Cold Case – Solved & Unsolved: Use of digital tools and data science techniques to facilitate cold case investigation

Tatjana Kuznecova
Dimitar Rangelov
Saxion University of Applied Sciences, Enschede[1]

Jaap Knotter
Saxion University of Applied Sciences, Enschede & Dutch Police Academy

Abstract
On average 125 murders take place in the Netherlands on an annual basis. However, not all such incidents can be solved. Currently there are more than 1700 unsolved homicide cases on the shelf at the National Police that classify as a 'cold case'. Investigation into these types of capital offenses takes a lot of time, money, and capacity. Applications of the current working method and available techniques are very labor-intensive and time-consuming. In addition, the pressure on the executive Police officers is high - from the Police organization, the Public Prosecution Service, the media, the next of kin, as well as society in general.
From an investigative point of view, it is relevant to provide direction in the criminal investigation and formulate and evaluate various case scenarios, while reducing a risk of 'tunnel vision'. From a scientific point of view, more research into homicide cases in the Netherlands is of eminent importance. Remarkably little has been written in scientific literature about this type of crime.
The project 'Cold Case: Solved & Unsolved' focused on the use of open, publicly available information sources to collect the data and gain more insight into homicide cases in The Netherlands. Applicability of various modern techniques, such as web-scraping, API software and Artificial Intelligence (AI) was explored to facilitate and automate data collection and processing tasks. A first concept of a 'smart' database was proposed, combining a web-based database platform with AI modules to filter and (pre-)process the data. With further development and training of AI modules, such a database might eventually support data-driven generation and/or prioritization of investigative scenarios. The data collected in the process was used in three scientific studies aimed at uncovering the relationships and patterns in the homicide data for The Netherlands.

Keywords: homicide, natural language processing, artificial intelligence, data science, open source data

1 Corresponding author's email: t.kuznecova@saxion.nl

Introduction

On average 125 murders take place in the Netherlands on an annual basis since 2015 (CBS, 2021). Investigation into these types of capital offenses takes a lot of time, money, and capacity. In addition, the pressure on the executive Police officers is high. Both from the Police organization, the Public Prosecution Service, the media, the next of kin, but also from society in general. The investigations by the Police are often followed closely.

Unfortunately, not all murder and manslaughter incidents can be solved. Currently there are more than 1700 unsolved murder and manslaughter cases that are on the shelf at the National Police that classify as "cold case" (see: National Police/Cold case infographic at Politie (2020)).

Within these types of unresolved files, there is often a lack of technologies and tools to deal with the crimes more effectively and efficiently. Applications of the current working methods and available techniques are very labor-intensive and time-consuming. It is crucial that a perpetrator is quickly identified and that he or she can be convicted for the offense committed.

From a scientific point of view, more research into murder and homicide in the Netherlands is of eminent importance. Still rather little has been written in scientific literature about this type of crime, especially compared to the United States (Liem *et al.*, 2013). Within the Dutch Police Academy, the insights from international studies are largely used (such as Adcock & Chancellor (2016) and Adcock & Stein (2017)). However, the question is whether that knowledge can apply to the Dutch situation one-to-one.

There is also a lack of consolidated datasets/databases that use an effective data structure, which may enable detection and exploration of patterns and relationships in the historical homicide cases in The Netherlands. One of the most widely used sources for international comparisons on homicide is the data about the cause of death, such as WHO Mortality Database[2], which only contains information on the number and characteristics of victims. One attempt at developing a unified homicide database framework is The European Homicide Monitor (EHM) that offers a dataset including 85 variables describing homicides in 2003-2006 in Sweden, Netherlands and Finland (Ganpat *et al.*, 2011; Liem *et al.*, 2013). However, a shortage of well-structured or labelled data still poses a serious limitation to the use of data-driven computational approaches, like machine learning, that often require structured datasets as an input.

Data-driven techniques are sometimes used in Police work, at least at a level of research and innovation development. An example of this are tools for predictive policing. For instance, CAS tool was developed for burglary prediction in The Netherlands (Mali, Bronkhorst-Giesen and den Hengst, 2017). However, a previously highlighted limited amount of structured data, as well as inconsistency in data collection and processing methods, create a significant bottleneck in implementing such methods and systems in practice. Given that data-driven techniques may eventually support scenario development and prioritization in crime investigation, it is beneficial to develop a database, where the variables and values are consistent and standardized, when possible. However, collection, pre-processing and analysis of large amounts of information on homicide cases can be very time-consuming and laborious. Therefore it is also paramount to explore what modern techniques and to what extent can be used to automate these processes and reduce data collection or digitization efforts.

Research group Technologies for Crime Investigation (formerly known as Advanced Forensic Technology) is a joint research group between Saxion University of Applied Sciences and Dutch Police Academy. The research group focuses on several pillar topics – CrimeBots (Robotics applications), Nano4Crime (Nanotechnology and sensing) and - a new research line - Data Science & Crime. The first project in the research line of Data Science & Crime is the project 'Cold Case: Solved and Unsolved'. One of the goals of this project was to investigate to what extent modern digital technologies and data science techniques can be used to collect, process, store and analyze data on homicide cases. The project also used a framework by de Kock (2014) as an inspiration to design the data structure for a dataset development. In this framework, twelve so-called Elementary Scenario Components (further referred to as ESC12) can be used to describe any storyline, including a crime scenario. De Kock proposed that ESC12 can be used to build a (smart) database combined with modern data science techniques that could facilitate

2 WHO Mortality Database: https://platform.who.int/mortality

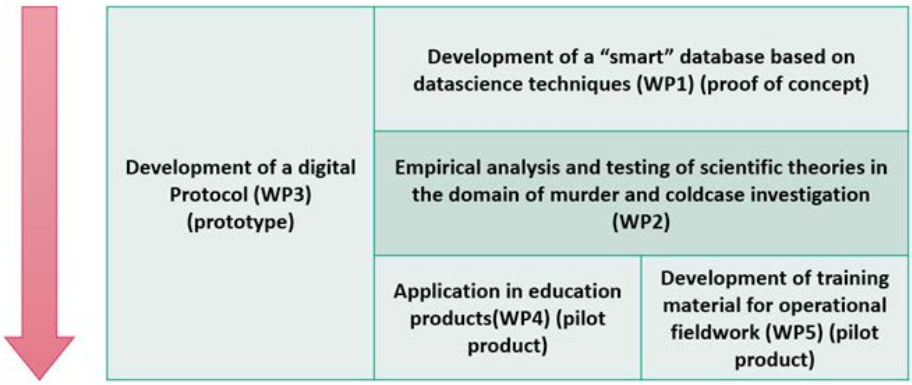
the analysis of the crime cases, make predictions, fill in the knowledge gaps by analyzing historical cases and comparing them with the current investigations (de Kock, 2014).

Furthermore, we focused almost exclusively on the use of open source data. While it could be highly beneficial to use confidential Police data or a combination of confidential and public data, access to such data was not possible in the frames of this project. It should be noted, however, that techniques discussed in this paper can apply to open source and confidential data alike.

Research scope

The project 'Cold Case: Solved & Unsolved' investigated to what extent modern digital tools, computational approaches and data science techniques can facilitate data collection and processing on homicide cases and help organize and direct the investigations. Another objective of the project was to integrate the developed knowledge and tools in the educational process and training materials for the forensic and police science students. The project consisted of five work packages, where each work package had its own objectives (Figure 1). This paper mostly covers the work done with in Work Packages 1 and 2 that focused on data collection, pre-processing, storage, and analysis of patterns and relationships.

Figure 1. 'Cold Case: Solved & Unsolved' project structure (source: authors)



Work Package 1 'Data collection and development of a 'smart' homicide database' mostly dealt with technology development to facilitate collection, pre-processing, and storage of the homicide data. It also covered the actual collection of the data from open sources using a combination of manual and (semi-)automated methods. As an important objective, various possibilities for automation of data (pre-)processing have been explored using Artificial Intelligence (AI) techniques, such as: a) automatically distinguish articles about homicides from other topics (pre-filtering); b) generate a short summary of an article; c) extract interesting information/components from an articles, such as ESC12. In the future, collected data and identified techniques are envisaged for the integration in a 'smart' database platform that can eventually facilitate crime analysis and investigation with data-driven methods.

The goal of Work Package 2 'Empirical analysis of the homicide data in The Netherlands' was to conduct scientific studies using the data collected in Work Package 1. It included application of various statistical methods and data modelling techniques to test theories and hypotheses. The results of this work package can reinforce theoretical understanding of the relationships and patterns occurring in the homicide cases specifically in The Netherlands, which can further be compared to the similar studies in other countries.

Main results and outputs

This section provides an overview on the main outcomes for the sub-objectives described in Chapter 2, concerning data collection, AI research and analysis of relationships and patterns in the homicide data.

Data collection

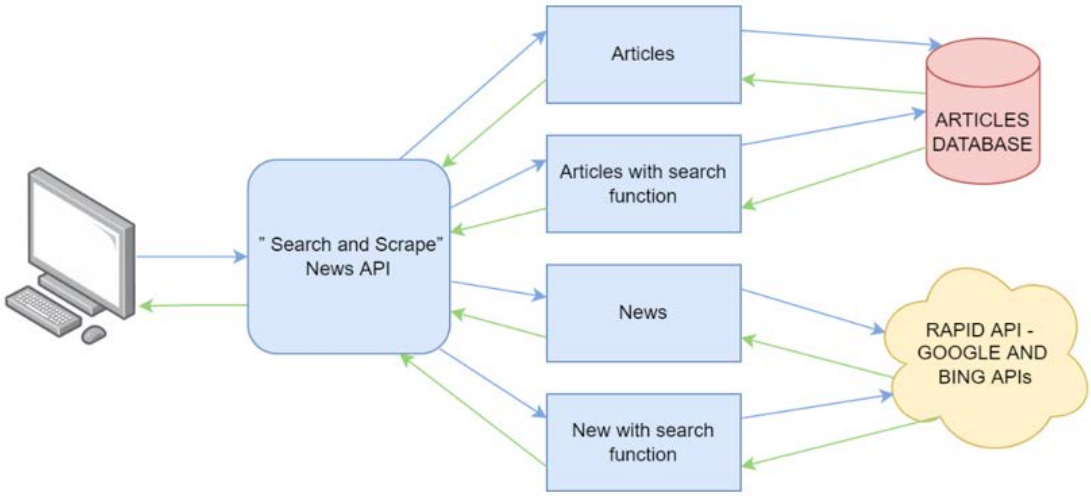
In this project, several types of data were collected using a variety of techniques. One stream of data consisted of a collection of articles on homicide-related topics. Another dataset was created manually by structuring information on homicide cases into a database template for further applications in data analysis and modelling studies. Lastly, a software tool (API) was developed to facilitate and automate collection of homicide-related articles from open sources.

The collection of articles about homicide cases was performed with the help of Python programming language. It was split in two flows. The first flow was a *'site scraper bot'*, which scraped the news, archives, and articles about homicide cases in a variety of websites. The second flow was based on *'Optical Character Recognition'* (OCR), which was used to process old newspaper archives from the Delpher portal[3]. The articles from these two approaches were filtered through the algorithm that decided whether articles concern homicide cases or other topics. With the combination of these two approaches, around 11 000 articles mentioning homicide cases were collected.

With the help of the students of Cold Case Minor course (year 2020) at Saxion University of Applied Sciences, another dataset was collected and processed in detail with manual techniques using information on around 300 homicide cases in The Netherlands for the period of 2006-2015. Each student was assigned a list of cases that he/she had to collect the data on. A template was developed and provided to the students that contained the desired data structure. In this project, a framework by de Kock (2014) on ESC12 was used as a basis for the dataset structure. For each case, students had to use at least five information sources (when possible). Further, students had to extract the necessary information and fill in the provided template. The resulting dataset was used in the studies on data analysis and modelling to uncover the patterns in the data.

It is paramount to have as much data on homicides as possible to make use of data-driven approaches. This data can be used: a) for empirical analysis and development of data models and analysis of relationships and patterns; b) and training and tuning of AI algorithms. Furthermore, the relevant data can come in different formats and from different sources. To facilitate the data collection process, an API (Figure 2) was developed that can be used as a standalone tool, or eventually in conjunction with other software (e.g., connected with a database platform). Several sources were pre-defined for the tool (such as BING News, Google News). The tool's architecture allows for an easy connection to a wider range of sources in the future.

Figure 2. 'Search and Scrape API schematic (source: authors)



3 Delpher newspaper archive: https://www.delpher.nl

Research and Development in AI

AI development was conducted in three directions:

1. distinguish homicide article from other topics
2. generate a summary of an article
3. automate extraction of interesting components, such as ESC12.

The work method is mainly focused on Natural Language Processing (NLP) family of algorithms. NLP is usually used to process, analyze, and extract information from natural human language data (such as texts). Given the complexity of human language, different neural network-based methods have been developed to date, among them Word2Vec, Sense2Vec and other. The main model used within the project is the Bidirectional Encoder Representations from Transformers (BERT) (Devlin *et al.*, 2019). Processing pipelines based on known methods were developed to tackle the described tasks.

The most challenging part of AI research was extracting the interesting components from text automatically (for instance, name of a victim or perpetrator, date or location of a crime, etc.). The AI model worked well in some cases, however, at the same time, it performed badly on some other difficult examples. Initially the pipeline was tested on Dutch articles translated to English via Google Translate. Therefore, retraining and testing algorithms with data in the Dutch language is required in order to further work with Dutch texts. Most of the mistakes in text processing come from the Coreference Resolution component. Coreference Resolution is a critical component of natural language understanding and higher-level NLP applications including information extraction, text summarization, and machine translation. It is the process of determining whether two expressions in natural language refer to the same entity in the world (Soon, Lim & Ng, 2001). The complexity of the human language processing and text interpretation still limits the effectiveness of the current AI algorithms. Data collection and labelling for AI training is an extremely time-consuming and laborious task, and the current project could not ensure the necessary resources for that. Therefore, the Coreference Resolution problem will require further work in the context of this research.

Summary generation algorithm worked quite successfully, as well as a filtering algorithm to classify homicide-related articles. For this task various methods were tested, with the most successful (BERT algorithm) reaching accuracy of more than 96% (Table 1).

Table 1. Test results of the algorithms to distinguish homicide articles from other topics (source: authors)

Methods	Accuracy
Bernoulli NB	87.42%
Naive Bayes	87.63%
MNB	87.97%
NuSVC	88.58%
Voted	93.06%
Linear SVC	94.85%
Logistic Regression	95.36%
SVC	95.82%
BERT	96.11%

Development of a (smart) database concept

The ultimate goal of this research is to eventually develop a homicide database equipped with 'smart' features driven by AI algorithms to support investigative process with data-driven insights. The first steps were taken towards that goal. A prototype of a web-based database platform was developed that can store the data (text articles) collected on homicide cases (Figure 3).

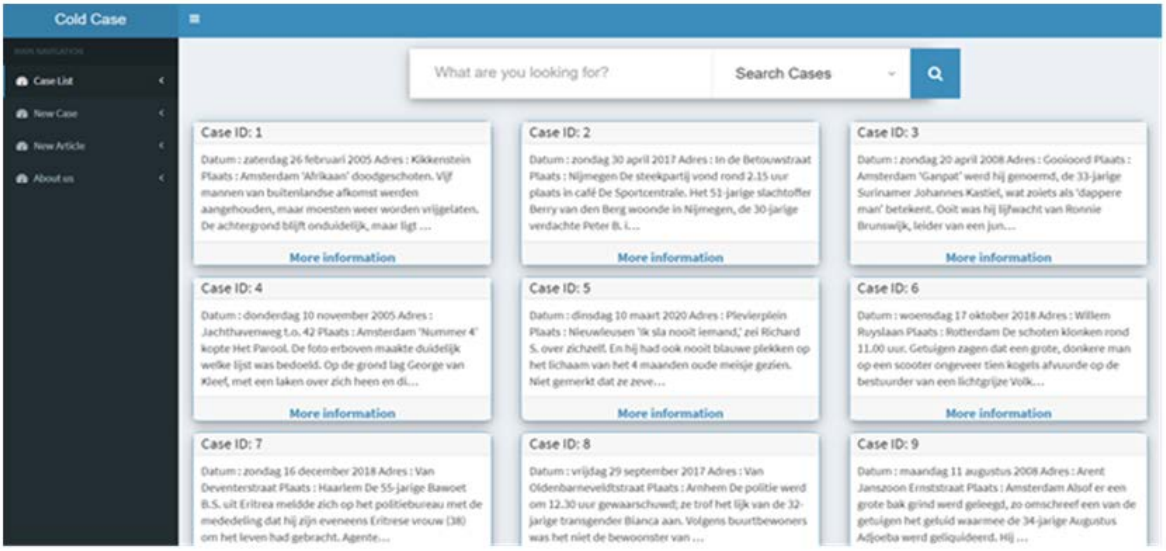
The database organizes articles on a case-by-case basis (so one homicide case can have multiple articles linked to it). Furthermore, a case description includes a ESC12 components structure based on the framework by de Kock (2014). The connection with AI engine was also set up and tested. We foresee that in the future it will be possible to integrate all the tools described in this paper to achieve a fully functional 'smart' database platform. The 'Search and Scrape API' will automatically take new information from the news portals or other sources (such as Police information systems, if data use permits allow) about homicide cases, which will be analyzed and filtered by an AI algorithm. The function of Artificial Intelligence to generate short summaries or overviews of articles or case descriptions can be implemented directly in the platform. On the other hand, the AI models for extraction of interesting components from the text can be re-trained with new data and developed further with other methodologies and approaches. While the AI module is not yet ready to re-

place the capabilities of humans for text interpretation, it may become a useful tool in the future as a kind of an assisting/recommendation feature for the users.

The current research was not able to tackle the actual generation or prioritization of crime scenarios. This part can be tested in the future, when sufficient amount of information on homicide cases is collected, thus enabling effective pattern mining by means of data-driven techniques.

The design of the platform and the user experience can be made to match the needs of Police officers involved in detecting homicide cases. This means embedding the platform with their workflow and fully or partially covering their needs.

Figure 3. Database platform (source: authors)



Empirical analysis of homicide data

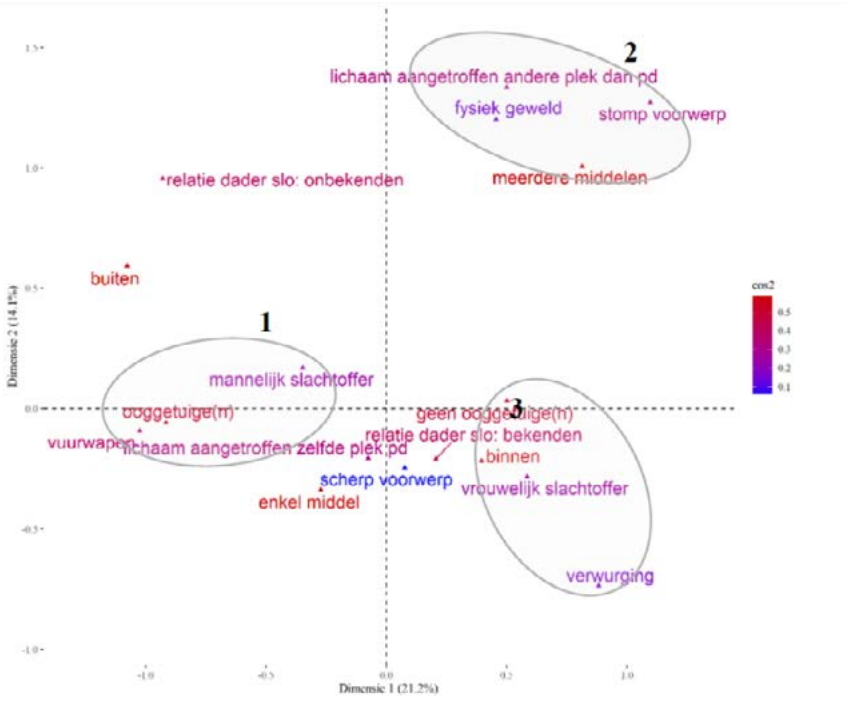
Data collected during the project was used to explore meaningful relationships and patterns in the homicide cases in The Netherlands. This can be considered the first step in determining the potential of (open source) data on homicide cases to be used in predictive or pre-

scriptive data-driven systems. Several scientific studies were conducted and final theses were developed by students in collaboration with University of Leiden, Dutch Police Academy and Amsterdam University of Applied Sciences (HvA).

The first research was conducted by Hanneke van de Mortel, University of Leiden (van de Mortel, 2020) as her MSc thesis. The research focused on predicting the relationship between the perpetrator and the victim on the basis of victim characteristics and the modus operandi. Such research is the first step to identify relevant relationships in the homicide cases that can help determine a direction of the investigation. This study was conducted using a dataset of ~300 homicide cases in The Netherlands for the period of 2006-2015 (manual data collection). The research method included bi-variate correlation analysis and predictive modelling with Logistic Regression method.

The second research was conducted by the student of Dutch Police Academy, Rob Schipperheyn, as his MSc graduation research (Schipperheyn, 2021). This scientific work focused on identifying the clusters of co-occurring variables in the homicide dataset and translating them into practical scenario-based investigation recommendations. The global objective of this research was to facilitate the development and use of more scientifically substantiated scenarios in police investigation using data science insights. The research methodology included: univariate analyses, bi-variate analysis, and multi-variate analysis in a form of Multiple Correspondence Analysis (MCA) method (Figure 4). Like the previous study, this research was based on the data on ~300 homicide cases for the period of 2006-2015.

Figure 4. Output of an MCA model (Schipperheyn, 2021)



The third study was conducted by the student of Amsterdam University of Applied Sciences, Izzy van der Veur, as his BSc graduation thesis (van der Veur, 2021). The research focused on the question: 'To what extent can the nature of a homicide be determined on the basis of social economical, geographical and demographic characteristics of the location of a homicide?'. In contrast to the first two studies of this work package, this research used the data scraped from a website Moordatlas[4], for the period of 2016-2020. This time period was chosen due to a higher degree of completeness of the available data, which makes the results more valid and reliable. The data scraped from Moordatlas was further processed to develop a data-

4 Moordatlas website: www.moordatlas.nl

set. Bi-variate correlation analysis was used to explore meaningful relationships, while geographic mapping was used to look at the spatial distribution of homicide cases in The Netherlands.

Conclusions

This paper described the outputs of the project 'Cold Case: Solved & Unsolved' completed in the research group of 'Technologies for Criminal Investigations' at Saxion University of Applied Sciences and Dutch Police Academy. The project explored and tested development and applications of various digital tools and data science techniques to facilitate and automate collection, pre-processing and analysis of open source data on homicide cases.

It was possible to achieve a certain degree of automation of some of the data collection and processing steps. For example, AI algorithm for classifying articles about homicides performed well (highest accuracy >96%). However, the complexity of the human language processing and text interpretation still limits the effectiveness of the current AI algorithms for more difficult tasks. Hence, development of a 'smart' database equipped with a fully functional and effective AI engine was not attainable in this project. Thus, further research is needed to achieve a more effective extraction and structuring of interesting information from the text. Current efforts were also limited by the lack of suitable data for training of the AI models. Furthermore, AI training and testing were restricted by the available computational capacity and some technical disruptions during the development stage.

The prototype of the web-based database platform currently allows for manual entry of the data, therefore it can be used without AI as well. Overall conclusion is that AI does not yet match the human capacity to interpret text, however with proper training AI has the potential to be used as an assisting or recommending tool in addition to experts' judgement.

Concerning the applicability of open-source data for homicide research (and eventually investigation), we can conclude that use of open-source data is associated with certain limitations and risks. Over- or under-representation of certain (groups of) cases is possible, especially due to differences in media attention to certain types of homicide cases: for instance, unusual, scandalous, or somewhat mysterious cases often receive more media coverage. Given that data collection is an extremely time-consuming and laborious process, in this project it was possible to only collect the data on a limited number of cases (around 300 cases). Furthermore, in those cases, not all the variables could be filled in, thus many variables became unusable in the analysis due to a large amount of missing values.

For further research, a number of reliable public sources may be pre-defined using a set of specific criteria. Research presented here aimed to explore usability of a wider range of sources, which might have led to inclusion of incorrect information in a dataset.

Relatively small set of cases included in a detailed structured dataset (~300 cases) and a large amount of missing values for many of the variables limited the exploration of relationships and patterns. However, some significant relationships could still be identified in the three scientific studies conducted. This suggests that potential to use predictive and prescriptive data modelling techniques in homicide research should be investigated further.

Discussion and recommendations

While working with the open source data for the homicide cases in The Netherlands, we found that the amount of detail in the open source data is limited. Moreover, the data may be biased or not trustworthy. This creates a serious limitation for data analysis using open source data. Further research could be devoted to working with Police data or a combination of Police files and open source data. This is, however, associated with significant difficulties of getting access to the confidential data.

AI can be a powerful tool in recognizing complex patterns. However, more research is necessary in order to automate the processing of (big) textual data. With the state-of-the-art of the technology and considering a sensitive nature of the forensic or criminological applications, it is not possible to completely replace a human expert with an AI algorithm. Further work is required concerning both data collection and AI development to enable the use of data-driven insights and/or predictive algorithms in the homicide investigation. We suggest that AI can be potentially used in combination with human judgement, as a recommendation tool. With sufficient data, AI-powered tools can even-

tually support scenario generation and prioritization, identify groups of similar cases in the historical database and compare historical records to an ongoing case.

It should be noted, however, that more research on ethical issues should be conducted, in order to avoid biases and ensure the correct use of the information generated by computer algorithms. As suggested by van Brakel (2016), big data and predictive tools can have benefits for policing, but such techniques may also bring disempowerment of individuals, groups, and society depending on implementation and intentions behind their use. Moreover, an open question still remains - how do we make sure that data-driven techniques actually help prevent a tunnel vision, instead of reinforcing it? With this concern in mind, carefully designed operational workflows and application strategies should be embedded in the Police practice to accommodate the correct use of data-driven techniques.

Data-driven research often requires good quality structured datasets. Our first step in that direction was development of a structured dataset with about 300 cases on homicides in The Netherlands in the period of 2006-2015, using information from published news and other open sources. With more than 200 variables in the dataset structure, this is an extremely time- and resource-consuming endeavor. Techniques for more efficient methods of information extraction and structuring from big (textual) data should be further explored. Another possibility may lie with collaborations with volunteers, universities or other organizations that might contribute to the task of data collection.

Closer collaboration with the Police (or other potential user groups) is crucial for the development of a relevant data-driven tool. We suggest that the future projects should fit in the development agenda of the National Police, and the tools should be developed with the input from the Police experts.

Acknowledgment

We cordially thank *Tech For Future* - programme for co-funding the project 'Cold Case: Solved & Unsolved' and making this research possible. We also thank our project partners - Pandora Intelligence, Icologiq, SDProject, Saxion University of Applied Sciences, Dutch Police Academy, Factor Veiligheid - for contributions in research and development. We thank University of Leiden and Amsterdam University of Applied Sciences for contributions to this research. Our sincere gratitude goes to researchers Dung Le and Dimitar Rangelov – for their work on AI and web-scraping methods; Hanneke van de Mortel, Izzy van der Veur and Rob Schipperheyn – for their applications of the collected data in empirical data analyses; 'Smart Solutions Semester' student Luuk Cloosterman – for his work on 'Search and scrape' API; all students of Cold Case Minor course – for their contributions to data collection and feedback on developed tools. We also thank volunteer organization Bureau Dupin for the support and knowledge exchange in the various steps of this research.

References

- Adcock, J.M. & Chancellor, A.S. (2016) *Death Investigations, The 2nd Edition*. 2nd edn. CreateSpace Independent Publishing Platform.
- Adcock, J.M. & Stein, S.L. (2015) *Cold cases: An evaluation model with follow-up strategies for investigators*. 2nd edn. CRC Press, Taylor & Francis Group. doi:10.1201/b10204.
- CBS (2021) *Minder moorden in 2020, wel meer jongeren vermoord*.
  Available at: https://www.cbs.nl/nl-nl/nieuws/2021/39/minder-moorden-in-2020-wel-meer-jongeren-vermoord#:~:text=Onder
- de Kock, P.A.M.G. (2014) *Anticipating criminal behaviour: using the narrative in crime-related data*. 1st edn. WLP.
- Devlin, J. *et al.* (2019) 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota, pp. 4171–4186. doi:10.48550/arxiv.1810.04805.
- Ganpat, D.S. *et al.* (2011) 'Homicide in Finland, the Netherlands and Sweden: A first study on the European Homicide Monitor Data', Homicide Studies 17(1) 75 –95

- Liem, M. *et al.* (2013) 'Homicide in Finland, the Netherlands, and Sweden: First Findings From the European Homicide Monitor', *Homicide Studies*, 17(1), pp. 75–95. doi:10.1177/1088767912452130.
- Mali, B., Bronkhorst-Giesen, C. & den Hengst, M. (2017) *Predictive policing: lessen voor de toekomst. Een evaluatie van de landelijke pilot*. Apeldoorn: Politieacademie.
- Politie (2020) *Nieuwe coldcasekalender ook in tbs-instellingen*.
  Available at: https://www.politie.nl/nieuws/2020/januari/17/00-nieuwe-coldcasekalender-ook-in-tbs-instellingen.html
- Schipperheyn, R. (2021) *Scenario's van moord en doodslag: Exploratief onderzoek naar samenhangende kenmerken van kapitale delicten*. Politieacademie.
- Soon, W.M., Lim, D.C.Y. & Ng, H.T. (2001) 'A machine learning approach to coreference resolution of noun phrases', *Computational Linguistics*, 27(4), pp. 521–544. doi:10.1162/089120101753342653.
- van Brakel, R. (2016) 'Pre-Emptive Big Data Surveillance and its (Dis)Empowering Consequences: The Case of Predictive Policing', in *Exploring the Boundaries of Big Data*. Amsterdam: Amsterdam University Press, pp. 117–141. doi:10.2139/ssrn.2772469.
- van de Mortel, M.E.J. (2020) *In hoeverre kan de (soort) relatie tussen de dader en het slachtoffer van dodelijk geweld binnen Nederland, worden voorspeld aan de hand van slachtofferkenmerken en de modus operandi?* Universiteit Leiden.
- van der Veur, I. (2021) *De mogelijke bepaling van de aard van een levensdelict aan de hand van de sociaaleconomische, geografische en demografische kenmerken van het plaats delict*. Hogeschool van Amsterdam.