International Journal of Security and Networks

ISSN online: 1747-8413 - ISSN print: 1747-8405
https://www.inderscience.com/ijsn

An ex-convict recognition method based on text mining

Mingyue Qiu, Xueying Zhang, Xinmeng Wang

DOI: 10.1504/IJSN.2022.10049089

Article History:

Received:	21 June 2022
Accepted:	27 June 2022
Published online:	03 April 2023

An ex-convict recognition method based on text mining

Mingyue Qiu*

School of Information Technology,
Nanjing Forest Police College,
Nanjing, 210023, China
Email: qiumy@nfpc.edu.cn
*Corresponding author

Xueying Zhang

Key Laboratory of Virtual Geographic Environment,
Nanjing Normal University,
Nanjing, 210023, China
Email: zhangsnowy@163.com

Xinmeng Wang

School of Information Technology,
Nanjing Forest Police College,
Nanjing, 210023, China
Email: wangxm@nfpc.edu.cn

Abstract: Currently, a large proportion of existing cases in the grassroots public security organisations were committed by ex-convicts. Grassroots police officers cannot directly and rapidly judge whether a suspect is an ex-convict who has committed a case. To solve this problem, an attempt is made to analyse the case report data in a branch bureau in 2021 through data mining. Using the brief case texts in the case report data as the data source, different models based on various algorithms were established to judge whether the ex-convict committed the case. Next, using different algorithms, the ex-convict in the database was ascertained based on the similarity degree results. Finally, the similarity results (the highest similarity reached 94.8) using different methods were calculated, added, and ranked in descending order to submit an ex-convict list to the grassroots police officers for further artificial judgement. Accordingly, grassroots police officers can conduct rapid recognition of ex-convicts when a case is reported. The present model is tested well in the actual applications in the local police stations, suggesting that the model can provide overwhelming support in the daily work of police stations, and with the mutual cooperation and gradual promotion among the police stations, large amounts of human and material resources can finally be saved.

Keywords: natural language processing; text mining; similarity analysis; recognition of people with previous conviction.

Reference to this paper should be made as follows: Qiu, M., Zhang, X. and Wang, X. (2023) 'An ex-convict recognition method based on text mining', *Int. J. Security and Networks*, Vol. 18, No. 1, pp.10–18.

Biographical notes: Mingyue Qiu is with Information Technology College of Nanjing Forest Police College, and the research direction is Public security information science. He is mainly engaged in the research and teaching of public security intelligence, data mining, data anti-smuggling, etc., and has accumulated a lot of knowledge and experience related to the theory of anti-smuggling information science and public security data modelling. The author actively explores the development direction of public security information science in the era of big data and studies how to integrate advanced big data technology into the theoretical research and practical work of public security information science.

Xueying Zhang is a Professor at Nanjing Normal University. She received her doctorate from Nanjing University of Science and Technology. His research interests include geographic big data, location intelligence, and big data GIS.

Xinmeng Wang is with Information Technology College of Nanjing Forest Police College, and the research direction is public security information science. He mainly engages in internet information analysis and judgement, data mining, modelling, and other research. As a teacher of

public security colleges, he also pays attention to practical combat, cultivates relevant professional ability and preliminary research ability suitable for public security work, and cultivates innovative, compound, and applied high-quality public security talents who meet the needs of the modernisation of public security work in the new era.

1 Research background

According to the public security sub-bureau data, the cases committed by the ex-convicts mainly show the following three characteristics. First, the ex-convicts occupy an astonishing 80% of the total perpetrators among the local cases that have been detected, who invariably undergo the characteristic transitions from static, individual, offline, and non-professional to fled, organised, online and professional. Second, among all the cases committed by the ex-convicts, the details of the cases are invariably quite similar to the history of the previous case. Third, despite the significant social activities changes, public security organisations' alarm-receiving numbers have increased exponentially. However, the grassroots policies show no signs of upgrading their systems for recognising the ex-convicts, thereby leading to a huge backlog of cases over time (Jia, 2021). Many criminal suspects continue to be increasingly fluky and unruly, violating the laws of the land and forming vicious circles.

Currently, local public security organs attach great importance to the control of ex-convicts. They have also established a database of ex-convicts, including all the information of ex-convicts mastered by public security organs. However, as for how to use the database, local public security organs always use it for inquiry. In reality, the criminal, in this case, cannot be found until the suspect is confirmed in the interrogation, which only plays the role of database query and has a hysteresis for identifying the criminal. In addition, to identify ex-convicts, grassroots police staff need a certain amount of work experience, such as the older resume police staff. The police staff with less work experience often do not have the ability. Therefore, grassroots police officers can only use traditional methods to identify ex-convicts. Among the cases investigated by the grassroots public security agencies such as local police stations, two noticeable features are noteworthy – generally, a small amount of money is involved, and repeated criminal behaviours are noticed. For the above reasons, this study is aimed to recognise such ex-convicts based on text mining. Specifically, when an illegal case is registered, the grassroots police can rapidly judge whether the criminals have such records and then lock the detailed persons to realise the innovations based on the traditional detection techniques.

The main contributions of this work are as follows:

- The studies cannot directly aid the grassroots police officers in solving the current cases by the ex-convict. Based on the previous analysis, the present study is aimed to achieve the fast recognition of ex-convicts.
- The fast recognition of ex-convict includes two steps: rapidly judging whether the ex-convict committed the case and then determining which ex-convict committed the case.
- Different models (namely, support vector machine, random forest, backpropagation neural network, decision-making tree, and extreme gradient boosting) based on various algorithms were established to judge whether the ex-convict committed the case.
- The ex-convict in the database was ascertained based on the similarity degree results using different algorithms (namely, one-hot-based encoding cosine similarity algorithm, Jaccard similarity degree, and Levenshtein similarity algorithm).
- The similar results were ranked in the descending order to submit an ex-convict list to the grassroots police officers for further artificial judgement.

This article is organised as follows. An extensive analysis of various types of investigation thoughts propounded on big-data analysis methods by scholars is presented in Section 2. The required data sources and data pre-processing methods for ex-convict recognition based on text mining are elaborated in Section 3. Based on five features, five text mining models are selected in Section 4 for judging whether the perpetrator of a given case has previous conviction records. In Section 5, based on the three proposed text similarity algorithms, the details of the present case and the past cases in the ex-convict database are analysed, and the similarity degrees of the present case with the previous records are illustrated. Section 6 summarises the present research results and actual applications of the conceptual model.

2 Literature review

Public security informatisation has been advancing in recent years. Classification algorithms based on machine learning have aroused the attention of public security organisations. Classification aims to categorise the data into different groups. The computer first calculates, seeks, and ascertains different classification logics in the database (Xia et al., 2020); to predict following certain rules when the texts to be predicted are input. Machine learning (ML)-based classification algorithms can find certain classification logics from a large number of disorderly text data and thus, have become a hot issue (Qiu, 2019). Currently, some commonly-used ML models include random forest, support vector machine (SVM), backpropagation neural network (BPNN), extreme gradient boosting (XGBoost), and

convolutional neural network (CNN) (Chen et al., 2021a; Xia et al., 2022).

Scholars from all over the world have conducted a great deal of research. Ferguson (2017) pointed out that the real-time capability and predictability of crime monitoring can be realised in big-data police affairs. Joh (2016) concluded that the police service big-data analysis method could eliminate the dependence on the individual experiences and seek the potential relation between the cases and the suspects based on the comprehensive data mining results with machine learning algorithms. Youngmin and Andrew (2019) analysed the related factors affecting the crimes by the homeless and suggested that the risk terrain model can offer effective prediction. Jesia et al. (2019) compared different algorithms (including some decision-making tree, random forest, AdaBoost algorithm, guided aggregation algorithm (bagging), and extreme random tree algorithm) in the prediction accuracy of criminal activities. Mai and Zhu (2019) analysed the investigative interrogation texts with text mining for automatic classification and tagging of the investigative interrogation texts, which can provide auxiliary support for linking the cases and criminal investigation of the same types of crimes. Hu and Zhang (2019) established an integrated text data mining system for investigation and interrogation based on big data technology in 2019, which can give full play to the core roles of police in text mining. Based on big-data thinking and data processing technique, Yang (2022) converted the judgement mode of inquiry records from passive to active analysis mode, thereby effectively relieving the shortage of investigation resources. Zhang et al. (2019) succeeded in recommitting the prediction of ex-convicts based on the big-data model, which can provide scientific reference for crime prevention and control in advance. In these studies, scholars mainly performed mining on the police texts based on big data technology to achieve the transformation from passive to active analysis; however, analysing the investigation and interrogation texts in most of the studies neglected a real problem in most of the texts in the investigation and interrogation (Shi, 2020). Because of the small amount of money involved in these cases at the grassroots police stations, the related models are difficult to apply at the grassroots public security organisation level, such as local police stations. In the case of the recognition of the ex-convicts, scholars mainly carried out the analysis in the forward direction and analysed whether the ex-convict will re-commit the crime (Xia et al., 2019; Duan and Xu, 2015; Cao and Su, 2020). Unfortunately, the studies cannot directly aid the grassroots police officers in solving the current cases by the ex-convict. Based on the above analysis, the present study is aimed to achieve recognition in the two steps: first, rapidly judging whether the ex-convict committed the case and then determining which ex-convict committed the case.

3 Data pre-processing based on brief case texts

3.1 Data source

Alarm receiving and data disposal are the most basic but have huge data volume in public security services (Huang, 2016). In this study, the data were sourced from the year-round alarm receiving and disposing data of a police station under the administration of a branch bureau, as well as the regional data of all the ex-convicts. By deleting the repeated and invalid alarms with a proportion of 23.3% of the total number, 23,678 valid data records were obtained. Since the present brief case texts were written in Chinese, the model cannot be directly analysed, and the brief case texts had to be pre-processed before the computation with the model.

3.2 Data pre-processing

Since the computer can recognise several items of text, the brief case text needs to be pre-processed before training which mainly includes the following aspects:

1 text segmentation and deletion of stop words

2 feature engineering

3 division of training set, test set and coding of object variable

4 definition of training function and prediction function.

3.2.1 Text segmentation and deletion of stop words

Text segmentation refers to re-organising a continuous sequence of words into word sequences following certain specifications. Deletion of stop words refers to deleting the words in the text with no practical significance but influences the classification results. After the deletion of stop words, the quality of textual features can be improved, the size of the text feature space, as well as unnecessary storage space and calculation time, can be reduced, and simultaneously, word frequency distribution can be balanced (Li, 2015). The stop word list discussed by grassroots police officers based on the stop word list released on Baidu.com was used as the word library in this study.

3.2.2 Feature engineering

After segmentation and the deletion of stop words, the texts can be treated as a set consisting of multiple words, which machine learning can hardly recognise. The current algorithms execute the calculation on vectors and values (Lisa et al., 2020). Therefore, the texts should be transformed into eigenvectors to conform to the format required for text mining with models.

This study selected five basic features with low computational load, commonly used in dealing with basic police affairs. These features can fall into three types – augmented features based on the text information, counting vectors, and the term frequency-inverse document

frequency (TF-IDF) vectors. Further, TF-IDF features include TF-IDF features of words, N-gram TF-IDF features, and part-of-speech (PoS) TF-IDF features.

After augmentation, the features obtain a relatively small amount of information, fully sourced from the texts rather than derivation based on other formulas. Therefore, augmented features occupy an important role in machine analysis. This study merged the features from the texts and the other four types of features obtained by other methods. Finally, it converted the brief case texts into four features as the input to the computer for model establishment. The counting vectors can restore the frequency of words in each segment of brief case text. The conversion mode can be closer to the original texts and directly transform the meanings of original texts into vectors. For the TF-IDF of each word, the inverse frequency of the word in the text can be calculated to convert the meaning of brief case text into vectors using different methods. By considering the sequence relation hidden between the adjacent words, N-gram TF-IDF combines and transforms the connected words in a sentence to restore the relation among different words in the text during the transformation. The PoS TF-IDF fully considers the PoS of the word in the text and takes the feature transformation of the text from the perspective of PoS, which reflects the word's PoS features so that the model can understand different logic (Mariana et al., 2019).

Considering the unique advantages of these different features, this study selected different features to establish the models rather than focusing on them and integrating them with the model to comprehensively and fully dig out the information from the brief case texts.

3.2.3 Division of training set, test set, and encoding of the object variables

After the above feature engineering, a dataset can be obtained. However, the model should first seek logic from the training set and then judge whether the logic is accurate with the test set. However, the obtained dataset is an integral whole, and the model cannot distinguish the training and the test samples. Therefore, the dataset should be divided into training and test sets.

To be specific, the training set is used for model training. The machine model can learn their relation with the input and the corresponding output. The test set is used for the final validation of the model performance. The test set includes the results of the input data with the established model, i.e., the output of the training model on the new input data. The test set cannot be used for training. By comparing the prediction results of the samples in the test set with the samples in the training set, the accuracy and the recall rate can be calculated to evaluate the model accuracy.

3.2.4 Definition of training function and prediction function

After completing the above procedures (the feature transformation on the brief case texts, the division of the training set and the test set, and the encoding of object variables), the obtained dataset now satisfy the requirements of the model recognition. Therefore, the model's training and prediction functions should be defined in the next step. The training function is used as the invocation model to seek the training set's logic. The prediction function can provide the prediction results when the grassroots police officers input the cases under investigation. In addition, the accuracy, recall rate, precision ratio, F1-score, and confusion matrix can be calculated to reflect the model's prediction accuracy.

For the confusion matrix, four terms – true positive (TP), true negative (TN), false positive (FP), and false-negative (FN) should first be defined. TP samples refer to the samples with both positive prediction and true values. TN samples refer to the samples with both negative prediction and true values. FP samples refer to the samples with positive predictions but true negative values. FN samples refer to the samples with negative predictions but positive true values. Table 1 shows the confusion matrix.

Table 1 Confusion matrix

Prediction category	*True category*	
	1	*0*
Positive	TP	FP
Negative	FN	TN

The definition of some evaluation indexes is described below.

The accuracy can be calculated as:

$$\frac{TP+TN}{TP+TN+FP+FN}*100\% \tag{1}$$

The precision is defined as the proportion of the samples with both prediction and true values of 1 and in all the samples with the values of 1, which can be calculated as:

$$\frac{TP}{TP+FP}*100\% \tag{2}$$

The recall rate is defined as the proportion of the samples with both prediction and the true value of 1 in all the samples with the prediction values of 1, which can be calculated as:

$$\frac{TP}{TP+FN}*100\% \tag{3}$$

F1-score is an index for measuring the accuracy of the established classification model, which considers both the accuracy and the recall rate of the model. F1 score can be regarded as a weighted average of the accuracy and the recall rate, with a maximum value of 1 and a minimum value of 0. The more favourable the model is when the F1 score is close to 1. F1 score can be calculated as:

$$F_1 = 2*\frac{Precision*Recal}{Precision+Recal} \tag{4}$$

4 Prediction on whether the suspects are ex-convicts

Based on pre-processing data results described in the above section, this section establishes the related model based on the SVM model, decision-making tree model, random forest model, XGboost model, and BPNN model to judge whether the suspects are ex-convicts. The models produce different results for different theories and calculation methods with the same input of eigenvectors. This study adopted five different models for calculating the eigenvectors. Accordingly, the brief case texts can be mined more comprehensively and thoroughly to obtain deep information. The above five models are featured by light computational burden but with high accuracy and, therefore, apply to grassroots public security organisations to predict whether the suspects of these cases are ex-convicts or not.

4.1 Judgement on whether the suspects are ex-convicts based on the SVM model

SVM attempts to map original data to a high-dimensional space and finds a hyper-plane that separates two types of samples to the full extent in the new feature space. Based on Vapnik-Chervonenkis (VC) dimension theory and the principle of structural risk minimisation, SVM aims to seek the optimal compromise between the complexity degree of the finite samples and the learning ability of samples to obtain the optimal generalisation ability.

In a linear classifier, the hyper-plane $f(x) = \omega x - b = 0$ acts as a classifier. If $f(x) > 0$, the point belongs to class 1; if $f(x) < 0$, the point belongs to class 1. The optimal partition hyperplane established by SVM can maximise the shortest distance from the point in class 1 to the hyperplane and the shortest distance from the point in class 1 to the hyperplane. To solve the following optimisation problem, the weight vector ω and the offset b can be finally calculated as:

$$\min \frac{1}{2}\|\omega\|^2 \tag{5}$$

s.t.

$$y_i\left(\omega * x_i + b\right) \geq 1, \quad i = 1, 2, \ldots, n \tag{6}$$

Based on the Lagrangian function method, the optimisation problem can be transformed into the following dual problem:

$$\min Q(\alpha) = -\sum_{i=1}^{i} \alpha_i + \frac{1}{2}\sum_{i=1}^{i}\sum_{i=1}^{i} \alpha_i \alpha_j y_i y_j i x_j \tag{7}$$

s.t.

$$\sum_{i=1}^{i} \alpha_i y_i = 0 \tag{8}$$

$$\alpha_i \geq 0 \tag{9}$$

Finally, the following expression can be obtained:

$$\omega^* = \sum_{i=1}^{i} \alpha_i^* y_i x_i, \; b^* = y_i - \sum_{i=1}^{i} \alpha_i y_i \left(x_i \bullet x_j\right) \tag{10}$$

4.2 Judgement on whether the suspects are ex-convicts based on the C4.5 model

C4.5 model is a decision-making tree algorithm. The decision-making tree is a tree structure similar to the flow chart, in which each internal node expresses the test on an attribute, each branch represents the test output, and each leaf node stores a class label. After establishing a decision tree, for a tuple in an unknown class, the path from a root node to a leaf node is traced, and the prediction of the tuple is stored in the leaf node. The advantage of the decision-making tree algorithm lies in seeking exploratory knowledge discovery without any domain knowledge or parameter settings. The decision-making tree algorithm is suitable for finding the generality of many texts.

4.3 Judgement on whether the suspects are ex-convicts based on a random forest model

The random forest model is an integrated method based on the decision-making tree algorithm. Some samples are randomly extracted from original samples in put-back sampling mode using random forest. Multiple sample sets can be generated by repeating the above procedures; afterward, each sample set can produce a decision tree. During the generation of each decision tree, some features can be randomly extracted during the branching process of each node, which is then involved in the branching of a decision tree. Next, recursive branching is performed. During each recursive branching process, some features are randomly extracted from the residual characteristics and then participate in branching (since the features involved in branching cannot appear in the nodes after the present node) to generate multiple decision trees. For the newly input sample, each tree can generate prediction results. Though some trees may produce extreme prediction errors for new samples, the majority of the trees can make an accurate prediction. Because of the randomness of samples and variables, the extreme wrongs may easily be neutralised, and the right prediction results can be highlighted according to the principle that the minority is subordinate to the majority. In this way, the prediction accuracy can be enhanced overall.

4.4 Judgement on whether the suspects are ex-convicts based on the XGboost model

Extreme gradient boosting (XGBoost) is a kind of gradient boosting algorithm and residual decision tree. The basic idea is to add each tree to the model gradually. After adding each decision tree, the overall model performance (with declined objective function) can be enhanced. The combined classifier can be constructed by using multiple decision trees (i.e., multiple single weak classifiers), and each leaf node is assigned a certain weight. Adding regular

terms in XGBoost explicit can effectively control the model complexity and avoid over-fitting, enhancing the model's generalisation ability. The use of first-order and second-order partial derivatives and the second-order can contribute to a more rapid and accurate gradient descent. By expressing the function as the second-derivative format of independent variables based on Taylor expansion, leaves can be divided and optimised only based on the input data values without selecting the loss function. In essence, the selection of the loss function can be separated from the optimisation and parameter-setting of the model algorithm. Thus, decoupling can enhance the applicability of the XGBoost algorithm so that it can select the loss function following the requirements. XGBoost can be used for both the classification and the regression.

The detailed calculation process can be described below:

$$\hat{y} = \phi(x_i) = \sum_{k=1}^{k} f_k(x_i) \tag{11}$$

where

$$F = \{f(x) = \omega_{q(x)}\}(q: R^m \to T, \omega \in R^T) \tag{12}$$

where $\omega_{q(x)}$ denotes the score of the leaf node q, and $f(x)$ denotes a tree that can be regressed.

4.5 Judgement on whether the suspects are ex-convicts based on the BPNN model

The back-propagation neural network (BPNN) mainly consists of an input layer, one or multiple hidden layers, and an output layer, which can deal with linear and nonlinear problems. The learning process includes two procedures – the forward propagation of signal and the backward propagation of error. BPNN employs an error back-propagation algorithm for iteration to reduce the error to an acceptable range.

4.6 Model implementation process and comparison of results among different models

Table 2 lists the accuracy values after the training on all data using different models. On the whole, the prediction results are favourable. Despite SVM, despite the low precision rate, the overall accuracy exceeds 50%, and the recall rates are high. Therefore, grassroots police officers can appropriately reduce the weight of the SVM model in artificial judgement. C4.5 model, random forest model, XGboost model, and BPNN all perform well in the prediction, with the precision rates of over 90% and even a maximum rate of 99%. Each model has its advantages. The prediction of ex-convicts using these five models can more fully and thoroughly dig out various logic from brief case texts, thereby enhancing the prediction accuracy. The grassroots police officers can either trust the accuracy of the model based on their own experiences or make overall consideration and judgement on whether the suspects are the ex-convicts by comprehensively considering the results of all models.

Table 2 Comparison of the training results using different models

Model name	*Feature engineering*	*Accuracy*	*Recall rate*	*Precision rate*	*F1-score*
SVM	Counting vector features	0.47	0.71	0.48	0.58
	TF-IDF features of a single word	0.53	0.72	0.55	0.69
	N-gram TF-IDF features	0.45	0.69	0.45	0.52
	POSTF-IDF features	0.61	0.73	0.55	0.65
Decision-making tree algorithm	Counting vector features	0.91	0.94	0.88	0.91
	TF-IDF features of a single word	0.92	0.95	0.89	0.91
	N-gram TF-IDF features	0.9	0.94	0.9	0.92
	POSTF-IDF features	0.92	0.93	0.91	0.9
Random forest algorithm	Counting vector features	0.94	0.98	0.9	0.94
	TF-IDF features of a single word	0.93	0.99	0.91	0.92
	N-gram TF-IDF features	0.95	0.99	0.91	0.93
	POSTF-IDF features	0.92	0.97	0.92	0.91
Xgboost	Counting vector features	0.95	0.98	0.92	0.95
	TF-IDF features of a single word	0.95	0.99	0.93	0.93
	N-gram TF-IDF features	0.96	0.96	0.93	0.95
	POSTF-IDF features	0.98	0.95	0.95	0.93
BPNN model	Counting vector features	0.97	0.98	0.96	0.97
	TF-IDF features of a single word	0.96	0.99	0.98	0.98
	N-gram TF-IDF features	0.97	0.98	0.96	0.97
	POSTF-IDF features	0.95	0.97	0.97	0.98

5 Similarity analysis of cases

After judging whether the suspects are ex-convicts, the grassroots police officers can only know whether the ex-convicts committed the crime. In the case of high probability, the ex-convict's specific information should be investigated further. It is worth noting that all the illegal cases committed by an ex-convict are almost similar to the previous ones. Accordingly, the similarity degrees between the present brief case and the previous cases of ex-convicts in the ex-convict database can be calculated with the similarity degree algorithm and then ranked in descending order. Finally, the grassroots police officers can make the judgement according to the ranking results to detect the case.

5.1 Application background of the text-similarity degree model

This study adopted three different methods, namely, cosine similarity based on one-hot coding, Jaccard similarity, and Levenshtein similarity, for calculating similarity degree. Different theoretical bases and computational methods can obtain different results from different methods on the same data source. To obtain the ex-convict list more fully and accurately, the best way is to adopt multiple algorithms for calculating the similarity degrees of the same data source. Considering that different algorithms have their advantages, the calculation results using different methods can be added and ranked in descending order to form an ex-convict list. Finally, the list can be submitted to the grassroots police officers for artificial judgement to determine the perpetrator's detailed information. Following the data above pre-processing procedures, the results after text segmentation and the deletion of the stop words were used for calculation.

5.2 Illustration of the advantages of similarity algorithm

The cosine similarity algorithm based on one-hot encoding shows the following advantages. First, the features can be expanded via one-hot encoding. Secondly, after the encoding, the continuous variables can change from one weight to multiple weights, enhancing the model's nonlinear ability. Thirdly, the algorithm can execute without normalising the parameters (since the data to be processed are restricted to certain ranges after the processing with a certain algorithm), which can avoid complex calculations. Finally, the feature with a higher weight can be divided into several features with different weights for management, reducing the effect of abnormal values on the model and enhancing the model's stability. Therefore, owing to low computational load and favourable stability, the cosine similarity algorithm based on one-hot encoding applies to grassroots public security organisations (Chen et al., 2021b).

Jaccard similarity algorithm relatively shows a simple calculation mode and gives the similarity degrees between two texts by calculating the union set and the intersection set and taking the division. Owing to its simplicity in the calculation, the Jaccard algorithm applies to the current condition of grassroots public security organisations. However, the shortcomings are also obvious since the algorithm only considers whether the word exists while neglecting the sequences of words. The shortcomings impose a slight effect on the brief cases. By taking the word 'knifepoint' as an example, the suspect committed the crime at knifepoint no matter how many times and where the word appears in the text. Therefore, ignoring the appearance times and sequence of the word almost imposes no effect. This study selected the Jaccard algorithm for text similarity calculation.

The advantages of the Levenshtein similarity algorithm are described below. Using the Levenshtein algorithm, the similarity can be measured by the number of transformation operations from one text to another. More operations indicate low similarity between the two texts (Chen et al., 2020). For brief case text, similar words, sentences, and expression methods exist in addition to different information, including clues and details. Too many operation numbers between two texts suggest different detailed information between the two cases, such as various types of details and clues. Accordingly, calculating the similarity degree between two brief case texts with the Levenshtein algorithm is accurate to a certain degree.

Overall, different similarity algorithms have different advantages. All algorithms apply to the current situation of grassroots public security organisations such as local police stations for analysing the cases with a small amount of money involved. The grassroots police officers can take comprehensive consideration in artificial judgement and believe one algorithm or add different results.

5.3 Comparison of prediction results using different models

Table 3 lists part results of the similarity degrees of a brief case with the corresponding cases by all ex-convicts in the database. The secret-related data were hidden and not shown in the table. Different similarity algorithms show their advantages and disadvantages, respectively. This study did not assign the weights to different algorithms but added the calculation results and ranked them in descending order.

Table 3 lists the calculated similarity degrees of the present brief case, with the cases against all the ex-convicts in the database. The higher place in the ranking results, such as
no. 7 (the highest similarity reached 94.8), suggests the greater similarity between the brief cases of the ex-convict with the present case. At that moment, the ex-convict list should be submitted to the grassroots police officers for judgement. The grassroots police officers can screen out ex-convicts with high similarity degrees for further artificial judgement. They can further determine whether the case was committed by the ex-convict based on their own experiences, video investigation and inquiry results. The

model was proved to be accurate in actual applications to the local police stations.

Table 3 Ranking results of different similarity models

No.	*Cosine similarity*	*Jaccard similarity algorithm*	*Levenshtein similarity algorithm*	*Results*
7	94.8	15.28	67.39	177.47
9	92.29	20.59	62.12	175.00
10	90.16	5.33	59.2	154.69
5	96.11	23.36	34.34	153.81
8	88.71	3.7	58.91	151.32
6	91.6	3.28	46.6	141.48
3	89.17	3.03	48.62	140.82
4	88.6	3.23	46.67	138.50
1	93.88	7.81	34.39	136.08
2	93.33	9.27	23.56	126.16

6 Conclusions

This study attempted to determine whether the case was committed by the ex-convict based on the brief case texts and screen out the ex-convict range for investigation. The similarity degrees between the present and the previous cases were analysed based on the prediction results and whether the ex-convict committed the case. Based on the advantages of different algorithms, this study took a comprehensive consideration, compared different results, and made predictions. Finally, an ex-convict list was submitted to the grassroots police officers for reference and the subsequent artificial judgement to determine the final perpetrators. Accordingly, the case detection efficiency was ensured in a labour-saving way. The present research can also aid grassroots police officers in solving criminal cases. In the future, we will add many other influencing factors of ex-convicts into the model to train the ex-convict's identification models.

Acknowledgements

This study is partially supported by Project on No. LGZD202304 the Fundamental Research Funds for the Central Universities; Outstanding Young Backbone Teacher of Jiangsu Universities 'Qinglan Project' 2021.

References

Cao, X.B. and Su, K. (2020) 'Investigation on affective influence under big-data background', *Journal of Zhejiang Police College*, Vol. 2020, No. 5, pp.42–47.

Chen, G., Shen, J.P., He, J., Dai, X. and Wang, W.Y. (2021a) 'An overall analysis method of urban road parking lots based on data mining', *International Journal of Security and Networks*, Vol. 16, No. 2, pp.105–111.

Chen, G., Shi, X.L., Chen, M. and Zhou, L. (2020) 'Text similarity semantic calculation based on deep reinforcement learning', *International Journal of Security and Networks*, Vol. 15, No. 1, pp.59–66.

Chen, G., Yao, R., Chen, G., Chen, J. and Li, T. (2021b) 'A smart urban management information public opinion analysis system', *International Journal of Security and Networks*, Vol. 16, No. 2, pp.92–97.

Duan, K.Y. and Xu, M.L. (2015) 'Investigation on police alert prediction model based on BP neural network', *Science & Technology Information*, Vol. 13, No. 18, pp.230–231.

Ferguson, A.G. (2017) *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement*, pp.92–93, New York University Press, New York.

Hu, X.Y. and Zhang, W. (2019) 'Data mining and analysis of investigation and interrogation documents based on big data', *Journal of People's Public Security University of China (Social Sciences Edition)*, Vol. 35, No. 6, pp.35–43.

Huang, S.H. (2016) 'Investigation of detection strategy of network crimes based on spatial behavioural analysis model', *Journal of China Criminal Police College*, Vol. 2016, No. 4, pp.49–53.

Jesia, Q.Y., Mahfil, Q.S., Zaisha, Z. and Khan, M.H. (2019) 'Predicting crime using time and location data', *Proceedings of the 7th International Conference on Computer and Communications Management (ICCCM 2019)*, pp.132–136.

Jia, M. (2021) 'Application of artificial intelligence in investigation and interrogation and its risks regulation', *Journal of the Armed Police Academy*, Vol. 37, No. 2, pp.32–36.

Joh, E.E. (2016) 'The new surveillance discretion: automated suspicion, big data and policing', *Harv. L. & Pol'y Rev*., Vol. 2016, No. 10, p.15.

Li, H. (2015) 'Review on word similarity algorithms', *Journal of Modern Information*, Vol. 35, No. 4, pp.172–177.

Lisa, S., Stewart, J. and Jamie, L. (2020) 'The usual suspects: prior criminal record and the probability of arrest', *Police Quarterly*, Vol. 24, No. 1, pp.31–54.

Mai, J.J. and Zhu, L.F. (2019) 'Police intelligence text mining analysis based on natural language processing', *Chinese Security & Protection*, Vol. 116, No. 9, pp.96–98.

Mariana, P., Montaño, S., Agudelo, K., Idárraga-Cabrera, C., Fernández-Lucas, J. and Herrera-Mendoza, K. (2019) 'Emotion recognition in young male offenders and non-offenders', *Physiology & Behavior*, Vol. 207, No. 1, pp.73–75.

Qiu, L.F. (2019) *Research on Social Safety Risk Analysis Based on Machine Learning*, A Master's thesis from the Public Security University of China.

Shi, S.C. (2020) *Mining of Crime Individual Patterns of Encroaching on Property and Identity Prediction*, A Master's thesis from the Public Security University of China.

Xia, X., Xiao, Y. and Liang, W. (2020) 'SAI: a suspicion assessment-based inspection algorithm to detect malicious users in smart grid', *IEEE Transactions on Information Forensics and Security*, Vol. 15, No. 1, pp.361–374.

Xia, X., Xiao, Y., Liang, W. and Cui, J. (2022) 'Detection methods in smart meters for electricity thefts: a survey', *Proceedings of the IEEE*, February, Vol. 110, No. 2, pp.273–319.

Xia, Z.Y., Ruan, K. and Liu, B.H. (2019) 'Establishment of police alert judgment and prediction system based on machine learning', *Modern Information Technology*, Vol. 3, No. 11, pp.77–81.

Yang, Q.P. (2022) 'Study on investigation and interrogation mode under big-data background', *Network Security Technology & Application*, Vol. 3, No. 12, pp.138–139.

Youngmin, Y. and Andrew, P. (2019) 'Using risk terrain modeling to predict homeless related crime in Los Angeles, California', *Applied Geography*, Vol. 109, No. 8, pp.1–12.

Zhang, L.H., Niu, H.T. and Wang, Z.N. (2019) 'Research on the construction of early warning model of criminals based on big data', *Netinfo Security*, Vol. 19, No. 4, pp.82–89.