive predictions were correct. This class had only 7 observations in test data. Those falsely classified observation were classified as burglary to business premises. At least model recognized those as burglaries.

ICCS class 050130 burglary of public premises had sensitivity of 25 per cent and precision of 20 per cent. So, we would have better chances by tossing a coin. This class had only 13 observations in train data and 4 in test data.

3.2 Results in original data

While R was used in Machine Learning, the SAS software is used in analysing the results and combining classification to original data. SAS is widely used at Statistics Finland and has server environment.

Besides those theft offences classified with ML model also some other offences were included to get better completeness level of ICCS 0501 and 0502 categories.

These included offences were unauthorized use (28:7-9) and stealing of a motor vehicle for temporary use (28:9a-9c). These offences were classified to correct ICCS categories (050211, 050212, 050290 and 050219) by simple keyword search.

In year 2019 there were total of 23.700 raised charges from theft, unauthorized use or stealing of motor vehicle offences. When Swedish text were dropped the final data had 23.400 charges. Only 8.800 of these were the principal offence of charges.

There were total of 2.200 burglaries in year 2019. And as principal offences total of 800 burglaries. This is something that we didn’t know before. In ICCS forced entry into fenced area or theft from basement lock-up are not burglaries but theft offences instead.

Tabel 5 Imputable offences in court and persons senteced by principal offence 2019

	Offences total	Principal offence
Total	23380	8790
0501 Burglary	2170	810
05011 Burglary of business premises	1150	430
05012 Burglary of private residential premises	920	340
050121 Burglary of permanent private residences	750	310
050122 Burglary of non-permanent private residences	160	30
05013 Burglary of public premises	70	20
05019 Other acts of burglary	30	10
0502 Theft	21220	7980
05021 Theft of a motorized vehicle or parts thereof	1350	410
050211 Theft of a motorized land vehicle	980	270
050212 Illegal use of a motorized land vehicle	180	70
050213 Theft of parts of a motorized land vehicle	60	20
050219 Other theft of a motorized vehicle or parts thereof	130	50
05022 Theft of personal property	3610	1120
050221 Theft of personal property from a person	1080	290
050222 Theft of personal property from a vehicle	690	140
050229 Other theft of personal property	1850	690
05023 Theft of business property	16110	6390
050231 Theft from a shop	13810	5630
050239 Other theft of business property	2300	760
05024 Theft of public property	60	30
05026 Theft of services	20	20
05029 Other acts of theft	70	20

One interesting finding was that burglaries are committed by younger persons: of convicted 15- to 20-year-olds, around 14 per cent were found guilty of burglary, whereas those aged 40 and over only 8 per cent were found guilty of burglaries.

Less than 45 per cent of those under 21 of age were found guilty of shoplifting, while the share was almost 70 per cent for those aged 40 and over.

Although the model doesn’t give a 100 per cent accuracy, it gives some new insight to data and opens new possibilities and future interests.

About 10 per cent of those sentenced from burglary of private residential premises (05012) were given a fine instead of a prison sentence. This might be because in ICCS access by deception with intend to commit theft is classified as burglary.

The high proportion of fines sentenced by courts of first instance can be explained by the fact that nearly 60 per cent of theft offences were petty thefts and the only sentence from petty theft is a fine.

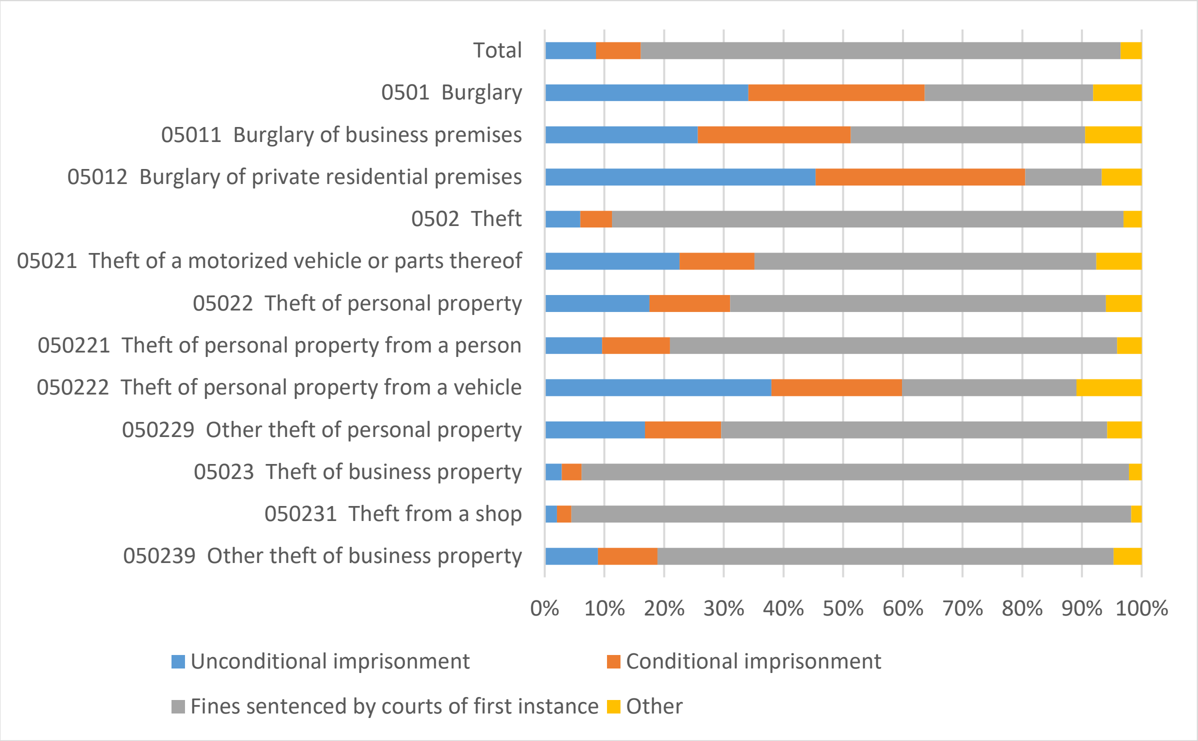

Figure 1 Sentences by principal offence 2019

Previously Statistics Finland was unable to provide the kind of data shown in figure 1. but with the use of machine learning technics, it is possible without manually reading all 23.000 texts every year.

Sentence here refers to the sentence from all offences in a particular court case and it is affected by total number of offences, including minor offences. Quite many of those sentenced being guilty of theft offences had committed more than one theft or other offences. Many had also committed minor narcotic offences.

Below is an example where 4 persons have first broken car's window and then ignition and driven the car under a bridge. There they have stolen petrol from the car. The car was found but the petrol not. There was also property damages. In this case there were also reports from damage of property and illegal use of motorized land vehicle.

" A, B, C ja D ovat yhdessä anastaneet E omaisuutta ottamalla ensin luvattomasti käyttöönsä osoitteessa Y pysäköitynä olleen E:n omistaman henkilöauton tunkeutumalla ajoneuvoon ikkunalasin rikkomalla ja rikkomalla virtalukon ajoneuvon käynnistämiseksi. Tekijät ovat kuljettaneet ajoneuvon Z sillan alle, jossa he ovat anastaneet ajoneuvosta polttoaineen. Anastettu omaisuus on jäänyt kateisiin ja teosta on aiheutunut murtovahinkoja. A on ollut teon tehdessään alle 18-vuotias."

4. Discussion

This model was done with bag-of-words principle, so only single words were used, and word order had no meaning. In the future, it would be interesting to use n-grams, (bi-grams, trigrams etc). This would increase the size of document-term-matrix and so more computational power would be required. Now only a basic office laptop was used, and it took more than 5 hours to find best parameters for Random Forest model.

Also, larger training data would be useful and probably would give better results since data has some class imbalance issues.

As mentioned before, NLP-models tend to be large and computationally heavy. A server environment would be beneficial in this case. However, special caution must be kept since this kind of data will have persons names, addresses and other GDPR-regulated data. This is especially the case with violent and sexual offences.

Use of servers or more efficient computer would also allow the use of deep learning algorithm like FinBert or Google Bert trained in Finnish.

This study shows that the use of an open-source software can yield good results and the building of ML models doesn't have to be expensive. Use of ML can also give some new insight from the data and thus help to produce better statistics, at least in the short run. In the long run it would of course be better to change the original classification and data recording system.

However, in shorter run, this method could be very useful to classify offences to the ICCS categories. This study was done to theft offences, but there are other interesting categories in the ICCS, for example robbery (04), property damage (0504) or acts against the natural environment (10).

5. References

Criminal code of Finland, Retrieved 10.5.2022, (https://www.finlex.fi/fi/laki/kaannokset/1889/en18890039_20150766.pdf

International Classification of Crime for Statistical Purposes, version 1.0 (2015) https://www.unodc.org/unodc/en/data-and-analysis/statistics/iccs.html

Tidytextminig, retrieved 10.5.2022, https://www.tidytextmining.com/tfidf.html#tfidf

TM package, retrieved 10.5.2022,

https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf 9.5.2022

TowardsDataScience, retrieved 10.5.2022, https://towardsdatascience.com/decoding-the-confusion-matrix-bb4801decbb

UDPIPE, retrieved 10.5.2022,

https://cran.r-project.org/web/packages/udpipe/index.html