Call for Witness: Analyzing Suspect Descriptions from Bavarian Police Press Statements

Sammie Douglas, Max Haag, Anton Lechuga, Alexander Sobieska, Dylan Paltra

Suspect Descriptions?

Example 1: [...] Male, 180-185 cm tall, slender figure; dark hoodie, dark pants, dark cap, black mask, and light-colored gloves. [...]

Example 2: [...] Male, approximately 1.70 meters tall, around 25 years old, slender, approximately 70 kg in weight, with an oriental appearance (orientalisch), fair-skinned, short hair, and a chin beard. He was dressed in a white T-shirt. [...]

Racialized Language

- **Racialized language**: linking certain characteristics to ethnicity or race
- **Limitation**: causality is not defined or analyzed here; multifactorial nature of problem
- **Focus**: appearances of searched suspects, other aspects of racialized language might be interesting as well

Many roads lead to Rome...

Supervised ML Approach

Screen Scraping

Dictionary Approach

API Sampling

Caffeine Approach

Named Entity Recognition

Data collection

Data Collection

Finding interesting data outside social media

1. Police press statements
2. Geospatial data (demographics) for statements
3. Combining these data sources into a coherent framework

Collecting press statements

1. Get a list of all press statements on the website

   Police website is scrapable yet annoying (search form), but there’s an API used by the website we can call directly to obtain metadata

2. Download the individual press statements
3. Separate statements and extract text

   Some statement pages contain more than one statements (non-uniform across departments)

Collecting Geospatial Data

Regionalatlas des Statistischen Bundesamtes

OpenData Bayern

- Shapefiles for Governorates and Districts
- Collecting Shapefiles with Demographics

Measurement

Named Entity Recognition

Identify the broader regions a crime has happened

Different approaches:

- Complete Texts
  - Adds a lot of noise, often several locations mentions
- First Sentence
  - Geospatial reductive approach
  - Not as fine-grained, but correct location

Geocoding of Crime Locations

Usage of Google API (thanks, Carsten) to map the identified location names to actual geospatial coordinates

- These coordinates are used to map the crimes
- USA-Based Service skewing coordinations

Measuring language in suspect descriptions

1. Identifying perp descriptions in statements
    a. Supervised learning
    b. String search
2. Measuring language
    a. Supervised learning
    b. Dictionary

Supervised learning to identify suspect descriptions

We labelled suspect descriptions in 350 press statements and trained a convolutional neural network (CNN) model with tok2vec using spaCy.

Sadly, it did not perform very well due to the greatly varying lengths of the tagged sequences.

Dictionary Approach

We built two dictionaries for identifying…

- suspect descriptions using text identifiers such as descript*, suspect*, …
- racialized language with two sub-categories
    - European: Slavic, East-European, …
    - Not-European: Oriental, *Südländisch* [Southern], North-African, …
    - major limitation and consideration, reproducing racialized language & concepts

Visualization

(don’t share, preliminary results):

https://sicss2022.slack.com/archives/C054N6XF6QL/p1691073161254289

Future Research Avenues

Geospatial fine-tuning

Classifying criminal offences

- In which criminal offences is racialized suspect descriptions more prevalent?

Combination with weather data

- More (& different) crimes in summer?

Normative stance: How do we describe suspects better?

Now it’s time for your questions!