RESEARCH

Open Access

Unsupervised identification of crime problems from police free-text data

Daniel Birks[1,2*], Alex Coleman[2] and David Jackson[3]

Abstract

We present a novel exploratory application of unsupervised machine-learning methods to identify clusters of specific crime problems from unstructured modus operandi free-text data within a single administrative crime classification. To illustrate our proposed approach, we analyse police recorded free-text narrative descriptions of residential burglaries occurring over a two-year period in a major metropolitan area of the UK. Results of our analyses demonstrate that topic modelling algorithms are capable of clustering substantively different burglary problems without prior knowledge of such groupings. Subsequently, we describe a prototype dashboard that allows replication of our analytical workflow and could be applied to support operational decision making in the identification of specific crime problems. This approach to grouping distinct types of offences within existing offence categories, we argue, has the potential to support crime analysts in proactively analysing large volumes of modus operandi free-text data—with the ultimate aims of developing a greater understanding of crime problems and supporting the design of tailored crime reduction interventions.

Keywords: Policing, Burglary, Unstructured data, Text mining, Machine learning

Background

When a crime is recorded by police, situational and behavioural information describing the incident are often captured in a free-text narrative account. These data, commonly referred to as Modus Operandi (MO) notes, are routinely used for both administrative and investigatory purposes. Yet beyond such function, these accounts also have the potential to increase understanding of specific crime problems in ways that support crime reduction efforts. The key challenge in realising this potential relates to the unstructured nature of MO notes. Put simply, free-text data are more challenging to analyse at scale than structured measurements of crime events such as quantity, location, and time—all of which immediately lend themselves to traditional analytical approaches such as measuring rates of offending, examining spatial distributions across neighbourhoods, or measuring hourly, weekly and monthly variations. These challenges dictate that potentially actionable insights into specific crime problems can be lost in the categorisation of crime events into tractable but restrictively homogenous groupings. In response to this problem, this paper proposes a novel method for automatically clustering specific crime problems within existing crime categories based on the application of unsupervised text-mining algorithms to narrative crime report data. This approach, we argue, has the potential to support police analyst decision making regarding specific criminal MOs and help identify both existing and emerging criminal behaviours in a systematic fashion that can support efforts to better understand, and ultimately reduce, victimisation.

Crime classifications

Crime events recorded by police within the United Kingdom classify incidents according to a hierarchical structure of UK Home Office codes, classes, sub-classes and offence categories. These categories reflect the

*Correspondence: D.Birks@leeds.ac.uk
[1] School of Law, University of Leeds, Leeds, UK
Full list of author information is available at the end of the article

© The Author(s) 2020. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Notifiable Offence List (Data.gov.uk, 2018)—which for each class maps to a relevant legislative act that is used to charge individuals suspected of committing a particular offence. To illustrate, an attempted residential burglary is recorded with code 28/3, class "Burglary", sub class "Residential Burglary", offence category "28F Attempted Burglary—Residential", its application associated with Section 9 of the Theft Act (1968).

The above classifications are devised to support a range of criminal justice processes and enable comparability of crime measurement within and between jurisdictions. Yet, while necessary for the administrative needs of the criminal justice system, this approach to classifying crimes does not necessarily capture ecologically appropriate descriptions of actual crime problems. Put simply, not all offences of a particular classification are created equal—and while characterising them as such makes sense for administrative purposes, it may not if one seeks to best understand the types of criminal opportunities and behaviours that manifest in a particular context to generate crime problems.

Business as usual in police free-text analyses

In modern policing, when a crime is recorded a wealth of information is collected and recorded across various systems. Collectively, these data include the offence classification, the time, date, locations of the offence, attending officers, and disposal types–to name but a few. In addition, most recorded crimes also include a narrative account of the incident in question —the modus operandi or MO. The MO is the digital equivalent of an officer's notebook—recording the observed or inferred method of operation associated with a specific criminal incident. These data provide context for the event and can subsequently be used both in investigatory work and disclosed in court to provide context to criminal justice practitioners. To illustrate, the following is an excerpt from the MO notes associated with two residential burglaries.

MO 1 "Attacked property is mid town house with driveway to the front along with gardens to both front and rear located within a residential area. At time stated person/s unknown go to front door and open letter box and using unknown instrument hook door key from a shelf in the porch. Use same keys to open front door and gain entry remove two sets of car keys from the porch area. Go to a XXXX parked on the drive gain access using keys. Make off at speed with both vehicles direction of travel towsrds XXXX having been disturbed by the occupant."

MO 2 "Attacked property is a large detached dwelling on a busy road. Property is surrounded by large fences, gates and bushes. Between times stated suspect approach rear patio doors at locus and attempt to gain entry by using mole grip type implement to snap lock. Lock snapped however unable to gain entry. Suspects then use molegrip type implement to snap lock on front porch door. Lock snapped, door opened and house alarm sounds. Suspects jump over wall at front of dwelling, get into vehicle parked opposite and make off down XXXXX in direction of XXXXX."

Reading the above accounts, two things are immediately apparent. First, MO notes can provide key insights into a given crime that are likely difficult to capture through closed response data fields. Second, while both offences are classified as residential burglaries, their description in this unstructured data indicates two substantively different criminal opportunities and offending behaviours.

The utility derived from this specificity in increasing the likelihood of detection has been well understood for over a century (Fosdick 1916). At the same time it is now well acknowledged that an understanding of offending processes is key in the effective design and implementation of crime reduction interventions (Cornish 1994); a premise that underlies prominent crime reduction paradigms including situational crime prevention (Clarke 1983), problem-oriented policing (Goldstein 1979; Braga 2008), intelligence-led policing (Ratcliffe 2016) and more broadly crime science (Laycock 2013). This body of research and practice has clearly demonstrated that an understanding of specific criminal opportunities, and the means by which they are exploited, enables the design and delivery of targeted crime reduction strategies. To illustrate, the identification of specific burglary MOs described in our previous examples, i.e. the targeting of car keys, or entry by snapping insecure UPVC door locks, should inform different crime prevention interventions–situational measures to prevent the use of 'hook and cane' methods, or the fitting of 'anti-snap' locks, respectively. It is under the rationale of identifying such insights, that these free-text accounts of crime are the focus of analyses in this paper.

Manual analyses of free-text crime reports

Analyses of MO notes are conducted for a number of distinct reasons—first and foremost, for investigatory purposes—here an analyst reads offence descriptions across a number of crimes to manually determine similarities. If there are only a small number of offences to be analysed, this can be effective, but still time consuming and is often utilised when one seeks to determine potential offences carried out by a specific suspect during criminal investigation proceedings. Through manual assessments of offence similarities, potential additional crimes can be identified that a specific known offender may have committed. These can then be presented as potential additional offences for prosecution. Additionally, in some cases analysts attempt to determine patterns across large volumes of offences to inform operational or strategic planning. Yet manually reading and retaining information from potentially hundreds of records is resource intensive, error prone, and as a result, highly impractical.

This combination of the insights that free-text data may provide into specific offences, and the difficulties associated with accessing them in a timely and systematic manner, has led to a number of analytical 'work arounds' in police information systems. The most common being the use of manual "keyword" searches across large data sets. This process depends either on predetermining keyword(s) to be looked for based on previous crime patterns identified, or dip sampling a smaller sub-set of data for analysis to determine possible patterns identifiable through specific searches. Both of these methods are problematic in determining ongoing or emerging trends: the initial process is completely reliant on previous knowledge and the second process is open to significant bias and can miss smaller groupings of MO characteristics that would only be evident within analysis of the full data set. A key problem for the standard analysis of MO and free-text information is that small clusters of offence characteristics are easily missed, especially if emerging patterns are referenced differently because no standardised descriptive phrasing has been developed.

In addition to these manual searches the process of recording can sometimes be augmented by the use of categorical fields in recorded crime data that seek to identify the presence or absence of keywords in the MO associated with a particular offence. Often referred to as *markers*; these fields aim to denote whether an offence is associated with a specific behaviour or context e.g. firearms, alcohol or drugs, domestic disputes etc. While popular, this approach to systematising MO analyses is also subject to a number of weaknesses. First, it requires that police are aware of, and consistently use, marker information when recording crime incidents. Second, it requires that police incrementally add marker fields to recording systems—a process that can rapidly become cumbersome and potentially undermine application by increasing burdens on the time police have in attending to the recording process. Third, in the same manner of post hoc keyword searches, the approach is wholly reactive in nature —such that, a marker for a particular characteristic of offences must be previously known (and presumably occur at sufficient regularity) to those who devise such fields to include it in standard data input forms in the first place.

Motivation

With the above in mind it is clear that while free-text data currently recorded and analysed by police offer great potential for increasing understanding of particular types of offending, a range of significant constraints dictate that the utility derived from such data is likely not in-step with the collective investment associated with its capture. It is against this backdrop that we explore a new approach to analysing MO free-text data within traditional crime classifications, with the goal of supporting operational police analysts in proactively identifying specific crime problems across large numbers of historic crime reports associated with a particular offence. And thus, developing means to create more ecologically valid and granular *within-crime* classifications that may increase the likelihood of identifying particular crime problems, and, by extension, aid in the conception and design of measures designed to reduce their occurrence.

Our approach

In moving the field forward, here we build on previous work that has sought to apply the text mining approach of topic modelling to analyse free-text data in domain specific contexts. Topic modelling is a statistical approach that aims to identify hidden semantic structures within a collection of unstructured textual data (corpus). The specific form of topic modelling applied in this work is Latent Dirichlet Allocation (LDA), a probabilistic topic model that assumes each document contains a mixture of topics, where topics are represented as a distinct probability distribution over all the words in the corpus (vocabulary) (Blei et al. 2003). The approach allows for documents to be categorised probabilistically capturing the heterogeneity of narrative events that may have overlaps in the described event. For instance, the word 'bank' may be used in topics about a river but also in topics associated with finance.

LDA has previously been applied to investigate a wide range of questions including performing literature reviews within research fields to identify emerging research trends (Moro et al. 2015), identifying trends in public policy debates (Benites-Lazaro et al. 2018) and in combination with sentiment analysis to determine controversial topics between communities (Panasyuk et al. 2014).

In the context of our explanatory efforts to identify crime problems in a way that could support applied decision making, we selected LDA for two primary reasons. First, due to its maturity relative to other emerging techniques that utilise deep learning approaches to topic modelling (Zhu and Xie 2018). LDA is robust, well understood, and as discussed above has previously been productively applied in a broad range of areas. Second, the intuition of LDA, that documents are made up of a mixture of topics and topics are groups of words that co-occur together with specific probabilities offers an accessible approach to understanding how the model produces given outcomes. This, we felt, was crucial given that algorithmic transparency, explainability and ultimately understandability are critical in engendering trust and consent in applications of computational methods within the criminal justice system (Babuta et al. 2018).

Related work

The application of topic modelling in a crime context is an emerging area of research. Recent work has explored the use of hierarchical topic models to determine nuances within narrative police data that could better inform understanding of the causal processes that underlie criminal behaviour (Kuang et al. 2017). Utilising non-negative matrix factorization (NMF) methods, Kuang et al. highlighted how topic modelling could help identify more ecologically valid relationships between broad crime types. Analysing over 5 years' worth of crime reports from the Los Angeles Police Department the authors construct hierarchies of topics that identify textual distinctions within reports such as between property and violent crime, firearm or knife/sharp weapon incidents. While the authors discuss the potential for this approach to aid with the automatic classification of crimes, which they feel their model is insufficient to accomplish, their work highlights the potential of topic modelling to capture narrative insights that may ultimately be useful for policing and crime-reduction decision making.

In subsequent research Pandey and Mohler (2018) compare two different topic model implementations in collectively clustering a corpus of crime reports associated with seven types of crime occurring in Los Angeles over a five-year period. In particular, they explore the performance of two topic model implementations–NMF and LDA—comparing the clustered crime classifications produced by these approaches to the seven traditional administrative crime categories. Classifications produced by the topic models are compared to traditional crime categories and initially evaluated in terms of their topic coherence score—a performance metric which measures the degree to which words within a topic are similar. Subsequently, building on the ubiquity of crime's spatial concentration, the authors test the hypothesis that offences grouped under the topic models will exhibit greater levels of spatial concentration than those grouped within traditional crime categories. Results of their analyses indicate that offence clusters derived from the application of LDA exhibit both the greatest coherence scores and levels of spatial clustering relative to both NMF based topic models and traditional crime categories.

In addition to these efforts to understand similarities *between* crime types, the notion of crime-linkage—where offences sharing very specific characteristics suggesting a shared perpetrator—has also seen some attempts at automation. Chohlas-Wood and Levine (2019) present analyses of unstructured police free-text data using their Patternizer software. Patternizer utilises a variety of features captured in a crime report including, but not limited to, narrative text to identify similar crime events. Narrative text features differ depending on crime type. In burglaries, a similarity score was calculated for words shared between the two reports; for robbery and grand larceny, a term-frequency inverse document frequency vector of each document is compared to determine a cosine similarity score. Finally, all crime types also include a feature representing the cosine similarity score for all unstructured text between two reports along with a feature counting rare words that do not occur in a training corpus. These features are then used in a decision tree along with other non-text features to identify highly similar crime events. Whilst this approach does not apply a topic model it does highlight an alternate approach that aims to capture specific details that exist only within police free-text reports and allow police to produce potentially actionable insights. Moreover, a range of other methods aimed at supporting crime-linkage have also been explored (see for instance Adderley and Musgrove 2003; Bennell et al. 2009; Bennell et al. 2012; Oatley, Ewart, and Zeleznikow 2006).

Beyond the analyses of free-text recorded by police, several studies have also explored how LDA topic models might be used to analyse crime related phenomena discussed in unstructured content posted online. Chen et al. (2015) combine LDA and collaborative representation classifiers in an attempt to detect 'criminal intention' in free-text extracted from online blogs. Relatedly, Gerber (2014) apply LDA topic models to geotagged tweets posted to the online platform Twitter in Chicago to generate neighbourhood dominant topics indicative of the ecology of a particular location. These classifications are then used to positively augment a kernel density estimation based crime prediction algorithm that seeks to estimate levels of 25 distinct offences in the immediate future.

In contrast to these previous applications, the present paper presents a machine learning framework for analysing police free-text data to gain insight into different crime problems that exist *within* a single crime classification—but that are not necessarily indicative of a single perpetrator. To illustrate this approach, we analyse MO data associated with residential burglary incidents

Table 1 Examples of narrative text from crime reports classified as Residential Burglary

Narrative text
Attacked premises are semi detached dwelling in residential area. Suspect unknown approach rear of premises and smash rear transom window with unknown instrument. Reach through and open casement window and climb through into bathroom. Property removed from inside and high value property also left behind egress as entry and make good escape unseen
Between times stated complainant is working at customers home address doing some interior design work. Whilst complainant and customer are working in living room on 1st floor of mid terrace 3 storey premise, suspect/s unknown have approached insecure front door to the premise opened it activating door chime which occupants do not hear, suspects have then entered hallway and picked up complainants bag and egressed as entry. Suspects have then made their escape good unseen and unchallenged
Between times and dates stated unknown suspect/s reach through the cat flap at the rear door at locus using unknown implement, remove the complainants jacket from the back of the dining room chair. The jacket contains the complainants wallet. Suspect/s make off with property unseen unheard in unknown direction
Attacked property is a privately owned end terrece multi occupancey dwelling. Between times stated suspect/s enter through insecure ground floor window. Tidy search conducted and vehicle keys removed from kitchen hooks. Suspect make their escape through same and leave stealing vehicles. Vehicle XXXXX found burnt out
Modus operandi summary…..Attacked property is a mid-terraced property located on a quiet residential street. Between times stated unknown suspect approaches the front of the property and with bodily force kicks open the basement window. Suspects gain entry to the property and untidy search in conducted. Suspects exit property with stolen items and make off in unknown direction

occurring in a major metropolitan borough of the UK. This framework was developed as an attempt to determine whether unsupervised topic modelling approaches could be applied to automatically cluster offences by specific MO described in crime reports. To accomplish this we apply Latent Dirichlet Allocation to probabilistically determine latent topics within the corpus which correspond to specific MOs. The most probable topic is then assigned to a crime report as the dominant topic in order to cluster reports together and assess whether this form of topic modelling is able to identify specific MOs without prior knowledge of them. Our primary goal here is to document a proof-of-concept use of topic modelling for understanding distinct criminal behaviours within an existing official crime category—with the ultimate aim of supporting crime reduction efforts. With this in mind, we also explore how the output of such models might be used to support crime analyst decision making—developing a prototype visualisation dashboard that could be deployed by police analysts in problem scanning and analysis (Eck and Spelman 1987; Weisburd et al. 2008).

The remainder of the paper is set out as follows—we begin by outlining the data and methods used to conduct our analyses, subsequently we describe our primary results. Next, we describe a prototype dashboard that combines topic modelling output with traditional spatial and temporal crime data to support crime analyst decision making. To conclude we discuss the limitations and implications of our work and outline several potential avenues for further research.

Data and methods

Data used in this study relates to 9621 incidents classified as residential burglary occurring over an unspecified, consecutive 2-year period in a major metropolitan region of the UK. These data contain a unique offence ID, the time and date the offence occurred, the location the offence occurred aggregated to UK postcode sector,[1] and the 'modus operandi notes' free-text account of the incident recorded by police personnel. For the majority of our analyses it is this free-text data that is analysed—and subsequently re-joined to the other data described above to provide further context. Due to recording practices, the narrative text was a document of a maximum of 11 ½ lines in length that were disclosable in court. These entries capture important behavioural or environmental information about the events surrounding or leading up to a residential burglary. Examples can be seen in Table 1. Prior to obtaining this data, all free-text was cleaned in order to ensure no records contained identifiable information. For the remainder of this section we now describe how this modus operandi text was prepared and analysed.

All work in this paper was performed using Python 3.6.8. Source code for the project can be found at https://github.com/QuantCrim-Leeds/Police-Free-Text-LDA-Dashboard.

Text pre-processing

Figure 1 depicts the steps undertaken to pre-process MO free-text documents for subsequent analyses. Initially,

[1] A postcode sector contains approximately 3000 address points. Our study area contains 150 post code sectors for ~ 21,000 postcodes.

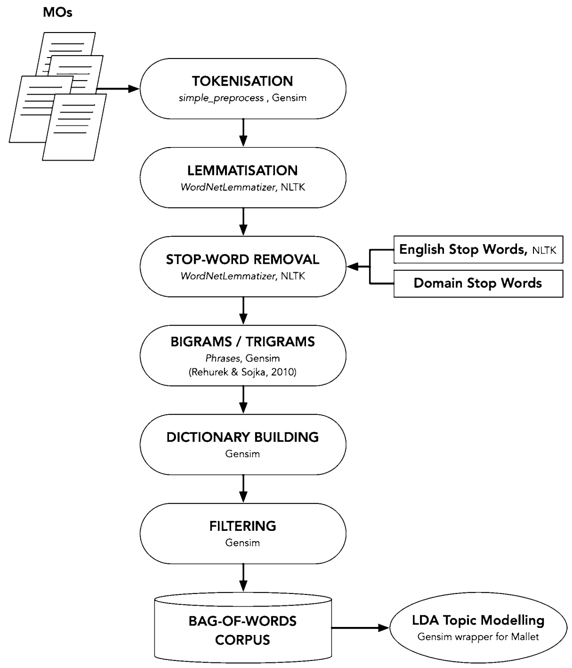


Fig. 1 Pre-processing workflow (including functions and packages used to implement each step)

the collection of all MOs (referred to subsequently as the corpus) were tokenised. Tokenisation converted individual MOs from single long character strings into lists of individual words strings. This process removed all punctuation, converted all text to lower case and removed accented characters. It was then necessary to use lemmatisation to reduce all inflected forms of words to their dictionary form (lemma)—thus ‘walking’ becomes ‘walk’–this process ensures that all forms of the same word are grouped together for the purposes of analyses. Next, stop words were removed from the corpus, these are words that convey little meaning about the subject of a sentence such as articles, adverbs and pronouns. A standard list of English stop words commonly used in LDA was used for this purpose. In addition, as is often the case in the application of LDA, the stop word list was extended with the following domain specific words that occurred frequently within reports but yielded little information for the topic model (‘suspect’, ‘victim’, ‘time’, ‘comp’, ‘complainant’, ‘stated’, ‘unknown’, ‘property’). The complete stop word list is shown in Table 2. Subsequently, commonly occurring word pairs (bigrams) and triplets (trigrams) within the corpus were identified and substituted with underscores replacing whitespace–for example ‘double glazed’ becomes ‘double_glazed’ allowing these common phrases to be captured in subsequent

Table 2 A complete list of stop words used

Stop words								
A	**Between**	**Few**	**How**	**My**	**Re**	**Theirs**	**Wasnt**	**Your**
About	Both	For	I	Myself	S	Them	We	Youre
Above	But	From	If	Needn	Same	Themselves	Were	Yours
After	By	Ft	In	Neednt	Shan	Then	Weren	Yourself
Again	Can	Further	Into	No	Shant	There	Werent	Yourselves
Against	Comp	Had	Is	Nor	She	These	What	Youve
Ain	Complainant	Hadn	Isn	Not	Shes	They	When	
All	Couldn	Hadnt	Isnt	Now	Should	This	Where	
Am	Couldnt	Has	It	O	Shouldn	Those	Which	
An	D	Hasn	Its	Of	Shouldnt	Through	While	
And	Did	Hasnt	Its	Off	Shouldve	Time	Who	
Any	Didn	Have	Itself	On	So	To	Whom	
Are	Didnt	Haven	Just	Once	Some	Too	Why	
Aren	Do	Havent	Ll	Only	Stated	Under	Will	
Arent	Does	Having	M	Or	Such	Unknown	With	
As	Doesn	He	Ma	Other	Suspect	Until	Won	
At	Doesnt	Her	Me	Our	Suspects	Up	Wont	
Be	Doing	Here	Mightn	Ours	T	Ve	Wouldn	
Because	Don	Hers	Mightnt	Ourselves	Than	Very	Wouldnt	
Been	Dont	Herself	More	Out	That	Victim	Y	
Before	Down	Him	Most	Over	Thatll	Victims	You	
Being	During	Himself	Mustn	Own	The	Was	Youd	
Below	Each	His	Mustnt	Property	Their	Wasn	Youll	

analyses. These steps were applied to the corpus as a whole prior to the construction of the bag-of-words dictionary.

Using the pre-processing corpus of documents represented by tokens, a dictionary was then created of string tokens to integer token IDs—this approach attributes each unique text string found within the corpus a numeric value. Subsequently, a filtering process was applied to the corpus. This entailed the removal of very rare (occurring in less than 10 documents) and very common (occurring in more than 70% of documents) words. Filtering is common in applications of natural language processing and seeks to both reduce the number of dimensions of a dataset to aid with computational efficiency, while also retaining meaningful words that are important in the corpus. It is often applied in LDA to encourage topic models to train on more meaningful words and thus generate meaningful topics. Given an absence evidence to direct parameter selection, thresholds for the filtering process were initially selected in a hope to aid with computational efficiency. Subsequently, a series of exploratory experiments using the MO dataset demonstrated that these thresholds reduced model runtime with minimal impact on the topics identified.

Collectively these processes generate a bag-of-words corpus–a representation of the corpus vocabulary (a list of unique words, bigrams, trigrams etc.), and the respective count of their occurrence within the corpus. Following the pre-processing steps described above yielded a bag-of-words corpus of 1857 unique tokens to be subsequently analysed with the topic models.

Topic modelling

The process of LDA based topic modelling involves identifying latent topics within text documents by iterating over the bag-of-words corpus to identify word co-occurrences. A topic is a distinct probability distribution over all the words in the vocabulary determined generatively by LDA through iterations through the corpus. It is possible to interpret topics by visualising the top seven to ten most probable words for each topic. Domain knowledge is then applied to label a topic (e.g. most probable words 'bank','river','duck' could be described as the topic 'riverside' compared to 'bank', 'mortgage', 'savings' might be described as 'finance'). Here, topics would be more probable terms identified from MO documents that would provide insight into a particular type of residential burglary offence.

Having prepared the bag-of-words corpus, LDA topic modelling was performed using Gibbs sampling for posterior approximation. The dictionary and bag-of-words corpus produced during text pre-processing are passed to the LDA model. The LDA model also has two additional hyper parameters that control the prior distributions for the topic probability distribution over documents (alpha) and words (beta). These parameters affect the model as follows: a higher alpha places more weight on documents being composed of more topics, whilst a low alpha places more weight on documents being composed of fewer topics. The beta hyper parameter controls similar prior probabilities for word-topic distributions, with a higher beta placing more weight on topics being composed of more words from the total corpus, whilst a lower beta places more weight on topics being composed of fewer words from the entire corpus. For our work these hyper parameters were unchanged from the implementation defaults of ($\alpha = 50$, $\beta = 0.01$) and the number of iterations was set to 1000. While an extensive discussion of parameter selection is beyond the scope of this study, the default values were selected given the exploratory nature of the research and a lack of further evidence to guide parameter selection. For a more detailed description of topic modelling, these parameters and their impacts on processing, readers are directed to (Blei and Lafferty, 2009).

Importantly, like the majority of topic modelling implementations, LDA requires researchers to select a number of expected topics that feature within the corpus to be analysed. Given our goal to discover previously unknown groupings of offences–it was clearly not possible to identify a priori the number of topics to use to topic model this dataset. In addressing this problem, previous exploratory applications of topic modelling have adopted different approaches to selecting the topic number such as working down from a perceived top-bound until a reasonable aggregation of topics is achieved (Röder et al. 2015) or arbitrarily selecting a topic number proportional to the number of documents used (Benites-Lazaro et al. 2018).

In this study we implement an alternate heuristic for selecting a suitable topic number that aims to consistently maximise the topic coherence score of the model. The topics subsequently described are LDA topics denoted as the top-ten words ordered by the probability of that word occurring in that topic. Topic coherence scores are measures for determining the user interpretability of topics produced by a topic model. They provide a measure of how well a set of topics generated by LDA describe the corpus of documents LDA has been trained on. Thus, in order to determine the optimum number of topics to derive the most insight from the MO corpus, a range of models with increasing numbers of topics were tested and the Cv coherence score evaluated for each model. In this case, Cv coherence was chosen because it has been shown out to best correlate with human rankings of LDA topics (Röder et al. 2015). The topic number

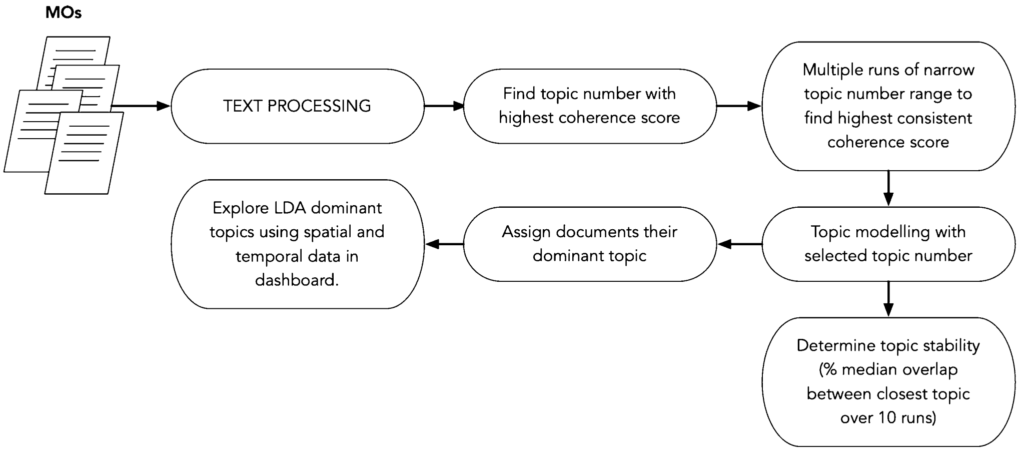

MOs
TEXT PROCESSING
Find topic number with highest coherence score
Multiple runs of narrow topic number range to find highest consistent coherence score
Topic modelling with selected topic number
Assign documents their dominant topic
Explore LDA dominant topics using spatial and temporal data in dashboard.
Determine topic stability (% median overlap between closest topic over 10 runs)

Fig. 2 Experimental work scheme for this paper

with the highest coherence score was then taken forward for subsequent dominant topic analysis.

A further consideration was how to ensure the generation of stable LDA topics. LDA is a non-deterministic method, representing documents and words as sparse Dirichlet distributions leads to different local maxima being reached each time the model is run. This is a well-known issue within the topic modelling field (Agrawal et al. 2018) and therefore steps were taken to determine over multiple runs how stable LDA topics were. This was determined comparing the top-ten words in topics from multiple runs and determining the percentage of topic-words shared by the most similar topics over different runs of LDA. Whilst this approach does not solve the issue of topic instability it does allow scrutiny of the stability of topics generated.

Once an LDA model was established with an optimal number of topics, documents were analysed using the model to determine their dominant topic. A dominant topic refers specifically to the topic that was assigned the highest probability for a given document. This was then used to cluster documents into dominant topic groups for subsequent analysis to determine whether LDA topics were able to group crime reports into meaningful clusters relating to similar MOs.

Results

A summary of our analytical approach described above is outlined in Fig. 2. It begins with the text pre-processing described above. Next, steps are performed to determine the optimum number of LDA topics using a topic coherence score that scores how well LDA topics describe the entirety of a corpus of documents. Finally, to ensure the high coherence score was reproducible, the topic number from the previous step was compared again across a narrower range of topic numbers over multiple LDA runs. The topic number with the highest median coherence score and the narrowest interquartile range was selected for further experimentation. LDA was then performed with this topic number to create the working model, from which all subsequent analyses were performed. This model was then tested to determine the stability of topics and used to assign documents with a dominant topic label (based on the most probable LDA topic for that document). These labelled documents were then used in a prototype dashboard to allow visualisation of the spatial and temporal distributions of identified topics.

Text processing

An overview of the most common words in the corpus is shown in Table 3. This highlights expected vernacular for narrative reports with domain specific abbreviations such as 'extp' and 'entp' (exit point and entry point respectively) and common descriptive features around properties: 'door', 'dwelling', 'house', 'window', 'detached', etc. Pre-processing yielded a tokenised corpus that was converted into a bag-of-words format that reduced the corpus by removing overly common or very rare words.

Topic number selection

Once text was processed into a bag-of-words format we then use the heuristic method described above to determine the number of topics for latent Dirichlet allocation to identify within the corpus. For this study the initial topic number from the wide coherence score process was 17 as shown in Fig. 3a, this was tested in the repeat coherence score process and 21 was selected (shown in Fig. 3b) as the final topic number based on the criteria outlined above.

Table 3 The top 25 most common words in the full unprocessed corpus of 358,949 words with exact counts and proportion of the word amongst all words shown

Word	Count	Proportion
Door	13,031	0.052
Dwelling	10,819	0.043
House	9217	0.037
Rear	8226	0.033
Window	6473	0.026
Entp	5714	0.023
Front	5473	0.022
Occupied	5336	0.021
Seen	4801	0.019
Detached	4097	0.016
Extp	3896	0.015
Suspect	3534	0.014
Unknown	3523	0.014
Property	3415	0.014
Semi	3070	0.012
Entry	3056	0.012
Lower	2918	0.012
Single	2767	0.011
Lock	2390	0.009
Attack	2328	0.009
Terraced	2302	0.009
Aentp	2261	0.009
Side	2017	0.008
Victim	1965	0.008
Kitchen	1879	0.007

Having identified the number of topics that maximises both topic coherence and stability, the working model was generated for subsequent investigation. This produced 21 LDA topics, each represented as a list of words that co-occur with a high probability within that topic. Table 4 depicts the top seven keywords (in keeping with research concerning human comprehension (Miller 1956)) occurring in each of the 21 topics and their associated probabilities. On initial inspection, these LDA topics already demonstrate interesting sets of co-occurring words within Residential Burglary MOs. With distinctions appearing between topics that capture information around specific objects i.e. ‘vehicle’, ‘keys’, ‘house’, and ‘car’, compared to topics that capture more descriptive words associated with the environment i.e. ‘area’, ‘premises’, ‘attacked’, ‘residential’.

Topic allocation

Next, documents were clustered using the LDA topic distribution to allow for the comparison of original narrative reports and to determine whether LDA topics identified specific crime MOs. LDA assigns each document a probability distribution over the LDA topics based on word use which we used to cluster the documents into dominant topic groups (most probable LDA topic). The size of these clusters is shown in Fig. 4 highlighting that dominant topics are not distributed uniformly. More importantly this clustering allowed us to compare narrative texts within each LDA topic and determine whether LDA is a suitable approach to identifying similar crime MOs.

Topic stability

As discussed above, the non-deterministic nature of LDA dictates that a serious analytical consideration with LDA topic modelling is the stability of topics produced. If LDA is to be useful in applied settings it must be able to consistently generate a stable set of topics from the same corpus of documents. This limitation is a current area of active research, with a number of technical approaches suggested to improve topic stability (see Agrawal et al. 2018, Mantyla et al. 2018). While such approaches are beyond the scope of this study, for transparency we have reported the stability of topics generated by the working model. In Table 5 we outline each topic, its top seven keywords, and the median % overlap between its most similar topic in repeated LDA runs. Performing stability analyses, the working model was used as a reference model and nine repeats of LDA were performed generating ten runs of a 21 topic model (including the reference model). The top seven words in the reference model topics were then compared to the most similar topic (highest number of shared top seven words) in the repeat models, and a percentage score was given for the number of shared words (number of shared words/number of potential shared words (7)). These scores were generated for each comparison between the working model and the nine replicate models and the median percentage overlap figure is reported.

Overall, 17 of the 21 topics showed a median keyword overlap of greater than 50% indicating that within the most probable words of that topic at least four out of seven words occurred consistently during multiple model runs. However, four topics achieved 42.9% median keyword overlap (three key words out of seven) suggesting that whilst the top three most probable words of these topics were consistent there was a greater level of redundancy in the remaining four most probable words. As discussed above, topic stability is a known issue in topic modelling applications. Given the applied nature of our research we propose that the importance of stability ultimately be tempered by the analytical value derived from the topic model to end-users—in this case, crime analysts.

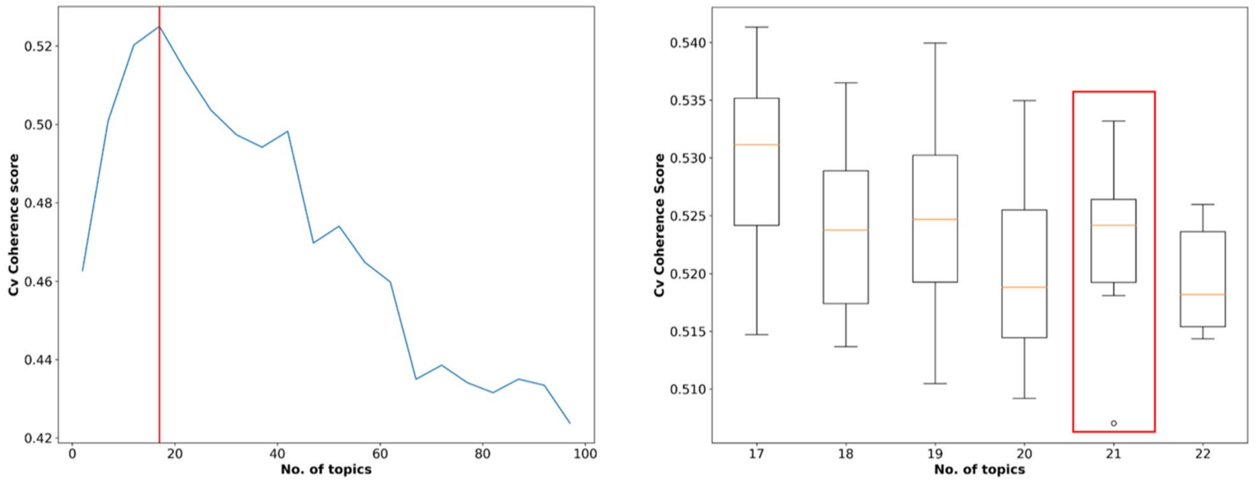


Fig. 3 Topic number determination through a two-step search process. **a** Long range search: a number of LDA models were generated using a range of topic (k) numbers (starting at 2, ending at 100, stepping by 5 each time) and coherence scores calculated and plotted. k = 17 was selected (shown by red line) as the topic number that yielded the highest Cv coherence score. b) Narrow range search: To ensure this topic number consistently gave a high coherence score LDA models were trained in multiple runs (× 10) for each topic number and coherence scores plotted as a box and whisker plot. During this stage the selected k number from the long range search was compared with models trained with topic numbers immediately within 6 steps greater than the long range selected k. From this narrow range search topic number 21 was identified as the highest and most consistent and selected for further experiments. These are plotted as box and whisker plots, with the median shown in yellow. The red box indicated the distribution of the final selected k topics number

Table 4 Top 7 keywords per topic and associated probabilities

Topic	Top 7 keywords and associated probabilities
1	Make (0.18); enter (0.15); remove (0.14); unseen (0.12); approach (0.06); insecure (0.05); direction (0.05);
2	Police (0.05); foot (0.04); male (0.03); witness (0.03); disturbed (0.02); occupant (0.02); hears (0.02);
3	Door (0.18); front (0.08); makes (0.08); enters (0.05); insecure (0.05); bed (0.04); approaches (0.04);
4	Door (0.2); force (0.1); open (0.06); bodily (0.04); bodily_force (0.04); wooden (0.04); garage (0.04);
5	Lock (0.18); handle (0.05); euro (0.04); entry (0.04); snap (0.04); profile (0.04); attack (0.03);
6	Area (0.16); premises (0.11); attacked (0.1); residential (0.09); dwelling (0.07); road (0.07); situated (0.05);
7	Attacked (0.07); premises (0.07); dwelling (0.07); quiet (0.04); detached (0.04); offender (0.04); sac (0.03);
8	Address (0.1); home (0.04); key (0.03); leaves (0.02); returns (0.02); whilst (0.02); aggrieved (0.01);
9	Vehicle (0.12); keys (0.11); house (0.06); car (0.06); locked (0.05); secure (0.05); driveway (0.04);
10	Entry (0.24); gain (0.17); make (0.14); direction (0.08); unseen (0.07); approach (0.07); times (0.05);
11	House (0.16); terraced (0.08); front (0.07); attacked (0.07); back (0.06); mid (0.06); terrace (0.06);
12	Damage (0.11); entry (0.09); causing (0.08); gained (0.07); implement (0.05); jemmy (0.04); frame (0.04);
13	Made (0.11); removed (0.11); entry (0.08); times (0.07); direction (0.07); means (0.07); approached (0.06);
14	Room (0.14); living (0.09); kitchen (0.08); living_room (0.07); bedroom (0.04); left (0.04); front (0.02);
15	Detached (0.18); semi (0.17); semi_detached (0.15); side (0.11); house (0.1); attacked (0.05); residential (0.04);
16	Floor (0.17); flat (0.14); ground (0.1); escape (0.06); good (0.06); good_escape (0.05); premises (0.03);
17	Door (0.32); locus (0.18); front (0.17); times (0.12); dates (0.03); leave (0.03); date (0.02);
18	Window (0.24); person (0.14); open (0.1); kitchen (0.06); rear (0.05); climb (0.04); bedroom (0.04);
19	Search (0.22); untidy (0.11); items (0.07); egress (0.07); tidy (0.07); rooms (0.06); removing (0.05);
20	Rear (0.11); window (0.11); smash (0.07); glass (0.06); alarm (0.04); large (0.04); reach (0.02);
21	Rear (0.24); garden (0.16); patio (0.07); doors (0.06); gate (0.04); access (0.03); side (0.03);

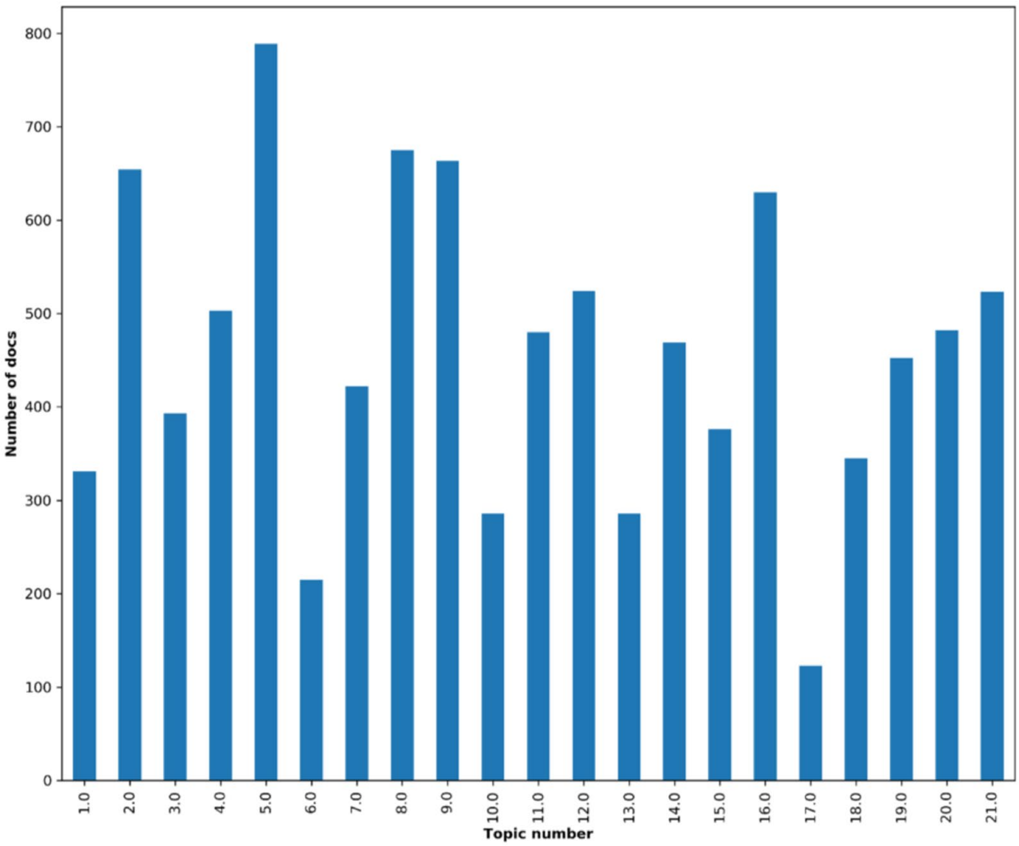


Fig. 4 The number of documents in each dominant topic

Topic inspection

Whilst topic keywords do not elucidate a recognisable MO, the clustering of reports into dominant topics (as described previously) aimed to, upon inspection, identify specific MOs. In Table 6, we can see examples of four topics, with their keywords, three examples of original crime reports within these LDA topics, and those document's dominant topic probability. On inspection it is clear that in each of the topics shown similar themes have been captured. In Topic 2 the three cases all describe incidents where successful burglaries have not happened but where witnesses have called the police either because they suspect a burglary may occur or because they have interrupted a burglary. Reports shown for Topic 9 describe incidents where suspects have entered a property and stolen the victim's car keys and then stolen the victim's car—a recognisable residential burglary MO well known to police analysts. Interestingly, two of the reports shown include a police operational tag (redacted for publication) which is used by the police to highlight burglaries that lead to a vehicle theft. The sample reports from topic 5 describe three burglaries where mole grips have been utilised by offenders in an attempt to gain entry to a property through UPVC doors—another recognisable MO. Reports from topic 18 all describe incidents where entry has been gained into a property via a window and where nothing has been taken. At the dominant topic level the approach has not made the distinction between the two accounts where the suspect has leveraged an insecure window and the report where the suspect has simply smashed the window. However, the fact LDA creates a probability distribution of topics over a document means the additional nuance of utilising an insecure window versus smashing a window may be discernible when accounting for the 2nd most probable topic (after the dominant topic).

Examining all topics (examples for which can be found in Appendix 1), it is clear that some topics produced by the model offer more meaningful clustering of offences with respect to what police analysts are likely to find useful than others. To illustrate, topic 16 has clustered crime reports relating to burglary offences that have occurred in properties separated into flats—something that may well be recorded through closed form responses associated with dwelling type traditionally recorded by police. Similarly, Topic 21 has clustered

Table 5 Topic stability comparison

Topic	Topic keywords	Median % overlap between closest topic
1	Make, enter, remove, unseen, approach, insecure, direction	64.3
2	Police, foot, male, witness, disturbed, occupant, hears	42.9
3	Door, front, makes, enters, insecure, bed, approaches	42.9
4	Door, force, open, bodily, bodily_force, wooden, garage	71.4
5	Lock, handle, euro, entry, snap, profile, attack	85.7
6	Area, premises, attacked, residential, dwelling	71.4
7	Attacked, premises, dwelling, quiet, detached, offender, sac	42.9
8	Address, home, key, leaves, returns, whilst, aggrieved	57.1
9	Vehicle, keys, house, car, locked, secure, driveway	71.4
10	Entry, gain, make, direction, unseen, approach, times	71.4
11	House, terraced, front, attacked, back, mid, terrace	85.7
12	Damage, entry, causing, gained, implement, jemmy, frame	57.1
13	Made, removed, entry, times, direction, means, approached	57.1
14	Room, living, kitchen, living_room, bedroom, left, front	85.7
15	Detached, semi, semi_detached, side, house, attacked, residential	71.4
16	Floor, flat, ground, escape, good, good_escape, premises	57.1
17	Door, locus, front, times, dates, leave, date	42.9
18	Window, person, open, kitchen, rear, climb, bedroom	71.4
19	Search, untidy, items, egress, tidy, rooms, removing	85.7
20	Rear, window, smash, glass, alarm, large, reach	57.1
21	Rear, garden, patio, doors, gate, access, side	71.4

Using the working model as a reference, 9 runs of LDA were performed and topic stability was assessed by counting shared topic keywords between the most similar topic. This was calculated as a percentage and the median was calculated over the 9 runs to give a representation of how stable a topic was given how often the words of the topic occurred together

reports together due to descriptions of the rear garden as the route of attack by the suspect(s). These examples demonstrate one potential limitation of LDA with the second report within this topic actually stating the 'suspect unknown has from direction unknown'. This report however has been included in this topic because of an earlier detailed description of the garden and additional access points to the property.

Nevertheless, given that the unsupervised algorithms applied here incorporate no prior information about the nature of these problems or the domain in which they are applied, in many cases the approach has clearly captured additional detail lost in administrative classification which may well have not been accessible through manual analyses—simply due to the amount of resources that would be required to categorise over 9000 crime reports. Overall, these examples support the idea that this and similar approaches could be utilised to capture more specific features of crimes, and in a number of cases such as the car key burglary and mole grip example above, a specific crime MO.

Mobilising insight

While the outputs of the topic model discussed above are insightful, we are also keenly interested in exploring the utility of combining them with other data, in support of the often time-poor crime analyst. To that end, a prototype dashboard application was developed which combined the clustering of crime reports into topics with traditional crime data, thus providing means to contextualise model outputs and explore any distinct spatial and temporal trends associated with particular crime report clusters as identified through the topic model.

This tool was developed to provide a user-facing platform that would allow for police analysts to rapidly interrogate spatial and temporal trends surrounding identified topics. It aims to provide a front-end for the topic model that allows police analysts to rapidly test their own hypotheses against the model, and act as an 'algorithm in the loop' providing additional insight and data to a domain expert who can then make informed judgements from this and other data sources about how best to

Table 6 A comparison of different topics and the narrative text clustered into those topics identifies specific modus operandi and ecology crime behaviours

Topic number and keywords	Dominant topic probability	Original text
Topic 2: *police, foot, male, witness, disturbed, occupant, hears*	0.353	Officers attend a call of two males in the rear of XXXX. The males made off from scene. A description of both males was passed and both males fitting that description in that immediate area at that time were arrested
	0.300	Between times stated suspects have been acting suspicious walking down driveways of properties on the street and looking through house windows. Witness calls the police as suspects are at the rear of a property on the street and not known what they are doing and not recognised. Witness calls the police. Police arrive at the scene and suspects located. Two suspects are detained at the scene other suspects make off on foot. Suspects detained and cannot account for why they are present. Suspects arrested for consiracy to commit burglary
	0.259	On date and between times stated the witness is at his home address and goes to his next door neighbours to borrow the lawn mower. As he goes to the garden shed he looks to his left and sights the suspect stood next to the window which leads into the garage. The witness sights the suspect with a screw driver and spanner undoing bolts which keeps the metal grill in place that leads into the garage. The suspects sights the witness and starts to run towards the witness. The witness who is in fear punches the suspect on the nose causing his nose to bleed. The witness and suspect have a wrestle and the witness sustains a grazed forearm and swelling to his right knuckles. The witness eventually detains the suspect and starts to call the police and the suspect pushes the witness causing him to stumble back and the suspects makes good his escape over the fields towards the XXXX
Topic 9: *vehicle, keys, house, car, locked, secure, driveway*	0.251	XXXXX Suspect/s have gained entry to the property via the ground floor kitchen windown forcing it open via unknown means. Once inside the suspect/s have removed a set of house keys that were in the kitchen door lock. This bunch of keys also had on it the ignition keys for XXXX vehicle that was parked up infront of XXXX property on the road side. The suspects have then used the vehicle keys to gain entry to it and start the engine. They have then drove off in the vehicle where they have then collided with a parked vehicle at the end of the same road and abandoned the stolen vehicle
	0.223	Suspects have entered insecure property via rear door. Suspects have removed car keys for a XXXX vehicle from kitchen work top. Using the keys suspects have then approached locked and secure motor vehicle and made off in same
	0.208	XXXXX Locus is semi-detached property. Between times and date stated victim believes that the front door was locked and secure, suspect/s approach front door and open the same using front door key and leaving the key in the lock on the outside of the door. (It is unknown how they have obtained the key) suspect/s enter into the hallway and remove car keys for vehicle XXXX from the victims coat pocket and egress through the front door to the victims car which is parked locked and secure on the roadside. Suspect/s enter the m/v using keys and make off in the vehicle from scene

Table 6 (continued)

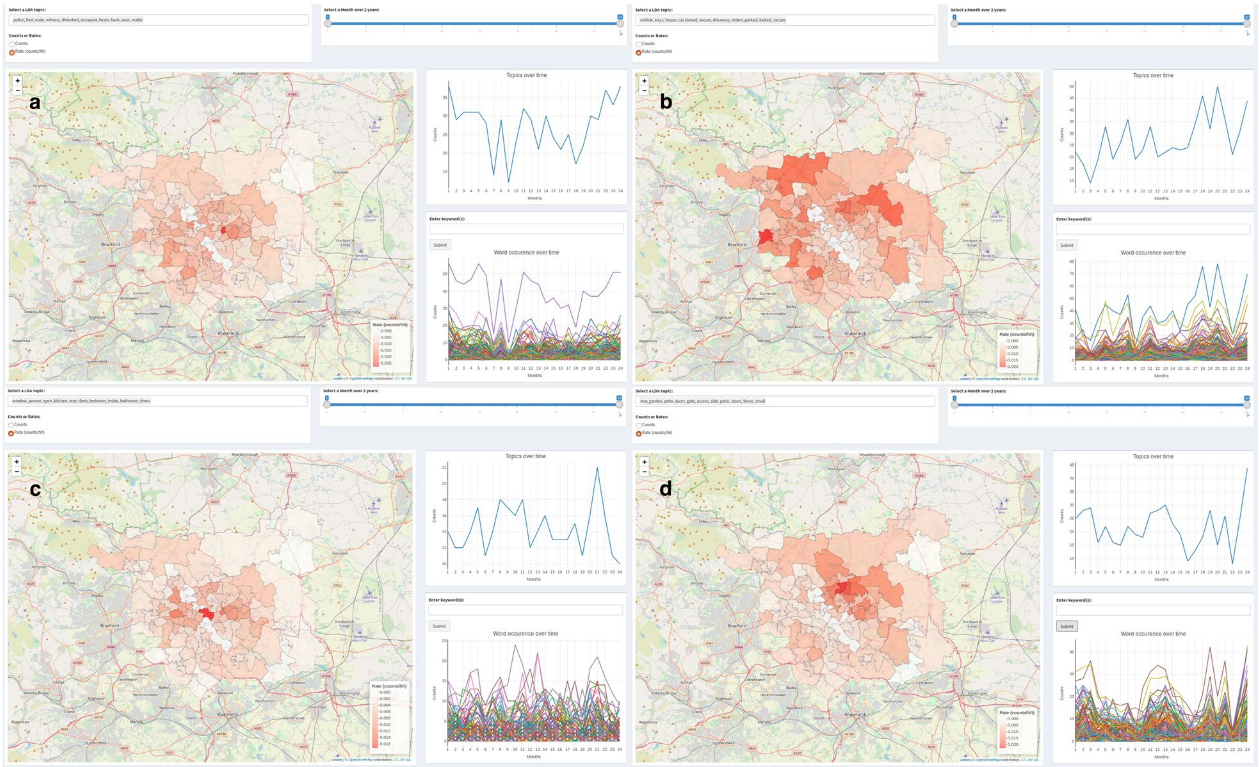
Topic number and keywords	Dominant topic probability	Original text
Topic 18: *window, person, open, kitchen, rear, climb, bedroom*	0.223	Attacked premises is a 3 bedroomed semi detached house off a main road, with a garden to the rear with a rear single storey extension and decking area and a block-paved driveway to the front and right side. Between times stated the family are all staying out for the night. Person(s) unknown approach rear of property, climb on the extension roof and jemmied the small top window open causing the window to come away from the frame and open. Person(s) have reached in and opened the larger left side window and climbed inside the premises. Person(s) make untidy search in upstaris bedrooms going through all jewellery boxes and ignoring electircal items. Persons egress as entry and make off unseen in unknown direction. (Believed nothing taken, tbc by victim once they have had chance to look properly)
	0.186	Attacked premise as a mid-terrace property that is set back from a main road by a front garden. During times stated the victim has been asleep in bed. He has deliberately left the living room window slightly open and closed the curtains. In the morning he has come downstairs to find the curtains open and the window pulled fully open. The bin outside has been slightly moved. Victim suspects that someone has used the bin to reach the window, open the curtains and look inside. Nothing stolen
	0.164	Premise is a ground floor flat. Between times stated bathroom window has been smashed with a brick. suspect has reached in and opened insecure transom window and climbs through window leaving footprint marks on sink and bath. No apparent search made of premise and egress through front door
Topic 5: *lock, handle, euro, entry, snap, profile, attack*	0.260	Between times and dates stated unknown suspect/s attend locus via unknown means and direction, then use mole grips or similar items to try and pull away the casing around the door handle to expose the euro profile lock, due to quality of the casing the casing and locks remain intact, suspects also try and jemmy above the euro profile by using screwdriver or similar article in attempt to remove or damage euro profile lock on the side door. No entry gained and no items stolen
	0.239	Mole grip: locus is a semi detached property on a residential street. Between times stated, suspect/s have entered rear garden and attack rear uvpc door by using mole grips to remove lock and barrel, also surrounding the lock are burnt marks and believe suspect/s also use blow torch on lock. Comp has anti snap locks fitted on all doors and suspect/s unable to gain entry. No property stolen
	0.218	Attacked property is a large detached dwelling on a busy road. Property is surrounded by large fences, gates and bushes. Between times stated suspect approach rear patio doors at locus and attempt to gain entry by using mole grip type implement to snap lock. Lock snapped however unable to gain entry. Suspects then use molegrip type implement to snap lock on front porch door. Lock snapped, door opened and house alarm sounds. Suspects jump over wall at front of dwelling, get into vehicle parked opposite and make off down XXXXX in direction of XXXXX

The topic number, topic keywords (top 7 most probable words), the dominant topic probability for that document, and the original crime report text

triage and maximise the effectiveness of crime reduction decisions.

To accomplish this, LDA dominant topic clusters were added as topic labels onto the original dataset provided and relevant columns were extracted and used as a data source for a prototype dashboard application built using the R language library *Shiny* (Chang et al. 2017). This utilised the four letter sector postcode location and month of offence data (24 months total) included for each free-text report to create maps presenting counts (or rates) of each LDA topic. Reports were then aggregated into medium super output areas (MSOAs) and rates calculated as the rate of residential burglaries per 1,000 households in the MSOA.

The temporal data allowed for the introduction of a slider that could be used to visualize the distribution



a

Select a LDA topic: police, foot, male, witness, disturbed, occupant, hears, back, sees, males

Select a Month over 2 years: 1 – 24

Counts or Rates:
○ Counts
● Rate (counts/hh)

b

Select a LDA topic: vehicle, keys, house, car, locked, secure, driveway, stolen, parked, locked_secure

Select a Month over 2 years: 1 – 24

Counts or Rates:
○ Counts
● Rate (counts/hh)

c

Select a LDA topic: window, person, open, kitchen, rear, climb, bedroom, inside, bathroom, times

Select a Month over 2 years: 1 – 24

Counts or Rates:
○ Counts
● Rate (counts/hh)

d

Select a LDA topic: rear, garden, patio, doors, gate, access, side, patio_doors, fence, small

Select a Month over 2 years: 1 – 24

Counts or Rates:
○ Counts
● Rate (counts/hh)

Fig. 5 Screenshots of the Shiny dashboard application interface rendered for the four example topics shown in Table 6 for the complete 24 month period of data provided with counts shown. Includes temporal plot of counts of topic over time as top-right chart and distribution of word use in selected topic over time in button-right. **a** Topic 2. **b** Topic 9. **c** Topic 18. **d** Topic 21

of crimes associated with a given topic from month to month or to show aggregate counts over a specific period. Examples of the four topics shown in the previous table rendered within the dashboard app can be seen in Fig. 5. These show distinct geographical separation of these LDA topics over the entire 24 months of data provided. On clicking on an MSOA the application also renders the MO text of the crime reports associated with the selected topic occurring in that MSOA—with the aim of allowing analysts to identify and extract these incidents for further scrutiny.

In addition, functionality was included to graphically visualise the counts of burglaries associated with each LDA topic over time (shown in the line chart in the top-right of each subfigure in Fig. 5). This allows crime analysts to rapidly examine the longitudinal trends in a particular topic—thus permitting identification of increasing, stable and declining occurrences of particular clusters of MOs as identified by the model. Another feature included to aid the work of police analysts was a searchable keyword over time chart (shown as the bottom-right chart in Fig. 5, populated with all words in selected topic as an example). This uses tokens produced for each report at the pre-processing stage and would allow police analysts to identify changes in word use that could relate to evolving criminal behaviours or recording practice.

Discussion

A significant amount of information is captured as unstructured free-text data when police record a crime. This data usually includes a narrative description of the events surrounding and including the incident in which a crime has occurred—the *modus operandi*. These data include unique behavioural and environmental information about the circumstances of a crime that may aid in the identification of specific crime problems, and ultimately support those who seek to devise problem-specific crime reduction efforts. Yet currently, due to a number of significant constraints, such unstructured text data are underutilised to proactively identify trends in offending.

In response to this problem, this paper set out to develop a systematic framework for analysing unstructured MO data using machine learning methods, in order to identify distinct crime problems traditionally aggregated within a single administrative crime category. While building on previous work that has applied related methods to identify the ecological similarities *between* crime types, this research represents the first use of topic modelling to identify latent topics from crime narrative text *within* a specific crime classification. To illustrate this approach, we applied dominant topic labelling to cluster residential burglary MOs into specific topics which sought to correspond to distinct types of offending. Subsequently, exploring the potential practical utility of this approach, clustered data was combined with traditional recorded crime features such as location and time of offence in a prototype dashboard devised to be used by crime analysts to understand more about how specific burglary problems were distributed both spatially and temporally.

Results of the example implementation of our framework highlight the potential of this approach in automating tasks that otherwise would require considerable crime analyst resources and would likely be subject to range of unavoidable biases. The techniques presented here demonstrate a practical pipeline for tapping into rich free-text data routinely collected by police, and supporting the identification and analyses of specific behavioural and environmental characteristics of crime. While such analytics cannot, and do not seek to, replace the judgement of human analysts, in the future they may provide means to rapidly 'pre-process' large quantitates of data into more tractable datasets. Most importantly, this approach should enable 'human algorithms' to devote their efforts to the more complex tasks of understanding specific problem behaviours and devising potential solutions to them.

In the future we foresee considerable potential for increasing the systematic proactive use of free-text data in policing. Thus, while our framework undoubtedly still requires refinement, we now briefly discuss several potential extensions that could improve, and capitalise on, the approach presented here. First, while our results demonstrate the potential utility of this approach in analysing residential burglaries MOs, it remains an empirical question if the techniques presented here can extract similar insights from reports associated with other types of offending. Burglary MOs are often recorded in a more procedural manner relative to other offences, and this may confer advantages in utilising natural language processing techniques. Consequently, further research is required to assess the applicability of the technique to free-text associated with other offences.

Second, while our model sought to classify all residential burglaries there is likely scope to develop more specific topic models that capitalise on contextual information currently available through the recording process. To illustrate, one might explore the impact of estimating separate models for offences that occur overnight and during daylight hours —or where offences take place in flats and houses (assuming dwelling type is routinely collected by an agency). This approach might increase the likelihood that topic models extract meaningful differences with respect to particular types of offending. Alternatively, such questions may be explored through the application of *structural* topic models (Roberts et al. 2014), an extension of LDA which allows for the incorporation of additional variables into the model (e.g. day/night), and quantification of the degree to which they covary with topic prevalence. Sadly, in this study such ancillary data were unavailable to test whether this was the case. Similarly, there are other more advanced NLP techniques, such as those which incorporate word embeddings—techniques which attempt to encode the semantic meaning of words in a document—that might be explored in this context and would enable the model to capture more semantic information, something that is not captured in our current bag-of-words LDA approach. Nevertheless, given the largely black-box nature of these approaches, we remain cognisant of the importance of explainability of outcomes in the current context.

Third, our approach taken to cluster MOs within a two year period was primarily made due to data availability (in terms of coverage both in space and time) and a desire to generate a sufficiently large corpus for analyses. A potentially important issue that requires further investigation is the degree to which MOs change over time—and, thus the temporal window over which topic models should be estimated to account for these potential changes. Approaches to topic modelling that account for temporal structure have been developed such as *dynamic* topic modelling (Blei and Lafferty 2006). This extension of LDA groups documents into time slices, allowing for topic-word distributions to evolve over time as word use changes. With data covering a longer time period this approach would certainly warrant exploration to determine how descriptions of MOs change over time. Nevertheless, the dynamic topic modelling approach is still constrained by having to select a fixed topic number, initially constraining the potential to identify emergent MOs that are not present from the initial starting time.

Fourth, following prominent research in the spatiotemporal analyses of volume crime, a natural advancement of the tools presented here would be to compare and contrast space–time structure of different types of burglary MOs as identified through the model. Following the underpinning of such research (Gill and Pease 1998), a number of questions spring to mind —namely, do residential burglaries that cluster in space and time also cluster in method of commission? Research that explores such similarities and differences in criminal behaviours would seem likely to offer significant insights for crime reduction efforts.

Finally, a timely extension of the approach would be to utilise the probability allocation of topics to documents in an attempt to identify new or emerging crime MOs from documents that are not well categorized by LDA when trained against an existing model. This approach might support the development of early warning systems that could identify emerging criminal behaviours to help police and their partners better adapt to the changing nature of crime problems seen in modern society.

With that said our approach also has a number of limitations that warrant discussion. First, in the applied context in which we have proposed these methods, it is important to reiterate that the practical utility derived from groupings of documents will be obviously be variable depending on their intended use. That is, that there may be MOs that are clustered together for reasons that are not of meaningful use to crime analysts. This limitation is of course inherent to the unsupervised approach—which incorporates no information about the subsequent use of such groupings. In response to this criticism, we would argue that without such an approach—weaknesses and all, proactive analyses of MO text at scale is at-best very rarely undertaken—simply due to the substantial resources it requires. Thus, while it would be optimal if all groupings of offences identified were practically useful, any that can provide insight to crime analysts represent added-value in the analytical process that would otherwise not be realised. Relatedly, it is important to note that the unsupervised nature of the clustering methods presented here do limit the degree to which a meaningful evaluation of their utility can easily be undertaken. Such an evaluation would require the labelling of known MOs within a dataset which could subsequently be used to test the validity of the model in clustering known similar reports. However, this process would be highly resource intensive requiring multiple domain experts to label a large quantity of data. Consequently, we suggest the most pragmatic method of evaluation is through providing the tools presented here to crime analysts—those who are best placed to identify their strengths and weaknesses in conducting the analyses they have been devised to support. This was a key rationale in our creation of the prototype dashboard.

Second, as previously described, LDA can suffer from topic instability which can hamper the technique's repeatability. While a number of potential solutions have been proposed (see Agrawal et al. 2018) this remains an ongoing research problem. Consequently, in this paper we took steps to report the stability of topics generated, with future work set to investigate if this has a significant effect on dominant topic clustering and more importantly the analytical utility derived from topic labels.

Third, our heuristic for selecting the number of topics existing within the corpus, while specifically optimised for human understandability may not be the most appropriate method for identifying all the relevant features of MOs within this dataset. In examining some of our clustered crime reports it is possible to see there may be further topics a report could be separated into. As such, additional work is required to determine how best to identify topic number before a model is trained. While attempts were made to utilise a topic modelling approaches that did not require a specified number of topics: hierarchal Dirichlet process (HDP), tests showed that the technique did not cluster reports into distinct recognisable MOs as were observed using LDA. Nevertheless, further experiments using HDP could be pursued as an alternate means of overcoming the topic number selection problem.

Relatedly, clustered topics in the prototype dashboard are currently presented by the most common words within a topic. In some applications of topic modelling the dominant topic is subsequently manually interpreted and labelled by a human analyst. For example, a topic with common words *vehicle, keys, house, car, locked, secure, driveway* might become 'car-key burglary'. This approach has both strengths, in that it may increase interpretability of patterns; and weaknesses—in that labelling can be highly subjective and ideally requires the consensus of multiple domain experts. Moving forwards, it may be beneficial to provide functionality within the dashboard for analysts to manually provide labels for topics as they are scrutinised.

Finally, and most significantly, the usefulness of the techniques presented here are all constrained by the significant caveat of how police file their reports. This includes word choice, use of professional terms, abbreviations and discretion classifying events into crime types which all have the potential to strongly impact on insights gained from this approach. While a unified vocabulary for crime reports is not required (and may produce unintending consequences, such system gaming through word choice), to maximise the insight they can deliver, natural language processing methods do obviously require those who record modus operandi to provide a sufficient and unique text report that captures the fulsome account of the crime (to the best of their knowledge). Whilst steps could be taken at the data processing step to account for specific nuances in word use or abbreviations it is still a crucial consideration that the quality of the crime reports being used will as ever correlate to the quality of the topic model.

In this regard, one might argue that demonstrating the utility of methods such as the one presented here may incentivise police to maximise their efforts in recording information concerning incidents.

Conclusion

The exploratory work presented here outlines a novel analytical workflow for how unsupervised natural language processing techniques can be applied to police free-text data to automatically provide insights into specific crime problems that exist within a single crime category. Currently, much of the unstructured data captured by police is not proactively analysed and the work described above attempts to set out a robust and reproducible method that could be utilised in the near future by those who routinely analyse free-text crime data. In an age of growing resource pressures, developing new techniques that can harness data routinely collected by police services seems a worthy goal in supporting police services deal with the myriad of demands they are faced with. While traditional crime classifications serve well to delineate crimes into recognisable types, such administrative classifications are often too coarse to derive crucial crime-specific details. By applying advances in machine learning and related methods, police and their partners may be able to proactively increase their understanding of specific crime problems in service of developing more tailored crime reduction responses. To this end, we hope that these initial efforts highlight the power of collaborating with crime reduction practitioners to develop innovative solutions to the diverse and changing crime problems faced by our communities.

Acknowledgements
Not applicable.

Authors' contributions
DB and AC designed the study, analysed the data and authored the article. DJ advised on the study design, provided the data, and co-authored the paper. All authors read and approved the final manuscript.

Funding
This work was supported by Wave 1 of The UKRI Strategic Priorities Fund under the EPSRC Grant EP/T001569/1, particularly the "Criminal Justice System" theme within that grant and The Alan Turing Institute'.

Availability of data and materials
Source code for the project can be found at https://github.com/QuantCrim-Leeds/Police-Free-Text-LDA-Dashboard.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] School of Law, University of Leeds, Leeds, UK. [2] Leeds Institute for Data Analytics, University of Leeds, Leeds, UK. [3] Safer Leeds, Leeds, UK.

Received: 10 February 2020 Accepted: 17 September 2020
Published online: 07 October 2020

References

Adderley, R., & Musgrove, P. (2003). Modus operandi modelling of group offending: a data-mining case study. *International Journal of Police Science & Management, 5*(4), 265–276.

Agrawal, A., Fu, W., & Menzies, T. (2018). What is wrong with topic modeling? And how to fix it using search-based software engineering. *Information and Software Technology, 98,* 74–88.

Babuta, A., Oswald, M., & Rinik, C. (2018). Machine learning algorithms and police decision-making legal, ethical and regulatory challenges. London: Royal United Services Institute for Defence and Security Studies. Retrieved from https://rusi.org/sites/default/files/201809_whr_3-18_machine_learning_algorithms.pdf.pdf

Benites-Lazaro, L. L., Giatti, L., & Giarolla, A. (2018). Topic modeling method for analyzing social actor discourses on climate change, energy and food security. *Energy Research & Social Science, 45,* 318–330.

Bennell, C., Jones, N. J., & Melnyk, T. (2009). Addressing problems with traditional crime linking methods using receiver operating characteristic analysis. *Legal and Criminological Psychology, 14*(2), 293–310.

Bennell, C., Snook, B., Macdonald, S., House, J. C., & Taylor, P. J. (2012). Computerized crime linkage systems: a critical review and research agenda. *Criminal Justice and Behavior, 39*(5), 620–634.

Blei, D.M. and Lafferty, J.D., (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning* (pp. 113-120).

Blei, D.M. and Lafferty, J.D., (2009). Topic models. In *Text Mining* (pp. 101-124). Chapman and Hall/CRC.

Blei, D.M., Ng, A.Y. & Jordan, M.I., (2003). Lafferty, John, ed. "Latent Dirichlet Allocation". Journal of Machine Learning Research. 3 (4–5): pp. 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993.

Braga, A. A. (2008). *Problem-oriented policing and crime prevention*. Monsey: Criminal Justice Press.

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2017). Shiny: web application framework for R. *R package version, 1*(5).

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L. and Blei, D.M., (2009). Reading tea leaves: How humans interpret topic models. In Advances in neural information processing systems (pp. 288-296).

Chen, S. H., Santoso, A., Lee, Y. S., & Wang, J. C. (2015). Latent dirichlet allocation based blog analysis for criminal intention detection system. In 2015 International Carnahan Conference on Security Technology (ICCST) (pp. 73-76). IEEE.

Chohlas-Wood, A. and Levine, E.S., (2019). A Recommendation Engine to Aid in Identifying Crime Patterns. *Interfaces.*

Clarke, R. V. (1983). Situational crime prevention: its theoretical basis and practical scope. *Crime and Justice, 4,* 225–256.

Cornish, D. B. (1994). The procedural analysis of offending and its relevance for situational prevention. *Crime prevention studies, 3,* 151–196.

Data.gov.uk (2018) Home Office counting rules for recorded crime, https://data.gov.uk/dataset/695f6775-3e51-4dd4-911a-19575638384c/home-office-counting-rules-for-recorded-crime

Eck, J., & Spelman, W. (1987). *Problem-solving: problem-oriented policing in newport news*. Washington, D.C.: Police Executive Research Forum. https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=111964.

Fosdick, R. B. (1916). Modus operandi system in the detection of criminals. *Journal of Criminal Law & Criminology, 6*(4), 560–570.

Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems, 61,* 115–125.

Gill, M., & Pease, K. (1998). Repeat robbers: Are they different?. In Crime at work (pp. 143-153). Palgrave Macmillan, London.

Goldstein, H. (1979). Improving policing: a problem-oriented approach. *Crime & Delinquency, 25*(2), 236–258.

Kuang, D., Brantingham, P. J., & Bertozzi, A. L. (2017). Crime topic modeling. *Crime Science, 6*(1), 12.

Laycock, G. (2013). Defining crime science. In *Crime science*(pp. 25-46). Willan.

Mantyla, M.V., Claes, M. & Farooq, U., (2018). Measuring LDA topic stability from clusters of replicated runs. In Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (p. 49). ACM.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63*(2), 81–97.
Moro, S., Cortez, P., & Rita, P. (2015). Business intelligence in banking: a literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. *Expert Systems with Applications, 42*(3), 1314–1324.
Moro, S., Pires, G., Rita, P. & Cortez, P., (2019). A text mining and topic modelling perspective of ethnic marketing research. Journal of Business Research.
Oatley, G., Ewart, B., & Zeleznikow, J. (2006). Decision support systems for police: lessons from the application of data mining techniques to "soft" forensic evidence. *Artificial Intelligence and Law, 14*(1–2), 35–100.
Panasyuk, A., Yu, E. S. L., & Mehrotra, K. G. (2014). Controversial topic discovery on members of congress with twitter. *Procedia Computer Science, 36*, 160–167.
Pandey, R. & Mohler G. O., (2018). Evaluation of crime topic models: topic coherence vs spatial crime concentration, *IEEE International Conference on Intelligence and Security Informatics (ISI)*, Miami, FL, 2018, pp. 76-78, https://doi.org/10.1109/isi.2018.8587384.
Ratcliffe, J. H. (2016). *Intelligence-led policing*. Chicago: Routledge.
Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.
Röder, M., Both, A. & Hinneburg, A., (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining* (pp. 399-408). ACM.
Theft Act (1968), http://www.legislation.gov.uk/ukpga/1968/60/contents
Weisburd, D., Telep, C.W., Hinkle, J. C., & Eck, J. E. (2008). *The effects of problem oriented policing on crime and disorder.* https://www.campbellcollaboration.org/media/k2/attachments/1045_R.pdf.
Zhu, S. & Xie, Y. (2018) *Crime incidents embedding using restricted Boltzmann machines* https://arxiv.org/pdf/1710.10513.pdf

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC