Anonymization and anonymized text data in statistical production

Matti Kokkonen, Katja Löytynoja and Henna Ylimaa

Nordic Statistical Meeting 2022, Iceland

22.8.2022-24.8.2022

Data anonymization

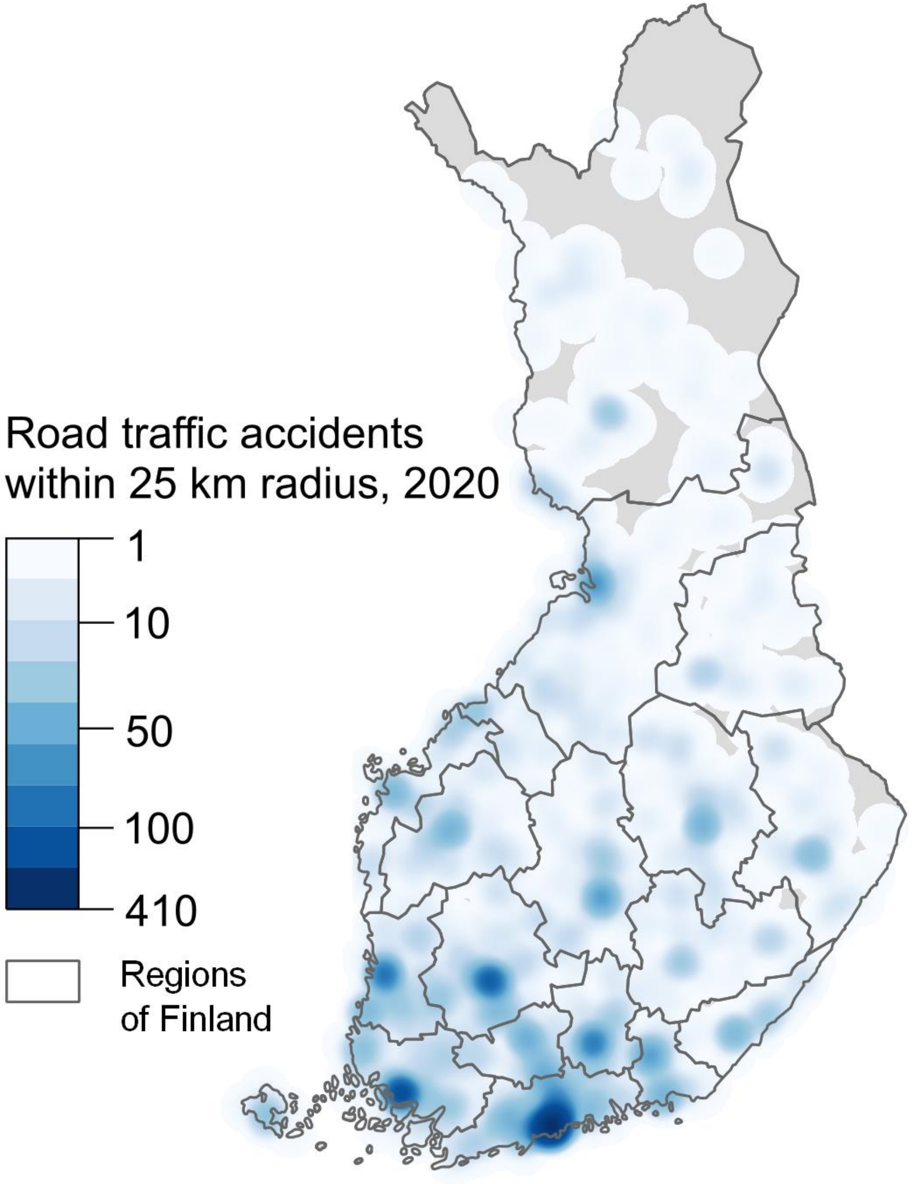
- According to article 5 GDPR: Principles relating to processing of personal data
  - Personal data shall be: (c) adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed (‘data minimisation’)
- We wanted to see if the Statistics on road traffic accidents would be possible to produce with the anonymized text data
  - Main data source for the statistics is the accident data from police including the written accident reports
  - Currently tabular accident data is supplemented from the text data
  - The text is read and interpret by the handler

Text data anonymization

- We produced a prototype tool for text data anonymization
  - Developed primary for simulation purposes
- Our goal was to
  - Simulate the production process with the anonymized text data
  - See if the Statistics on road traffic accidents would be possible to produce with anonymized text data
  - Study the impact of anonymization to the statistics and the production process
- We focused purely on anonymizing names within the text
  - Personal codes and other identifiers in a specified form are more trivial to anonymize

Statistics on road traffic accidents in Finland

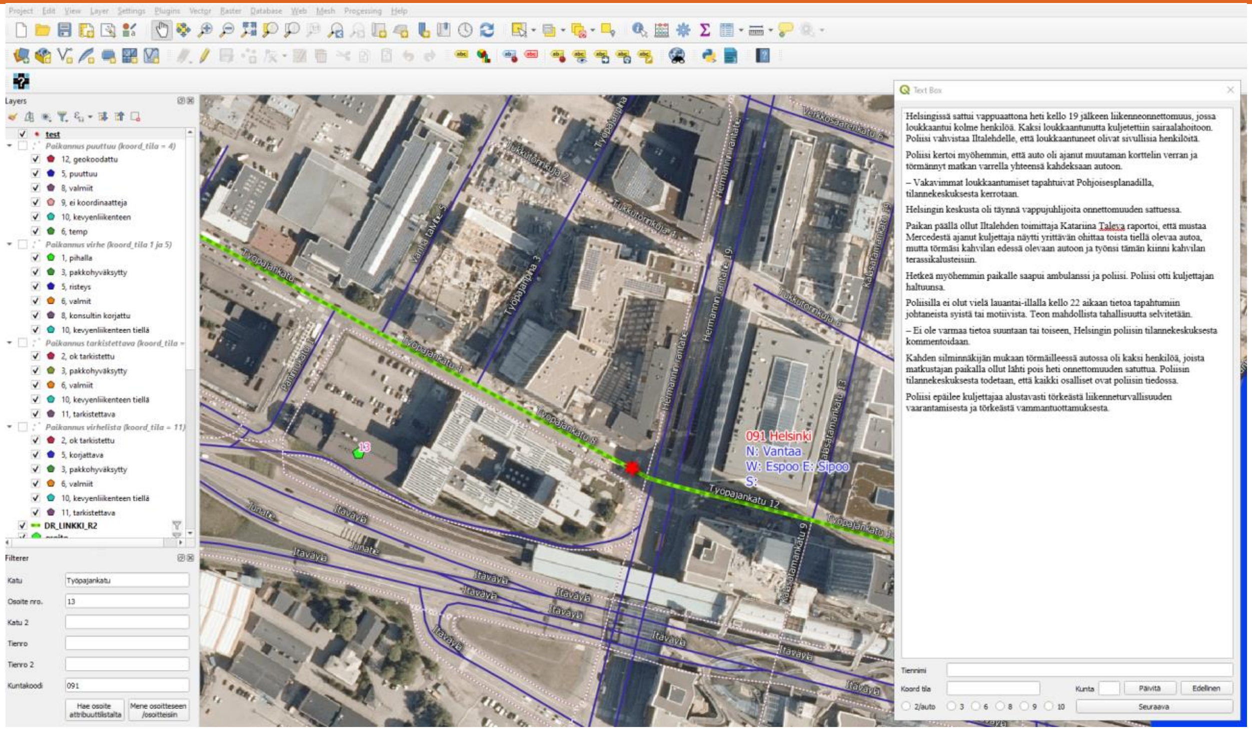
- Statistics Finland produces the official road traffic accident statistics in Finland
- The statistics contains
  - Accidents that have led to personal injuries
  - Number of deaths and injuries
  - Comprehensive information of the people and vehicles that have been involved in the accidents
  - The references to the individuals and links and relationships between the individuals and vehicles are important to maintain in the anonymized text data



Accident report usage

- In the production process the written accident report is used in selected cases for:

1. Geopositioning:
   - GIS-application (QGIS) case by case
   - Handler benefits from all the geospatial information (addresses, business names, known places, and landmarks
2. Supplementing and correcting tabular data on accidents and participants
   - Interpreting the text data and transforming the information into tabular form
   - Handler has to be able to identify the vehicles and individuals and place the individuals in the correct vehicle
3. Text data is also examined with selected keywords
   - Necessary keywords should not be anonymized



Filterer

Katu: Työpajankatu

Osoite nro.: 13

Katu 2:

Tienro:

Tienro 2:

Kuntakoodi: 091

Hae osoite attribuuttilistalta | Mene osoitteseen /osoitteisiin

Helsingissä sattui vappuaattona heti kello 19 jälkeen liikenneonnettomuus, jossa loukkaantui kolme henkilöä. Kaksi loukkaantunutta kuljetettiin sairaalahoitoon. Poliisi vahvistaa Iltalehdelle, että loukkaantuneet olivat sivullisia henkilöitä.

Poliisi kertoi myöhemmin, että auto oli ajanut muutaman korttelin verran ja törmännyt matkan varrella yhteensä kahdeksaan autoon.

– Vakavimmat loukkaantumiset tapahtuivat Pohjoisesplanadilla, tilannekeskuksesta kerrotaan.

Helsingin keskusta oli täynnä vappujuhlijoita onnettomuuden sattuessa.

Paikan päällä ollut Iltalehden toimittaja Katariina Taleva raportoi, että mustaa Mercedestä ajanut kuljettaja näytti yrittävän ohittaa toista tiellä olevaa autoa, mutta törmäsi kahvilan edessä olevaan autoon ja työnsi tämän kiinni kahvilan terassikalusteisiin.

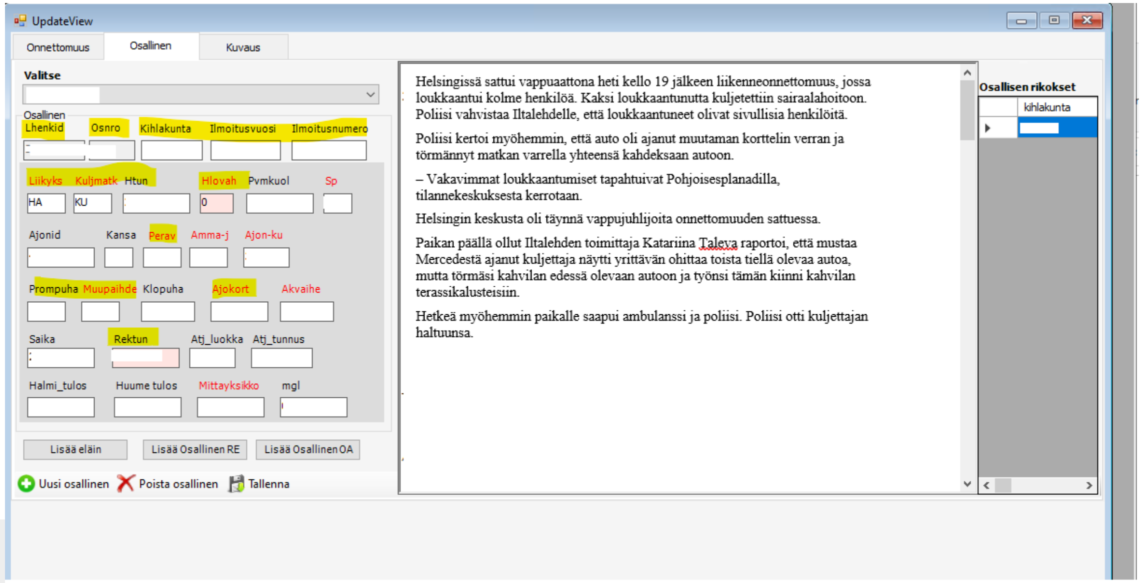
Hetkeä myöhemmin paikalle saapui ambulanssi ja poliisi. Poliisi otti kuljettajan haltuunsa.

Poliisilla ei olut vielä lauantai-illalla kello 22 aikaan tietoa tapahtumiin johtaneista syistä tai motiivista. Teon mahdollista tahallisuutta selvitetään.

– Ei ole varmaa tietoa suuntaan tai toiseen, Helsingin poliisin tilannekeskuksesta kommentoidaan.

Kahden silminnäkijän mukaan törmäilleessä autossa oli kaksi henkilöä, joista matkustajan paikalla ollut lähti pois heti onnettomuuden satuttua. Poliisin tilannekeskuksesta todetaan, että kaikki osalliset ovat poliisin tiedossa.

Poliisi epäilee kuljettajaa alustavasti törkeästä liikenneturvallisuuden vaarantamisesta ja törkeästä vammantuottamuksesta.



Osallinen

Lhenkid: | Osnro: | Kihlakunta: | Ilmoitusvuosi: | Ilmoitusnumero:

Liikyks: HA | Kuljmatk: KU | Htun: | Hlovah: 0 | Pvmkuol: | Sp:

Ajonid: | Kansa: | Perav: | Amma-j: | Ajon-ku:

Prompuha: | Muupaihde: | Klopuha: | Ajokort: | Akvaihe:

Saika: | Rektun: | Atj_luokka: | Atj_tunnus:

Halmi_tulos: | Huume tulos: | Mittayksikko: | mgl:

Helsingissä sattui vappuaattona heti kello 19 jälkeen liikenneonnettomuus, jossa loukkaantui kolme henkilöä. Kaksi loukkaantunutta kuljetettiin sairaalahoitoon. Poliisi vahvistaa Iltalehdelle, että loukkaantuneet olivat sivullisia henkilöitä.

Poliisi kertoi myöhemmin, että auto oli ajanut muutaman korttelin verran ja törmännyt matkan varrella yhteensä kahdeksaan autoon.

– Vakavimmat loukkaantumiset tapahtuivat Pohjoisesplanadilla, tilannekeskuksesta kerrotaan.

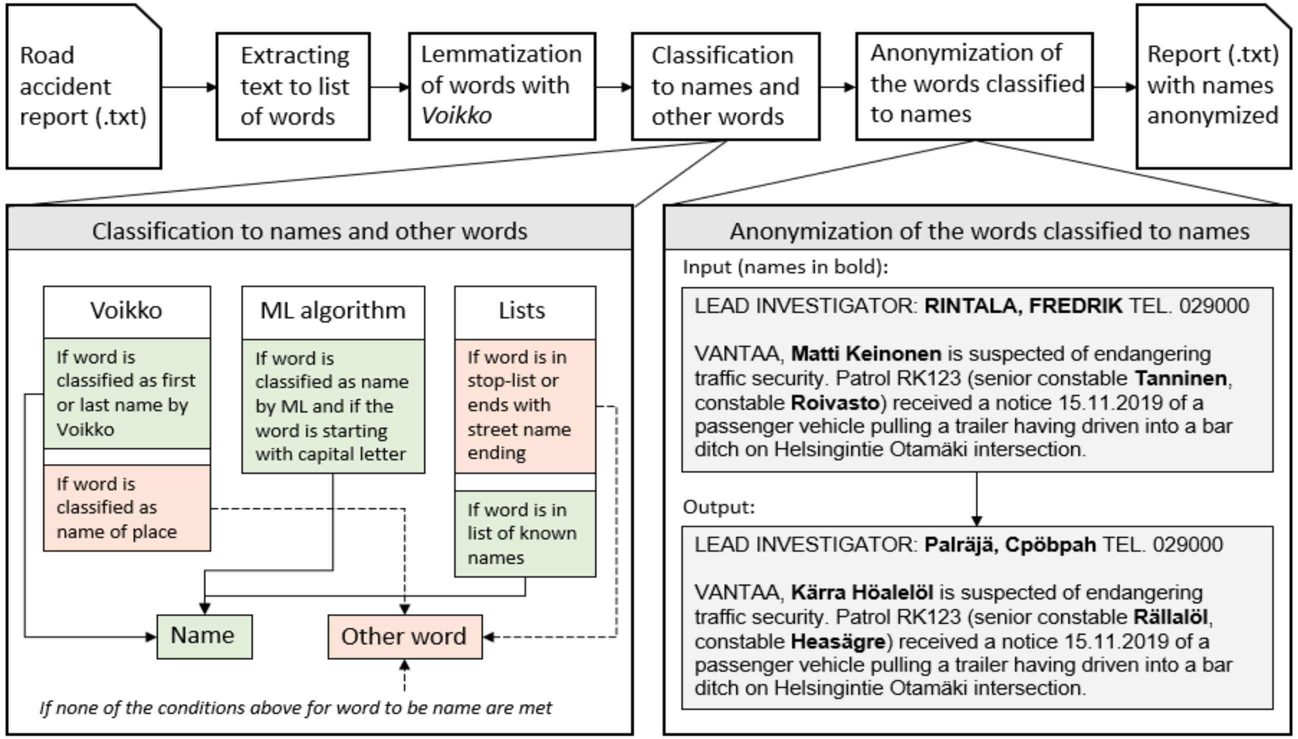
Helsingin keskusta oli täynnä vappujuhlijoita onnettomuuden sattuessa.

Paikan päällä ollut Iltalehden toimittaja Katariina Taleva raportoi, että mustaa Mercedestä ajanut kuljettaja näytti yrittävän ohittaa toista tiellä olevaa autoa, mutta törmäsi kahvilan edessä olevaan autoon ja työnsi tämän kiinni kahvilan terassikalusteisiin.

Hetkeä myöhemmin paikalle saapui ambulanssi ja poliisi. Poliisi otti kuljettajan haltuunsa.

NameFinder –tool for anonymization

- For our simulation purposes we developed an anonymization tool that:
  - Classifies individual words into two categories: Names and other words → does not consider the context
  - Anonymizes (pseudomizes) the names
- NameFinder -tool is in its final form a four-step program
  1. NLP-tool “Voikko” for classification and lemmatization
  2. Classification machine learning algorithm
  3. Stop-list for words that are not names in this context
  4. Acceptance-list for words that are considered names
- Anonymizes (pseudonymizes) names by changing individual characters to random character (same within the document)
  - Thus, “automatically” maintaining the connection between references to an individual


Road accident report (.txt)
Extracting text to list of words
Lemmatization of words with *Voikko*
Classification to names and other words
Anonymization of the words classified to names
Report (.txt) with names anonymized
Classification to names and other words
Voikko
If word is classified as first or last name by Voikko
If word is classified as name of place
ML algorithm
If word is classified as name by ML and if the word is starting with capital letter
Lists
If word is in stop-list or ends with street name ending
If word is in list of known names
Name
Other word
If none of the conditions above for word to be name are met
Anonymization of the words classified to names

Input (names in bold):

LEAD INVESTIGATOR: **RINTALA, FREDRIK** TEL. 029000

VANTAA, **Matti Keinonen** is suspected of endangering traffic security. Patrol RK123 (senior constable **Tanninen**, constable **Roivasto**) received a notice 15.11.2019 of a passenger vehicle pulling a trailer having driven into a bar ditch on Helsingintie Otamäki intersection.

Output:

LEAD INVESTIGATOR: **Palräjä, Cpöbpah** TEL. 029000

VANTAA, **Kärra Höalelöl** is suspected of endangering traffic security. Patrol RK123 (senior constable **Rällalöl**, constable **Heasägre**) received a notice 15.11.2019 of a passenger vehicle pulling a trailer having driven into a bar ditch on Helsingintie Otamäki intersection.

Simulations with the NameFinder

- We performed two production simulations with the anonymized text data
- 1. simulation with the base version of NameFinder
  - Native finnish names were identified well
  - Lemmatization errors produced false-positives
  - Names that are used for both humans and places resulted in place names being anonymized same thing happened with animal species names
- 2. simulation with the improved version of NameFinder
  - Tool was improved especially regarding place name identification
  - Words that Voikko regocnized as place names were not anonymized
  - Stop lists were used and for example words that ended ”road” or ”street” were not anonymized

508	42	~	~	0
Accidents	Problem with geopositioning	Problem with participant supplementation	Problem with participant corrections	Problem with examination with keywords

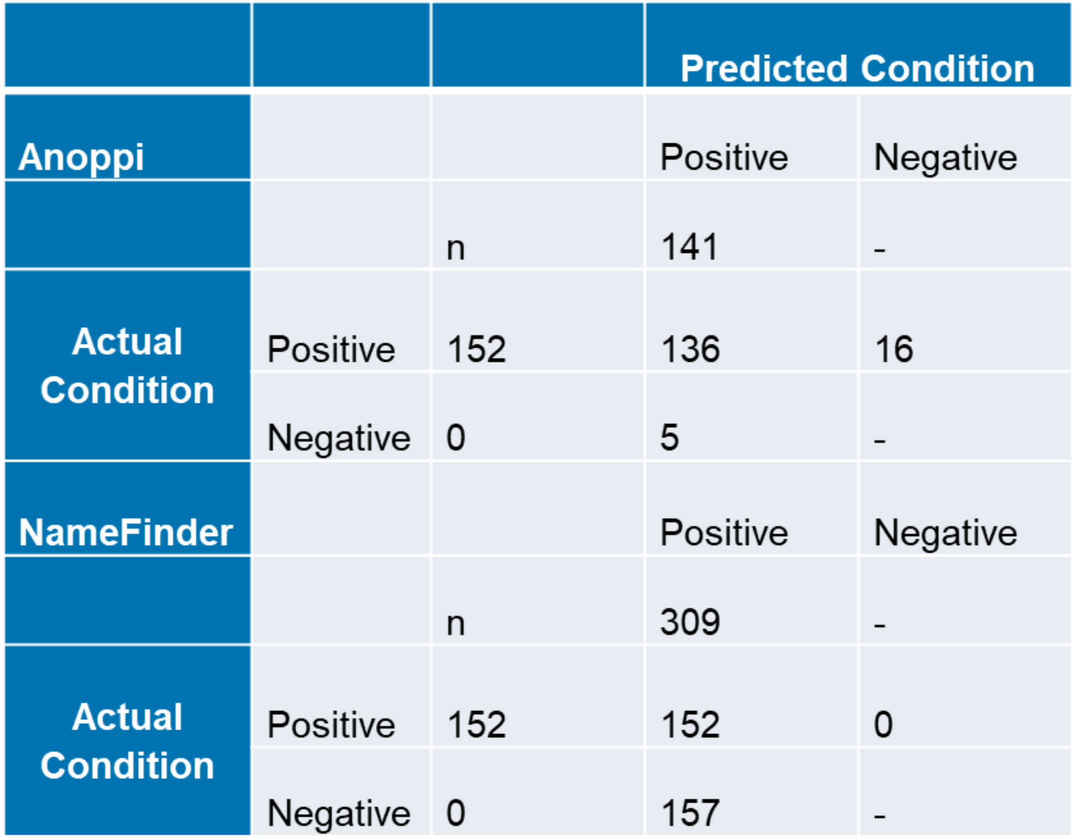
220	10	~	~	0
Accidents	Problem with geopositioning	Problem with participant supplementation	Problem with participant corrections	Problem with examination with keywords

Anonymization impact to statistics and the production process

- Anonymization affects mainly geopositioning
  - In some cases information to determine exact place is anonymized
  - Affects the information on the characteristics of the road
  - The effect is seen on the microdata level, but the final figures in the statistics are not greatly affected by these anonymization errors
- Readability for the handler
  - Possible interpretation errors with multiple individuals and vehicles
  - Keeping track of the entities
- Performance times for anonymization prosecesses can be rather long
  - Optimization is needed

Comparing external tool Anoppi with NameFinder

- We had a change to test an external anonymization tool “Anoppi”
  - Anoppi is developed in a project led by Ministry of Justice
  - Tool for automated court decision anonymization (preprosessing)
- Anoppi produces a list of persons and business entities within a document
  - Keeps track of the entities and makes possible to maintain the relationship between the individuals
  - Would suit to our use case
- Anoppi has API which we used for a controlled comparison test
  - Simulated data was used due to data protection issues
    - Personal pronouns were replaced by names



			Predicted Condition	
Anoppi			Positive	Negative
		n	141	-
Actual Condition	Positive	152	136	16
	Negative	0	5	-
NameFinder			Positive	Negative
		n	309	-
Actual Condition	Positive	152	152	0
	Negative	0	157	-

Future plans

- We would like to install the Anoppi-tool to our local server
  - Further tests with actual road traffic accident data set
  - Implementation to future text data anonymization processes?
- There might be changes in the main source data
  - Might offer possibilities for easier text data anonymization
- Changes to the Statistics on road traffic accidents
  - Text data interpretation might become redundant with the use of broader range of data sets

Contact information and further reading

Matti Kokkonen
matti.kokkonen@stat.fi
+358 29 551 3770

Katja Löytynoja
katja.loytynoja@stat.fi
+358 29 551 3537

Henna Ylimaa
henna.ylimaa@stat.fi
+358 29 551 3832

- Statistics on road traffic accidents (stat.fi)
- Anoppi-project (Ministry of Justice)