Unstructured Malay Text Analytics Model in Crime

Rosmayati Mohemad¹, Nazratul Naziah Mohd Muhait¹, Noor Maizura Mohamad Noor¹ & Zulaiha Ali Othman²

¹Faculty of Ocean Engineering Technology & Informatics, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, MALAYSIA.
²Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, MALAYSIA.

rosmayati@umt.edu.my

Abstract. The concern in crime prevention to realize a safe environment for public’s welfare is one of the central agenda to government authorities. The increasing volume of crime reports with unstructured data that are heavily loaded with text presents the biggest challenge to human experts in analysing the data efficiently as it contains lots of hidden and valuable information such as modus operandi and common crime patterns. The capacity limitations in human information processing demand an effective data analytics model to uncover these hidden and potentially meaningful information from large-scale of data especially in Malay language for improving analysis and supporting decision-making in crime investigation. The objective of this paper is to propose Unstructured Malay Text Analytics Model to support serial crime investigation based on modus operandi patterns. The significance output of this paper is to assist experts especially Royal Malaysia Police in crime investigation based on modus operandi patterns analysis as one of the efforts in reducing rate of crime index.

1. Introduction

The concern in crime prevention to realize a safe environment for public’s welfare is one of the central agenda to government authorities in many nations. The government of Malaysia, under a dynamic plan known as Government Transformation Program (GTP) has lined a target to reduce the rate of crime index or reported crime as one of the national key areas [1]. According to Royal Malaysia Police as reported in [2], despite of the decrement in total crime index in Malaysia from year 2009 to 2012, the property offences crimes collectively contribute to approximately 80% to 82% of all reported index crime. The highest rate of crime against properties has generated considerable fear within community, making public perceptions towards safety as a challenge issue that demanding critical attention at both local and national level [2][3][4].

Crime reports are documents detailing a critical information on a crime or a problematic incident to investigators to start an investigation and lead to a number of criminals arrested and cases cleared [5]. The increasing volume of crime reports with unstructured data that are heavily loaded with natural language text presents the biggest challenge to human experts in analyzing the data efficiently as it contains lots of hidden and valuable information such as modus operandi patterns. Modus operandi is the set of habits and characteristics of criminals, as well as patterns of crime operation which become stereotyped [6][7]. Unlike the structured data stored in a database, unstructured data written in free text is difficult to understand by mature database techniques and such rigid manner is insufficient to reveal
a great deal of potentially useful information [8][9]. Relying on manual analysis to discover implicit information buried in the vast amount of such crime reports is a tedious, complex, time and labor intensive [5][9][10]. Thus, the capacity limitations in human information processing and database demand an effective data analytic model to uncover these hidden and potentially meaningful information from large-scale and high-dimensional of data especially in Malay language for improving analysis and supporting decision-making in crime investigation.

Several data analytics model when dealing with unstructured text have shown significant potential to support intelligence analysis and decision-making in many domains such as healthcare crime investigation [11], education [12], manufacturing [13], banking [14] and others by integrating proper techniques in natural language processing, text mining and visualization. Even though studies relating to text-based processing for English language are extensive, insufficient studies have been conducted to work on Malay language due to its morphological variants. The Malay language is rich in colloquial, idiomatic expressions and literary allusions. Although the Malay language is widely used within the South East Asia region by more than 200 million people, it has become one of the less resourced languages in the world. Due to this, a limited number of computational linguistic researches were found related to Malay language. In crime investigation itself, there is no analytics tool available for processing Malay text. Thus, unstructured Malay text analytics model is necessary to develop an automatic approach to accurately analyze the inherent structural properties of modus operandi and crime patterns from Malay text crime reports which are essential for both crime investigation and the development of operational strategies to prevent crime.

The remaining of this paper is organized as follows. Section two discusses about the state-of-the-art of crime analysis pattern and the arising issues in the domain. The subsequent section outlines the research methodology framework. Meanwhile, the proposed unstructured Malay text analytics model is presented in the next section. Finally, the last section concludes with a summary and future research directions.

2. Related work
Crime analysis is implemented within law enforcement organizations that referring to a set of systematic and analytical processes for identifying and analyzing patterns and trends in crime data that valuable to police agencies and communities [15][16]. It can be achieved through identifying the relevant information such as modus operandi patterns, hence useful to support criminal investigation and crime prevention. Modus operandi is a unique way a criminal acts during a crime that can be classified into set of habits or behavior of working, spatial and temporal behavior and method of operating or functioning to successfully accomplish a crime [17]. Analyzing and mapping modus operandi patterns from unstructured data such as crime reports and records are important factors in crime investigation when the investigators attempt to find correlation between cases. Series of studies have been done to link spatial and temporal behavior similarity [18][19][20], criminal behavior similarity [21] and crime process [7]. However, such manual approach that requires the involvement of crime experts’ domain knowledge is a challenging and time consuming with the increasing volume of unstructured text. In such a way, effective solutions using information technologies and computational approaches are needed to automate the processes.

In the past decade, the exponential increase in the creation of text analytics tools for supporting crime analysis has shown concerted efforts in handling of public-safety systems. With text analytics tools, the massive volume of unstructured crime data can be extracted, visualized and analyzed for further decision-making processes. Text analytics is the process of applying algorithms in order to analyze sets of data and extract useful and unknown patterns, relationships and information [22]. Furthermore, text analytics are used to extract previously unknown useful, valid and hidden patterns and information from large data sets as well as to detect important relationships among the stored variables. Therefore, text analytics have a significant impact on current research development especially in the crime analysis [23]
[8]. However, these text analytics tools were designed for processing English-like data. There are limited number of researches for processing Malay language.

Malay is not only a native language for Malaysia but also one of the languages used in Indonesia, Brunei, Singapore, and southern Thailand. The Malay language possesses its own unique structure and grammar [24]. A limited number of computational linguistic researches that was found on Malay text analytics. There are several studies have been conducted in text summarization [24][25], sentiment analysis [26][27][28], text normalization [29] and information retrieval [30]. Most of these studies suffer from a public Malay corpus that serves as the source of data, which currently distributed and unintegrated.

3. Research methodology framework

The overall research methodology framework is depicted in Figure 1. It consists of several phases, including theoretical study, data collection, identification of modus operandi indicator, extraction of modus operandi and classification and association. These phases contain methodological details that enables the formulation of Malay text analytics model.

![Figure 1. Research Methodology Framework](image)

3.1. Theoretical study

Above of all, literature study is performed as the theoretical study. In this phase, the past and current literatures related to text analytics are collected, comprehended and analysed for identifying the area of concern and problems. Comprehensive literature study on existing text analytics model is conducted to find out about the purposes, gaps and concrete justification of designing unstructured Malay text analytics model.

3.2. Data collection

The data used in this study is collected from real-world house-breaking crime reports of Malay Royal Police Department in between year 2010 to 2013. The corpus contains 100,383 crime reports in Malay language. The details distribution of data is summarized in Table 1. The format of data is originally in excel documents.
### Table 1. Summary of house-breaking crime reports dataset

| Year | Number of reports |
|------|-------------------|
| 2010 | 26,777            |
| 2011 | 27,650            |
| 2012 | 24,122            |
| 2013 | 21,834            |

3.3. **Identification of modus operandi indicator**

The list of modus operandi indicator is crucial to indicate the relevant information that potentially need to be extracted. Thus, relevant modus operandi indicator need to be identified and constructed as a corpus. There are two methods that have been applied to recognize the potential modus operandi patterns, which are in-formal interview and literature search.

#### 3.3.1. Interview

The main purpose of conducting an extensive interview is to understand the basic modus operandi indicators that are frequently used in crime investigation, which leads to the criminal detection. Therefore, in-depth information is gathered by consulting domain experts of Malaysia Royal Police Department. Potential related key phrases are identified and linked based on the input given by the experts.

#### 3.3.2. Literature search

This method is carried out to identify potential modus operandi indicators in previous researches. Related journals and articles in crime domain are reviewed, focusing on house-breaking crime investigation.

3.4. **Extraction of modus operandi**

The process of extracting information that potentially relate to modus operandi is done automatically based on the identified modus operandi patterns. Therefore, analysis on relevant information extraction algorithms for Malay text are studied and tested with the collected data.

#### 3.4.1. Analysis of data

The collected data is observed in order to interpret its types, format and morphology. In this step, the data is also prepared for pre-processing. Several pre-processing techniques for Malay text such as tokenization, stop word and special characters removal and stemming are studied and tested with the collected data.

#### 3.4.2. Analysis of processing rules

This step involves the analysis on information extraction algorithms for Malay text using rule-based, statistical approach and machine learning. Latent semantic analysis is also studied. All these algorithms are compared, developed and tested in order to identify its strength and weakness. Based on these existing algorithms, an improved algorithm is designed for processing Malay text.

#### 3.4.3. Analysis of required output

The possible output presentation is observed and determined. In this step, the evaluation methods are also identified to measure the accuracy and efficiency.

3.5. **Association of modus operandi**

Potential algorithms to associate and link the extracted modus operandi patterns and for the classification of similar modus operandi patterns are studied and analysed. Among potential algorithms are apriori,
logistic regression, support vector machine, k-nearest neighbor, neural network and random forest. Rapid Miner tool will be using to process the extracted data.

4. The proposed unstructured Malay text analytics model
In this section, the proposed unstructured Malay text analytics model is presented. The model as shown in Figure 2 contains four components; pre-processing, modus operandi feature extraction, modus operandi association and classification, and visualization dashboard.

![Diagram of the proposed unstructured Malay text analytics model.](image)

**Figure 2.** The model of unstructured Malay text analytics in crime.

4.1. Text pre-processing
Text pre-processing is one of the crucial component to clean and transform the text data for quantitative analysis. Among of the multi-step processes involved are tokenization, stop word removal and stemming.

4.2. Modus Operandi feature extraction
List of potential modus operandi is extracted based on the identified modus operandi indicators. Modus operandi features are categorized into named entity, crime behavior and crime process. Named entity aims to recognize nouns that related to names of person, number of criminals, tools used by the criminal, missing properties, spatial and temporal behavior. Meanwhile, criminal behavior denotes on how the criminal acts during the crime such as criminal disguise, the way criminal breaks through the premise, the way criminal acts towards victim and the way criminal rob the property. Crime process is represented by the series of actions performed and the object characteristics involved during crime.
4.3. **Modus Operandi Association & Classification**

This component aims to extract plausible associations between crimes based on modus operandi patterns, grouping similar crimes based on similar criminal behavior and crime process, eliciting co-criminal chain and potential list of suspects based on spatial-temporal behavioral similarity. Here, machine learning algorithms that are extensively employed for association and classification and fit to the problems are applied.

4.4. **Visualization Dashboard**

Visualization dashboard is used to display various types of visualization for further analysis in order to assist investigators for better decision-making.

5. **Conclusion**

Malaysia Ministry of Home Affairs Strategic Plan 2015-2020 and Government Transformation Program (GTP) Roadmap under the National Key Result Area (NKRA) have initiated a plan to reduce the rate of crime index at 5% annually by 2020. Therefore the proposed unstructured Malay text model is expected to provide actionable information for the police management to make a better decision in crime investigation and assist domain experts (Royal Police Department) to develop and implement more effective law enforcement policies and crime prevention programs. Next, further analysis on algorithms for Malay information extraction and algorithms on how to associate the extracted modus operandi to assist decision-making is expected to be done in order for designing an improved algorithm. In addition, prototype of the proposed model is expected to be developed.

**Acknowledgement**

This research is supported by Ministry of Education Malaysia, under Fundamental Research Grant Scheme (FRGS) with vote number 59541. The authors would also like to acknowledge Royal Police Department Malaysia for their full support to this research.

**References**

[1] Prime Minister Department, *Government Transformation Programme - The Roadmap - Executive Summary*. Putrajaya, Malaysia: Performance Management and Delivery Unit, Prime Minister Department, 2010.

[2] Y. K. Teh, “The Royal Malaysia Police has got its strategy wrong: Laws alone do not bring down crime rates,” *Int. J. Police Sci. Manag.*, vol. 17, no. 1, pp. 9–16, 2015.

[3] M. A. Mohit and H. M. H. Elsawahl, “Crime and Housing in Kuala Lumpur: Taman Melati Terrace Housing,” *Asian J. Environ. Stud.*, vol. 2, no. 2, pp. 53–63, 2017.

[4] S. Ishak, “Perceptions of People on Police Efficiency and Crime Prevention in Urban Areas in Malaysia,” *Economics*, vol. 4, no. 5, pp. 243–248, 2016.

[5] C.-H. Ku and G. Leroy, “A decision support system: Automated crime report analysis and classification for e-government,” *Gov. Inf. Q.*, vol. 31, no. 4, pp. 534–544, 2014.

[6] C. Bannel and D. V Canter, “Linking commercial burglaries by modus operandi: tests using regression and ROC analysis,” in *Applications of Geographical Offender Profiling*, D. Canter and D. Youngs, Eds. 2017.

[7] Y.-S. Li and M.-L. Qi, “An approach for understanding offender modus operandi to detect serial robbery crimes,” *J. Comput. Sci.*, vol. 36, p. 101024, 2019.

[8] R. N. Zaeem, M. Manoharan, Y. Yang, and K. S. Barber, “Modeling and analysis of identity threat behaviors through text mining of identity theft stories,” *Comput. Secur.*, vol. 65, pp. 50–63, 2017.

[9] Y.-H. Tseng, Z.-P. Ho, K.-S. Yang, and C.-C. Chen, “Mining term networks from text collections for crime investigation,” *Expert Syst. Appl.*, vol. 39, no. 11, pp. 10082–10090, 2012.

[10] W. Jentner, D. Sacha, F. Stoffel, G. Ellis, L. Zhang, and D. A. Keim, “Making Machine Intelligence Less Scary for Criminal Analysts: Reflections on Designing a Visual Comparative Case Analysis Tool,” *Vis. Comput.*, vol., no., pp. 1–17, 2018.
[11] T. S. Yange, H. A. Soriyan, and O. O, “A Schematic View of the Application of Big Data Analytics in Healthcare Crime Investigation,” *J. Heal. Informatics Africa*, vol. 4, no. 1, pp. 32–41, 2017.

[12] A. R. Cano, B. Fernández-Manjón, and Á. J. García-Tejedor, “GLAID: Designing a Game Learning Analytics Model to Analyze the Learning Process in Users with Intellectual Disabilities,” in *Serious Games, Interaction and Simulation*, C. V. de Carvalho, P. Escudeiro, and A. Coelho, Eds. Springer, Cham, 2017, pp. 45–52.

[13] S.-J. Shin, J. Woo, and S. Rachuri, “Predictive Analytics Model for Power Consumption in Manufacturing,” *Procedia CIRP*, vol. 15, pp. 153–158, 2014.

[14] N. Sun, J. G. Morris, J. Xu, X. Zhu, and M. Xie, “iCARE: A framework for Big Data-based Banking Customer Analytics,” *IBM J. Res. Dev.*, vol. 58, no. 5/6, pp. 4:1–4:9, 2014.

[15] D. V. Rohini and P. Isakkki, “Crime Analysis and Mapping through Online Newspapers: A survey,” in *2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE’16)*, 2016, pp. 1–4.

[16] J. B. LeBlanc, J. Elder, C. W. Bruce, and R. B. Santos, “Definition and Types of Crime Analysis,” 2014.

[17] J. E. Douglas, A. W. Burgess, A. G. Burgess, and R. K. Ressler, *Crime Classification Manual*, Second Edi. Jossey-Bass A Wiley Imprint, 2006.

[18] D. Johnson, “The space/time behaviour of dwelling burglars: Finding near repeat patterns in serial offender data,” *Appl. Geogr.*, vol. 41, pp. 139–146, 2013.

[19] L. Markson, J. Woodhams, and J. W. Bond, “Linking Serial Residential Burglary: Comparing the Utility of Modus Operandi Behaviours, Geographical Proximity, and Temporal Proximity,” *J. Investig. Psychol. Offender Profiling*, vol. 7, no. 2, pp. 91–107, 2010.

[20] K. J. Bowers and S. D. Johnson, “Who Commits Near Repeats? A Test of the Boost Explanation,” *West. Crim. Rev.*, vol. 5, no. 3, pp. 12–24, 2004.

[21] H. Chi, Z. Lin, H. Jin, B. Xu, and M. Qi, “A decision support system for detecting serial crimes,” *Knowledge-Based Syst.*, vol. 123, pp. 88–101, 2017.

[22] A. Ittoo, L. M. Nguyen, and A. van den Bosch, “Text analytics in industry: Challenges, desiderata and trends,” *Comput. Ind.*, vol. 78, pp. 96–107, 2016.

[23] N. Qazi and B. L. W. Wong, “An interactive human centered data science approach towards crime pattern analysis,” *Inf. Process. Manag.*, vol. 56, no. 6, p. 102066, 2019.

[24] N. Zamin and A. Ghani, “A Hybrid Approach for Malay Text Summarizer,” in *Proc. Int. Multi-Conference Eng. Technol. Innov.*, 2010.

[25] S. Alias, M. G. K. H. Siti Khaotijah, and T. Tan Ping, “A Malay Text Corpus Analysis for Sentence Compression using Patterngrowth Method,” *J. Teknol.*, vol. 78, no. 8, pp. 197–206, 2016.

[26] A. Alsaffar and N. Omar, “Integrating a Lexicon Based Approach and K Nearest Neighbour for Malay Sentiment Analysis,” *J. Comput. Sci.*, vol. 11, no. 4, 2015.

[27] T. Al-Moslmi, S. Gaber, A. Al-Shabi, M. Albareed, and N. Omar, “Feature Selection Methods Effects on Machine Learning Approaches in Malay Sentiment Analysis,” in *The Proceeding of 1st ICRIL-International Conference on Innovation in Science and Technology (IICIST 2015)*, 2015, pp. 444–447.

[28] N. I. Zabha, Z. Ayop, S. Anawar, E. Hamid, and Z. Z. Abidin, “Developing Cross-lingual Sentiment Analysis of Malay Twitter Data Using Lexicon-based Approach,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 1, pp. 346–351, 2019.

[29] M. A. Saloot, N. Idris, and R. Mahmud, “An architecture for Malay Tweet normalization,” *Inf. Process. Manag.*, vol. 50, no. 5, pp. 621–633, 2014.

[30] H. M. Hanum, Z. A. Bakar, N. A. Rahman, M. M. Rosli, and N. Musa, “Using Topic Analysis for Querying Halal Information on Malay Documents,” *Procedia - Soc. Behav. Sci.*, vol. 121, pp. 214–222, 2014.