ISSN 2723-6471

Adaptive Decision-Support System Model for Automated Analysis and Classification of Crime Reports for E-Government

Taqwa Hariguna [1,*], Athapol Ruangkanjanases [2,*]

[1] Department Information System, Universitas Amikom Purwokerto, Indonesia
[2] Chulalongkorn Business School, Chulalongkorn University, Thailand
[1] taqwa@amikompurwokerto.ac.id; [2] athapol@cbs.chula.ac.th
* corresponding author

(Received June 22, 2023; Revised July 19, 2023; Accepted August 20, 2023; Available online September 25, 2023)

Abstract

This study explores the potential of text analysis and classification techniques to improve the operational efficiency and effectiveness of e-government, particularly within law enforcement agencies. It aims to automate the analysis of textual crime reports and deliver timely decision support to policymakers. Given the increasing volume of anonymous and digitized crime reports, conventional crime analysts encounter challenges in efficiently processing these reports, which often lack the filtering or guidance found in detective-led interviews, resulting in a surplus of irrelevant information. Our research involves the development of a Decision Support System (DSS) that integrates Natural Language Processing (NLP) methods, similarity metrics, and machine learning, specifically the Naïve Bayes' classifier, to facilitate crime analysis and categorize reports as pertaining to the same or different crimes. We present a crucial algorithm within the DSS and its evaluation through two studies featuring both small and large datasets, comparing our system's performance with that of a human expert. In the first study, which encompasses ten sets of crime reports covering 2 to 5 crimes each, the binary logistic regression yielded the highest algorithm accuracy at 89%, with the Naive Bayes' classifier trailing slightly at 87%. Notably, the human expert achieved superior performance at 96% when provided with sufficient time. In the second study, featuring two datasets comprising 40 and 60 crime reports discussing 16 distinct crime types for each dataset, our system exhibited the highest classification accuracy at 94.82%, surpassing the crime analyst's accuracy of 93.74%. These findings underscore the potential of our system to augment human analysts' capabilities and enhance the efficiency of law enforcement agencies in the processing and categorization of crime reports.

Keywords: Adaptive Decision-Support System; Automated Analysis; Classification of Crime Reports; E-Government

1. Introduction

In the rapidly evolving landscape of modern governance, E-Government initiatives have gained prominence as a means to enhance the efficiency, transparency, and responsiveness of public services. Among the multifaceted challenges faced by government agencies, the effective management and analysis of crime reports hold particular significance [1]–[3]. In this context, security issues are becoming increasingly complex and demand a more effective response from the authorities [4]. The increasing number of reported crimes is a major challenge for the government in managing resources and formulating appropriate policies [5]. Therefore, an adaptive and automated decision support system is needed to analyze and classify crime reports. With this system, it is expected that the government can respond to crime reports faster and more accurately, thereby improving public safety.

In addition, the development of information and communication technology has changed the way crime reports are received by the government [6]. People can now easily report crimes through various digital platforms. Therefore, a decision support system model is needed that is able to integrate data from various sources and produce a more comprehensive analysis. Crime report classification is an important stage in crime data processing that enables a better understanding of the types of crimes that occur [7]. In the context of this research, the use of artificial intelligence and data analytics to classify crime reports is an innovative step that has a significant impact.

The classification of crime reports allows governments and law enforcement to identify trends and patterns of crimes that may occur in an area [8]. This can assist in more efficient resource allocation and law enforcement prioritization.

For example, by analyzing crime reports by crime type or location, authorities can take more appropriate preventive measures. Crime report classification can also be used for real-time crime monitoring [9]. With a system that can automatically classify crime reports, authorities can respond to crime incidents more quickly. This can improve public safety and provide more effective responses to emergency situations. The classification of crime reports can help in legal investigations and prosecutions. With more accurate and detailed analysis of crimes, evidence can be better collected, and criminals can be identified more easily. This can increase the effectiveness of law enforcement and ensure that criminals are dealt with more fairly in accordance with applicable laws. As such, the classification of crime reports is a key step in improving the law enforcement system and overall community safety.

In addition to its practical benefits, this research also has significant academic relevance. The development of a decision support system model for crime report analysis and classification can be an important contribution to the field of artificial intelligence and information technology. The results of this research are expected to serve as a foundation for further development in the field of data analysis and security in the country. Government agencies are tasked with efficiently responding to the growing influx of digitized text information and databases. One effective solution involves the integration of a Decision-Support System (DSS) with text mining and classification techniques, offering significant benefits such as aiding crime analysts in their investigations and allowing citizens to access e-government programs to monitor neighborhood crime promptly.

This research delves into the utilization of Natural Language Processing (NLP) techniques in conjunction with similarity metrics and classification methodologies to streamline and automate crime analysis. Specifically, the study focuses on the essential task of sifting through crime reports and identifying those that pertain to the same or similar criminal incidents. Identifying reports related to the same crime can amplify the available information for apprehending suspects or enhancing preventive measures. Furthermore, pinpointing similar crimes holds paramount significance in comprehensively analyzing crime patterns, gang activities, and optimizing the allocation of law enforcement resources.

The proposed approach employs similarity metrics and classification methods to pinpoint reports detailing similar or identical criminal incidents. The efficacy of this algorithm is rigorously assessed, juxtaposed against the performance of a trained analyst. To validate the practical applicability of our Decision-Support System in a real-world context, we conducted a comprehensive experiment employing datasets of varying sizes, encompassing a diverse range of crime reports. Through this evaluation, we gauged the algorithm's performance and compared it to the crime analyst's classification abilities, thereby establishing its practical utility and reliability.

2. LITERATURE REVIEW

2.1. Adaptive Decision-Support System

Adaptive Decision-Support System (DSS) is a computer system designed to assist decision makers in dealing with complex and often unstructured situations [10]–[13]. DSS integrates various data sources, analysis techniques, and computational tools to provide relevant information and support better decision making. One of the key features of a DSS is its ability to adapt to different situations and understand user preferences and needs. The main advantage of Adaptive DSS is its ability to learn and change over time [13]. It can utilize historical data and existing patterns to produce increasingly accurate recommendations. In other words, the longer a DSS is used, the better it becomes at providing advice that is appropriate to the changing context.

In addition, Adaptive DSS also has the ability to interact with users [14]. This means that the DSS can receive input from the user and customize recommendations based on the preferences or priorities expressed by the user. This capability allows decision-makers to have more control over the decision-making process, while still utilizing assistance from the system. Adaptive DSS is also highly relevant in the context of current research on crime report analysis and classification. Given the changes in crime patterns that may occur over time, Adaptive DSS can assist authorities in responding quickly. It can improve the ability to classify diverse and changing crime reports according to the current situation.

Overall, Adaptive Decision-Support System is an important innovation in decision-making that allows users to utilize increasingly sophisticated data and analytics, while retaining control over the decision-making process. In the context of the current research, the use of Adaptive DSS can assist in automatically and adaptively analyzing and classifying crime reports, improving the efficiency and effectiveness of crime response.

2.2. Automated Crime Report Analysis

Automated Crime Report Analysis is a technological approach used to automate the process of evaluating, sorting, and understanding crime reports that enter the police system or relevant government agencies [1], [15]. The main objective of this method is to improve efficiency and effectiveness in crime data management, as well as enabling authorities to respond to crime events faster and more accurately. Automated analysis of crime reports involves using technologies such as artificial intelligence and natural language processing to extract key information from incoming crime reports [16]. This includes identifying the type of crime, location, time of occurrence, and reported perpetrator. In addition, the system can also classify reports into different categories, such as street crime, theft, domestic violence, and so on.

The main advantage of automated analysis of crime reports is its ability to manage large volumes of crime data quickly and accurately. This allows authorities to track crime trends, identify possible patterns, and take more effective preventive or enforcement actions. In addition, with this automation approach, crime reports can be integrated with e-Government systems, thus allowing the public to access crime information more transparently and efficiently [17].

However, there are several challenges that need to be overcome in the development of automated analysis of crime reports. One of them is ensuring the accuracy and validity of the data used, as errors in analysis can have a serious impact on law enforcement. In addition, the protection of privacy and data security are also important aspects that need to be considered so that sensitive information in crime reports is preserved. Nonetheless, the potential benefits of automated analysis of crime reports in supporting better law enforcement and public services make it an interesting and relevant area of research and development.

2.3. Classification of Crime Reports

Classification of Crime Reports is an important process in the analysis and management of crime-related data. In this context, a crime report is written or digital information that includes various details about the crime incident, such as the type of crime, date and time of the incident, location, witnesses, and description of the incident [18]–[21]. The purpose of crime report classification is to categorize these reports into appropriate groups based on their characteristics, thus facilitating efficient analysis, monitoring, and action by the authorities.

One of the main benefits of classifying crime reports is its ability to aid investigation and law enforcement. By categorizing reports by crime type, authorities can more easily identify possible crime patterns, as well as common offenders or modus operandi. This can increase the effectiveness of law enforcement efforts in apprehending criminals and preventing similar crimes from occurring in the future. Crime report classification also plays an important role in crime trend analysis. By looking at the history of classified crime reports, authorities can identify changes in crime patterns over time. This information can help in planning security policies and allocating resources more wisely. Proper classification also allows the government to respond to crime incidents more efficiently, especially in emergency situations.

Crime report classification can also provide benefits to the general public. For example, with a system in place that allows citizens to access and monitor crime reports in their area, they can be more aware of potential risks and take appropriate precautions. Thus, the classification of crime reports is not only beneficial to the authorities, but also provides benefits to the public in maintaining their security and safety. In a digital and e-Government era, crime report classification also plays an important role in optimizing the use of technology for better security and law enforcement purposes.

2.4. Crime Data Automation

Crime data automation refers to the use of technology, specifically artificial intelligence (AI) and automated computer systems, to efficiently collect, analyze, and manage crime data. This approach aims to improve the processing of crime reports, enabling law enforcement and government agencies to respond to crimes more quickly and accurately. One of the key aspects of crime data automation is the system's ability to automate the collection of crime reports from various sources, including reports coming in through various digital platforms and emergency calls. With the help of technologies such as NLP, the system can recognize and extract important information from the reports, such as the type of crime, location, time, and other details. This makes it possible to group similar or related crime reports, which can help in crime pattern analysis.

Crime data automation also focuses on utilizing artificial intelligence to automatically classify crime reports. This involves using machine learning algorithms and models to identify the type of crime contained in the report. In this way, law enforcement can quickly determine the actions that need to be taken based on the type of crime reported. Furthermore, crime data automation systems can assist in the monitoring and analysis of crime trends. By collecting and analyzing data continuously, these systems can help in identifying crime patterns that may develop, such as a spike in crime at a certain time or in a certain location. This can help law enforcement in smarter resource allocation.

3. Methodology

3.1. Research Model

To compare and contrast a large set of text reports, crime relevant information must be extracted and similarities between crime reports must be measured. This research has developed algorithms that extract crime-associated entities such as crime scenes, people, weapons, vehicles, and types of crimes based on our hierarchical lexicon, measure document similarity between crime reports, and classify the reports into the same and different crimes. We report here on our next generation document similarity algorithms. They differ from our previous version since they combine our original rule-based approach with a new depth-of-nodes and expert weighting approaches. In addition, a new, complete experiment that uses small and bigger, realistic datasets with more crime reports and several different crimes was conducted with a logistic regression and Naïve Bayes' approach for the crime report classification. We developed a three-layer DSS which consists of an information processing layer, similarity-computation layer, and text-classification layer (see figure 1).

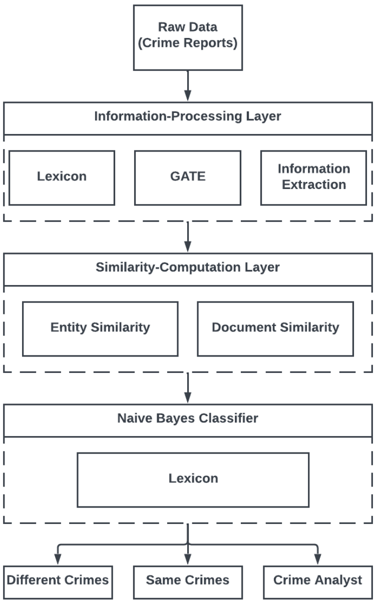

Figure.1. Framework of decision support system.

3.2. Similarity-Computation Layer

The similarity-computation layer is composed of two main components: the entity and document similarity algorithms. Our algorithms combine the two most common similarity measures Jaccard (a component of the entity similarity algorithm) and Dice coefficient (a component of the document similarity algorithm) and the semantic similarity measure. We have evaluated both algorithms earlier with good results and found 87-92% classification accuracy [22], [23] for the document similarity algorithm to identify reports describing the same crime. Therefore, we limit this section to a short overview shown in Tables 2 and 3. Our new generation similarity algorithms include two weighting approaches: the depth-of-nodes (a component of the entity similarity algorithm) and expert weighting (a component of the document similarity algorithm) shown in Table 4. The usefulness of a depth-of-nodes weighting to compute similarity scores has been demonstrated by others, but to our knowledge has never been tested for crime related information. For example, [24] used the depth in WordNet to compute semantic similarity. Didit [25], [26] used a weighting approach, depth-relative scaling, for word sense disambiguation. Wu and Palmer [27] used the depths of two concepts in a tree to compute conceptual similarity for verb translations between English and Indonesian. In our earlier study, we compared our entity similarity algorithm with two WordNet-based similarity algorithms: [28], [29] because both similarity measures are also based on a hierarchical lexicon to compute similarity scores and use depth-of-nodes in a tree to refine the similarity scores. To compare different similarity measures, a gold standard for the comparison with human ratings was developed. The gold standard is based on similarity evaluations by fifteen human evaluators of word pairs [30]. The original gold standard contained 179 entities pairs. However, since pronouns are ignored by the new algorithm, we eliminated them from the gold standard resulting in 143 entity pairs. The reliability and agreement of ratings among fifteen raters were measured by interclass correlation coefficient (ICC) analysis. The result shows that all Cronbach's Alpha values are higher than .9, which indicates that the rater scores are reliable. Our weighted entity similarity measure showed the strongest correlation (r = .783, ← b .001) with 15 human raters (a gold standard), followed by the Leacock and Chodorow similarity measure (r = .679, ← b .001), and the Wu & Palmer similarity measure (r = .573, ← b .001).

Table.1. Components of information-processing layer

Components	Description
Tokenizer	Segments the text into individual tokens such as words and punctuations.
Sentence splitter	Identifies boundaries of sentences in text reports.
POS tagger	Assigns parts of speech to each word such as noun, verb, and adjective.
Stemmer Gazetteer	Identifies the main part of tokens. For example, the stem of 'attacks' and 'attacked' is 'attack'.
Ortho-matcher	Each gazetteer is a list of words used to locate entities such as type of weapons and a suspect's age. The gazetteer lists have been collected and organized into a domain-specific, hierarchical lexicon. Our new lexicon in this study contains 20 semantic trees including 38,000+ words and phrases. Each tree has one root node and several levels of child nodes. Root and child nodes are the main classes and subclasses of the classification. Recognizes uppercase letters such as a brand of automobile Toyota'.
Noun phrase chunker	Extracts noun phrases such as 'a nine-inch knife' in text.
JAPE rules	JAPE rules have been developed to extract entities such as addresses, locations, and people's ages and names.

ISSN 2723-6471

Information filtering	A process to remove stop words such as 'a', 'an', and 'the', remove duplicate entities, and keep relevant entities.

Table.2. Components of entity similarity algorithm

Components	Description
Entity similarity algorithm	Entity Similarity (c1, c2) = $\frac{sem(c1,c2)+Jaccard(c1,c2)}{2} \times \frac{(depth(c2)+depth(c2))/2}{maxdepth(c)}$
Semantic similarity	Calculates the shortest distance between nodes in a tree. A shorter distance between nodes represents higher similarity. Sem (c1, c2) represents a semantic score between two concepts c1 and c2.
Jaccard coefficient	Jaccard coefficient is used to measure overlapping tokens between entities. Jaccard (c1, c2) represents a Jaccard coefficient between concepts c1 and c2.
Depth-of-node weighting	More specific information is weighted more heavily. Max depth (c) represents the maximum depth of a tree. depth (c1) and depth (c2) represent the depth of c1 and c2 in a tree.

Table. 3. Components of document similarity algorithm

Components	Description
Document similarity algorithm	Document Similarity = (average weighted entity similarity + root-node Dice Coefficient + Weighted child-node Dice Coefficient) / 3.
Entity similarity	See Table 2.
Dice coefficient	Measures overlapping information for higher level concepts between documents. Dice coefficient between documents D1 and D2 is defined as: $\frac{2\|D1 \cap D2\|}{\|D1\|+\|D2\|}$ The numerator, |D1 ∩ D2], represents the nodes common to both documents while the denominator represents the combined D1 + D2 number of nodes in both documents. 2 D1 D2
Weighted Dice coefficient	Weighted Dice Coefficient (D1, D2) = $\frac{2\|D1 \cap D2\|}{\|D1\|+\|D2\|}$ × Expert Weighting
Expert weighting	Emphasizes the importance of specific information such as specific race and people's name for crime analysis. Table 4 shows an overview of the adjusted weights for such nodes in our hierarchical lexicon with higher importance. Expert weighting = Sum of weightings for all overlapping child nodes / Max weighting.

Table. 4. Expert weighting - nodes with increased importance

Weightings	Nodes
Triple weighting	• Specific race and ethnicity • Specific weight, height, and age • Specific people's names

Double weighting	• Specific brands and types of vehicles • Specific types of weapons • Type of crimes • Specific locations and stores • Specific electronic devices • Jewelry • Specific date and time
Equal weighting	• All other general information, e.g., general people, clothing, and personal belongings.

3.3. Text-Classification Layer

To identify whether crime reports discuss the same crime, there are multiple approaches possible to classify crime reports ranging from completely manual to automated approaches [31], [32]. We opted for a semiautomated and automated approach. The text-classification layer includes two excellent candidates for making this classification: binary logistic regression (a semi-automated approach) and Naïve Bayes' classifier (an automated approach). Logistic regression is a statistical technique that uses a logistic function to classify cases into categories. Binary logistic regression is commonly used to predict dichotomous results from predictor variables. In our system, the predicted variable is a function of the probability that a similarity score will be in one of two categories - the same crime or different crime. We use the SPSS binary logistic regression tool to conduct such classification and identify cutoff values by the author. A cutoff value can be used to distinguish between high and low similarity. If two reports describe the same crime, their content will be very similar which will result in a high similarity score. If two reports describe completely different crimes, their content will be different and similarity scores will be low. By training datasets, we can locate the best probability cutoff value that leads to the highest percentage of correct classification results.

When a significant amount of training dataset is available, automated approaches can also be used to classify crime reports to minimize human involvement and training processes. Naïve Bayes' classifier is based on a probabilistic learning method. Google Collab, an open-source data mining toolbox, was used to train and apply the Naïve Bayes' classifier. We split crime reports into training and test datasets and the unused report pairs, 13,915 pairs, were used as the training dataset for Naïve Bayes' classifier. The evaluation focuses on our weighted document similarity with different sizes of datasets and classification methods.

3.4. Weighted Document Similarity Study Design

Two studies were conducted with the small and big datasets where both crime analyst and system performance were compared. The first study used small datasets containing 10 crime reports and up to 5 different crimes per set, while the second study used two big datasets containing 40 and 60 crime reports respectively and 16 different crimes per set. The second study differs in the much larger number of crime reports and the number of different crimes in each dataset. This allows us to evaluate our approach in a realistic setting when the number of reports and crimes increases because a crime analyst's effort and time required will increase, and accuracy may decrease. We expect that with increasingly large datasets, the usefulness of our system will also increase.

3.5. Crime Report Collection

In the process of assembling crime reports, we initially enlisted 40 volunteers, all aged 18 or older and without any law enforcement experience. We employed 17 distinct video clips sourced from police training materials, online surveillance footage, and commercial films, with an average clip duration of 2 minutes, ranging from 1 to 5 minutes. Each participant observed four video clips, each viewed a total of 10 times, and was instructed to record their observations. For our initial study, we utilized 100 out of 170 generated crime reports. In our subsequent investigation, focusing on larger datasets, we engaged 30 additional participants from different schools. The same video clips were employed, watched 7-8 times each, yielding 120 crime reports. For this study, we utilized 100 out of the 120 crime reports.

3.6. Research Dataset

In the context of crime video labeling, two researchers reviewed a set of seventeen video clips, obtained from various sources such as police training videos, commercial movies, surveillance footage, and TV programs. These video clips were categorized into five different types of crimes, namely burglary, robbery, assault, theft, and assault, with varying degrees of violence (low, moderate, and high), based on definitions from authoritative sources such as the National Criminal Justice Reference Service, Bureau of Justice Statistics, and Federal Bureau of Investigation. Additionally, the researchers created multiple datasets, labeled from Dataset 1A to Dataset 1J, by randomly selecting 2-5 crime reports from each of the seventeen video clips, resulting in ten datasets with an average report length of 87 words. Furthermore, two larger datasets, referred to as Dataset 2A and Dataset 2B, were compiled by selecting 2-3 and 3-4 reports, respectively, from each of the sixteen different crime video clips, resulting in 40 and 60 crimes for Dataset 2A and Dataset 2B, respectively, with average report lengths of 96 and 113 words, respectively (as summarized in Tables 5 and 6).

Table. 5. Small dataset - video set selection

Datasets (N reports)	Number of different crimes	Video labels (N reports)	Average number of words in reports
Dataset 1A (10)	2	CV11 (5), CV01 (5)	111
Dataset 1B (10)	2	CV02 (5), CV12 (5)	77
Dataset 1C (10)	2	CV13 (5), CV03 (5)	67
Dataset 1D (10)	3	CV04 (3), CV14 (3), CV08 (4)	89
Dataset 1E (10)	3	CV09 (4), CV15 (3), CV05 (3)	104
Dataset 1F (10)	4	CV06 (2), CV04 (3), CV15 (3), CV16 (2)	86
Dataset 1G (10)	4	CV05 (3), CV07 (2), CV10 (3), CV17 (2)	71
Dataset 1H (10)	4	CV06 (3), CV08 (2), CV01 (2), CV16 (3)	107
Dataset 11 (10)	5	CV07 (2), CV09 (2), CV14 (2), CV11 (2), CV02 (2)	96
Dataset 1J (10)	5	5: CV10 (2), CV03 (2), CV12 (2), CV13 (2), CV017 (2)	68
Total (100)	34		87

Table. 6. Big dataset - video set selection

Datasets (N reports)	# of different crimes	Video labels (N reports)	Average number of words in reports
Dataset 2A (40)	16	CV01 (2), CV02 (2), CV03 (2), CV06 (2), CV13 (2), CV14 (2), CV15 (2), CV17 (2), CV04 (3), CV05 (3), CV07 (3), CV08 (3), CV09 (3), CV10 (3), CV12 (3), CV16 (3)	97
Dataset 2B (60)	16	CV01 (2), CV12 (3), CV17 (3), CV02 (3), CV03 (3), CV04 (4), CV05 (4), CV06 (4), CV07 (4), CV08 (4), CV09 (4), CV10 (4), CV13 (4), CV14 (4), CV15 (4), CV16 (4)	112
Total (100)			107

3.7. Research Variable

In this study, we examined various methods, including binary logistic regression, Naive Bayes' classifier, and human analysis by a crime analyst, to determine whether reports discussed the same crime. To measure our results, we employed different dependent variables for two separate studies. In the first study, we utilized receiver operating characteristic (ROC) curves and grouping accuracy, while the second study focused on grouping accuracy and the time expended. Additionally, we conducted a ROC curve analysis to illustrate the strengths and limitations of our algorithm, defining accuracy as the ratio of true positives plus true negatives to the sum of true positives, false positives, false negatives, and true negatives.

3.8. Research Dataset

Two studies were conducted on different dates with the same crime analyst. The study with smaller datasets had been conducted first. The second study with the larger datasets was then conducted after 6 months. The crime analyst met with the researcher at his local Police Department in California and was shown the crime reports in a paper format only (dataset are available on https://www.kaggle.com/datasets/agilesifaka/vancouver-crime-report). Each dataset of crime reports was given each at a time. First, a set of paper-based crime reports were given. The crime analyst was asked to group the reports that, according to him, discussed the same crime and label each group with type-of-crime information (see Fig. 2). The number of different crimes and the label information were not provided to the crime analyst. Upon completion of the first task, the crime analyst was given the next dataset. The same grouping process was required to complete the task for both studies. Each set of crime reports was also processed by the system and grouping accuracy was measured. The weighted document similarity shown in Table 3 was used.

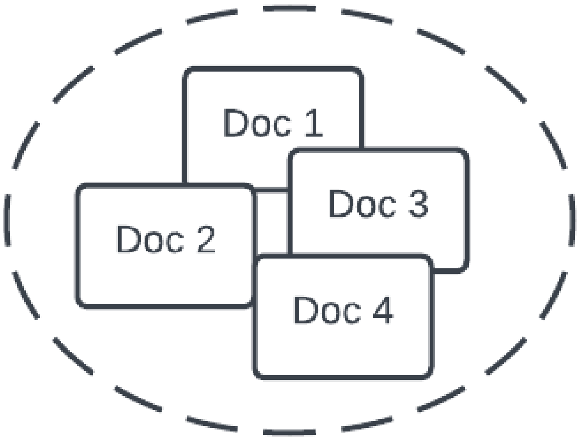

Figure. 2. Grouping crime reports with a type-of-crime label

4. Result Analysis

4.1. Analysis

The ROC curve and the area under the ROC curve (AUC) are commonly used to visualize the trade-offs between sensitivity (true positive rates) and specificity (false positive rates) for possible cutoff values. The closer the ROC curve is to the upper left corner, the better the performance. The AUC is used to measure accuracy. The higher the AUC value, the better the overall performance. Fig. 3 shows the performance of the weighted document similarity algorithm. The ROC curve is far above the reference line (the worst case scenario) and close to the upper left corner (most accurate). The AUC of the weighted document similarity algorithm shows the approach to be significantly better than chance, with the AUC (AUC = .912, → b 0.001) significantly above 0.5. The lower bound of AUC is .879 while the upper bound is .945 with an asymptotic 95% confidence interval.

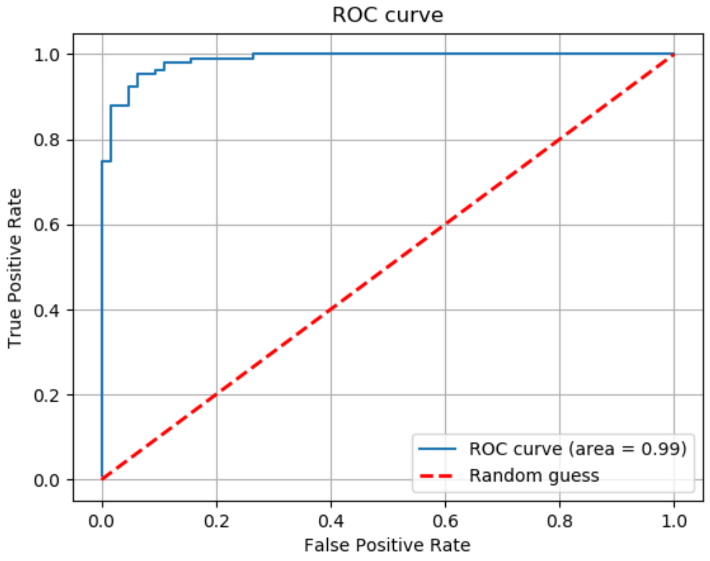

Figure. 3. The ROC curve for all 450 crime report pairs

4.2. Accuracy Analysis

Table 7 presents the detailed results for the 10 datasets for the binary logistic regression, Naïve Bayes' classifier, and the manual approach. The crime analyst achieved the highest accuracy of 96%. The second best score, 88.89% accuracy, was obtained with the semi-automated classification approach (binary logistic regression) and a slightly lower accuracy of 86.67% was obtained for the automated classification approach (Naïve Bayes' classifier). A closer look at individual datasets reveals that the accuracy was in some cases very high, e.g., above 95% in Dataset 1B with binary logistic regression and in Datasets 1D and 1H with Naïve Bayes' classifier, while sometimes as low as 73.33% accuracy in Dataset 1A (Naïve Bayes' classifier). Table 8 shows the average accuracy based on the number of different crimes. When three different crimes were tested, the highest average accuracy (N91%) was found for all classification approaches. Surprisingly, when only two different crimes were tested, the lower average accuracy (80%-94%) was found for the three classification approaches.

Table. 7. Small datasets - accuracy for each dataset

Datasets (N)	Nr. of different crimes	Binary logistic regression cutoff	Naïve Bayes' classifier accuracy	Crime analyst accuracy	Accuracy
Dataset 1A (10)	2	0.254	81.00%	74.33%	100%
Dataset 1B (10)	2	0.244	96.56%	92.11%	100%
Dataset 1C (10)	2	0.255	85.44%	78.78%	84%
Dataset 1D (10)	3	0.275	94.33%	96.56%	100%
Dataset 1E (10)	3	0.246	94.33%	87.67%	100%
Dataset 1F (10)	4	0.240	92.11%	87.67%	100%
Dataset 1G (10)	4	0.252	81.00%	83.22%	100%
Dataset 1H (10)	4	0.256	92.11%	96.56%	87%
Dataset 11 (10)	5	0.250	85.44%	89.89%	100%

Dataset 1J (10)	5	0.242	94.33%	89.89%	91%
Total (100)	34		90.89%	88.67%	96%

Table. 8. Small datasets - average accuracy based on the number of different crimes

Nr. of different crimes	Binary logistic regression average accuracy	Naïve Bayes' classifier average accuracy	Crime analyst average accuracy
2	87.87%	81.62%	95.81%
3	94.33%	92.11%	99.00%
4	87.96%	89.46%	96.56%
5	90.47%	89.89%	96.56%

Table. 9. Big datasets - accuracy for each dataset and time spent by the crime analyst

Datasets (N)	Time spent by the crime analyst (min)	Naïve Bayes' classifier accuracy	Crime analyst accuracy
Dataset 2A (40)	87	96.26%	95.74%
Dataset 2B (60)	62	93.63%	94.35%
Total (100)	149	95.82%	94.72%

To simplify the classification process and reduce human involvement, Naïve Bayes' classifier was selected to process the large datasets (batch processing evaluation). Table 9 presents the results of the batch processing evaluation, including the time spent by the crime analyst and the accuracy of both Naïve Bayes' classifiers. Surprisingly, the system scores with the Naïve Bayes' classifier achieved higher accuracy (94.82%) than the crime analyst's grouping accuracy (93.76%). For Dataset 2A with 40 crime reports, the system achieved the highest accuracy (95.26%), while the crime analyst's grouping accuracy (94.74%) was slightly lower. For Dataset 2B with 60 crime reports, the system achieved the highest accuracy (94.63%), while the crime analyst's grouping accuracy (93.33%) is slightly lower. The crime analyst spent more time on Dataset 2A (87 min) than on Dataset 2B (62 min). For the user experience, the crime analyst rated Dataset 2A as the most difficult to group crime reports describing the same crimes together, rated "slightly easy" for grouping the increasing crime reports, and rated "slightly confident" for grouping crime reports correctly.

5. Discussion

5.1. Contributions to E-Government and Law Enforcement

This paper presents how NLP techniques can be used to increase government agencies' efficiency Shoopik et al. [1] without sacrificing the quality of text report analysis. Automated text report analysis is attractive because it can enhance and accelerate the discovery and analysis process and consequently shorten government agencies' response time to the public sector. A DSS can be useful in processing and analyzing crime information efficiently and generating decision support information necessary to solve crimes. Such a DSS can alleviate information overload, reduce the time to search, process, and analyze crime information, and thus reduce crime analysts' workload and save the precious police resources for more important tasks, typically when budget cuts are confronted by law enforcement agencies.

5.2. Technical Contributions

The information extraction algorithm and domain-specific lexicon were developed to extract crime-related entities from unstructured text reports. The lexicon we constructed focuses on a single domain so we can ignore the words in other domains to reduce errors. The lexicon and extracted entities are reusable, since high precision and recall were achieved for the proposed algorithm. They can be used to measure entity, string, sentence, and even document similarity and to highlight words and phrases. We provided an integrated solution including a domain-specific lexicon, NLP techniques, similarity measures specially tailored for crime analysis, and an automatic classification approach to automate the process of analyzing crime reports. For example, the experimental results show how the similarity scores can be combined with an automated classification approach, i.e., Naïve Bayes' classifier and achieved high accuracy when the number of crime reports and of different crimes significantly increases. Furthermore, the similarity scores generated by the algorithms we developed are reusable, since high classification accuracy was obtained for automated and semi-automated classification approaches. They can be used for crime report classification, clustering, and even information visualization.

6. Conclusions

The law enforcement agencies have adopted online crime reporting systems, resulting in a surge of digital crime reports. While these systems provide valuable information, they also present the challenge of analyzing and classifying the growing volume of reports, making it a time-consuming task for crime analysts. The complexity of crime report analysis no longer lies solely in similarity measures; it now demands domain-specific lexicons and specialized weighting approaches explicitly designed for automated crime analysis. To address this challenge, we have developed a Decision Support System (DSS) capable of identifying reports related to similar or identical crimes.

This study approach uses a document similarity algorithm to generate a similarity score, which is then integrated with automated methods, specifically a Naive Bayes classifier, to identify reports that discuss the same criminal incident. To maximize the effectiveness of criminal analysts without increasing financial costs or manpower requirements, our algorithm is fine-tuned using hierarchical depth and expert knowledge encoded as rules. Evaluation results show high accuracy, justifying the feasibility of this approach.

The DSS contributions extend to the realms of e-government and law enforcement. By showcasing the potential of Natural Language Processing (NLP) techniques, we illustrate how e-government agencies can improve efficiency without compromising the quality of text report analysis. Automated text report analysis offers the advantage of speeding up the discovery and analysis process, which in turn reduces government response times. DSS proves invaluable in efficiently processing and analyzing crime information, generating decision-support information critical to crime resolution. Such systems reduce redundant information, reduce the time required for data search and analysis, and consequently ease the workload of crime analysts, saving valuable police resources, especially during budget constraints.

The technical contributions include the development of information extraction algorithms and domain-specific lexicons designed to extract crime-related entities from unstructured text reports. This lexicon, which is tailored to one domain, minimizes errors by ignoring words from other domains. The extracted lexicons and entities exhibit high levels of precision and recall, so they can be reused for a variety of purposes, including entity, string, sentence, and document similarity measurement, as well as word and phrase highlighting. We provide an integrated solution that includes domain-specific lexicons, NLP techniques, specialized similarity measures for crime analysis, and automated classification approaches, which simplifies the crime report analysis process.

Although our work shows substantial progress, there are two main limitations. First, the system sometimes classifies two reports describing different crimes as the same due to high similarity scores, which often arise due to overlap in the reports (e.g., similarities between suspects or stolen items). NLP, text mining and advanced classification techniques can help distinguish such cases. Second, our study involved only one expert participant.

To improve the practicality and usability of the system for law enforcement agencies, future efforts should involve collaboration with multiple experts and stakeholders, as well as additional user studies.

In the future, we plan to improve the Information Extraction (IE), similarity, and classification algorithms. For the IE algorithm, our goal is to incorporate lexical resources such as Wikipedia and Urban Dictionary to improve precision and recall in entity extraction. In terms of similarity algorithms, we intend to explore various weighting schemes and data sets to improve the performance and applicability of the system.

References

[1] F. Skopik, G. Settanni, and R. Fiedler, “A problem shared is a problem halved: A survey on the dimensions of collective cyber defense through security information sharing,” *Comput. Secur.*, vol. 60, pp. 154–176, 2016, doi: 10.1016/j.cose.2016.04.003.

[2] N. Shonhadji and A. Maulidi, “Is it suitable for your local governments?,” *J. Financ. Crime*, vol. ahead-of-p, no. ahead-of-print, Jan. 2020, doi: 10.1108/JFC-10-2019-0128.

[3] N. Archer, “Consumer identity theft prevention and identity fraud detection behaviours,” *J. Financ. Crime*, vol. 19, no. 1, pp. 20–36, Jan. 2012, doi: 10.1108/13590791211190704.

[4] U. M. Butt, S. Letchmunan, F. H. Hassan, M. Ali, A. Baqir, and H. H. R. Sherazi, “Spatio-temporal crime hotspot detection and prediction: a systematic literature review,” *IEEE access*, vol. 8, pp. 166553–166574, 2020.

[5] M. Usmonov, “Legal Legislative Basis for Detection of Information Crime.,” *Sci. Acad. Pap. Collect.*, 2021.

[6] T. Vo *et al.*, “Crime rate detection using social media of different crime locations and Twitter part-of-speech tagger with Brown clustering,” *J. Intell. Fuzzy Syst.*, vol. 38, no. 4, pp. 4287–4299, 2020.

[7] A. S. Adeniji and M. Emekayi, “Information Technology and Crime Detection and Fighting in Nigeria and India: A Comparative Analysis,” *J. Glob. South Res. Secur. Dev.*, vol. 1, no. 1, 2022.

[8] V. Mahor, R. Rawat, S. Telang, B. Garg, D. Mukhopadhyay, and P. Palimkar, “Machine learning based detection of cyber crime hub analysis using twitter data,” in *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)*, IEEE, 2021, pp. 1–5.

[9] G. Jerry *et al.*, “Camera-Based Crime Behavior Detection and Classification,” *Available SSRN 4480362*.

[10] N. S. Muhlis Tahir, “Penerapan Algoritma Fp-Growth Dalam,” *J. Ilm. NERO*, vol. 6, no. 1, pp. 56–63, 2021.

[11] J. Ahamed, R. N. Mir, and M. A. Chishti, “Industry 4.0 oriented predictive analytics of cardiovascular diseases using machine learning, hyperparameter tuning and ensemble techniques,” *Ind. Robot Int. J. Robot. Res. Appl.*, vol. ahead-of-p, no. ahead-of-print, Jan. 2022, doi: 10.1108/IR-10-2021-0240.

[12] M. Pourabbasi and S. Shokouhyar, “Unveiling a novel model for promoting mobile phone waste management with a social media data analytical approach,” *Sustain. Prod. Consum.*, vol. 29, pp. 546–563, 2022, doi: https://doi.org/10.1016/j.spc.2021.11.003.

[13] D. S. Abdelminaam, F. H. Ismail, M. Taha, A. Taha, E. H. Houssein, and A. Nabil, “CoAID-DEEP: An Optimized Intelligent Framework for Automated Detecting COVID-19 Misleading Information on Twitter,” *IEEE Access*, vol. 9, no. December 2019, pp. 27840–27867, 2021, doi: 10.1109/ACCESS.2021.3058066.

[14] A. Rizal and W. K. Raharja, “Pemilihan Pegawai Penerima Beasiswa Pendidikan Menggunakan Algoritma TOPSIS Dan ISO 9126 Pada Badan Nasional Penanggulangan Bencana (BNPB),” *J. Teknol. Inf.*, vol. 6, no. 1, 2022, [Online]. Available: http://jurnal.una.ac.id/index.php/jurti/article/view/2575%0Ahttp://jurnal.una.ac.id/index.php/jurti/article/viewFile/2575/2015

[15] C. V. Amasiatu and M. H. Shah, “The management of first party fraud in e-tailing: a qualitative study,” *Int. J. Retail Distrib. Manag.*, vol. 47, no. 4, pp. 433–452, Jan. 2019, doi: 10.1108/IJRDM-07-2017-0142.

[16] O. E. Akinbowale, H. E. Klingelhöfer, and M. F. Zerihun, “An innovative approach in combating economic crime using forensic accounting techniques,” *J. Financ. Crime*, vol. 27, no. 4, pp. 1253–1271, Jan. 2020, doi: 10.1108/JFC-04-2020-0053.

[17] N. Omar, Z. ‘Amirah Johari, and M. Smith, “Predicting fraudulent financial reporting using artificial neural network,” *J. Financ. Crime*, vol. 24, no. 2, pp. 362–387, Jan. 2017, doi: 10.1108/JFC-11-2015-0061.

[18] M. Abdellatif *et al.*, “A taxonomy of service identification approaches for legacy software systems modernization,” *J. Syst. Softw.*, vol. 173, no. 6, pp. 110–131, 2021, doi: 10.1016/j.jss.2020.110868.

[19] G. Boskou, E. Kirkos, and C. Spathis, “Classifying internal audit quality using textual analysis: the case of auditor selection,” *Manag. Audit. J.*, vol. 34, no. 8, pp. 924–950, Jan. 2019, doi: 10.1108/MAJ-01-2018-1785.

[20] J. P. B. Saputra and R. P. Bernarte, “The Naive Bayes Algorithm in Predicting the Spread of the Omicron Variant of Covid-19 in Indonesia: Implementation and Analysis,” *Int. J. Informatics Inf. Syst. Vol 5, No 2 March 2022DO - 10.47738/ijiis.v5i2.131* , Mar. 2022, [Online]. Available: http://ijiis.org/index.php/IJIIS/article/view/131

[21] Z. Jiang, B. Gao, Y. He, Y. Han, P. Doyle, and Q. Zhu, “Text classification using novel term weighting scheme-based

improved tf-idf for internet media reports," *Math. Probl. Eng.*, vol. 2021, pp. 1–30, 2021.

[22] H. Abbasimehr and M. Shabani, "A new methodology for customer behavior analysis using time series clustering: A case study on a bank's customers," *Kybernetes*, vol. 50, no. 2, pp. 221–242, 2021, doi: 10.1108/K-09-2018-0506.

[23] M. T. Lwin, M. M. Aye, and others, "A modified hierarchical agglomerative approach for efficient document clustering system," *Am. Sci. Res. J. Eng. Technol. Sci.*, vol. 29, no. 1, pp. 228–238, 2017.

[24] R. Endsuy, "Sentiment Analysis between VADER and EDA for the US Presidential Election 2020 on Twitter Datasets," *J. Appl. Data Sci.*, vol. 2, no. 1, pp. 8–18, 2021, doi: 10.47738/jads.v2i1.17.

[25] V. Cohen-Addad, V. Kanade, F. Mallmann-Trenn, and C. Mathieu, "Hierarchical clustering: Objective functions and algorithms," *J. ACM*, vol. 66, no. 4, 2019, doi: 10.1145/3321386.

[26] D. Suhartono and K. Khodirun, "System of Information Feedback on Archive Using Term Frequency-Inverse Document Frequency and Vector Space Model Methods," *IJIIS Int. J. Informatics Inf. Syst.*, vol. 3, no. 1, pp. 36–42, 2020, doi: 10.47738/ijiis.v3i1.6.

[27] H. Fei, S. Wu, Y. Ren, and D. Ji, "Second-Order Semantic Role Labeling with Global Structural Refinement," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 1966–1976, 2021, doi: 10.1109/TASLP.2021.3082706.

[28] G. Ramaswami, T. Susnjak, A. Mathrani, J. Lim, and P. Garcia, "Using educational data mining techniques to increase the prediction accuracy of student academic performance," *Inf. Learn. Sci.*, vol. 120, no. 7/8, pp. 451–467, Jan. 2019, doi: 10.1108/ILS-03-2019-0017.

[29] S. S. Khan and B. Taati, "Detecting unseen falls from wearable devices using channel-wise ensemble of autoencoders," *Expert Syst. Appl.*, vol. 87, pp. 280–290, 2017, doi: 10.1016/j.eswa.2017.06.011.

[30] J. John and R. Thakur, "Long term effects of service adaptations made under pandemic conditions: the new 'post COVID-19' normal," *Eur. J. Mark.*, vol. 55, no. 6, pp. 1679–1700, Jan. 2021, doi: 10.1108/EJM-08-2020-0607.

[31] A. Faizah, "Implementation of the Convolutional Neural Network Method to Detect the Use of Masks," *IJIIS Int. J. Informatics Inf. Syst.*, vol. 4, no. 1, pp. 30–37, 2021, doi: 10.47738/ijiis.v4i1.75.

[32] I. Nordat, B. Tola, and M. Yasin, "The Effect of Work Motivation and Perception of College Support on Organizational Commitment and Organizational Citizenship Behavior in BKPSDM, Tangerang District," *Int. J. Appl. Inf. Manag.*, vol. 2, no. 3, pp. 37–46, 2022, doi: 10.47738/ijaim.v2i3.36.