Data & Knowledge Engineering 69 (2010) 303–315

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

A Language Modelling approach to linking criminal styles with offender characteristics

R. Bache [a,*], F. Crestani [b], D. Canter [c], D. Youngs [c]

[a] University of Glasgow, Glasgow, Scotland, United Kingdom
[b] University of Lugano, Lugano, Switzerland
[c] International Research Centre for Investigative Psychology, University of Huddersfield, England, United Kingdom

ARTICLE INFO

Article history:
Available online 13 October 2009

Keywords:
Offender profiling
Information Retrieval
Language Modelling
Investigative psychology

ABSTRACT

The ability to infer the characteristics of offenders from their criminal behaviour ('offender profiling') has only been partially successful since it has relied on subjective judgments based on limited data. Words and structured data used in crime descriptions recorded by the police relate to behavioural features. Thus Language Modelling was applied to an existing police archive to link behavioural features with significant characteristics of offenders. Both multinomial and multiple Bernoulli models were used. Although categories selected are gender, age group, ethnic appearance and broad occupation (employed or not), in principle this can be applied to any characteristic recorded. Results indicate that statistically significant relationships exist between all characteristics for many types of crime. Bernoulli models tend to perform better than multinomial ones. It is also possible to identify automatically specific terms which when taken together give insight into the style of offending related to a particular group.

1. Introduction

Since the earliest criminological studies it has been clear that broadly speaking criminals have characteristics that distinguish them from the general population. There have also been attempts to demonstrate that certain classes of crime are typically committed by people who have similar characteristics. It has also been claimed that what may be called the 'style' of the crime or the pattern of behaviour, typical of any set of crimes, relates directly to subsets of characteristics of offenders. This process of making inferences about significant features of an offender on the basis of the kinds of people who commit crimes in that style has often been called 'offender profiling'. In general such 'profiles' are drawn from the subjective judgement and experience of putative experts with little empirical basis for their claims. In a few studies, most notably Canter and Fritzon's study of arson [5], it has been demonstrated that there are empirically sound relationships between, inter alia, the age, psychiatric background and personal relationships of offenders and dominant features of the crime, such as the nature of the target and whether there was more than one linked incident. Farringdon and Lambert [7] used structured data to explore pairwise statistical relationships between offenders of characteristics and various identified crime features stored typically as categorical data.

Where crime the description is textual, the few empirical studies that have been carried out to develop models relating offence style to offender characteristics have relied on intensive content analysis procedures that derive categories from open-ended police and related data sources. The labour intensity of the work as well as problems of access has meant that

* Corresponding author.
E-mail addresses: bache@dcs.gla.ac.uk (R. Bache), fabio.crestani@unisi.ch (F. Crestani).

0169-023X/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.datak.2009.10.009

only limited archives have been examined. Furthermore the procedure of deriving content categories and assigning cases to those categories requires a high level of expertise and even with extensive training the content analysis can suffer from the unreliability inherent in subjective judgments. If the behavioural actions are stored solely as structured data this problem is removed. However, the price paid is that only those behavioural actions deemed in advance to be of interest can be considered. The nature of free text is that the police officer or other individual recording the crime can describe any information that might be considered relevant. There has been work to analyse free text within police databases by Chen et al. [6] but this identifies only named entities using natural language processing (NLP) to find links between crimes rather than for offender profiling.

Information Retrieval (IR) techniques offer an automatic method of analysing text and relating documents in a way that does not require parsing the natural language text under consideration. Fundamental to IR is the creation of a bag-of-words index where the frequencies of words in a document are counted but the word order and sentence structure are ignored. In recent years the advent of Language Modelling has impacted fundamentally on IR since language models are essentially generative. Instead of linking two documents (and in the context of classic document retrieval we consider the query to be just another document), we assume that there is some underlying stochastic process that generates documents. We shall argue that the behaviour of the offender generates some description of the crime and thus generative models have a natural psychological interpretation.

One consequence of a bag-of-words index as the input to any predictive model is the 'curse of dimensionality.' Since each separate item of vocabulary has a count, the number of dependent variables will run into hundreds if not thousands. This will often exceed the number of data points being used to train any model and thus making statistical regression impossible. Although it is possible to reduce the dimensionality of the input space by techniques such as Latent Semantic Indexing [12], this will then lose any direct relationship between the specific items of vocabulary and the characteristic in question. It is important here to understand the context in which the profiling technology proposed here will be used. It is not sufficient for a computer system merely to make an inference that a particular offender has certain characteristics, it will also be necessary for the system to give some indication as to why. Thus the models will require some degree of transparency. Since Language Modelling assigns probabilities to certain words it is possible also to identify which items of vocabulary and thus underlying behavioural actions that the model associates with each group.

Language Modelling has been used to link solved crimes to unsolved crimes with the purpose of prioritising suspects from a list of known offenders that have committed crimes of the same type [1]. Such techniques rely on the words and structured data used in the descriptions relating to features of the offenders' behaviour. The question we wished to address in this study was whether Language Modelling could identify common patterns of behaviour in groups of offenders and thus relate offender characteristics to criminal styles. In mathematical terms, this is in many ways similar although from a psychological perspective it is considered a quite different problem. We therefore looked at four crucial characteristics of an offender: age, gender, ethnic group and occupation, to determine if automated analysis of free text could identify differences between the terms used to describe crimes committed by offenders in different groups. Of course, in principle, this approach can be extended to any other characteristics of an offender that may be of interest, such as known usage of drugs, how far the offender has travelled to the crime site, a history of committing crimes of a different type (e.g., sex offenders with or without a history of property crime), marital status, sexual orientation or even personality characteristics. The limitation on such possible studies is a function of the availability of the criterion data about the characteristics of an offender, which must be derived from solved crimes where such information is recorded.

Age, gender and ethnic group are readily available in a reasonably (although it must be admitted not totally) reliable form in all police records and are of great theoretical and tactical interest and were thus a useful starting point for exploring the applicability of Language Models to the 'offender profiling' problem. Occupation is often recorded as a free-format string and is perhaps less reliable since it is not possible to corroborate readily by appearance and in many cases is not recorded at all. Nevertheless is was possible here to determine with some accurancy whether the offender worked or not.

There were four purposes of the investigation. Firstly, we wanted to see whether differences existed between groups defined by the four characteristics of age, sex ethnic group and employment status. Secondly, we wished to show which type of language model would perform better for this data since there are two types that may be applied to this kind of data: multinomial and multiple Bernoulli (explained in Section 3). Thirdly, it would be useful to determine if the language models could give us insight as to what the differences in style between the groups were. Fourthly, a possible use of this approach is attempting to predict the characteristics of an offender from the description of an unsolved crime. Although eye-witness descriptions of offenders can be unreliable they are unlikely to mistake sex or very broad age group. Determining ethnic group is more difficult although a white European is unlikely to be described as 'black'. However, there are areas where such a predictive model would be useful, for example where no witnesses were present such as certain burglaries, thefts and acts of vandalism. Since occupation (whether working or not) cannot readily be determined by appearance alone, this would potentially be applicable to more crimes. Thus we wished to establish if the models had any predictive power.

The rest of the paper is organised as follows: Section 2 describes the nature of the police digital archives and of the data to be analysed. In Section 3 we show how Language Modelling may be used to characterise criminal behaviour and provide a novel justification for its use. We also explain the two types of language model used here. Section 4 describes how the crime descriptions, recorded as a mixture of structured data and free text, are indexed for the purposes of analysis. In Section 5 we apply the models to 8 datasets representing different crimes types and show that there is a statistically significant difference between groups defined their characteristics. Section 6 looks more specifically at the features associated with each group.

This allows us to answer questions such as "How does a male assaulter differ from a female one?" and makes the models transparent, thus taking us beyond a traditional classification problem. In Section 7, we offer some conclusions.

2. Criminal data used for analysis

The data used in this study were extracted from a police distributed digital archive containing crimes committed over a four-year period in an inner city district. Since the data was not collected with this type of analysis in mind, certain processing of the data was necessary before analysis could take place. For each reported crime, details were recorded about how that crime was reported, known as the *features* and many of these relate to the offenders' behaviour. Where a crime was solved, details of each of the culprits were also stored. Where these details relate to the type of person concerned they are known as the *characteristics* of the offender.

2.1. Crime features

The three fields that gave information about how the crime was committed:

- Free text describing the method employed by the offender in an informal note form, typically consisting of 20 words.
- Zero or more feature codes (typically 1 or 2) where each code represents the presence of a specific aspect of behaviour from a predefined list.
- A single allegation code describing the type or sub-type of offence.

Henceforth we shall consider this information to be a single conceptual document. Crimes were grouped according to allegation code with similar offences placed in the same set. For example, there are eight possible allegation codes which describe different sub-types of burglary and these offences were grouped as burglaries.

By far, most of the information contained in a crime description was in free text form. Furthermore, the nature of free text is that it is entirely flexible in what the user, a police officer, can record subject to a limit of the size of the text field. This contrasts with the structured data in the form of allegation and feature codes where the information that can be stored must be determined in advance of the data being collected. Thus the behavioural features that correlate with the offender's can emerge in any analysis of this data. There were no initial hypotheses about which features may relate to which characteristics.

2.2. Offender characteristics

Only solved crimes were considered for the study and then only offences with a single offender were used. For each crime therefore there was information about the age, gender, ethnic appearance and occupation of the offender. Gender and age were recorded in an obvious way. Ethic appearance was stored as one of several codes which relate to predefined categories, e.g., *white European, dark European, Afro-Caribbean, Arab/Egyptian.* For some of the groups, the number of offenders was low, reflecting, inter alia, the ethnic composition of the area. The two groups identified with sufficient datapoints to permit analysis were white European and Afro-Caribbean. The occupation was stored as a free text field. Problems arose in the very large number of descriptions used – some of which were either abbreviated or mistyped. We therefore decided to focus on two large groups: those in work denoted by a profession and those who were unemployed, identified as *unemployed* or *unemployed joiner*, etc. A further group that could be subjected to analysis were students (including those still at school). A small number were identified as retired, a patient or a prisoner. In approximately 10% of the cases no information was available – the field was blank.

3. Language models of behaviour

The rationale for being able to distinguish groups of offenders from their actions rests on there being significant behavioural differences between these groups and these differences being revealed in the vocabulary used to record the crime within the criminal records. We argue in the rest of this section that Language Modelling provides a theoretically principled way of relating a document describing an offence to the group most likely to contain the culprit. However, first we briefly consider both an approach previously employed to analyse textual descriptions of crimes and an approach to profiling reliant information about the crime being entirely structured data.

3.1. Content analysis

In previous studies [3,4], features have been identified by means of content analysis of the text. Typically, a researcher reads all the documents and identify a set of features. Then, for each document in the collection, each feature is marked either present or absent. There is therefore an indexing vocabulary corresponding to the set of features which typically will

run into tens of terms. The dimensionality of the problem is reduced further by smallest space analysis. This gives a sufficiently small set of independent variables to perform multivariate regression.

Such a procedure necessarily requires human analysis of each crime description and is both time consuming and requires a degree of expertise by the indexer. It is also essentially a pre-coordinate form of indexing in that the researcher has to determine in advance what the features of interest are. The standard approach to indexing used in IR is to reduce the document to a bag of words which is post-coordinate in that nothing other than the stopwords are thrown away. Structured data such as the feature codes can be added to as extra 'words' to the bag thus allowing both forms of data to be combined. However, two problems arise. Firstly, the number of indexing terms now runs to hundreds or thousands. Secondly it is no longer true that each term represents an identifiable behavioural feature.

3.2. Using only structured data

Farrington and Lambert [7] performed a study on burglars and violent offenders in which all the information used to describe the crime was either categorical or numerical. No free text was automatically analysed. They sought to link characteristics of offenders and various identified features of the crimes by pairwise comparison. They were able to identify a number of statistically significant results between such pairs. As such it is a useful contribution to the research. However such a form of analysis has three drawbacks that we address here. Firstly, it requires that identifiable features are recorded as structured data when in many cases such information will be expressed as free text. This would specifically be the case where witness statements were being used in place of police reports. Secondly, only those features defined in advance can be, by definition, linked to characteristics and there is therefore no way that various features contained in the text, which correlate with particular characteristics, can emerge. Thirdly, there is no obviously way that a set of crime features can be combined in a theoretically principled way to make an inference about a particular offender characteristic.

3.3. Features and terms from free text

The relationship between automatically derived terms and behavioural features that would be identified by humans is complex and still not fully understood. There is certainly no one-to-one correspondence. Many terms in the document give no information about the offender's behaviour and thus do not relate to any identifiable feature, e.g., *suspect, unlawfully, address*. There will also be sets of terms that may all relate to the same feature since identical behaviour can be recorded in different words. For example, where an assault has the feature of being racist in nature the terms *racist, racial, racially* or indeed quoted terms of racist abuse would all indicate the presence of this feature. No automatable technique could be found to conflate the synonyms and remove the irrelevant terms. Furthermore the effort required to perform this task on the hundreds of terms by hand would be comparable to content analysis since it is difficult to identify what is and what is not a synonym out of context. For example, in the case of burglary the following four words may appear superficially related: *electricity, electric, electronic, electrical*. However the first two qualify the type of coin-operated meter raided for cash and the last two describe the type of goods stolen. This could not be known without some familiarity with the free text.

By using a Language Modelling approach the underlying relationship between terms and behavioural features can be exploited. This was demonstrated by linking individual offenders to unsolved crimes from past offending history [1] and provides evidence that vectors of terms contain feature information. Indeed, if there were not some identifiable relationship between the language used to describe actions and those actions then language would have little utility. The task, though, is to demonstrate that specific, automatable Language Modelling approaches do actually generate fruitful and reliable distinctions.

3.4. Document generation

The (unintended) consequence of an offender committing a crime is that, if reported to the police, a document will be created to describe it. Such a document will be the result of some investigation and will describe features of the offender's behaviour. We therefore argue that the offender's behaviour generates a document even though the putative author is actually a police officer. In fact the police officer is only partially free in the description s/he gives. Training will have limited the language on which the police officer can draw as will the particular data management system that s/he will be required to use. The police officer is thus an 'author' in a very restricted sense, but the document is definitely initiated and its content greatly influenced by what the offender does.

Human behaviour (criminal or otherwise) is not deterministic and so even if we knew the precise circumstances in which an offence was committed (which generally we do not) then it would be impossible to know exactly what the offender would do. Yet individuals and groups do exhibit patterns of behaviour in that they are more likely to perform one action more than another. The mechanism by which the actual events surrounding a crime are transmitted to the crime description is a noisy channel in that certain facts will be not be known and witnesses' recollection is never perfect. Given the richness of natural language, there will be any number of ways a set of circumstances may be recorded in writing even within the restricted vocabulary used by police officers. Thus, at best, we can see the generation of the document from the behaviour of one member of a group as a stochastic process; the relationship between the group of interest and the words used to

describe offences committed by them is probabilistic. For this reason a model based on probabilities is necessary to capture the complexity of the problem situation.

A language model [10,16] is a process that randomly emits terms (e.g., words) drawn from a predefined vocabulary. Each model will emit terms with differing frequencies and indeed a language model can be wholly characterised by the probability assigned to each term in the vocabulary. We argue that an offender generates behavioural features when committing an offence. Thus we can model an individual offender as having probabilities associated with each feature. Categories of offenders too will have probabilities associated with them although differences may be less pronounced. This will then be reflected in the probability of occurrence of terms in the documents generated by criminal activity.

Assigning crime reports to offender categories is a form of document classification. Researchers [2,14,15] have previously used language models for document classification and such an approach was essentially Bayesian. We too adopt a Bayesian approach but, in common with most IR applications, apply models that are *unigram* in that they consider each term independently and do not take account of the preceding tokens. The fact that we are using a mixture of terms derived from free text and structured data means that *n*-gram models that consider preceding terms would not be appropriate. Furthermore *n*-gram models require considerable training data both in terms of number and length of the documents which is not the case here.

The language model is thus a stochastic process that creates some vector of terms that corresponds to that created by automatic indexing. Strictly speaking, therefore, we should say that the language model generates the *index* of the document, but for brevity we shall say that we are generating the document. For each category of offender we will have one language model and for an unsolved crime we calculate which of these models was most likely to have generated it.

There are two types of unigram language model: multinomial and multiple Bernoulli. The former considers the number of occurrences each term has in a document and produces a non-negative integer-valued vector. The latter considers only whether a term is present of absent and so yields a Boolean vector where entries are 0 or 1. The original application of Language Modelling to IR applied the Bernoulli model [16] although multinomial models have become more frequently used in IR [10]. There are a number of reasons for believing *a priori* that Bernoulli models more closely match the properties of these data.

1. Terms derived from codes are either present or absent.
2. Terms from free text rarely appear more than once except stopwords (e.g., the, by, and) which were removed from the analysis anyhow.
3. Losada [11] shows that for question answering (QA), Bernoulli models work better. QA deals with very small sub-documents of a size not unlike the ones that are the subject of this analysis.

However, McCallum and Nigam [13] show that multinomial models start to outperform Bernoulli in document classification where the vocabulary is over 1000. The vocabulary sizes vary between datasets (see Table 1) some being greater and others less than 1000. Therefore we applied both types of model for the purposes of comparison.

3.5. Probability of generation

Without loss of generality we consider categorisation by gender. Thus we will have two language models: male (m) and female (f). For multinomial models, the probability of a given document, d, being created given the male model is:

Table 1
Details of datasets by crime.

Crime type	Number of crimes			Median age		Vocabulary size
Theft from vehicles	317			26		418
Other theft	380			25		808
Shoplifting	2381			29		1057
Assault	2230			31		1576
Criminal damage	934			26		1183
Damage to vehicles	253			27		471
Burglary	1292			27		1226
Robbery	352			18		632
	Crimes by group					
	Male	Female	White European	Afro-Caribbean	Employed	Unemployed
Theft from vehicles	244	4	177	65	35	96
Other theft	260	66	182	118	47	117
Shoplifting	1399	661	1157	641	208	1159
Assault	1691	382	1131	726	443	461
Criminal damage	725	124	493	289	134	199
Damage to vehicles	193	27	153	54	57	45
Burglary	1060	66	684	367	113	530
Robbery	252	11	59	182	24	96

$$p(d|m) = \prod_{t \in d} p(t|m)^{c(t,d)} \tag{1}$$

where $p(t|m)$ is the probability that the male model will emit term t and $c(t,d)$ is the count of occurrences of t in document d. For multiple Bernoulli models

$$p(d|m) = \prod_{t \in d} p(t|m) \times \prod_{t \notin d} (1 - p(t|m)) \tag{2}$$

since we consider not only the probability of a term being present but also the probability of a term being absent. We actually want to calculate the probability that an offender is male given we have seen the description of that crime, so we use Bayes' theorem

$$p(m|d) = \frac{p(m) \cdot p(d|m)}{p(d)} \tag{3}$$

where $p(m)$ is the prior probability that the offender is male. For most crime types, males are much more likely to offend than females. So, this can be estimated from the solved crimes as simply the proportion of offences committed by males. The denominator $p(d)$ is the probability that the document could have been generated at all. It can be eliminated by requiring that the sum of all probabilities is unity.

It is worth noting here that for other characteristics, the prior probability is based on the historic proportion of crimes committed by offenders with a given characteristic. In the case of age, since the median has been used to split the crimes into (roughly equal) sets, the prior will be very close to 0.5 – it will not be exact since the age is recorded as whole years. For occupation and ethnic appearance, the prior is based on the proportion of the crimes committed by members of each group. It is, of course, not possible to infer that all individuals from these groups are more or less likely to commit offences of a specific type without knowing the ethnic or occupational make up of the entire population. This information was not available and pursuit of such an investigation is, in any case, beyond the scope of this paper.

The solved crimes are used to calibrate the language models. We can estimate the probability that a language model, say the male one, will emit a given term t from the frequency with which it was emitted when generating all the solved male crimes. Thus for a multinomial model we have:

$$p(t|m) = \frac{\sum_{d \in M} c(t,d)}{\sum_{d \in M} |d|} \tag{4}$$

where M is the set of offences known to have been committed by males and $|d|$ is the size of document d. For a Bernoulli model

$$p(t|m) = \frac{|S_t|}{|M|} \tag{5}$$

where $S_t \subseteq M$ is the subset of documents containing the term t. These formulae define the *unsmoothed* language models, which if used to calibrate the language models, will result in poor performance because of the Zero Probability Problem (ZPP) as now explained.

3.6. Smoothing

The ZPP was identified as a problem in the application of Language Models [10,13] to IR. It refers to the problem of any term not appearing in the training data for a given category. The unsmoothed model will calculate the probability of any such term being emitted to be zero, and thus assign a probability of zero to any document containing it. For language models of behaviour, this has a clear interpretation.

A fundamental assumption of Investigative Psychology is that behaviour observed in the past is more likely to be observed in the future. But, if it has not been observed in the past, it cannot be assumed that it will never occur in the future – it is just less likely. The unfortunate consequence of an unsmoothed model is that it forbids novel behaviour. Furthermore, given the imprecise relationship between terms and features, the same behavioural feature might be described differently with new terms and this would mean a language model would also have a zero probability of generating it.

The standard solution to this problem is smoothing. This means adjusting the language model so that no term has a zero probability. There are numerous smoothing strategies [17]. Jelinek–Mercer [9] smoothing was shown previously to work well with crime data [1] and this also has a natural interpretation. We define one extra language model – a universal model that models the behaviour of every offender. This is calibrated using the entire dataset, including both solved and unsolved crimes. Thus the probability of any term being emitted will, by definition, never be zero. So, by mixing the past behaviour of male offenders with those of the universal offender we create a model that implies any future behaviour is possible, although past male behaviour is more likely. This can be expressed as:

$$p_{smoothed}(t|m) = (1 - \lambda) \cdot p(t|m) + \lambda \cdot p_{universal}(t) \tag{6}$$

where λ is the smoothing parameter set to 0.5.

4. Automatic indexing

Rendering free text into a bag-of-words index is an activity performed routinely in any IR application; text is tokenised, stemming is applied and stopwords are removed. However, here there were certain issues that are addressed here.

As we see in Section 6, the actual items of vocabulary are analysed and so the terms used should be readable by humans. In many IR applications the index is not actually viewed by the user so the fact that a stemmer, which applies a set of rules for removing endings, will mutilate the words and make some of them difficult to interpret is of no consequence. Here we used a lemmatiser rather than a stemmer – each word is reduced to its lemma, the form that appears in a dictionary. Not only does this give rise to a term that is a meaningful word but also can cope with the irregular strong verbs that feature very commonly in police text. WordNet [8] was used as a dictionary to look up irregular forms and ensure that when standard endings such as *ing* or *es* are removed the resulting stem is indeed a valid word.

The structured data (feature and allegation codes) appears in the police archive as two letter codes. Adding the codes directly to the bag-of-words would lead to three problems. Firstly an allegation code and feature code may be identical but will convey an entirely different meaning. Secondly, many of the two letter codes are valid two-letter words which tend to be removed as stopwords. Thirdly, the analysis of vocabulary would only be accessible to those with a good understanding of the (completely arbitrary) codes. Thus the codes were mapped to meaningful hyphenated phrases and preceded by a non-alphabetic character ($) so that they could be identified from terms which originated from the free text, e.g., $*victim-harassed*.

There was also a problem of gender and age specific vocabulary used in the description. In many cases the age and sex of the offender would be known when the document was created, e.g., if the offender was caught while committing the offence or a clear description was given by an eye witness. If there were words in the text which identified the age or sex of the offender, the language models would pick up on this so that would seriously bias the model. Therefore, sex and age specific vocabulary were either removed or converted to gender and age neutral proxies, e.g., *girlfriend* and *boyfriend* became *partner* whereas *youth, child, man* and *woman* became *person*, etc. Other words such as *young* were added to the list of stopwords.

5. Experimental procedure and results

The approach described here will work for any finite number of categories. Clearly when using gender as a characteristic, there will only be 2. For age there could, in principle be any number although to create one for each year of age would spread the training data too thinly but 3 or 4 categories would be sensible. Nevertheless it was decided that for the purposes of the experiment there would be just 2, above and below the median age henceforth known as *older* and *younger*. For the ethnic appearance data there was only sufficient data to allow the analysis in two categories: white European and Afro-Caribbean. For occupation analysis was carried out for two categories: employed and unemployed. Although there would have been sufficient data to analyse a third group, *student and school pupil*, this would correlate very highly with age and so this was not attempted.

Within each crime set there were a number of serial crimes – two or more crimes by the same offender. Including these in the analysis would create a possible source of bias. If, for example, a prolific female offender committed a number of crimes with a distinct Modus Operandi, this might be seen as defining female features when in fact it was just the behaviour of one female. We term this the serial crime problem.

An adaptation of the Leave-One-Out (LOO) method was adopted to validate the models. Usually, this would involve removing each data point in turn and using all others to train the models, simulating a scenario where there was already an archive of solved crimes and one new unsolved crime. However, to avoid the serial crime problem, we remove all the crime committed by each offender in turn and use all crimes not committed by that offender to train the models. Table 1 shows the size of the eight datasets with categorisation of the crimes.

5.1. ROC analysis

This form of analysis is applicable for any continuous variable being used to predict an item belonging to one of two groups. Here, we have the estimated probability that the offender for a given crime is either male, above median age, etc. By setting some cut-off point, the offences can be partitioned into two sets: upper and lower. Again, using classification by gender, for each possible cut-off point between 0 and 1, we expect more males to be in the upper set. A ROC graph shows the sensitivity of the model, the proportion of males in the upper set, against the specificity, the proportion of females in the lower set. A straight line implies that the model has no predictive power and is no better than random. A curve bowed above the 45° line implies that the model does have predictive power. A measure of the bowedness of this curve is AUC (Area Under the Curve) and values above 0.5 imply some predictive power. Fig. 1 shows the ROC curves for criminal damage for the age and gender models respectively using the Bernoulli formulation (the multinomial version would appear very similar). It demonstrates that whereas there are distinct behavioural differences based on age, these are weaker when considering male and female offenders. Tables 2–5 shows AUC for both models applied to the eight datasets. For two datasets there were not sufficient crimes committed by females to make analysis meaningful. Note that if the results were purely random, we would expect AUC to be below 0.5 as often as above it. Applying the Wilcoxon Ranked Sign Test to each of the categorisations allows

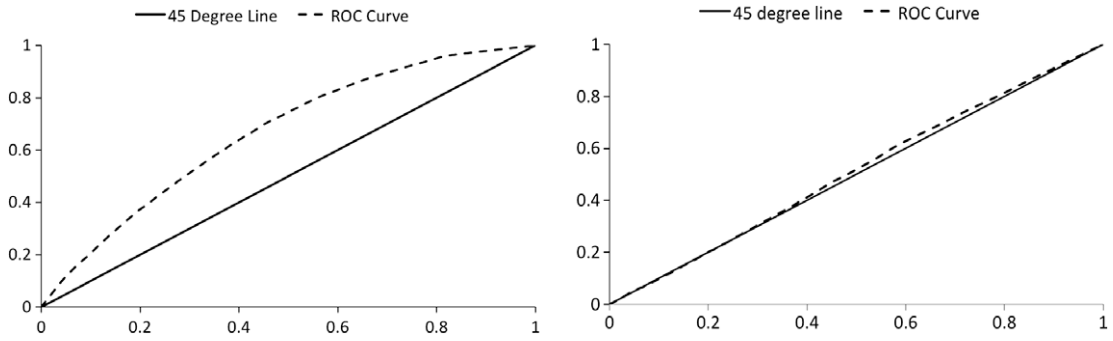


Fig. 1. ROC curve for age model (left) and sex (right) applied to criminal damage.

Table 2
Results of ROC analysis and Chi-squared test for gender.

Crime type	ROC analysis – AUC		Chi-squared test significance *p*	
	Multinomial	Bernoulli	Multinomial	Bernoulli
Other theft	0.558	0.564	0.483	0.269
Shoplifting	0.653	0.655	**<0.001**	**<0.001**
Assault	0.697	0.705	**<0.001**	**<0.001**
Criminal damage	0.557	0.558	0.219	0.455
Damage to vehicles	0.453	0.474	0.141	0.199
Burglary	0.584	0.588	**<0.001**	**<0.001**

Table 3
Results of ROC analysis and Chi-squared test for age.

Crime type	ROC analysis – AUC		Chi-squared test significance *p*	
	Multinomial	Bernoulli	Multinomial	Bernoulli
Theft from vehicles	0.638	0.623	**0.015**	**0.002**
Other theft	0.569	0.574	**0.046**	**0.022**
Shoplifting	0.586	0.587	**<0.001**	**<0.001**
Assault	0.564	0.561	**<0.001**	**<0.001**
Criminal damage	0.586	0.609	**<0.001**	**<0.001**
Damage to vehicles	0.586	0.610	**0.006**	**0.002**
Burglary	0.527	0.525	0.155	0.081
Robbery	0.677	0.690	**<0.001**	**<0.001**

us to conclude a relationship with age to 1% significance and with gender to 5% significance. For ethnic appearance and occupation it is to 5% significance using the multinomial model but 1% when using multiple Bernoulli. There are no striking differences between the performance of both types of models. However, if we use AUC as a measure of the performance of each model then we can use the two-sided Ranked Sign Test to determine if there were significant differences in performance of the models taking all crimes and all categorisations. We can conclude with 1% significance that Bernoulli models work better than multinomial ones. We also note that unlike McCallum and Nigam's study [13] there is no evidence that a larger vocabulary makes the Bernoulli model less effective.

5.2. Statistical significance of predictions

The AUC measure indicates the strength of the relationship between characteristics and features and the application of the ranked sign test indicates the models are detecting that these relationships are not random. However the AUC measure cannot determine whether the relationship is statistically significant for a given crime type and specific characteristic. For this we require a statistical test for significance that may be applied on one such instance. This will obviously take into account the size of the sample since, as the sample size grows, the probability of a difference of a given magnitude being due to randomness reduces. Thus we attempted to use the model to perform actual classification by choosing the optimal category – this being the one that maximises the posterior probabilities. Given that the models estimate the probability that the offender will be in one of two categories, we can set the cut-off point at 0.5. So if a crime shows a probability of more than 0.5 of being a male we assume the model predicts it to be a male. Otherwise we assume that we are predicting a female. A

Table 4
Results of ROC analysis and Chi-squared test for ethnic appearance.

Crime type	ROC analysis – AUC		Chi-squared test significance p	
	Multinomial	Bernoulli	Multinomial	Bernoulli
Theft from vehicles	0.648	0.648	**0.023**	0.164
Other theft	0.578	0.578	0.104	**0.029**
Shoplifting	0.525	0.527	**0.017**	**0.048**
Assault	0.543	0.540	0.171	0.133
Criminal damage	0.473	0.545	0.241	**0.029**
Damage to vehicles	0.654	0.652	**0.001**	**<0.001**
Burglary	0.573	0.574	**0.011**	**0.012**
Robbery	0.533	0.552	0.243	**0.019**

Table 5
Results of ROC analysis and Chi-squared test for occupation.

Crime type	ROC Analysis – AUC		Chi-squared test significance p	
	Multinomial	Bernoulli	Multinomial	Bernoulli
Theft from vehicles	0.501	0.524	0.099	0.122
Other theft	0.571	0.569	0.104	0.161
Shoplifting	0.561	0.559	0.355	0.283
Assault	0.558	0.557	**0.003**	**0.004**
Criminal damage	0.557	0.555	**0.041**	**0.033**
Damage to vehicles	0.627	0.632	**0.038**	**0.026**
Burglary	0.439	0.453	0.895	0.937
Robbery	0.544	0.526	0.285	**0.041**

Table 6
Statistical significant results for the multiple Bernoulli model.

Crime type	Gender	Age	Ethnic appearance	Occupation
Theft from vehicles		X		
Other theft		X	X	
Shoplifting	X	X	X	
Assault	X	X		X
Criminal damage		X	X	X
Damage to vehicles		X	X	X
Burglary	X		X	
Robbery		X	X	X

similar technique may be used for other characteristics. We can compare the predictions with the known category by using the Chi-squared test. Tables 2–5 show the level of significance with bold indicating significance at 5% for a one sided Chi-squared test.

Table 6 summarises the pairs for which a significant relationship was found using the Bernoulli model; this appears to be the more sensitive of the two models. We can see that all crime types are affected by at least one characteristic but no crime type appears to be affected by all characteristics. All characteristics affect at least three crime types.

The fact that some datasets do not show significance has two possible explanations. It may be that there was insufficient data and a larger dataset would pick up a (probably weak) correlation. However it may also be that there are no observable behavioural differences between the groups in the first place.

6. Analysis of vocabulary

Within traditional IR language models are used in a black-box fashion in that the actual probabilities assigned to particular words are not themselves of particular interest. Here we open up the language models in order to provide further analysis of what behavioural actions are associated with particular groups. This provides a degree of transparency to the models that will increase their credibility when used in the field. It also provides insights of a qualitative nature into how different groups differ in their commission of offences.

For each term in the vocabulary, we can determine to what extent it is associated with each category. Terms with a behavioural significance can be used to signify features more commonly associated with particular characteristics of offenders. So, we can define a measure of the maleness of a term as:

Table 7
Most common terms associated with gender characteristics.

Rank	Assault	
	Most male-related	Most female-related
1	punch	$common-assault
2	$victim-punched	parent
3	$actual-bodily-harm	slap
4	duspect	victim
5	sssault	$attack
6	head	accuse
7	partner	scratch
8	$victim-pushed	hit
9	$victim-threatened	face
10	$assault-section-eighteen	injury

Table 8
Most common terms associated with gender characteristics.

Rank	Criminal damage	
	Most older-related	Most younger-related
1	door	stone
2	victim	$criminal-damage-not-high-value
3	smash	cause
4	front	$graffiti
5	arrest	graffito
6	entry	property
7	make	spray
8	spouse	break
9	drive	wall
10	suspect	accuse

Table 9
Most common terms associated with ethnic appearance.

Rank	Damage to vehicles	
	Most white European	Most Afro-Caribbean
1	person	$attended vehicle
2	$vehicle-secure	suspect
3	door	bus
4	break	window
5	mirror	damage
6	wing	involve
7	smash	decamp
8	driver	write
9	car	van
10	arrest	house

$$p(t|m) - p(t|f) \tag{7}$$

where these are the probabilities of a given term being emitted from the unsmoothed language models of either type. A measure of maturity can be defined similarly using the younger and older models. Measures of affinity with ethnic appearance and occupation can be identified in the same way. We can then rank the terms by this measure and link the terms most associated with individuals with a given characteristic. Tables 7 and 8 show the top 10 terms associated with offender characteristics for assault with gender and criminal damage with age using here the multinomial model. Each model type will rank vocabulary slightly differently but the results are broadly similar. Note that some terms are proxies to avoid gender-specific terms. Examples are *parent*, *partner* and *spouse*. Note also that terms beginning with a $ sign are feature or allegation codes. All others come from the free text.

Inspection of the terms alone shows some distinct patterns. A typical style of assault by a male will involve *punching* and injury to the *head* as well as pushing and threatening, whereas females *scratch* and *slap*. Males are more likely to assault a girlfriend (*partner*) whereas for females the violence is inter-generational. It is not clear from the terms

Table 10
Most common terms associated with occupation characteristics.

Rank	Damage to vehicles	
	Most employed-related	Most unemployed-related
1	damage	$vehicle-secure
2	$attended vehicle	person
3	kick	unknown
4	cause	nearside
5	incident	$sharp instrument
6	road	front
7	rage	rear
8	$blunt-instrument	window
9	drive	unattended
10	mirror	smash

whether the assaults are of it parents or by *parents* although further inspection of the actual text reveals the latter is actually more likely. Male assaults appear to be more serious – *section 18 assault* and *actual bodily harm* are more serious allegations than *common assault*. For criminal damage, typical younger behaviour appears to be throwing *stones* and *spraying walls* with *graffiti*. Adults are more likely to *smash windows* and *kick* in *doors*. The damage appears to be directed at a specific *victim*. Whereas an assault must, by definition, have a victim this is not necessarily true of criminal damage. Smashing up a bus shelter is not a crime directed at an individual. However, criminal damage can be directed at a particular person and this appears to be a feature of older offenders. The presence of *spouse* implies that these are often domestic incidents.

Tables 9 and 10 show the top 10 terms associated with offender characteristics for damage to vehicles with ethnic appearance and occupation using here the Bernoulli model. It appears that white Europeans prefer to *break mirrors* on *cars* which are *secured* implying that they are parked. Afro-Caribbeans instead tend to target *attended vehicles* and are more likely to vandalise *buses* and *vans*. One possible explanation for this is that the different actions are part of a local culture developed amongst racially-homogeneous gangs. It would be necessary to obtain other datasets to determine if this pattern repeated elsewhere. When considering occupation, a different distinction emerges. Clearly those with jobs will have more money and therefore be more likely to run cars. It is therefore as drivers that they vandalise another car in a *road rage incident* which may involve *kicking* the *attended vehicle* of the other driver with whom the dispute has arisen. The unemployed, on the other hand, prefer to attack *secured* vehicles with a *sharp instrument*.

Clearly a more detailed analysis of all the terms, of which there are several hundred, over all the datasets would reveal further patterns. For example, older male shoplifters are more likely to steal spirits evidenced by words such as *bottle, whisky* and *vodka*. Also this is, of course, only one dataset from one locality and does not mean that these patterns are universal.

7. Conclusions

The results show that Language Models are capable of identifying differences in criminal styles between groups of offenders defined by some characteristic such as gender, age, ethnic appearance and broad occupation. It thus shows the power of an approach that could be applied more widely to other categories. Bernoulli models outperform multinomial models although the difference in performance is not great. For offences where there may be no eye-witnesses such as criminal damage, burglary and damage to vehicles the approach can indicate the likely age, sex or ethnic appearance of the offender. The one characteristic which is not revealed by appearance, that is whether the culprit is in employment, is applicable for crimes where the offender may or may have not been seen. Furthermore the models can be used to characterise styles of behaviour. The analysis of the most dominant terms in each language model does reveal interesting characterisations of behavioural styles. This takes the use of language models beyond mere classification in that it can shed light on the material differences in behaviours between categories of offender.

References

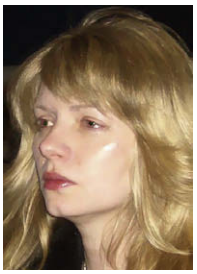
[1] R. Bache, F. Crestani, D. Canter, D. Youngs, Application of language models to suspect prioritisation and suspect likelihood in serial crimes, in: International Workshop on Computer Forensics, 2007.
[2] J. Bai, J. Nie, F. Paradis, Text classification using language models, in: Asia Information Retrieval Symposium, Poster Session, Beijing, 2004.
[3] D. Canter, Offender profiling and criminal differentiation, Legal and Criminal Psychology 5 (2000) 23–46.
[4] D. Canter, C. Bennell, A. Laurance, Differentiating sex offences: a behaviorally based thematic classification of stranger rapes, Behavioral Sciences and the Law 21 (2003) 157–174.
[5] D. Canter, K. Fritzon, Differentiating arsonists: a model of firesetting actions and characteristics, Legal and Criminal Psychology 3 (1998) 73–96.
[6] H. Chen, W. Chung, J. Xu, G. Wang, Y. Qin, M. Chau, Crime Data Mining: A General Framework and Some Examples, IEEE Computer Society, 2004.

[7] D. Farrington, S. Lambert, Predicting offender profiles from offense and victim characteristics, in: R. Kocsis (Ed.), Criminal Profiling: International Theory, Research and Practice, Humana Press Inc., Totowa, New Jersey, 2007.
[8] C. Fellbaum (Ed.), WordNet – An Electronic Lexical Database, MIT Press, Cambridge, MA, 1998.
[9] F. Jelinek, R. Mercer, Interpolation estimation of Markov source parameters from sparse data, in: Workshop on Pattern Recognition in Practice, The Netherlands, Amsterdam, 1980.
[10] J. Lafferty, Z. Cheng-Xiang, Probabilistic relevance models based on document and language generation, in: W.B. Croft, J. Lafferty (Eds.), Language Modeling for Information Retrieval, Kluwer Academic Publishers, Dordrecht, 2003.
[11] D. Losada, Language modeling for sentence retrieval: a comparison between multiple-Bernoulli models and multinomial models, in: Information Retrieval Workshop, Glasgow, 2005.
[12] C. Manning, H. Schütze, Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, MA, 1999.
[13] A. McCallum, K. Nigam, A comparison of event models for Na Bayes text classification, in: Proceedings of the AAAI/ICML-98 Workshop on Learning for Text Categorisation, AAAI Press, 1998, pp. 41–48.
[14] F. Peng, D. Schuurmans, Combining Na Bayes and *n*-gram language models for text classification, in: Twenty-Fifth European Conference on Information Retrieval Research (ECIR'03), Pisa, Italy, 2003.
[15] F. Peng, D. Schuurmans, S. Wang, Augmenting Na Bayes classifiers with statistical language models, Information Retrieval 7 (3) (2003) 317–345.
[16] J.M. Ponte, W.B. Croft, A language modeling approach to information retrieval, in: Proceedings of the Twenty First ACM-SIGIR, Melbourne, Australia, ACM Press, 1998, pp. 275–281.
[17] C. Zhai, J. Lafferty, A study of smoothing methods for language models applied to ad hoc information retrieval, in: Proceedings of SIGIR, 2001, pp. 334–342.

Richard Bache obtained his Batchelor's degree in Mathematics and Economics from the London School of Economics in 1986 and a Ph.D. in Software Engineering from South Bank Polytechnic in 1991. He has worked in both industry and a number of academic institutions such as City University (London), University of Paisley and University of Strathclyde. He has just completed an internship at the University of Glasgow examining measures of findability in patent data.

Fabio Crestani is a Full Professor at the Faculty of Informatics of the University of Lugano in Switzerland since 2007. Before that he was Professor of Computer Sciences of the University of Strathclyde (UK) in 2000-06 and Assistant Professor at the University of Padova (Italy) in 1992-97. In between he held research fellowships at the Rutherford Appleton Laboratory (UK), the International Computer Science Institute in Berkeley (USA), and the University of Glasgow (UK). He is an internationally recognised researcher in Information Retrieval, Text Mining and Digital Libraries having published over one hundred refereed publications on both theoretical and experimental aspects. Finally, since 2008 he is the editor-in-chief of the journal Information Processing and Management, published by Elsevier.

David Canter is the director of the International Research Centre for Investigative Psychology and is widely recognised as the founder of Investigative Psychology. He has carried out a great deal of research, has published 20 books and over 150 papers in learned professional journals, lectured around the world on various aspects on criminal behaviour and police investigation and has contributed to over 100 police and court cases.



Donna Youngs is Associate Director International Research Centre for Investigative Psychology in the School of Human and Health Sciences at the University of Huddersfield UK where she directs a series of recently awarded research projects looking at a variety of crimes and criminals. These studies explore a range of Investigative topics from the Geographical Profiling of Burglary, to Street Robbery, Youth Crime and Antisocial behaviour, Fraudulent Crime Reporting, Insurance Fraud and the Social Networks of Prolific Offenders.