IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 23 October 2023, accepted 2 February 2024, date of publication 7 February 2024, date of current version 15 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3363879

RESEARCH ARTICLE

CAMELON: A System for Crime Metadata Extraction and Spatiotemporal Visualization From Online News Articles

SIRIPEN PONGPAICHET, BOONYAPAT SUKOSIT, CHITCHAYA DUANGTANAWAT, JIRAMED JAMJONGDAMRONGKIT, CHANCHEEP MAHACHAROENSUK, KANTAPONG MATANGKARAT, PATTADON SINGHAJAN, THANAPON NORASET, AND SUPPAWONG TUAROB, (Member, IEEE)
Faculty of Information and Communication Technology, Mahidol University, Salaya 73170, Thailand

Corresponding author: Suppawong Tuarob (suppawong.tua@mahidol.edu)

This research project is supported by Mahidol University (Fundamental Fund: fiscal year 2023 by National Science Research and Innovation Fund (NSRF)).

This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by Mahidol University Central Institutional Review Board under Certificate of Exemption Nos. MU-CIRB 2022/253.1909 and MU-CIRB 2022/297.0211, and performed in line with the International Guidelines for Human Research Protection.

ABSTRACT Crimes result in not only loss to individuals but also hinder national economic growth. While crime rates have been reported to decrease in developed countries, underdeveloped and developing nations still suffer from prevalent crimes, especially those undergoing rapid expansion of urbanization. The ability to monitor and assess trends of different types of crimes at both regional and national levels could assist local police and national-level policymakers in proactively devising means to prevent and address the root causes of criminal incidents. Furthermore, such a system could prove useful to individuals seeking to evaluate criminal activity for purposes of travel, investment, and relocation decisions. Recent literature has opted to utilize online news articles as a reliable and timely source for information on crime activity. However, most of the crime monitoring systems fueled by such news sources merely classified crimes into different types and visualized individual crimes on the map using extracted geolocations, lacking crucial information for stakeholders to make relevant, informed decisions. To better serve the unique needs of the target user groups, this paper proposes a novel comprehensive crime visualization system that mines relevant information from large-scale online news articles. The system features automatic crime-type classification and metadata extraction from news articles. The crime classification and metadata schemes are designed to serve the need for information from law enforcement and policymakers, as well as general users. Novel interactive spatiotemporal designs are integrated into the system with the ability to assess the severity and intensity of crimes in each region through the novel Criminometer index. The system is designed to be generalized for implementation in different countries with diverse prevalent crime types and languages composing the news articles, owing to the use of deep learning cross-lingual language models. The experiment results reveal that the proposed system yielded 86%, 51%, and 67% F1 in crime type classification, metadata extraction, and closed-form metadata extraction tasks, respectively. Additionally, the results of the system usability tests indicated a notable level of contentment among the target user groups. The findings not only offer insights into the possible applications of interactive spatiotemporal crime visualization tools for proactive policymaking and predictive policing but also serve as a foundation for future research that utilizes online news articles for intelligent monitoring of real-world phenomena.

The associate editor coordinating the review of this manuscript and approving it for publication was Walter Didimo.

INDEX TERMS Crime monitoring, online news articles, spatiotemporal information, crime metadata extraction.

I. INTRODUCTION

The economic ramifications of crimes and accidents can be substantial, impacting not only individuals but also entire communities and nations [1], [2]. Although there have been reports of a decrease in crime rates in several developed nations [3], [4], it is still evident that developing countries continue to experience a high incidence of serious crimes and accidents, leading to personal injuries and fatalities, as well as hindering economic growth [5], [6], [7], [8]. Timely observation of the progression of crimes can be advantageous for law enforcement, policymakers, and other pertinent parties in their proactive and enduring efforts to prevent and address the root causes of these adverse incidents. For example, in the event that policymakers detect a surge in theft within a tourism-dependent city and identify that poverty contributes as a major root cause, it may be feasible to introduce incentives to generate employment opportunities in the tourism sector for individuals living in impoverished conditions. This approach could potentially lead to a decrease in instances of theft, an increase in tourist traffic, and the promotion of overall safety and economic well-being. Additionally, this information can furnish individuals with the ability to anticipate and prepare for potential criminal activity in the vicinity. For instance, if a surge in road accidents during a festive season is detected, the local police may inquire into the underlying factors and distribute suitable medical provisions to the impacted region. In the event that drunk driving is identified as a significant contributing factor, law enforcement agencies may consider deploying checkpoints or patrols in the vicinity of alcohol-selling places. Additionally, policymakers may opt to impose more stringent fines, limit the hours during which alcohol can be sold, and stimulate economic growth in the region by offering affordable transportation services to discourage individuals from driving while under the influence.

Many ideas have been proposed to implement criminal activity monitoring systems [9], [10]. Nevertheless, the majority of these systems are reliant on historical crime reports, which may suffer from delays and infrequency. To address these concerns, online news articles have been recognized as a trustworthy and timely alternative for obtaining information on crime incidents [11], [12], and have been integrated into contemporary crime monitoring systems [13], [14]. However, these news-based crime monitoring systems merely categorize news articles into suitable crime types and subsequently display them on the map according to the extracted geolocations. While these previously proposed systems have enabled the visualization of criminal activities across various regions, local law enforcement and national policymakers require more comprehensive insights into the trends and nuanced structural details of criminal activities to effectively manage crime and make informed policy decisions. In addition, general people conducting research on potential travel destinations, business investment opportunities, or places to reside would benefit from concise and informative data regarding the prevalence and severity of criminal activity in particular areas. Thus, it is imperative to develop a novel crime monitoring system that can effectively process real-time data from online news articles and provide comprehensive insights regarding detailed information, overall trends, and regional and national perspectives to cater to the diverse requirements of various stakeholders mentioned above. Therefore, this paper proposes *CAMELON*, an intelligent system for collecting, processing, and interactively visualizing spatiotemporal *C*rime and *A*ccident *M*onitoring information from *E*xtensive *O*nline *N*ews articles. Specifically, the system routinely collects news articles and classifies them into seven finer-grained crime types, including gambling, murder, sexual abuse, theft/burglary, drug, battery/assault, and accident. Each crime article is further processed to extract common crime metadata, including the criminal, victim, involving police, date/time, location, evidenced items, action, worth (or damage), root cause motivation, and trigger. Deep learning cross-lingual models are validated and deployed in both the classification and metadata extraction tasks. The best configuration of the models yielded 86%, 51%, and 67% F1 in the crime type classification, overall metadata extraction, and closed-form metadata extraction tasks.

The extracted information is stored in a database where the front-end system retrieves and further processes to interactively visualize the crime information in a spatiotemporal manner at the vicinity, regional, and national levels. The usability evaluation and qualitative survey confirm that the main functionalities of the proposed system are well received by the target users. The proposed system is highly generalizable to different contexts or countries with different sets of prevalent crime types since the processing pipeline is not specific to the proposed crime classification scheme. In addition, the utilization of cross-lingual models in the system enables it to generalize toward the processing of news articles composed in various languages without necessitating extra annotated datasets in the target languages.

Concretely, the key contributions of this paper are as follows:

- A survey was conducted among prospective target users to ascertain their current and preferred means of receiving crime-related information. The survey results revealed that a significant proportion of the participants wanted the ability to navigate crime information at the vicinity level. However, the conventional sources of obtaining such information, such as newspapers and television, were not deemed as effective by the majority of the respondents. This identified need gap necessitates the development of a novel crime information system that enables users to easily access and explore crime

data across various temporal scales and geographical locations.

- We proposed using deep learning methodologies for extracting structural crime data from digital news articles. The aforementioned data pertains to the classification of highly specific criminal activities and the retrieval of crime-related metadata. A novel scheme for crime classification and metadata labels was proposed where rigorous evaluations revealed that a cross-lingual model (i.e., XLMR) performed best in both tasks.
- We proposed the *CAMELON* system, designed to facilitate the intelligent collection, analysis, and interactive visualization of spatiotemporal crime information extracted from extensive online news articles. The system is equipped with interactive gadgets that display the trends of various crime types at both regional and national levels. Additionally, it provides detailed information on crime incidents on the map as well as a Criminometer that evaluates the severity and intensity of criminal activity in each respective area. A usability evaluation was conducted that quantitatively and qualitatively ascertained the usefulness of the proposed novel functionalities.

The remainder of this article is organized as follows. Section II provides an overview of the related work on crime monitoring systems, emphasizing those fueled by online news sources. Section III reports the preliminary survey on the crime information consumption of the target users. Then, Section IV discusses the proposed methodology in detail, of which the corresponding experiment results and discussions are provided in Section V. Section VI brings into attention societal implications and ethical issues should the proposed system be implemented for real-world applications. Finally, Section VII concludes the paper.

II. RELATED WORK

In recent years, there has been a growing interest among research communities in the field of predictive analysis and visualization pertaining to proximity-based criminal activities [15], [16], [17], [18]. Established online news outlets have emerged as viable alternative sources of timely and reliable real-world information that are easily comprehensible and dependable [19]. One of the reasons is that the utilization of standard language in news articles, with only minor deviations in linguistic styles, has resulted in satisfactory accuracy for many contemporary machine-learning language models [20]. Moreover, given that the business model of news publishers is predicated on the provision of factual and informative content, the statements contained in news articles are frequently subjected to validation procedures prior to dissemination, thereby circumventing the credibility challenges that are associated with information obtained from social media platforms [21]. This section discusses related work on mining news data for crime analysis purposes, emphasizing crime categorization, crime metadata extraction, and crime monitoring tasks.

A. AUTOMATIC CATEGORIZATION OF CRIME NEWS ARTICLES

Since news articles comprise a variety of topics, recent research has developed methods to discern crime-associated articles from those found in online news sources, frequently conceptualizing the issue as a binary (crime versus non-crime) or multiclass classification problem. The study conducted by Kalmegh [22] aimed to assess the effectiveness of REPTree, Simple Cart, and RandomTree algorithms in classifying Indian news articles into seven distinct categories using bag-of-words representation. Furthermore, Magnusson et al. [23] proposed using a basic conjugate Bayesian model to detect potential news leads that may pique the interest of journalists. The effectiveness of their proposed model was validated with a collection of reported offense news. Recently, the problem of identifying crime news was also addressed by Ghankutkar et al. [13] through a binary classification task where SVM, Naive Bayes, and Random Forest were validated on the TF-IDF representations of news articles.

However, the mere ability to classify news articles as crime or non-crime may not provide significant additional value, as numerous reputable news sources already have sections dedicated to crimes. Therefore, an additional research area pertaining to the extraction of crime information from news articles has adopted the multiclass classification task to classify a crime news article into one of the more detailed categories for subsequent meticulous analyses. Rajapakshe et al. [24] employed a dataset consisting of crime news articles gathered from Sri Lanka in 2018 and validated Decision Tree, Random Forest, and Support Vector Machine (SVM) algorithms for their ability to classify each article into one of nine crime categories, namely: murder, kidnapping, robbery, drug dealing, accident, rape, assault, and burglary. Umair et al. [25] conducted a study in which 900 news articles pertaining to crime were collected from eight prominent news outlets in Pakistan dated from 2011 to 2019 and categorized into eight distinct types of crime and accidents, namely robbery, accident, blast, kidnapping, murder, shot, suicide, and arrest. They employed an n-gram representation for each document and conducted experiments using k-Nearest Neighbors (kNN) and Random Forest algorithms for the purpose of multiclass classification. In addition, it is worth noting that every document underwent geo-coding through GeoPy[1] to enable its representation on the map. Recently, Thaipisutikul et al. [26] introduced a multi-class classification algorithm for categorizing news articles in Thailand into five more specific crime categories, namely burglary, accident, corruption, drug, and murder. The authors utilized TF-IDF features to represent individual news articles and validated various conventional machine learning classification algorithms, including Multinomial Naive Bayes, Gradient Boosting Machine, Random Forest,

[1] https://geopy.readthedocs.io/en/stable/

kNN, Multinomial Logistic Regression, and Support Vector Machine.

In addition to employing conventional n-gram-based and classification algorithms, the emergence of deep learning has facilitated the analysis of criminal activities in news articles. The study conducted by Rollo et al. [12] involved utilizing Word2Vec embeddings to represent Italian news articles and evaluating the efficacy of various traditional machine learning classification algorithms in categorizing these documents into one of the 13 crime categories, namely theft, drug dealing, illegal sale, robbery, aggression, scam, murder, kidnapping, mistreatment, evasion, sexual violence, money laundering, and fraud. In addition, the matter of data imbalance was thoroughly examined and addressed using SMOTE. The issue of fine-grained crime classification in Google News was tackled by Deepak et al. [27], who employed Fuzzy c-Means clustering to facilitate data labeling. Subsequently, GloVe was used to derive word embeddings, which were utilized to train a BiLSTM classifier to classify a news article into one of the 14 distinct crime types. Recently, the study conducted by Khan et al. [11] aimed to assess the efficacy of a Bangla BERT-based model in comparison to conventional deep learning models such as LSTM and BiLSTM in their ability to classify Bangla crime news headlines into six distinct categories, namely terrorism, murder, corruption, harassment, drug, and robbery. The study involved manual labeling of 7,897 news headlines in the Bangla language, which revealed that the Bangla-BERT-Base model exhibited superior performance.

B. CRIME METADATA EXTRACTION FROM NEWS ARTICLES

The ability to classify crime news articles into respective finer-grained categories, as reviewed in the previous section, could prove crucial for the automatic selection of relevant news articles for monitoring specific crime types. In addition, the ability to further extract specific common details about a crime, such as criminals, victims, actions, weapons, and root causes, could enable interesting applications in predictive policing and crime-related policymaking purposes, including criminal profiling [28], criminal tracking [29], and criminal motif analysis [30]. Given that crimes are among the most frequently covered topics by news media, existing research has been directed toward devising methods for extracting significant insights from news articles pertaining to crime events.

Primary crime metadata information, such as names, locations, and dates, are relatively easy to spot in a news article, where a set of rules or patterns could be constructed for the extraction. Ku et al. [31] were among the first to establish the problem of crime metadata extraction from text and proposed a rule-based method to extract crime-related information such as weapons, vehicles, time, persons, clothes, and location. Their method first utilized a noun phrase chunker to extract noun phrases, which were then passed to predefined JAPE (Java Annotations Pattern Engine) patterns and classified into their respective types. Rahem and Omar [32] proposed a rule-based algorithm using a set of patterns to extract drug crime information from online news articles, including location, nationality, drug names, quantity, and prices. The extracted locations were also linked to the gazetteer to develop a system for predicting where and how drugs were hidden, identifying the dealers' nationalities, and evaluating the drugs' prices in the market. Srinivasa and Thilagam [33] constructed a knowledge base for crimes by mining crime-related entities and relations from online newspapers. Rule-based and semantic similarity-based approaches were utilized to identify untagged and incorrectly tagged entities. Recently, Rahma and Romadhony [34] created a set of patterns that combine dependency parsing and part-of-speech tagging techniques to extract crime metadata attributes from news articles composed in the Indonesian language. These attributes include crime type, victim, criminal, location, and time.

These aforementioned studies relied on human-defined patterns and rules to extract crime information attributes, which could be effective if the rules have wide coverage of different variants of patterns that characterize the desired information. However, these language-specific pre-defined rule-based methods often face limitations in generalization to other linguistic patterns and unforeseen samples [35]. Therefore, to mitigate these issues, data-driven approaches have been proposed to extract crime metadata from online news articles. Arulanandam et al. [36] proposed a machine-learning method for extracting locations of theft crimes in newspaper articles. Their method first utilizes a Conditional Random Field (CRF) model to detect sentences having location information using a set of hand-crafted features, referred to as crime location sentences (CLS). Then, named entity recognition algorithms were deployed to identify location entities in each CLS. Later, Dasgupta et al. [37] proposed to extract crime-related entities and events from published news articles. Such information includes the criminal's name, victim's name, type of crime, location, date/time, and action taken against the criminal. Their method involves applying a named entity recognition algorithm to extract standard named entities, which are then categorized into respective crime entity types using a Support Vector Machine (SVM) classifier. In addition, Sedik and Romadhony [38] extracted the locations and dates from Indonesian crime news articles by first identifying sentences containing crime scene information using SVM. Then, a named entity recognition algorithm via SpaCy was applied to each crime scene sentence to extract locations and dates. Most studies utilizing machine learning approaches to extract crime metadata framed such a problem as a named entity recognition (NER) problem where conventional NER tools could be applied. However, such an approach is limited to extracting standard named entities such as persons, locations, dates, and times. In predictive policing, certain crime-related information, such as criminal actions, damages, the worth of evidence or stolen items, criminal's background motivation, and criminal's

motives to commit crimes, has been deemed useful [39] but would not be captured by standard NER tools due to being free-text, therefore, requiring more semantic understanding approaches. Therefore, a novelty of our proposed system is the ability to extract such free-text crime attributes in addition to those already investigated in the previous literature.

C. MONITORING SYSTEMS FOR CRIMES

Existing monitoring systems for crime activities using online news articles as the information source usually support a variety of analysis tools to classify, estimate, and visualize relevant incidents, as crime information presented in news media is rich in relevant factual information. Since the main source of crime information in these systems comes from newspaper articles that comprise a wide variety of topics, primary monitoring systems seek to automatically identify crime-related articles and then simply visualize them. For example, Wajid and Samet [40] proposed CrimeStand, an extended version of the NewStand system [41], a map-query interface for monitoring news sources, by adding a new ''Crime'' layer. This layer filtered and displayed only crime-related news along with types of crime. Their approach is to leverage StandfordNER for name and entity extraction and an SVM classifier to detect crime-related news. Furthermore, Ghankutkar et al. [13] and Chowdhury et al. [14] proposed systems that classify and visualize news articles into crime and non-crime types using machine learning models enhanced by incorporating news sentiments.

Besides online news data sources, some crime monitoring systems have also incorporated crime reports by authorized organizations. Since most crime reports often include the locations of the incidents, certain crime monitoring systems seek to utilize such geolocations to visualize these incidents on the map, where further analyses could be performed. Gorr and Lee [42] created an early warning system by estimating chronic hot spots using kernel density smoothing analysis. [43] introduced a PREVNET desktop application with many visual analyses, such as similar node features, collaborating clusters, and sub-cluster analysis, using network visualization and trend analysis. Similarly, Tatale and Bhirud [44] incorporated crime data from a police station and predicted the crime hot spots using a data mining technique. Sukhija et al. [45] used a statistical analysis tool (SaTScan) to identify crime hot spots by determining the crime clusters in statistical terms. Recently, Garcia-Zanabria et al. [46] used a narcotics dataset to perform spatiotemporal analysis to determine crime patterns at the street level. While the aforementioned systems have utilized the extracted geolocations of crime incidents for useful analyses that provide a better landscape of criminal activities in different areas, users in law enforcement and policy-making domains could further benefit from detailed crime metadata extracted from news content. Furthermore, the ability to assess the overall crime intensity and severity in each region could prove useful to general users in seeking safe destinations for their travels, residents, and businesses.

This study examines the limitations presented by the current crime monitoring systems reviewed above and proposes a novel system called *CAMELON*. This intelligent and comprehensive system is designed to cater to the needs for crime information of local law enforcement, national-level policymakers, and the general people. The objective of the *CAMELON* system is to effectively utilize the entirety of the information contained in crime news reports. This involves the extraction of comprehensive crime metadata from news content, utilizing the extracted geolocations and time of incidents for multi-level spatiotemporal visualization and analysis, summarizing regional criminal intensity and severity through the implementation of a novel Criminometer index, and developing a user-friendly web-based system for easy navigation. Furthermore, utilizing state-of-the-art deep learning cross-lingual language models, the system is designed to be generalizable to other crime classification schemes and linguistic contexts governing different criminal landscapes and languages in different countries. The results of the user-based usability evaluation indicate that the proposed system has the potential to be beneficial for the intended target users. However, the evaluation also highlights areas for improvement that could better meet the specific needs of different user cohorts. Table 1 presents a comparative analysis of the existing crime monitoring systems and the proposed *CAMELON* system.

III. PRELIMINARY STAKEHOLDER SURVEY

Before developing the proposed system, a survey was conducted among a diverse group of potential users, including individuals from various age groups and occupations. The survey primarily targeted those working in fields related to law enforcement, policymaking, students, and the general public. The aim of the survey was to evaluate the existing methods and objectives of obtaining crime-related information while identifying the gaps in the information needs. The inquiry was categorized into three distinct groups, namely, participants' background information, current methods of obtaining criminal activity updates, and requisites for location-based crime information. The survey was administered through the distribution of online questionnaires to the designated user cohorts in Thailand, allowing them to partake in an entirely confidential manner or abstain from participation.

A. BACKGROUND INFORMATION

A total of 119 participants responded to the survey whose distribution of age ranges is illustrated in Figure 1. The majority of the participants are at least 36 years old (66.4%), followed by the age ranges 19-26 years (22.7%), 27-35 years (9.2%), and not older than 18 years (1.7%). As suggested by the aforementioned age range distribution, the majority of the participants tend to already have stable jobs or are currently in college. Figure 2 shows the distribution of the participants' occupations. A total of 79 participants were government officers, followed by 29 students, six business

TABLE 1. Comparison between existing crime monitoring systems and the proposed *CAMELON* systems.

Publication	Data Source	Area of Implementation	Crime Classification	Metadata Extraction	Spatiotemporal Analysis	Web Based App
Modeling Machine Learning for Analysing Crime News (Ghankutkar et al., 2019)	Online News Aticles	USA	Machine Learning	No	No	No
Crime Monitoring from Newspaper Data based on Sentiment Analysis (Chowdhury et al., 2019)	Headlines of online newspaper in Twitter	World	Sentiment Analysis	No	No	No
CrimeStand: Spatial Tracking of Criminal Activity (Wajid & Samet, 2016)	RSS News	Worldwide	Machine Learning	No	Map Query Interface	Yes
Early Warning System for Temporary Crime Hot Spots (Gorr & Lee, 2015)	Part 1 Violent Crime Data	Pennsylvania, USA	Not Required	Using Metadata from Crime Data	Chronic Crime Hot Spots	No
A Tool for Analysis and Visualization of Criminal Networks (Rasheed & Wiil, 2015)	Chicago Narcotics Dataset	Chicago, USA	Not Required	Using Metadata from Crime Data	Network, Temporal, Composite Visual Analytics	Windows Based App
Spatial Visualization Approach for Detecting Criminal Hotspots (Sukhija et al., 2017)	District-wise Crime Dataset	Haryana, India	Not Required	Using Metadata from Crime Data	SATScan to Detect Crime Hot Spots	Google Earth
Criminal Data Analysis in a Crime Investigation System using Data Mining (Tatale & Bhirud, 2016)	Crime Data at the Police Station	Pune, India	Not Required	Using Metadata from Crime Data	Predictive Analysis on Crime Hot Spots	Yes
Mirante: A visualization tool for analyzing urban crimes (Garcia-Zanabria et al., 2020)	Authorized Crime Dataset	Sao Paulo & Sao Carlos, Brazil	Not Required	Using Metadata from Crime Data	Spatio-temporal Crime Patterns at Street Level	Yes
***CAMELON* (Proposed System)**	**Online News Articles**	**Thailand**	**Deep Learning**	**Deep Learning**	**Map Views, Timeline, Visualization Tools**	**Yes**

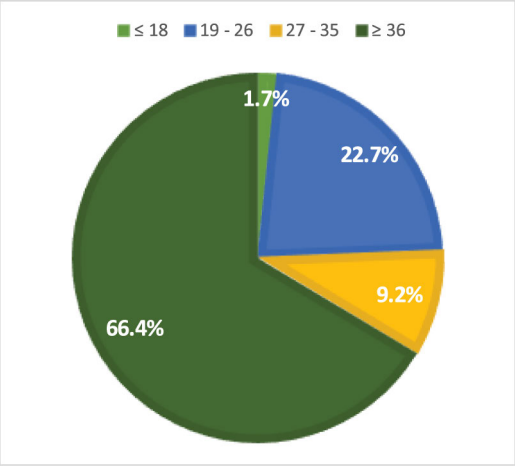

FIGURE 1. Age ranges of the survey population.

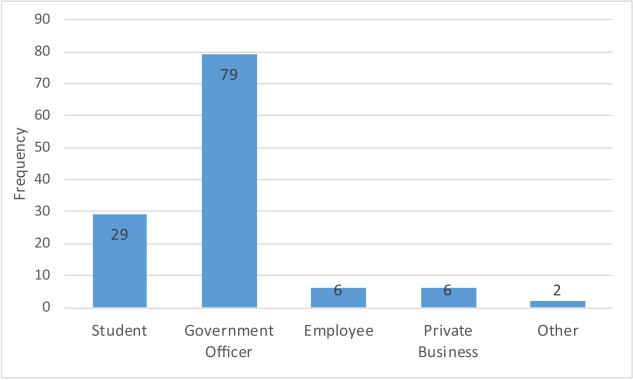

FIGURE 2. Current occupations of the survey population.

employees, six business owners, and two other occupations. Note that the government officer group also includes law enforcement officers and those working at the policymaking level. Furthermore, the sum of the numbers in Figure 2 is greater than 119 since a participant can have more than one job. For example, one can be a full-time government officer while pursuing a part-time graduate degree as a student.

B. CURRENT CRIME NEWS INFORMATION

The second part of the questionnaire inquired about the current means of receiving crime information from the news.

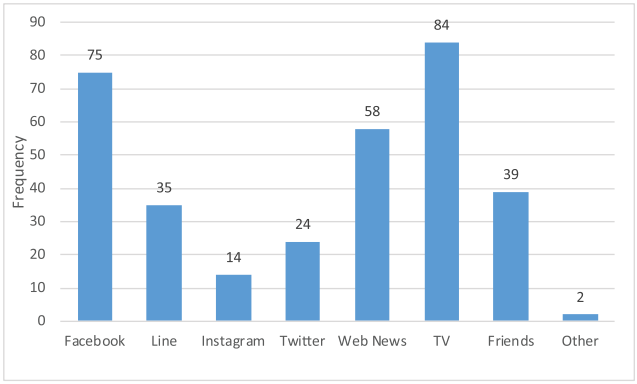

FIGURE 3. Sources for receiving crime news information.

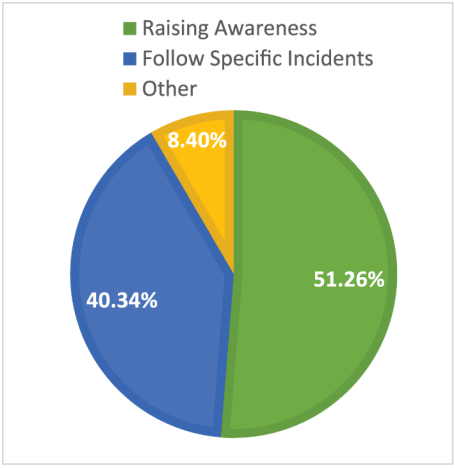

FIGURE 4. Purposes of receiving crime news.

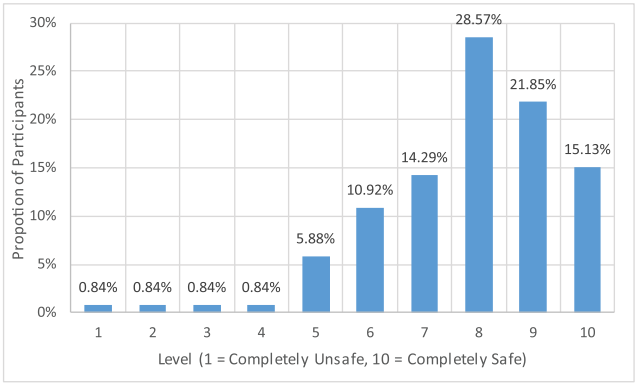

FIGURE 5. Perception of current safety.

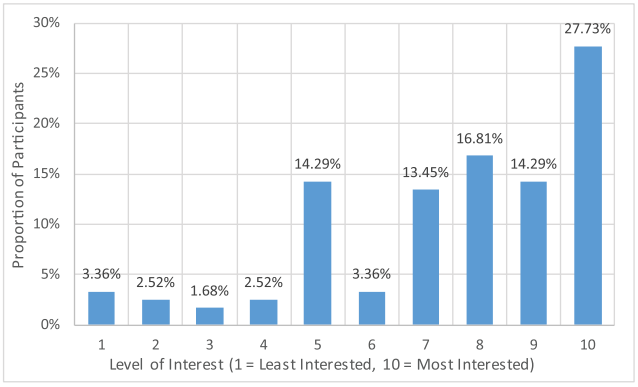

FIGURE 6. Level of interest for vicinity-level crime information.

Each participant was asked about the current channel for receiving crime news, where one could choose more than one option. Figure 3 summarizes the responses. The majority of the participants opted for television (n = 84), followed by Facebook (n = 75) and online news from the publishers' websites (n = 58).

Next, each participant was asked about the purposes of keeping abreast of crime news, and the responses were provided as free text. We summarized the responses to this question and categorized them into three categories: raising awareness (to prevent similar crimes from happening to oneself), following updates of specific crime incidents (especially major crimes that trigger public concerns or controversy), and others. Figure 4 visualizes the responses, where the majority followed crime news to learn from past incidents to prevent or to protect themselves from getting involved in ones. In summary, the survey responses in this part indicate that there is still a persisting need for timely and convenient information on crime activity.

C. NEEDS FOR SPATIAL CRIME INFORMATION

In Thailand, there has not been an established crime monitoring system that can process and visualize crimes in a spatiotemporal fashion. Only closely similar existing system in Thailand merely uses quantitative crime statistics from Thailand's Ministry of Interior[2] to visualize the frequency of different crime types at a coarse-grained regional level.[3] However, the crime statistics used in their system's visualization date back to 2019 without a further update. Furthermore, the spatial information is quite coarse-grained and does not provide a clear landscape of current criminal activities that local police can use to make an informed decision. Therefore, the questionnaire in this part was designed to understand the target groups' need for timely spatial information on criminal activities at the vicinity level.

The first question asked each participant whether they felt safe from everyday crimes around their vicinity. Figure 5 depicts the distribution of the responses ranging on a Likert scale from 1 (completely unsafe) to 10 (completely safe). The distribution appears left-skewed, indicating that the majority of participants felt comfortably safe in the proximity of their living.

The next question then inquired each participant about their interest in the ability to navigate crime information in a spatial manner at the vicinity level (both their own and

[2]http://edw-opendata.moi.go.th/dataset/page/5e9fb64e35a3945ea/521caba5cc1e2e915ed575168900

[3]https://github.com/KittapatR/Tumruat

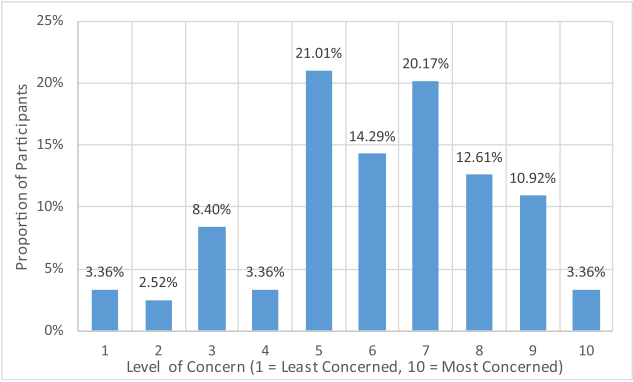

FIGURE 7. Level of concern for crime-related safety when traveling or relocating to other places.

others) as opposed to conventional ways of perceiving crimes traditionally ranked by recency and importance alone (i.e., highlighted crimes selected by news publishers). Figure 6 shows the distribution of the Likert responses ranging from 1 (least interested) to 10 (most interested). The distribution appears to be left-skewed, indicating that most of the participants yearned for the ability to navigate area-based crime incidents at the vicinity level.

The last question in this part specifically targeted users who were general people, inquiring how concerned they were about crimes when traveling or relocating to different places. The participants' responses were measured using the Likert scale, which ranges from 1 (least concerned) to 10 (most concerned). The distribution of these responses is visually represented in Figure 7. The distribution exhibits a left-skewed pattern, indicating that a significant proportion of respondents reported a degree of apprehension regarding criminal incidents occurring in the areas they intended to visit or reside in. It is noteworthy that the existing crime information systems in Thailand lack the capability to assimilate and furnish exhaustive analysis regarding criminal activities in individual locations, thus rendering them inadequate in addressing the concerns of the intended users with regard to crimes in close proximity. This evidence further underscores the importance of having a spatiotemporal crime activity visualization that is user-friendly and easily accessible to diverse users.

IV. METHODOLOGY

Prior research has proposed implementing crime monitoring systems that rely on news reports. However, these systems focus on classifying crime reports into finer-grained categories and geographically plot crime occurrences based on the extracted geolocations and gazetteers. As proposed, the *CAMELON* system introduces novel supplementary functionalities that involve extracting and visualizing crime metadata from individual news articles. Moreover, users can examine the patterns of particular criminal activities within each locality nationwide. The proposed system includes the novel Criminometer index that measures the aggregate level and severity of criminal activity in each region. Implementing such a system could potentially yield advantages not only for regional law enforcement and high-level governmental decision-makers but also for individuals seeking secure destinations for their purposes.

Figure 8 illustrates the high-level diagram of the proposed *CAMELON* system. First, online news articles are routinely collected. Each article is parsed for the publish date/time and textual information such as the title, introduction, and body text. Each article is then classified into finer-grained crime types, where non-crime articles are discarded. Each crime article is then parsed for metadata attributes. The crime type and metadata information are stored in a central database, where the front-end system retrieves relevant information for further processing and visualization. The proposed system exhibits a high degree of generalizability, as its methods are not limited by crime types, metadata schemes, languages, and countries, rendering it applicable in any geographical context where predominant sets of crime types are different and news articles are published in diverse languages. The next subsections delve into individual components in more detail.

A. DATA COLLECTION AND PROCESSING

The proposed *CAMELON* system is fueled by publicly accessible online news articles. Once reputable news outlets are identified, automated mechanisms could be programmed to collect news articles routinely in accordance with each news publisher's regulations. The frequency of collection can be determined based on the availability of computing resources. Each news article is collected in a raw HTML format. Furthermore, different news publishers present their news articles in different formats. Therefore, an HTML parser must be implemented for each news outlet with the aim of extracting commonly available information, namely publication date/time, title, introduction, and body text. The extracted information is stored in the database for further digesting and processing.

B. CRIME TYPE CATEGORIZATION

While most news outlets have a dedicated category for crime reports, such a dichotomy classification is often too coarse-grained for downstream analyses of criminal activities. Furthermore, publishers may pre-categorize some crime reports into non-crime categories. For example, "Local News" may also report crime and accident incidents from less populated regions within the country that are relatively less sensational compared to those that make it to the main "Crime" category. Therefore, relying on the publishers' categories is inherently insufficient for downstream tasks, especially those requiring analyzing crime activities at the vicinity level.

To mitigate the limitations, the proposed system incorporates an automatic categorization of news articles into their respective finer-grained crime types. Crime types considered in this research include gambling, murder, sexual abuse,

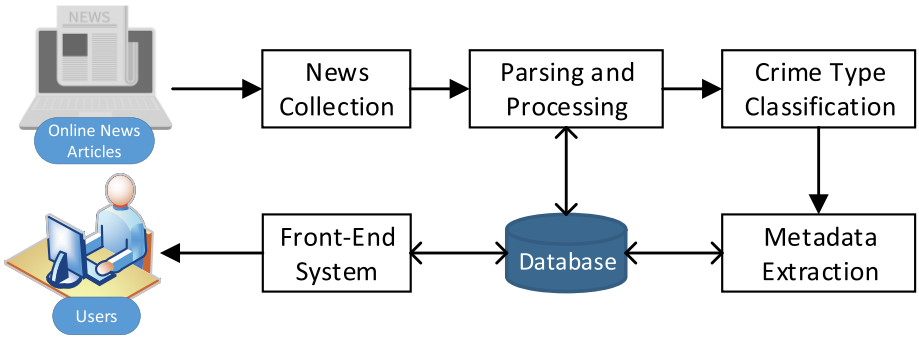

FIGURE 8. High-level diagram describing the proposed *CAMELON* system.

theft/burglary, illegal drugs (both dealing and consumption), battery/assault, and accidents. Articles that do not fall within the aforementioned crime types are classified as non-crime and neglected by the system. Note that categorizing accidents as criminal activities is a topic of debate [47]. However, this study chooses to delve into accidents, particularly those occurring on roads, due to the violent nature of such incidents that are instigated by human actions and lead to loss of life and property, akin to criminal acts. Consequently, the capacity to analyze and represent accident data may assist law enforcement personnel and policymakers in disrupting and averting such violence at its underlying causes. Furthermore, we observed that a crime news article could fall into multiple crime types. For example, a news report of a drug addict getting caught for beating up civilians while being influenced should be categorized as drug and battery/assault crimes. Therefore, the crime type categorization problem is framed as a multi-label text classification task, where a news article, represented by its textual content, is labeled with at least one crime type. A dataset for developing the automatic crime categorization was constructed. Three human annotators manually labeled a subset of news articles, where the majority votes were used to resolve the final labels.

Many state-of-the-art deep learning algorithms for document classification can be modified to serve the multi-label classification task by adjusting the last component of the fully connected layer to output independent probabilities associated with each class. This research considered four deep learning models, including Bidirectional Long Short-Term Memory (BiLSTM) [48], WangchanBERTa, Multilingual BERT (MBERT) [49], and XLM-RoBERTa (XLMR) [50]. WangchanBERTa [51] is a RoBERTa-based model pre-trained with Thai corpora. Pre-training these models does not rely on specific language knowledge and can be easily generalized to other languages as long as the text can be tokenized. The MBERT and XLMR models were pre-trained with parallel copula with diverse languages; therefore, these two models could support multiple languages by default.

We fine-tuned these models on the labeled dataset. The specifications of the models and the training were as follows. First, we used the binary cross-entropy loss to train the models using Adam optimizer [52]. The BiLSTM utilized the pre-trained Thai2Vec word embeddings with 300 dimensions [53] and was trained for 100 epochs using the learning rate of $1e-3$. The transformer-based models (WangchanBERTa, MBERT, and XLMR) were fine-tuned for four epochs [49] using the learning rate of $2e-5$, and the weight decay of 0.01.

We reported the performance of the classification models using a standard 10-fold cross-validation protocol. For each fold, the news articles were stratified and divided into training, validation, and testing sets in the ratio of 80:10:10. Standard classification evaluation metrics, including precision, recall, and F1-score of each class, were reported. We used the F1-score as the main evaluation criteria.

C. CRIME METADATA EXTRACTION

Given a crime news article, the ability to extract important crime information is useful in crime monitoring and analysis applications [33]. For example, learning the accumulative worth of theft incidents could allow policymakers to estimate the severity of such problems that can be used to investigate the population poverty in the target region as a potential root cause. Furthermore, upon observing a surge in road accidents from drunken drivers during a festival period, local law enforcement may opt to augment its patrol presence in the vicinity of recreational establishments or furnish transportation services for inebriated drivers. In addition, it is possible for policymakers at the local level to increase the penalties associated with instances of driving under the influence of alcohol.

The crime metadata extraction problem is framed as a named entity recognition (NER) task. Specifically, the crime metadata extractor parses a crime news article and identifies spans comprising sequences of same-class tokens that constitute different types of metadata of interest. Different labels, types, and examples of crime metadata attributes used in this research are listed in Table 2. The selection of these metadata labels is based on the previous studies reviewed in Section II such as criminal, victim, place, and date/time, with additional labels from consultation with experts in law enforcement

TABLE 2. Types of crime metadata as well as their descriptions and contextual examples. Certain named entities are anonymized.

Label	Type	Description	Example
Criminal	Person	Name or nickname of the person who commit crime	"...มีหยดเลือดไหลยาวเป็นทางจากบริเวณประตูหน้าบ้าน ส่วนผู้ก่อเหตุชื่อนายชื่อ***นายสมมุติ*** หรือ***ชื่อเล่น*** อายุ 18 ปี หลานชายแท้ๆ ซึ่งชาวบ้านช่วยกันควบคุมตัวไว้ที่หน้าบ้าน..." "... Drops of blood ran down the path from the front door. As for the perpetrator, ***Mr. Firstname*** or ***Nickname***, 18 years old, his real nephew, whom the villagers helped detain in front of the house. ..."
Victim	Person	Name or nickname of the direct victim	"... เจ้าหน้าที่กู้ภัย เร่งช่วยเหลือ ***นายสมมุติ นามสกุล*** อายุ 18 ปี หรือ "***อายุ***" ส่ง รพ.พัทลุง หลังถูกคู่อริขับรถยนต์ตามประกบ และใช้อาวุธปืนไม่ทราบชนิดและขนาด ยิงใส่รถเก๋ง ..." "... The rescue team rushed to help ***Mr. Firstname Lastname***, 18 years old, or "***Nickname***", sent to Phatthalung Hospital after being spied on by an enemy driving a car and using a firearm of unknown type and size to shoot at the car. ..."
Police	Person	Name of the police officer directly involved in the crime event	"... วันที่ 17 พ.ค. 65 ***ร.ต.อ.ชื่อตำรวจA นามสกุลตำรวจA*** รอง สว.สอบสวน สภ. XXX อ.เมืองXXX ได้รับแจ้งเหตุ ... จึงพร้อมด้วย ***พ.ต.อ.ชื่อตำรวจB นามสกุลตำรวจB*** ผกก.สภ.เมืองXXX นำกำลังตำรวจป้องกันและปราบปราม ตำรวจสืบสวน ..." "... On May 17, 2022, ***Pol. Capt. PoliceNameA PoliceLastNameA***, deputy inspector of the XXX Sub-Police, Mueang XXX District, received notice that ... along with ***Pol. Col. PoliceNameB PoliceLastnameB***, Superintendent of City XXX Police Station, led the police force to prevent and suppress the situation. ..."
Date/ Time	Date	Date and time when the crime incident takes place	"... เมื่อ***วันที่ 9 ก.ค.63*** มีชาวบ้านในพื้นที่หมู่ 2 ต.XXX อ.เมือง จ.XXX แจ้งว่า มีคนร้ายแอบเข้าไปทุบเจดีย์ ..." "... On ***July 9, 2020***, a villager in the area of Moo 2, XXX Subdistrict, Muang District, XXX Province reported that a villain sneaked into the pagoda. ..."
Location	Location	Location of the crime	"... คนร้ายแอบเข้าไปทุบเจดีย์บรรจุกระดูกคนตาย ภายใน***วัดควนมะพร้าว*** พังเสียหายจำนวนมาก ..." "... The villain sneaked in to break the pagodas containing the bones of the dead at ***Khuan Maphrao Temple***. Many pagodas have been damaged. ..."
Item	Object	Evidential items in the crime scene, such as drugs (drug crime), stolen items (theft crime), weapons (muder or battery/assualt crime), vehicles (accident), etc.	"... พร้อมของกลาง***มีดทำครัวยาว 5 นิ้ว*** ที่คนร้ายได้นำไปล้างและเช็ดคราบเลือดผู้ตายออกแล้ว ..." "... with a ***5-inch long kitchen knife*** that the villain has washed and wiped off the blood stains of the deceased ..."
Action	Free Text	Criminal action's toward the victim	"... จนกระทั่งวันนี้โดยสารรถตุ๊กๆ มาลงที่วัด แล้วก่อเหตุ***ใช้อาวุธมีดแทงพระจนบาดเจ็บสาหัส*** เบื้องต้นทราบว่า คนร้าย กับ พระองค์นี้ ไม่เคยมีเรื่องบาดหมางกันมาก่อน ..." "... Until today, the perpetrator took a tuk-tuk to the temple and ***pierced a monk with a knife, causing serious injury***. Initially, it was known that the perpetrator and the monk had never had any conflict before. ..."
Worth	Free Text	Damage of the criminal acts, such as financial worth (drug, gambling, and threft crimes), mortality (murder crime), physical injury (battery/assault crime), etc.	"... ตำรวจเข้าตรวจสอบจับกุมและตรวจยึดของกลางยาไอซ์ 1,500 กก. มูลค่าเบื้องต้นกว่า ***500 ล้านบาท*** แต่หากหลุดไปถึงประเทศที่ 3 จะมีราคาสูงหลายพันล้านบาท ..." "... The police inspected, arrested, and seized 1,500 kg of ICE drug, initially worth more than ***500 million baht***. But if it goes to the third country, it will cost several billion baht. ..."
Root Cause	Free Text	Criminal's root cause or background motivation that may have indirect effect on the dicision to commit crime such as criminal financial status, mental disorder, and addiction	"...นางไก่ กล่าวอีกว่า ที่ลูกสาวทำแบบนี้เพราะ***ไม่มีเงินซื้อของประทังชีวิตให้ลูกๆ*** จึงตัดสินใจขโมยของภายในร้านสะดวกซื้อ..." "... Mrs. Kai said that her daughter did this because she ***didn't have money to buy things and provide for her children***. So she decided to steal stuff from the convenience store. ..."
Trigger	Free Text	Motives that trigger the decision to commit crime such as extreme hunger, provocation, revenge, jealousy, etc.	"...พบรถเก๋งฟอร์ด เฟียสต้า สีขาว ไม่ติดหมายเลขทะเบียน ขับปาดหน้า ตนจึง***บีบแตรไล่ และเปิดกระจกด่าทอ*** แล้วขับรถต่อ โดยไม่ได้คิดอะไร จนกระทั่งขับมาถึงทางเบี่ยงออกคู่ขนาน พบรถคันดังกล่าวขับตามมา ลดกระจกยิงปืนใส่รถตนหลายนัด..." "... A white Ford Fiesta sedan without a registration number cut ahead of his car. He then ***honked, opened the mirror, scolded*** the other driver, and continued driving without thinking. Once the other driver reached a parallel detour, he reduced the mirror and shot a gun at the victim's car many times. ..."

and public policies. Note that these metadata attributes are shared among diverse types of crimes. Since the target case study comprises Thai news articles, a comparable English translation is also provided. The ***underlined bold-italic texts*** are the spans that should be extracted according to their corresponding label. Criminals, victims, and police are of the person type, which can be identified with names and their prefixes. Date/time, locations, and items are of the date, location, and object types, respectively. Note that the aforementioned metadata attributes are standard named entities commonly defined in a typical NER task. Therefore, it appears that a natural solution for most of the previous crime metadata extraction methods is to utilize existing NER tools to extract standard named entities (e.g., person, location, and date/time) first before classifying them into crime-related types [37], [38].

In addition to standard named entities, we also propose extracting free-text responses as crime metadata, including actions, worth, root causes, and triggers. An *action* describes the criminal acts performed by the perpetrators. The aggregate trends of criminal actions could shed light on the legal loopholes or weak enforcement that necessitates policy-level remedy [54]. For example, weapon control policies can be devised upon seeing heightened trends of murder and assault involving gun or knife fights. The *worth* information refers to damage or value as a result of the corresponding criminal action, such as the worth of stolen properties (theft crime) or damages (accident), the value of the evidenced drugs (drug crime), and injuries or mortality (murder, sexual abuse, and assault crimes). Such *worth* information could enable additional quantitative crime analysis pertaining to the collective damages of certain crime types in different regions [55]. The *root cause* information refers to the background condition or motivation that leads to the tendency to commit a crime, such as poverty, alcohol influence, belief, and mental condition. It should be noted that a root cause does not directly contribute to the decision-making process of committing crimes. Rather, it serves as a pre-existing condition that increases the perpetrator's chance to make the decision to engage in criminal behavior. Studying the root causes of crimes has gained attention in criminology [56]. While these root causes are not direct excuses for committing crimes, if evidence shows that a crime rate has risen because of certain root causes, then policymakers could inject preventive measures to combat or suppress these root causes before they escalate into uncontrollable criminal acts. For example, upon learning that theft crimes are prevalent in an area where people are poor, governmental agencies responsible for public labor could implement activities that involve furnishing people with the necessary skills for the job markets. Finally, *trigger* information or proximity risk factor [57] involves the causes that urge perpetrators to decide to commit crimes, especially violent and intentional ones. Triggers are often spontaneous circumstances, such as emotional temptation, hallucinations, and deception. What differentiates a root cause from a trigger is that a root cause can be directly and effectively addressed by implementing appropriate policies, while a trigger is more individual and circumstantial.

While the addition of the aforementioned novel free-text crime metadata attributes is crucial for crime monitoring and policy-level analysis purposes, extracting them can prove difficult because standard NER tools do not recognize these free-text attributes. Therefore, we propose to build a crime metadata extractor from scratch following the NER methodology. First, samples of crime news articles were annotated using Doccano's named entity recognition annotation tool,[4] as depicted in Figure 9. After the annotation was finished and verified, each article was exported from the annotation tool, where the text was tokenized. Since the case study involved Thai articles, PyThaiNLP's Newmm[5] (default) tokenizer was used; however, open-source tokenizers are available for various languages should the proposed system be adopted in a different linguistic setting. Figure 10 illustrates an example of annotated tokens from a sentence.

Several traditional machine learning and deep learning algorithms were explored for this task, including Conditional Random Field (CRF), Bidirectional Long Short-Term Memory with Conditional Random Field (BiLSTM-CRF), WangchanBERTa, and XMLR. The CRF model was optimized using the L-BFGS algorithm with a maximum of 500 iterations. The BiLSTM-CRF model initialized the word embeddings using the pre-trained Thai2Fit [53] with a hidden size of 400. The model was trained using the Adam optimizer with a recurrent dropout rate of 0.5, batch size of 32, and training epochs of 20. For transformer-based models (WangchanBERTa and XLMR), the models were fine-tuned using the Adam optimizer for four epochs using the learning rate of $2e-5$ and the weight decay of 0.01. Since the length of a news article may exceed the maximum capacity of deep learning models (i.e., 512 tokens), each article was first segmented into sentences, where the metadata extractor is run on each sentence instead of a whole document.

The assessment of the crime metadata extraction task's effectiveness involves using standard evaluation metrics such as precision, recall, F1, MCC, and accuracy. The annotated news articles were first document-wise divided into three stratified sets, namely training, validation, and testing sets, with a ratio of 80:10:10, respectively. The training process for each model involved utilizing the training set, with hyperparameters being optimized on the validation set and subsequent evaluation on the testing set.

D. SYSTEM IMPLEMENTATION

We implemented a web-based application to demonstrate and promote the proposed system's benefits. First, we analyzed the preliminary surveys, which guided us to the four analytical tasks addressed by *CAMELON*, including crime pattern identification, spatiotemporal crime navigation, crime metadata visualization, and regional overall crime assessment. Then, we designed the system architecture using a three-tier architecture approach [58]. Finally, a variety of visualization tools were selected to accomplish the aforementioned tasks.

1) SYSTEM TASKS

From the survey results in Section III, the proposed system is designed to offer users the following four primary tasks.

- Task #1 - Identify crime patterns at the national, regional, and province levels: Showing crime trends and patterns at different fine-grained spatial levels can better help estimate crimes, which benefits law enforcement officers in making preemptive decisions.

[4]https://github.com/doccano/doccano

[5]https://pythainlp.github.io/docs/2.0/api/tokenize.html

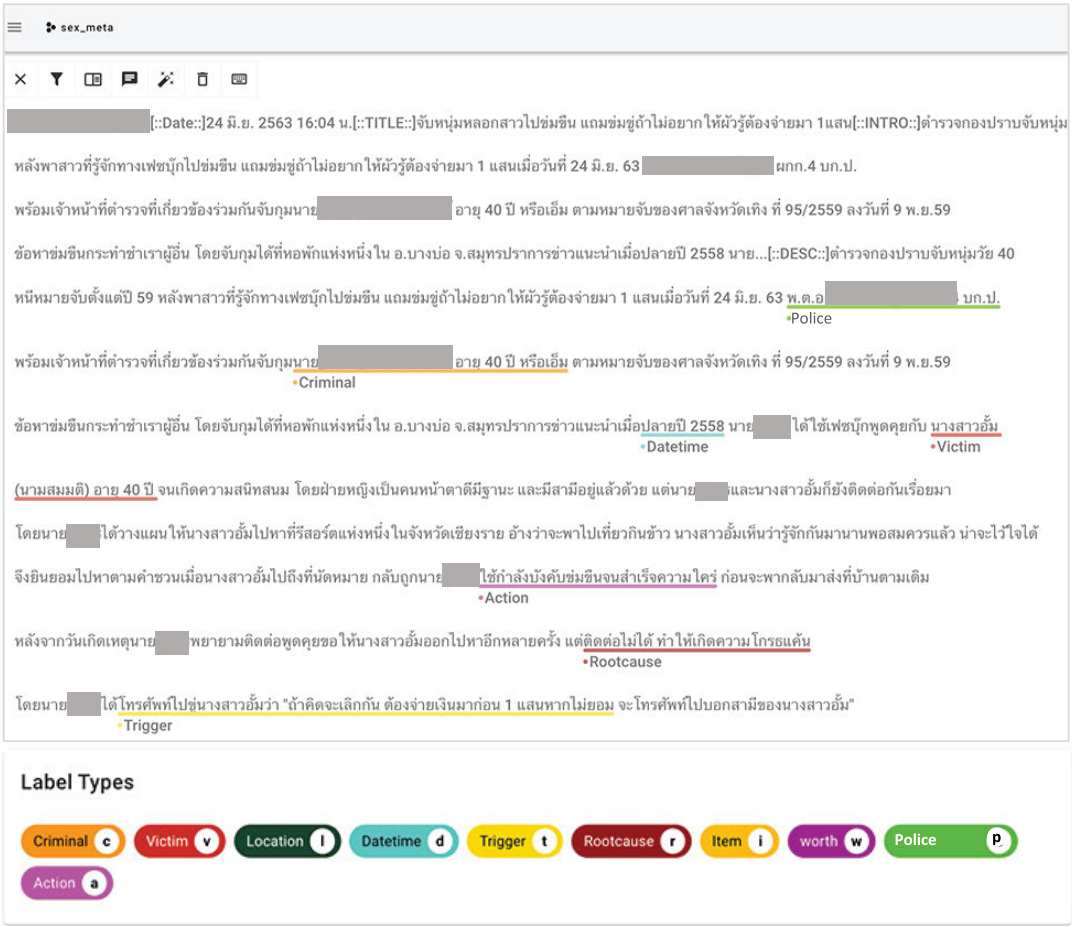

FIGURE 9. Example of Doccano's snapshot of labeling session for the crime metadata extraction task. Certain named entities are anonymized.

Token	เมื่อ	วัน	ที่	19	มิ	.	ย	.	ร	.	ต	.	ท	
NER Tag	O	B-DateTime	I-DateTime	I-DateTime	I-DateTime	I-DateTime	I-DateTime	I-DateTime	B-Police	I-Police	I-Police	I-Police	I-Police	I-Police
Token				รอง	สารวัตร	สอบสวน	รับ	แจ้ง	เหตุ	คน	ถูก	ฆ่า	เสีย	ชีวิต
NER Tag	I-Police	I-Police	I-Police	O	O	O	O	O	O	O	O	B-Action	B-Worth	I-Worth
Token	คา	บ้าน	เช่า	แห่ง	หนึ่ง	ใน	ต	.	ปา	เส	มัส	อ	.	สุ
NER Tag	I-Worth	I-Worth	I-Worth	O	O	O	B-Location	I-Location	I-Location	I-Location	I-Location	I-Location	I-Location	I-Location
Token	ไหง	โก	-	ลก	จ	.	นราธิวาส	ที่	เกิด	เหตุ	พบ	ศพ	น	.
NER Tag	I-Location	I-Location	I-Location	I-Location	I-Location	I-Location	I-Location	O	O	O	O	O	B-Victim	I-Victim
Token	ส	.				อายุ	56	ปี						
NER Tag	I-Victim	I-Victim	I-Victim	I-Victim	I-Victim	O	O	O						

FIGURE 10. Example of annotated tokens in a sentence. Certain named entities are anonymized.

- Task #2 - Interactive selection of spatial, temporal, and crime's type aspects of interest: Since we collected numerous new articles from more than ten years and offered them in aggregated and individual value, the ability to highlight the areas and temporal periods of interest to navigate crimes could prove to benefit users.
- Task #3 - Visualization of crime metadata extracted by the metadata extractor: Local police and policymakers can investigate the important structural details of each

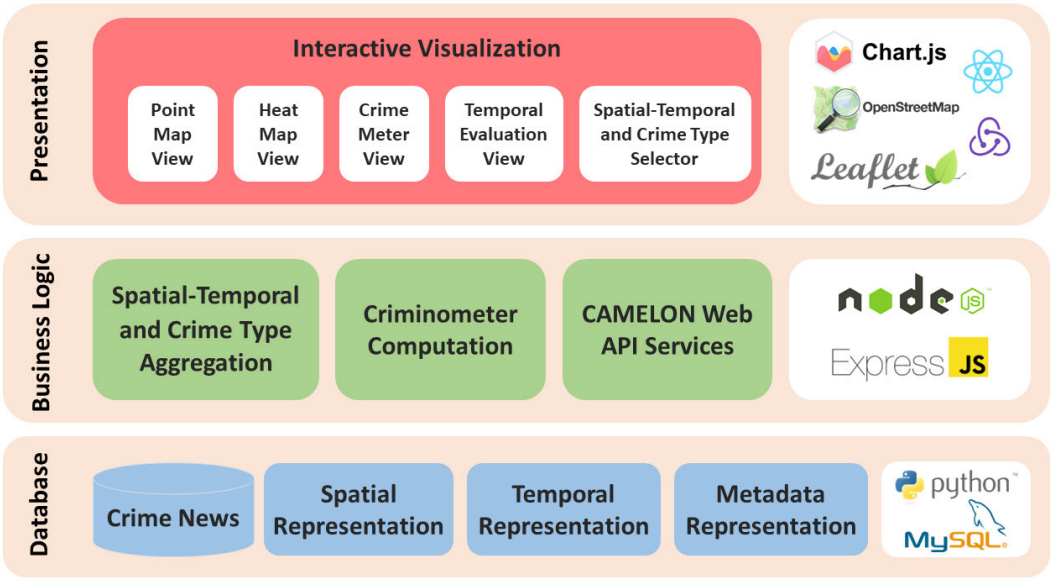

FIGURE 11. High-level diagram of the *CAMELON*'s 3-Tier system architecture.

crime incident without tediously going through all the news articles themselves.

- Task #4 - Crime assessment via the Criminometer index at the province level: Such a mechanism provides detailed information on crime incidents on the map as well as Criminometer indexes, constructed by well-established weights of different crime types and used for the evaluation of the severity and intensity of criminal activity in each area to promote safety awareness.

2) SYSTEM ARCHITECTURE

As shown in Figure 11, we used the three-tier architecture in designing the *CAMELON* system. The first tier is the database layer, which contains the data storage and data model used in the system. The second tier is the business logic layer, which handles the core logic and functionality of the system. Finally, the third tier is the presentation layer, which focuses on visual representation and user interaction. The detailed components and technology stack used in each layer are explained in the following section.

3) DATABASE LAYER

A Python program was implemented to pre-process the raw news articles and store the extracted data in a MySQL database, which is a robust and reliable database solution. These data are separated into spatial, temporal, and metadata representations.

a: SPATIAL REPRESENTATION

To analyze crime incidents and present them on the map, fetching accurate and precise locations of crimes is non-trivial. We extracted the location from each news article at the level of the nation, province, and specific geographical coordinates (latitude and longitude) using Google Geocoding API[6] to translate the extracted textual location to a specific geolocation. Furthermore, when the specific geolocation cannot be extracted or is unavailable, the coarse-grained location at the province level is used. Such a provincial location is extracted using PyThaiNLP's `tag_provinces` function[7] where it is mapped to the representative geolocation of the corresponding province using a gazetteer.

b: TEMPORAL REPRESENTATION

In addition to geographical coordinates, the temporal information pertaining to the commission of crimes and the publication of related news articles were extracted and subsequently stored within the database. Temporal data is typically obtainable at a granular level of date and sometimes a time of day.

c: METADATA REPRESENTATION

All news articles were categorized into appropriate crime types based on our proposed classification model. The crime categories were the fundamental and required metadata. In addition, other crime metadata (in Table 2) were extracted using our proposed model and stored in the database.

4) BUSINESS LOGIC LAYER

This layer is the interface between the database and presentation layers. It was designed to provide reliable, scalable, and efficient services for retrieving and processing crime-related data according to client-side requests. Three main components were implemented.

[6]https://developers.google.com/maps/documentation/geocoding/overview

[7]https://pythainlp.github.io/docs/2.1/api/tag.html

a: SPATIAL-TEMPORAL AND CRIME TYPE AGGREGATION

To reduce the computational time on processing a large volume of news documents, we aggregated the number of crimes by location, date, and crime types in various combinations, for example, the total number of articles that fall within each crime category during each month at the national level.

b: CRIMINOMETER COMPUTATION

This research proposes Criminometer as the crime index that represents the severity and intensity of the crime rate in each area at the province level. The Criminometer index value of the province x is calculated as follows:

$$Criminometer_x = \frac{\sum_{i \in I} C_i * W_i}{P_x} \quad (1)$$

where x represents a particular province, $i \in I$ represents the type of crime in the set of all crime types I, C_i is the number of crimes of the crime type i that occurred in the province x, W_i is the weight reflecting the violence associated with the crime type i, and P_x is the population size of the province x.

Since different types of crime may affect people's concerns and awareness at different levels. For example, knowing that a murder incident happens in the neighborhood would cause one to be more concerned about their safety than a gambling crime. Therefore, a different weight is used for each crime type in the calculation of the Criminometer index that reflects the level of intentional violence. In our research, the capital penalties for each crime type defined by the Thai Criminal Code 1956 [59] published by the Royal Thai Government was used as the referenced weights. Specifically, the crime type with higher legal punishment has a higher weight score as follows: murder ($W_1 = 4$), theft/burglary ($W_2 = 3$), sexual abuse ($W_3 = 2$), battery/assault ($W_4 = 2$), drug ($W_5 = 2$), gambling ($W_6 = 1$), and accident ($W_7 = 1$). However, future work adopting the concept of the proposed Criminometer index can use a different established weighting scheme.

Moreover, as mentioned by Cote [60], when comparing these indexes across different regions, one should keep in mind that the basis of each quantity may differ for each area unit. For example, it is intuitive that a capital province with a denser population would have higher crime incidents compared to rural ones. However, such big cities are also equipped with more stringent law enforcement and better crime management systems. Therefore, using the weighted frequency of crime incidents alone to constitute the crime index would be biased towards more rural areas as safe destinations than urban ones. Therefore, a way to cope with this problem is to normalize the weighted crime activities with some basis of the areas such as population. Implementing this concept, the proposed Criminometer index then normalizes the number of weighted crimes with the target province's population size, as reflected in Equation 1.

c: CAMELON WEB API SERVICES

This backend service is built using Node.js,[8] a popular and efficient server-side JavaScript runtime. These scalable and high-performance services can handle a large volume of requests from the client-side application.

5) PRESENTATION LAYER

This layer contains various interactive data visualization components. This frontend web application is built using React as its foundation, with various libraries supporting the desired functionalities. For example, *Chart.js* is a library for data visualization that currently supports eight chart types: bar, line, area, pie, bubble, radar, polar, and scatter. *OpenStreetMap* library contains geographical data presented on the map such as roads, trails, points of interest, etc. *Leaflet* is a well-known JavaScript library that supports many interactive map features such as tile layers, drag panning, and scroll wheel zoom. This allows users to select and filter the area of interest.

a: PIN MAP VIEW

Illustrated in Figure 12, this component displays crime incidents or other relevant data as markers or pins on a map. It allows users to see the precise locations of incidents along with extracted metadata such as the crime type, criminal's name, victim's name, the action, incident's time, and location. Users can also click on the link to read the original news articles if needed. Symbols and colors are used to represent each type of crime on the map. To find the area of interest, users can either select a province, district, and/or sub-district from the drop-down list at the top of the map, or they can simply move around the map area directly. Additionally, at the bottom of the pin map, users can specify the time frame to display pins or criminal events happening during such a period. The number of crimes is shown on the timeline to provide additional context. This visualization is designed to achieve the System's Task #3.

b: HEAT MAP VIEW

The heat map visualizes the density of crime incidents using color gradients, as depicted in Figure 13. It helps in identifying areas with higher or lower crime rates. The crime hotspot areas can be easily identified. Similar to the pin map, users can select specific areas by choosing province, district, and sub-district from the list, or they can move around the map. This heat map is designed to support the System's Task #1.

c: CRIME METER VIEW

As mentioned in the business logic layer, the crime index called *Criminometer* is computed at the province level. A choropleth map was chosen to visualize such Criminometer

[8] https://nodejs.org/en

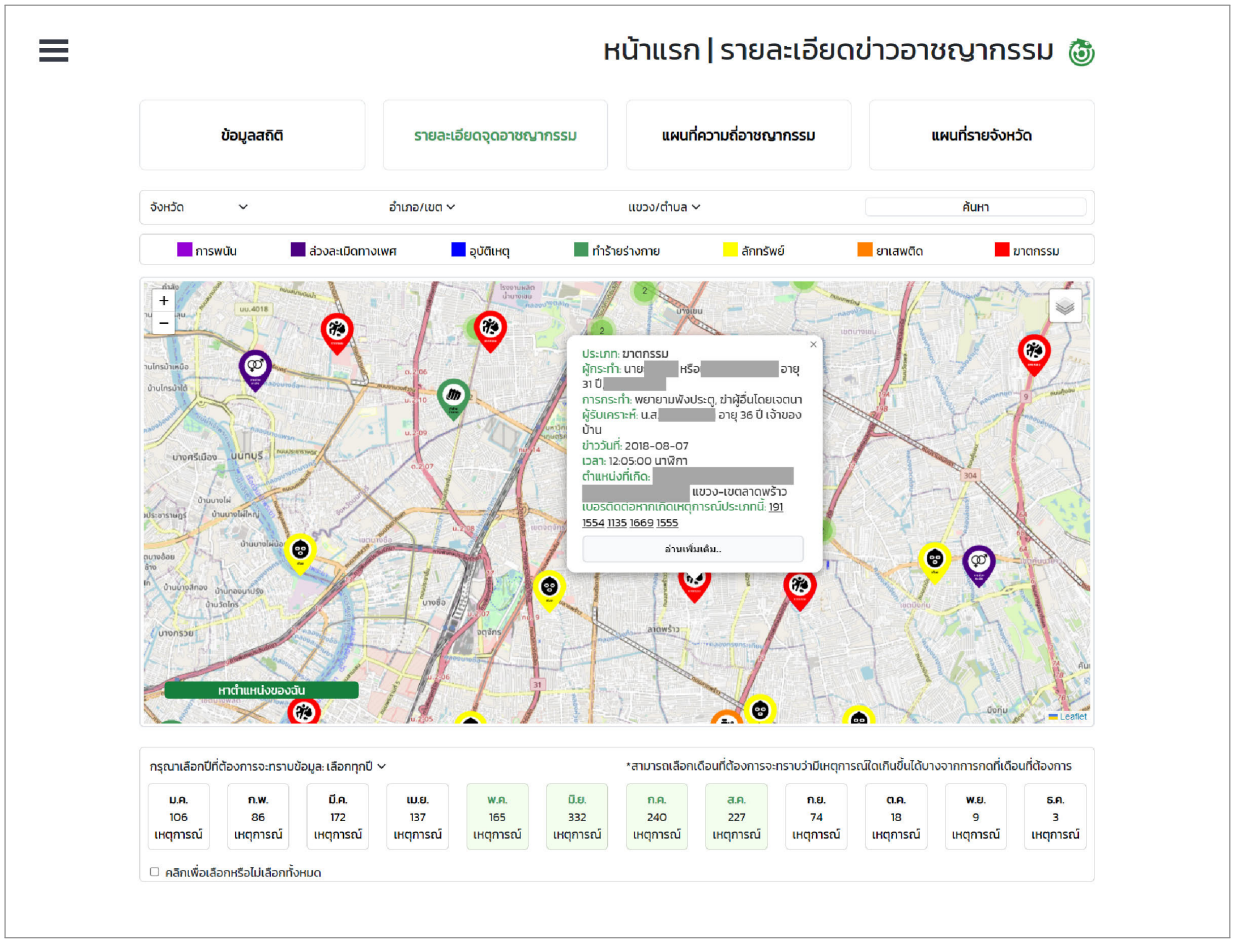

FIGURE 12. Example snapshot of the *Pin Map View* component with extracted crime metadata. Identifiable and personal information is censored.

indexes. This map uses different shades to represent different index values in specific geographic areas, which allows users to easily identify areas with higher or lower crime indexes, aiding in their understanding of crime hot spots and helping them make informed decisions. The Criminometer value was re-scaled into the range of 1 to 100 and then divided into five levels. Each level was represented with a different color, as shown in Figure 14. In addition, the top ten provinces with the highest Criminometer indexes are listed to urge stakeholders of the urgent crime management in these areas. By providing a comprehensive overview of crime rates at the provincial level, the choropleth map contributes to the System's Task #4.

d: TEMPORAL EVOLUTION VIEW

As illustrated in Figure 15, this analytics component provides analytical insights and summaries based on the processed data. It includes the following two sub-components. First, the *Show Crime Statistics Over Time* sub-component presents total crime rates in the country using line or bar graphs to highlight key statistics, trends, or patterns over time. Users can select a specific time range and crime type from the drop-down list. Second, the *Display Crime Type Chart* sub-component presents the total crime rate categorized by the crime type over time. The radar chart is used to make the data more understandable and interpretable. This temporal analysis is presented to achieve the System's Task #1.

e: SPATIAL, TEMPORAL, AND METADATA SELECTOR

Figure 15 also presents the selector component, which allows users to refine their data based on specific criteria. It includes the following two sub-components. First, the *Input Location* sub-component enables users to input a specific location to filter the data, such as provinces, districts, and subdistricts. It restricts the displayed or analyzed data to the selected location. Second, the *Input Date and Time* sub-component allows users to specify a date and time range to filter the data. It helps in narrowing down the data to a specific time period for analysis. This component ensures that the System's Task #2 is supported.

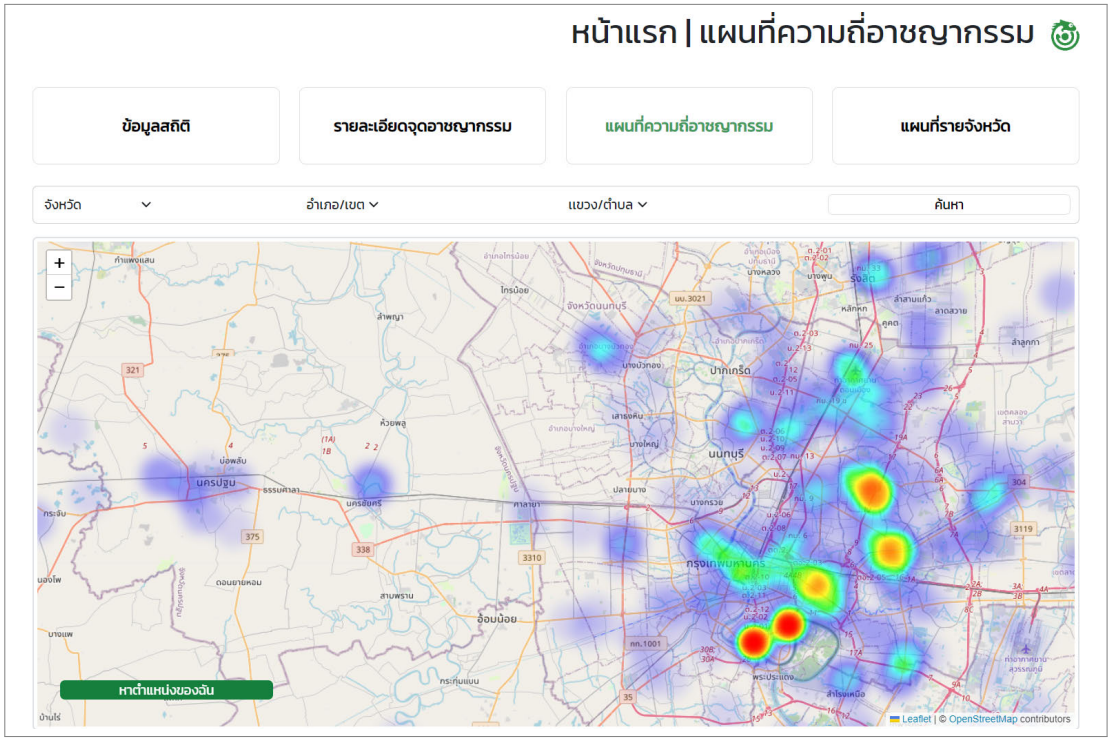

FIGURE 13. Example snapshot of the *Heat Map View* component.

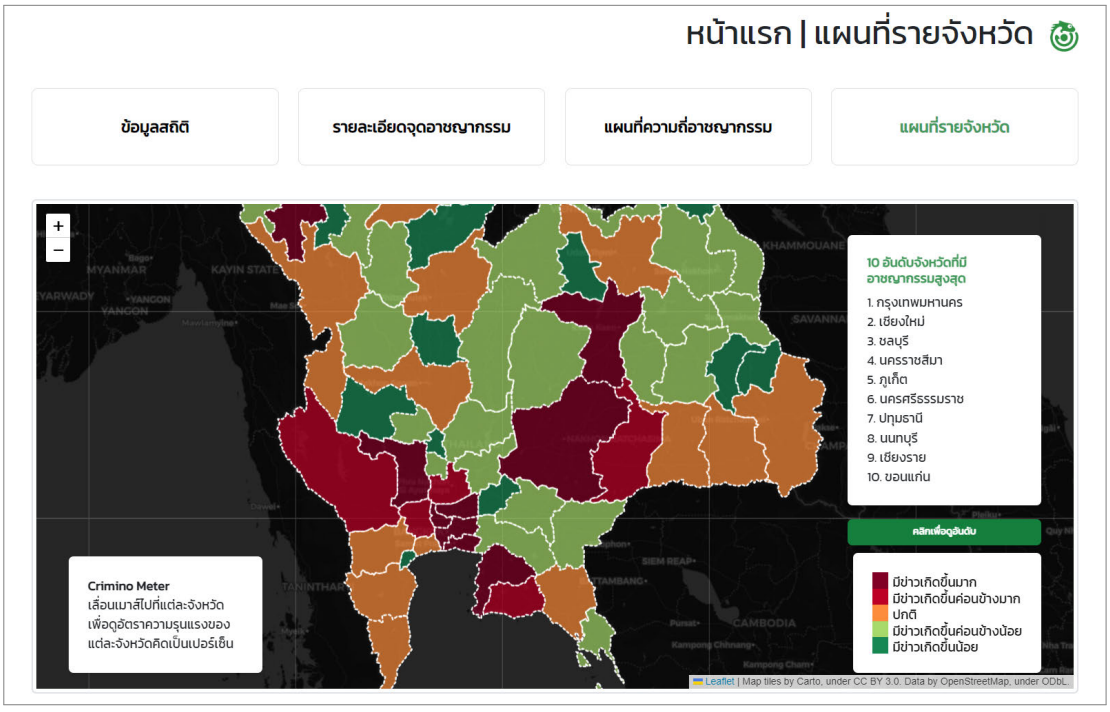

FIGURE 14. Example snapshot of the *Criminometer Map View* component.

V. EXPERIMENTS, RESULTS, AND DISCUSSION

This section discusses the datasets, experiments, and highlighted evaluation results pertaining to the crime type classification, crime metadata extraction, and the system usability evaluation.

A. CRIME TYPE CLASSIFICATION

The crime type classification task entails the development of a document multi-label classifier that assigns a news article to its corresponding crime categories. It should be noted that, according to the multi-label classification

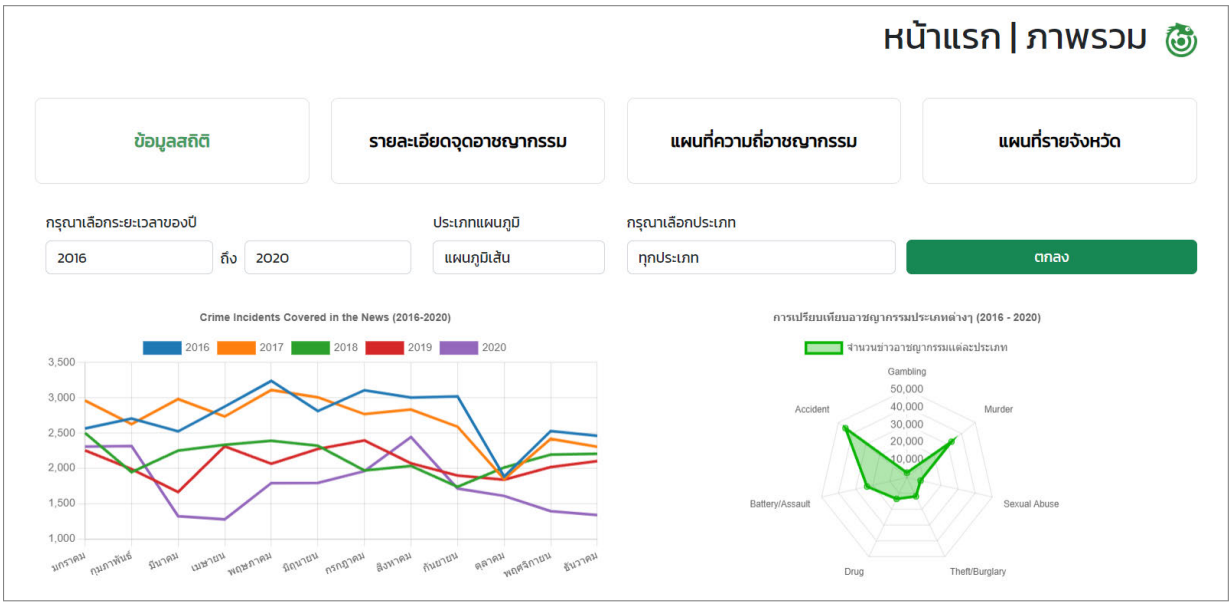

FIGURE 15. Example snapshot of the *Temporal Evaluation View* component.

TABLE 3. Statistics of the annotated news articles for the crime type classification task. Note that one article can be annotated with multiple crime types.

Class	# Articles	Proportion
Gambling	249	2.91%
Murder	2,557	29.85%
Sexual Abuse	673	7.86%
Theft/Burglary	774	9.03%
Drug	1,039	12.13%
Battery/Assault	1,889	22.05%
Accident	721	8.42%
Non-Crime	1,406	16.41%
All	8,567	100.00%

approach, a single article may be categorized under multiple crime types. A range of cutting-edge algorithms for document classification was assessed in terms of their efficacy, and the optimal classifier was selected for incorporation into the proposed system. This section provides a comprehensive analysis of the aforementioned selection process.

Table 3 summarizes the statistics of the 8,567 news articles annotated into seven crime and non-crime types. It should be noted that due to the possibility of an article being categorized under multiple crime types, the cumulative count of articles across all crime types exceeds the overall number of articles. The three most prevalent types of crimes with high frequency are murder, battery/assault, and drug-related offenses, respectively. It is also worth noting that the incidence of gambling-related crimes is relatively low when compared to other categories of criminal activity.

Table 4 highlights the classification performance of the different classifiers in terms of class-wise F1. The XLMR model demonstrated the highest F1 score of 0.860, which was superior to the second-best performing classifier, WangchanBERTa, by a margin of 2.02%. Moreover, XLMR exhibited superior performance in nearly all crime categories, with the exception of gambling crime, in which WangchanBERTa outperformed it. Given that the average F1 score served as the primary criterion for selection, the XLMR model for classifying crime types was chosen for incorporation into the system pipeline.

B. CRIME METADATA EXTRACTION

The crime type classifier validated in the previous section was used not only to categorize crime news reports into their finer-grained crime types but also to screen out non-crime articles. The crime metadata extractor next parsed each crime article to distill important crime-related information. This section reports the annotated data and evaluation results of the crime metadata extraction task. Additionally, pertinent discussion on sensitivity analyses of parameters is also presented.

The statistics pertaining to the 4,650 annotated news articles utilized for training and validating metadata extraction models, categorized according to various types of crimes, are presented in Table 5. In addition, the distribution of the number of entities, number of tokens, and average entity length (in terms of tokens) across various metadata labels are presented in Table 6. The average length of each entity ranges from 5 to 13 tokens, with location entities exhibiting the longest average length. The reason for this is that Thai crime news reports frequently include comprehensive location details, encompassing complete addresses that incorporate place names, road names, sub-districts, districts, and provinces. Notwithstanding their free-text nature, the trigger attributes exhibit the lowest average length. One possible explanation for this phenomenon is that trigger

TABLE 4. Classification performance comparison of the crime type classification task in terms of F1.

Classifier	Gambling	Murder	Sexual Abuse	Theft	Drug	Battery/ Assault	Accident	Non-Crime	Average
BiLSTM	0.707	0.833	0.711	0.611	0.750	0.606	0.724	0.755	0.712
WangchanBERTa	***0.888***	0.905	0.889	0.778	0.907	0.720	0.845	0.809	0.843
MBERT	0.008	0.854	0.753	0.658	0.849	0.641	0.733	0.789	0.661
XLMR	0.887	***0.917***	***0.904***	***0.818***	***0.916***	***0.753***	***0.846***	***0.839***	***0.860***

TABLE 5. Statistics of news articles, divided by relevant crime types, used in the annotation process for the crime metadata extraction task.

Crime Type	# Articles	Proportion
Gambling	232	4.99%
Murder	1,659	35.68%
Sexual Abuse	474	10.19%
Theft	516	11.10%
Drug	578	12.43%
Battery/Assault	924	19.87%
Accident	267	5.74%

TABLE 6. Statistics of annotated crime metadata entities.

Label	# Entities	# Tokens	Avg. Entity Length
Criminal	3,775	37,958	11.47
Victim	2,697	27,003	11.20
Police	3,894	31,784	8.38
Date/Time	2,062	13,634	6.74
Location	2,858	32,821	12.94
Item	3,211	31,329	10.38
Action	4,467	26,271	6.26
Worth	3,140	17,977	6.08
Root Cause	793	4,693	6.00
Trigger	1,357	7,982	5.94

information is typically only available for violent crimes, which tend to involve well-established risk factors such as alcohol consumption and emotional provocation.

1) MODEL SELECTION

The study employed the NER formulation to train and evaluate the performance of CRF, BiLSTM-CRF, WanchanBERTa, and XLMR models in extracting various crime metadata types. The problem at hand was formulated as a multiclass token classification task, wherein the objective was to classify each token present in the text as belonging to a crime metadata entity or not. The best extractor would be chosen for incorporation into the proposed system's pipeline.

Table 7 reports the evaluation results on the crime metadata extraction task of different NER models. We reported using per-token evaluation metrics, including precision, recall, F1-score, MCC, and accuracy. On average, XLMR performed best, with an average F1-score of 0.51, while WanchanBERTa had the second-best performance, with an average F1-score of 0.50. Comparing these two NER models, the XLMR model performed better on the criminal, victim, date/time, action, and root cause labels, while the WanchanBERTa model outperformed the other on the police, location, and worth

TABLE 7. Comparison of the performance among different classification algorithms of the crime metadata extraction task.

Model	Label	Precision	Recall	F1	MCC	Accuracy
CRF	Criminal	0.63	0.58	0.60	0.60	0.98
	Victim	0.52	0.39	0.45	0.45	0.99
	Police	0.46	0.53	0.49	0.48	0.98
	Date/Time	0.47	0.57	0.51	0.51	0.99
	Location	0.71	0.65	0.67	0.67	0.99
	Item	0.57	0.68	0.62	0.62	0.98
	Action	0.09	0.10	0.09	0.08	0.98
	Worth	0.55	0.37	0.44	0.45	0.99
	Root Cause	0.30	0.11	0.16	0.17	0.99
	Trigger	0.23	0.04	0.06	0.09	0.99
	Average	0.45	0.40	0.41	0.41	0.99
BiLSTM -CRF	Criminal	0.66	0.62	0.64	0.63	0.99
	Victim	0.61	0.52	0.56	0.56	0.99
	Police	0.49	0.61	0.54	0.54	0.98
	Date/Time	0.52	0.23	0.32	0.34	0.99
	Location	0.82	0.61	0.70	0.70	0.99
	Item	0.66	0.64	0.65	0.64	0.99
	Action	0.18	0.03	0.05	0.07	0.99
	Worth	0.61	0.32	0.42	0.44	0.99
	Root Cause	0.33	0.02	0.05	0.09	0.99
	Trigger	0.45	0.00	0.01	0.04	0.99
	Average	***0.53***	0.36	0.39	0.40	0.99
Wangchan BERTa	Criminal	0.65	0.65	0.65	0.64	0.99
	Victim	0.58	0.61	0.60	0.59	0.99
	Police	0.50	0.70	0.59	0.58	0.97
	Date/Time	0.54	0.64	0.58	0.58	0.99
	Location	0.78	0.88	0.81	0.80	0.99
	Item	0.55	0.84	0.67	0.67	0.98
	Action	0.23	0.21	0.22	0.21	0.97
	Worth	0.49	0.51	0.50	0.49	0.98
	Root Cause	0.31	0.15	0.20	0.20	0.99
	Trigger	0.33	0.16	0.21	0.22	0.99
	Average	0.50	***0.54***	0.50	0.50	0.98
XLMR	Criminal	0.66	0.66	0.66	0.65	0.99
	Victim	0.59	0.64	0.62	0.61	0.99
	Police	0.50	0.69	0.58	0.57	0.98
	Date/Time	0.58	0.69	0.63	0.63	0.99
	Location	0.73	0.88	0.80	0.80	0.99
	Item	0.56	0.84	0.67	0.68	0.98
	Action	0.25	0.21	0.23	0.22	0.98
	Worth	0.49	0.49	0.49	0.48	0.98
	Root Cause	0.33	0.16	0.22	0.22	0.99
	Trigger	0.33	0.15	0.21	0.22	0.99
	Average	0.50	***0.54***	***0.51***	***0.51***	0.99

labels in terms of F1-score. Since XLMR has a marginally better performance than WanchanBERTa and can support diverse languages due to being cross-lingual, facilitating applications that need multilingual information access and processing [61], it was chosen for integration into the system. It is interesting to note that while BiLSTM-CRF has been a popular choice for NER tasks in low-resource language settings [62], [63], [64], this has been proved otherwise for the Thai language, especially in our case study. Furthermore, the performance wielded by the BiLSTM-CRF model is quite

TABLE 8. Performance of XLMR on the crime metadata extraction task, using only closed-form entity types.

Label	Precision	Recall	F1	MCC	Accuracy
Criminal	0.66	0.65	0.66	0.65	0.99
Victim	0.58	0.58	0.61	0.60	0.99
Police	0.50	0.72	0.59	0.59	0.97
Date/Time	0.58	0.71	0.64	0.64	0.99
Location	0.73	0.90	0.81	0.81	0.99
Item	0.59	0.85	0.70	0.70	0.98
Average	0.61	0.74	0.67	0.67	0.99

on par with the traditional CRF model, with an average F1 score of roughly 0.40. It is also worth noting that the training times of CRF, BiLSTM-CRF, WangchanBERTa, and XLMR were 37, 479, 88, and 86 minutes, respectively. CRF had a modest training time, 57% faster than XLMR, due to not utilizing the deep learning mechanism (i.e., does not rely on GPUs for computation), while securing the performance on par with BiLSTM-CRF in terms of average F1-score, and was only 20% worse than that of XLMR. Hence, it can be inferred that although XLMR exhibited superior performance, in cases where the platform for implementation has restricted computational resources or constraints on model training time, the CRF model may prove to be a more viable and efficient alternative.

2) CLOSED-FORM METADATA EXTRACTION

In this research, closed-form metadata entities refer to those that are not free-text, including criminal, victim, police, date/time, location, and items. Each such entity often forms a contiguous chunk of noun phrases with clear boundaries in the text and, therefore, relatively easy to spot by humans and machines. The previous section reports the performance of the extractor models when trained to perform a multiclass classification task that also includes the free-text labels, such as action, worth, root cause, and trigger, which could potentially hinder the model's overall performance. Therefore, a natural question would arise as to how the model would perform if it focused on only closed-form metadata. To answer this research question, the best model from the previous section, XLMR, was chosen to retrain and evaluate with only the closed-form labels aforementioned defined.

Using the same evaluation protocol as the previous section, Table 8 summarizes the experiment results. As expected, the model's performance became better when focusing on only the closed-form metadata, as the average F1-score increased from 0.51 to 0.67 (31.37% improvement); however, with the sacrifice of the ability to extract the free-text metadata, which could be additionally valuable for crime analysis purposes. Therefore, the adopting platform would need to weigh out the pros and cons of integrating these novel free-text metadata attributes into their system. In addition, future work could seek to develop a dedicated model for free-text metadata extraction using the question-answering (QA) model training paradigm [65].

TABLE 9. Comparison between text segmentation modes (chunk-wise vs. sentence-wise) on the closed-form crime metadata extraction task.

Mode	Label	Precision	Recall	F1	Δ F1	MCC	Accuracy
Chunk	Criminal	0.66	0.70	0.68	-	0.67	0.99
	Victim	0.62	0.62	0.62	-	0.62	0.99
	Police	0.44	0.65	0.52	-	0.52	0.97
	Date/Time	0.52	0.83	0.64	-	0.65	0.99
	Location	0.63	0.80	0.71	-	0.70	0.99
	Item	0.50	0.85	0.63	-	0.64	0.98
	Average	0.56	0.74	0.63	-	0.63	0.99
Sentence	Criminal	0.66	0.65	0.66	-2.94%	0.65	0.99
	Victim	0.58	0.58	0.61	-1.61%	0.60	0.99
	Police	0.50	0.72	0.59	13.46%	0.59	0.97
	Date/Time	0.58	0.71	0.64	0.00%	0.64	0.99
	Location	0.73	0.90	0.81	14.08%	0.81	0.99
	Item	0.59	0.85	0.70	11.11%	0.70	0.98
	Average	***0.61***	***0.74***	***0.67***	5.53%	***0.67***	0.99

3) IMPACT OF THE TEXT SEGMENTATION METHODS

Currently, the metadata extraction module first segments a news article into sentences and processes them as a sequence of small documents. However, most deep-learning language models have a larger capacity to handle a document whose length is longer than a sentence. For example, the XLMR model can handle a document with up to 512 tokens at a time. A natural research question would arise as to how the model performance would differ if it were to process larger chunks of text. To address this research question, we performed another experiment where a news article was segmented into chunks of sentences of at most 500 tokens. The XLMR model was trained and validated using the evaluation protocol on the closed-form metadata, similar to the previous section.

Table 9 summarizes and compares the extraction performance between training the XLMR model with chunks and sentences of text. On average, training the model with a sequence of sentences yielded the best performance in terms of F1 of 0.67, outperforming the chunk-wise method by 4.69%. The enhancements in performance are particularly evident in the labels pertaining to location, police, and items, which exhibit F1 improvements of 14.08%, 13.46%, and 11.11%, respectively. This phenomenon may be attributed to the characteristic of closed-form entities, which are clearly defined on their own, enabling the model to make contextual decisions within the sentence. Consequently, the inclusion of multiple sentences within a single chunk may introduce extraneous and unhelpful information, thereby impeding the model's learning capacity.

C. SYSTEM USABILITY EVALUATION

A set of user studies was conducted to gauge the proposed system's ability to satisfy target users' requirements and needs. The usability evaluation was divided into two phases based on the testing methodology and target groups: 1) Intensive Usability Test, a comprehensive usability test and interview with a small group of participants, and 2)

Satisfaction Survey, a system's features satisfaction survey with a broader and larger group of participants.

1) INTENSIVE USABILITY TEST

We devised a comprehensive set of five test case scenarios aimed at evaluating the user's ability to accomplish tasks efficiently and determining the ease of use of various functions within the system. Each participant was given five tasks to perform, each of which was compared with the estimated time reference. Such an estimated reference of time usage in each case indicates the high usability of the system, calculated from the average time spent by the developers multiplied by three to account for the developers' familiarity with the system. This evaluation approach aims to provide users with ample time to interact with our system. The five test case scenarios are as follows:

1) *Display temporal analysis:* In the temporal evaluation view page, users were asked to display only the information of a specific crime type from the year 2016 until 2021. [Expected time: 15 seconds]
2) *Show crime metadata:* In the pin map page, users were asked to use the spatial selector to go to a specific region (province, district, and sub-district), find a pin representing a specific crime type, and then click to show the crime metadata detail. [Expected time: 40 seconds]
3) *Apply interactive selection:* On the pin map page, users were asked to display only pins representing a particular crime type during a specific month and year. [Expected time: 40 seconds]
4) *View Criminometer:* In the crime meter view page, users were asked to find the top 10 provinces with the highest crime indexes and find a way to show the Criminometer index of a given province. [Expected time: 25 seconds]
5) *Analyze crime hotspot area:* In the heat map page, find the largest crime hot spot in Thailand. [Expected time: 15 seconds]

These carefully constructed test cases encompassed a wide range of user interactions to gather meaningful insights into the system's usability and identify potential areas for improvement.

All the system testers were older than 18 years old. They could comprehend Thai news and had digital literacy in browsing the internet and using personal computers. The time spent on each task by nine participants (testers) is reported in Figure 16. The results indicate a positive achievement, demonstrating that the majority of participants were able to complete the assigned tasks within the expected time limit successfully. The overall outcome suggests that our system exhibits a high level of user-friendliness, as testers were able to navigate its functionalities with relative ease. However, upon closer examination, it was observed that a few tasks exceeded our established baseline completion time, especially for Tester 2. He mentioned that the search button was too small, so he took a while to locate it.

2) POLICE FEEDBACK

In addition to the usability test, we also conducted a qualitative interview with five target users who were police officers from a local police station in Naknon Pathom province. Two of them were between 19-26 years old, the other two were between 27-25 years old, and the last one was older than 36 years old. We let them use the system for about five to ten minutes and then inquired about their opinions. We received very positive and high-value feedback. Specifically, we found that the police liked the pin map view the most. In addition, while the police officers were analyzing the crimes using the system, the metadata automatically extracted by the system was found to be useful and could save their time. Furthermore, the users revealed that they were not aware of any crime-related systems having key features similar to our proposed system. They also mentioned that the system looked practical and better than many of the applications currently used. A useful suggestion was made that the integration of police crime reports into the system would complement the reported statistics and inherently make the system more reliable and trustworthy. Meanwhile, another suggestion was made that the system could be improved by also increasing both the spatial and temporal granularity of the data when more data sources from all police stations in Thailand are accessible on a daily basis. However, some metadata of crime should be anonymous to protect personal data. Finally, they wished that the proposed system could be adopted at a national scale to not only assist police officers in keeping abreast of criminal activities in the neighborhood but also make people aware of the potential crimes in their areas.

3) SATISFACTION SURVEY

In addition to the useability evaluation with a small group of target users discussed in the previous section, we also conducted a general survey on a larger audience with a variety of occupations. The objective of this survey is to evaluate the system by various groups of users on more specific implementations and the usefulness of each feature. This finding can highlight the need for future improvements and refinements to achieve better user satisfaction. In total, we received responses from 56 participants: 18 government officers, 20 private sector owners/employees, and 18 university students. Each participant was given a URL to test-use the system via the web browser and asked to complete an online survey. The results were separated into two categories, including satisfaction with the system functionality and usefulness of the system features. It is important to note that the survey utilized a rating scale with a maximum score of 5, indicating the highest level of satisfaction/usefulness.

The results shown in Table 10 indicate that, on average, the participants were satisfied with all system functionalities by giving scores of more than 4 out of 5. In general, university

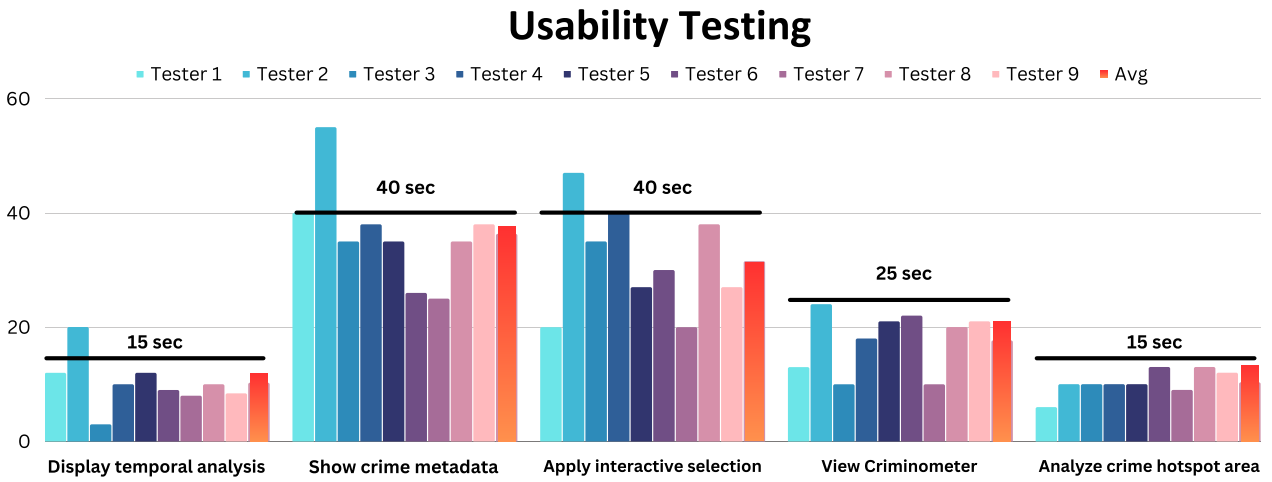

FIGURE 16. Time spent on the five test case scenarios by nine testers and average time spent.

TABLE 10. Users' satisfaction with the system functionality (Max score = 5 points.).

Satisfaction on Functionalities	Occupation		
	Goverment Officers	Private Sector Owers/Officers	University Students
Ease of Comprehension	4.00	4.25	4.50
Suitable Symbols	4.06	4.45	4.61
Fast Response	4.06	4.50	4.67
Ease of Use	4.22	4.55	4.72
Precision and Coverage	4.06	4.30	4.50

TABLE 11. The usefulness of each visualization component of the system (Max score = 5 points.).

Usefulness of Visualization	Occupation		
	Goverment Officers	Private Sector Owers/Officers	University Students
Graph	4.06	4.35	4.50
Pin Map	4.33	4.40	4.56
Heat Map	4.22	4.30	4.11
Crimenometer	4.17	4.45	4.33

students gave higher scores than the other two groups. This could be because students often use online resources and platforms, so they would be more familiar with the design of the web application and inherently more adaptive to using other online web services. The functionality with the highest satisfaction score among all participants is the ease of use. The students, who were mostly teenagers, gave the highest score of 4.72, while the private sector employees and government officers gave 4.22 and 4.55, respectively. The ease of use is an important factor in ensuring that all users can access and analyze crime data in an efficient and effective manner. The user-friendly design and interface are particularly important for users who are not technically inclined. On the other hand, the functionality with the lowest satisfaction score is the ease of comprehension. This topic represents whether the contents and information presented on the system are easy to read and understand. In this topic, the university students scored the highest, with 4.50, while the private sector employees and government officers gave slightly lower scores of 4.25 and 4.00, respectively. The survey findings pointed us in the direction where the system could be improved on how the digested information is presented and visualized in a readily comprehensible manner to specific groups of users.

Table 11 summarizes the usefulness evaluation of each implemented visualization in the system from the same participant cohort. Interestingly, different user groups had different opinions about the useful functions of the system. For example, the government officers found the pin map visualization to be the most useful, with an average score of 4.33, because they needed to see the detailed metadata of individual crime incidents. As a result, the graph and chart analyses were less popular for this group, with an average score of 4.06 out of 5. With the nature of their job responsibility, it is likely that these officers work with detailed data as part of their routine job functions; therefore, the summary or aggregated data might seem less useful to them. For the second group, the private sector owners/officers found the Criminometer map view to be the most useful, with the highest average score of 4.45. This could be because the business owners are familiar with various key performance indexes as part of measurements for running businesses. Thus, at a high level, the Criminometer values presented in the choropleth map view provide an interesting way to easily comprehend the overall criminal activity in each area. On the other hand, they found that the heat map view was the least useful compared to the other visualizations. This might be because, at the national level, the crime density seems to be high in big cities such as Bangkok. As a result, the information gained from this map at the high level may not be relevant to their particular needs or decision-making processes. But if we look into a specific province or district, this visualization can be useful in finding crime hotspot areas. Finally, the last group comprises university students. In agreement with the government officer group, the students also thought that the pin map view was the most useful, with an average score of 4.56, and the second most useful visualization was the graph and chart

analysis, with a similar score of 4.50. However, they found that the heat map view was the least useful feature. It is interesting that students found the pin map view as the most useful visualization, similar to the government officers, while the heat map was perceived as the least useful feature, similar to the private sector owners/officers. Regardless, all the visualization functionalities implemented in our system received an overall score higher than 4 out of 5, indicating that the system was perceived as very useful by the three groups of participants.

D. LIMITATIONS

While the evaluation results of the proposed system are promising, there is ample room for further enhancement. First, the current implementation of the system only has one access mode that serves all the different groups of users with diverse needs of crime knowledge. Specifically, certain distilled knowledge may be inappropriate or spurious for general users, such as victims' names and specific locations of the crime sites, especially in residential areas. Therefore, a separate panel could be implemented for general users, presenting only filtered relevant information that directly accommodates their intended uses.

Furthermore, the current system solely retrieves and displays crime-related data from digital news outlets. Although considered reliable, it is possible for the system to overlook certain criminal activities that fail to attract the attention of the news audience, such as minor robbery and battery incidents. Hence, in order to enhance the comprehensiveness of the system concerning criminal activities, it is possible to introduce a mechanism that enables users to report and authenticate crimes. In addition, the system could integrate the capability to establish a connection with the police database, thereby facilitating the automatic synchronization of supplementary reported crimes. However, such an information fusion mechanism would require the capacity to distinguish and consolidate crime reports pertaining to identical criminal occurrences, thereby presenting an event disambiguation challenge [66], [67].

VI. ETHICAL ISSUES AND SOCIETAL IMPLICATIONS

The *CAMELON* system is designed to offer a thorough and prompt representation and evaluation of spatiotemporal criminal occurrences across the country. The use of news articles from well-established publishers as sources of knowledge not only provides timely access to information from any location but also ensures reliability compared to the information obtained from social networking platforms [68]. Nonetheless, the utilization of openly accessible data that can effectively depict real-world phenomena may present both advantages and disadvantages that require cautious navigation. Hence, in the event that an organization decides to implement the proposed system, it is anticipated that there will be societal ramifications and ethical considerations that will need to be addressed. This section discusses some of the aforementioned issues.

A. PRIVACY

Although online news articles may be available to the general public, they often contain sensitive personal information, including but not limited to names, addresses, and vehicle license plates, that can be used to identify individuals. Apart from disseminating such data through publicly available news sources, the assimilation and analysis of such information may give rise to apprehensions among the impacted parties, which could potentially escalate into a privacy issue of a significant magnitude. For example, providing information about a particular site of criminal activity may elicit public scrutiny of said location, potentially causing disruption to the surrounding community. In addition, disclosing the identities of perpetrators, witnesses, or victims in news media may potentially infringe upon their personal data privacy. Though one may argue that such personal data is already in the news content and, therefore, already exposed to the public, the government implementing this system should take the high road and tread carefully when handling such sensitive and personal information.

B. SOCIETAL COSTS

The realization of the proposed system necessitates financial backing for data access, hardware, development, and upkeep. The provision of financial support, in the event that the government adopts the proposed system, is anticipated to be derived from tax revenues, with the extent of such support being contingent upon the project's scope. Moreover, the adoption of the proposed systems would entail indirect costs that arise from governmental responses to escalated criminal activities. These costs may include heightened patrol dispatches, the implementation of employment programs to address poverty as a root cause of crime, and the establishment of weapon control regulations. Although it is not the primary aim for individuals to assume this responsibility, the adopting organization must evaluate the direct and indirect expenses linked to the implementation of the proposed system for law enforcement and policymakers in comparison to the advantages that people will receive.

C. ECONOMIC IMPACTS

One of the identified user cohorts of the proposed system encompasses tourists seeking secure destinations for travel. While these tourists can use the system to explore criminal activities and via the proposed Criminometer indexes in specific locations before making their travel decisions, such functionality could prove to have both positive and negative economic impacts on certain regions. For example, regions that are considered relatively safe from crime tend to draw visitors, leading to the emergence of tourism-related enterprises and subsequent economic growth. Nonetheless, areas that experience a comparatively elevated level of criminal incidents would also encounter a negative impact resulting from a decline in tourism. Moreover, the system could be utilized by investors to investigate potential investment prospects in urbanization, including the development of

commercial centers, landmarks, or properties. An area that is perceived to have a high incidence of criminal activity would inevitably deter potential investors, leading to a missed opportunity for economic advancement.

D. ABUSE

Although the intended purpose of the proposed system is to serve the betterment of society, it is possible that individuals with malicious intentions may exploit its functionalities for personal gain. The utilization of publicly accessible online media sources to power the proposed system implies that identical data can be acquired by any individual, including those with criminal intentions. In the event that incidents of theft and homicide plague a city, unscrupulous merchants may take advantage of the people's apprehension by profiting from the sale of fraudulent surveillance equipment. Drug traffickers may also reap advantages from a city with reduced criminal incidents, thereby decreasing the necessity for police scrutiny in order to conceal and convey their illicit substances. Consequently, it is imperative that the government overseeing the implementation of the proposed system maintains meticulous records of its utilization and restricts authorization solely to relevant user cohorts to prevent potential misuse.

VII. CONCLUSION

This paper proposed *CAMELON*, an intelligent system designed for extracting and visualizing crime metadata from vast corpora of online news articles. First, a survey was undertaken to demonstrate the disparity between the current methods by which stakeholders obtain crime-related information and the necessity for a spatiotemporal and contemporaneous framework for the monitoring of crime at the vicinity, regional, and national levels. Compared with existing related work, our novel contributions are situated two-fold. First, the proposed crime metadata extraction system performs two primary functions: categorizing news articles into more specific crime types and extracting significant crime-related attributes from each article. A novel set of free-text crime metadata was also introduced to provide more comprehensive information about a crime incident. Second, the information that has been extracted is subjected to processing by the front-end system, which then presents it in an interactive spatiotemporal format. This enables users to observe the patterns of particular types of criminal activity across both regional and national domains. Furthermore, the system has the capability to display the incidents and associated metadata on a map, thereby enabling the examination of criminal activity at the local and community level. Finally, we proposed the novel Criminometer index that utilizes established criteria for criminal sentencing to measure both the magnitude and intensity of criminal activity within specific geographic areas. The validation of relevant experiments on crime type classification and crime metadata extraction was conducted using a case study of two reputable online news sources in Thailand. Additionally, an assessment of user experience regarding the usability of the proposed system was carried out, revealing a significant level of user contentment with all primary functions of the system. *CAMELON* has the potential to significantly enhance the effectiveness of local law enforcement agencies in formulating tactics aimed at mitigating the proliferation of particular categories of criminal activities. Additionally, it could assist national policymakers in developing appropriate measures to tackle the underlying factors that contribute to criminal behavior in a proactive manner. In addition, the system could potentially be used by general people to seek secure locations for tourism, investment, and residential purposes. Although the target users have expressed positive feedback regarding the proposed system, it is crucial to acknowledge its limitations and utilize them as a means of enhancement. Currently, the system exclusively displays the prevailing patterns of criminal activities, with the temporal proximity being ascertained by online news publications. It would be advantageous for all stakeholders if the system could predict crime trends, enabling them to respond proactively to significant shifts in anticipated criminal activity through appropriate planning and actions. Our future work will focus on exploring this direction.

REFERENCES

[1] C. Detotto and E. Otranto, "Does crime affect economic growth?" *Kyklos*, vol. 63, no. 3, pp. 330–345, Jul. 2010.
[2] P. Sharkey and G. Torrats-Espinosa, "The effect of violent crime on economic mobility," *J. Urban Econ.*, vol. 102, pp. 22–33, Nov. 2017.
[3] G. Farrell, A. Tseloni, J. Mailley, and N. Tilley, "The crime drop and the security hypothesis," *J. Res. Crime Delinquency*, vol. 48, no. 2, pp. 147–175, May 2011.
[4] H. Elonheimo, "Evidence for the crime drop: Survey findings from two Finnish cities between 1992 and 2013," *J. Scandin. Stud. Criminol. Crime Prevention*, vol. 15, no. 2, pp. 209–217, Jul. 2014.
[5] M. Natarajan, "Crime in developing countries: The contribution of crime science," *Crime Sci.*, vol. 5, no. 1, pp. 1–5, Dec. 2016.
[6] R. Suphanchaimat, V. Sornsrivichai, S. Limwattananon, and P. Thammawijaya, "Economic development and road traffic injuries and fatalities in Thailand: An application of spatial panel data analysis, 2012–2016," *BMC Public Health*, vol. 19, no. 1, p. 1449, 2019.
[7] K. Seresirikachorn, P. Singhanetr, N. Soonthornworasiri, A. Amornpetchsathaporn, and T. Theeramunkong, "Characteristics of road traffic mortality and distribution of healthcare resources in Thailand," *Sci. Rep.*, vol. 12, no. 1, p. 20255, 2022.
[8] S. Chokprajakchat, W. Techagaisiyavanit, D. Mulaphong, T. Iyavarakul, A. Kuanliang, and C. Laosunthorn, "Tracking violence in Thailand: The making of violent crime index," *Secur. J.*, Mar. 2023. [Online]. Available: https://doi-org.ejournal.mahidol.ac.th/10.1057/s41284-023-00369-2
[9] H. K. R. ToppiReddy, B. Saini, and G. Mahajan, "Crime prediction & monitoring framework based on spatial analysis," *Proc. Comput. Sci.*, vol. 132, pp. 696–705, Jan. 2018.
[10] S. S. Kshatri, D. Singh, B. Narain, S. Bhatia, M. T. Quasim, and G. R. Sinha, "An empirical analysis of machine learning algorithms for crime prediction using stacked generalization: An ensemble approach," *IEEE Access*, vol. 9, pp. 67488–67500, 2021.
[11] N. Khan, M. S. Islam, F. Chowdhury, A. S. Siham, and N. Sakib, "Bengali crime news classification based on newspaper headlines using NLP," in *Proc. 25th Int. Conf. Comput. Inf. Technol. (ICCIT)*, Dec. 2022, pp. 194–199.
[12] F. Rollo, G. Bonisoli, and L. Po, "Supervised and unsupervised categorization of an imbalanced Italian crime news dataset," in *Information Technology for Management: Business and Social Issues* (Lecture Notes in Business Information Processing), vol. 442, E. Ziemba and W. Chmielarz, Eds. Cham, Switzerland: Springer, 2022. [Online]. Available: https://doi-org.ejournal.mahidol.ac.th/10.1007/978-3-030-98997-2_6

[13] S. Ghankutkar, N. Sarkar, P. Gajbhiye, S. Yadav, D. Kalbande, and N. Bakereywala, "Modelling machine learning for analysing crime news," in *Proc. Int. Conf. Adv. Comput., Commun. Control (ICAC3)*, Dec. 2019, pp. 1–5.

[14] S. M. M. H. Chowdhury, Z. N. Tumpa, F. Khatun, and S. K. F. Rabby, "Crime monitoring from newspaper data based on sentiment analysis," in *Proc. 8th Int. Conf. Syst. Model. Advancement Res. Trends (SMART)*, Nov. 2019, pp. 299–304.

[15] H. Zhanjun, Z. Wang, Z. Xie, L. Wu, and Z. Chen, "Multiscale analysis of the influence of street built environment on crime occurrence using street-view images," *Comput., Environ. Urban Syst.*, vol. 97, Oct. 2022, Art. no. 101865.

[16] X. Zhang, L. Liu, M. Lan, G. Song, L. Xiao, and J. Chen, "Interpretable machine learning models for crime prediction," *Comput., Environ. Urban Syst.*, vol. 94, Jun. 2022, Art. no. 101789.

[17] C. J. Joubert, A. Saprykin, N. Chokani, and R. S. Abhari, "Large-scale agent-based modelling of street robbery using graphical processing units and reinforcement learning," *Comput., Environ. Urban Syst.*, vol. 94, Jun. 2022, Art. no. 101757.

[18] J. Wu, S. M. Abrar, N. Awasthi, and V. Frías-Martínez, "Auditing the fairness of place-based crime prediction models implemented with deep learning approaches," *Comput., Environ. Urban Syst.*, vol. 102, Jun. 2023, Art. no. 101967.

[19] N. Newman, "Mainstream media and the distribution of news in the age of social media," Dept. Politics Int. Relations, Reuters Inst. Study J., Univ. Oxford, Tech. Rep., 2011.

[20] R. Hauser, J. Vamvas, S. Ebling, and M. Volk, "A multilingual simplified language news corpus," in *Proc. 2nd Workshop Tools Resour. Empower People REAding DIfficulties (READI) 13th Lang. Resour. Eval. Conf.*, 2022, pp. 25–30.

[21] Z. Zhang and B. B. Gupta, "Social media security and trustworthiness: Overview and new direction," *Future Gener. Comput. Syst.*, vol. 86, pp. 914–925, Sep. 2018.

[22] S. Kalmegh, "Analysis of weka data mining algorithm reptree, simple cart and randomtree for classification of Indian news," *Int. J. Innov. Sci., Eng. Technol.*, vol. 2, no. 2, pp. 438–446, 2015.

[23] M. Magnusson, J. Finnäs, and L. Wallentin, "Finding the news lead in the data haystack: Automated local data journalism using crime data," in *Proc. Comput.+ Journalism Symp.*, 2016, pp. 1–4.

[24] C. Rajapakshe, S. Balasooriya, H. Dayarathna, N. Ranaweera, N. Walgampaya, and N. Pemadasa, "Using CNNs RNNs and machine learning algorithms for real-time crime prediction," in *Proc. Int. Conf. Advancements Comput. (ICAC)*, Dec. 2019, pp. 310–316.

[25] A. Umair, M. S. Sarfraz, M. Ahmad, U. Habib, M. H. Ullah, and M. Mazzara, "Spatiotemporal analysis of web news archives for crime prediction," *Appl. Sci.*, vol. 10, no. 22, p. 8220, Nov. 2020.

[26] T. Thaipisutikul, S. Tuarob, S. Pongpaichet, A. Amornvatcharapong, and T. K. Shih, "Automated classification of criminal and violent activities in Thailand from online news articles," in *Proc. 13th Int. Conf. Knowl. Smart Technol. (KST)*, Jan. 2021, pp. 170–175.

[27] G. Deepak, S. Rooban, and A. Santhanavijayan, "A knowledge centric hybridized approach for crime classification incorporating deep bi-LSTM neural network," *Multimedia Tools Appl.*, vol. 80, no. 18, pp. 28061–28085, Jul. 2021.

[28] P. Singh, S. Gupta, V. Gupta, P. Kuchhal, and A. Jain, "Face recognition–based surveillance system: A new paradigm for criminal profiling," in *Digital Forensics and Internet of Things: Impact and Challenges*. Wiley, 2022, pp. 1–18.

[29] W. P. Rey and G. V. Rolluqui, "Mobile automated fingerprint identification system (MAFIS): An Android-based criminal tracking system using fingerprint minutiae structure," in *Proc. 5th Int. Conf. E-Soc., E-Educ. E-Technol.*, 2021, pp. 62–68.

[30] T. Davies and E. Marchione, "Event networks and the identification of crime pattern motifs," *PLoS ONE*, vol. 10, no. 11, Nov. 2015, Art. no. e0143638.

[31] C. H. Ku, A. Iriberri, and G. Leroy, "Natural language processing and e-government: Crime information extraction from heterogeneous data sources," in *Proc. Int. Conf. Digital Government Res.*, 2008, pp. 162–170.

[32] K. R. Rahem and N. Omar, "Drug-related crime information extraction and analysis," in *Proc. 6th Int. Conf. Inf. Technol. Multimedia*, Nov. 2014, pp. 250–254.

[33] K. Srinivasa and P. S. Thilagam, "Crime base: Towards building a knowledge base for crime entities and their relationships from online news papers," *Inf. Process. Manag.*, vol. 56, no. 6, 2019, Art. no. 102059.

[34] F. Rahma and A. Romadhony, "Rule-based crime information extraction on Indonesian digital news," in *Proc. Int. Conf. Data Sci. Appl. (ICoDSA)*, Oct. 2021, pp. 10–15.

[35] Z. Khalid and O. U. Sezerman, "ZK DrugResist 2.0: A TextMiner to extract semantic relations of drug resistance from PubMed," *J. Biomed. Inform.*, vol. 69, pp. 93–98, May 2017.

[36] R. Arulanandam, B. T. R. Savarimuthu, and M. A. Purvis, "Extracting crime information from online newspaper articles," in *Proc. 2nd Australas. Web Conf.*, vol. 155, 2014, pp. 31–38.

[37] T. Dasgupta, A. Naskar, R. Saha, and L. Dey, "CrimeProfiler: Crime information extraction and visualization from news media," in *Proc. Int. Conf. Web Intell.*, Aug. 2017, pp. 541–549.

[38] R. R. Sedik and A. Romadhony, "Information extraction from Indonesian crime news with named entity recognition," in *Proc. 15th Int. Conf. Knowl. Smart Technol. (KST)*, Feb. 2023, pp. 1–5.

[39] M. Kaufmann, S. Egbert, and M. Leese, "Predictive policing and the politics of patterns," *Brit. J. Criminology*, vol. 59, no. 3, pp. 674–692, Apr. 2019.

[40] F. Wajid and H. Samet, "CrimeStand: Spatial tracking of criminal activity," in *Proc. 24th ACM SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, 2016, pp. 1–4.

[41] A. Abdelkader, E. Hand, and H. Samet, "Brands in newsstand: Spatio-temporal browsing of business news," in *Proc. 23rd SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, 2015, pp. 1–4.

[42] W. L. Gorr and Y. Lee, "Early warning system for temporary crime hot spots," *J. Quant. Criminology*, vol. 31, no. 1, pp. 25–47, Mar. 2015.

[43] A. Rasheed and U. K. Wiil, "A tool for analysis and visualization of criminal networks," in *Proc. 17th UKSim-AMSS Int. Conf. Modeling Simulation (UKSim)*, Mar. 2015, pp. 97–102.

[44] S. Tatale and N. Bhirud, "Criminal data analysis in a crime investigation system using data mining," *J. Data Mining Manag.*, vol. 1, no. 1, pp. 1–13, 2016.

[45] K. Sukhija, S. N. Singh, and J. Kumar, "Spatial visualization approach for detecting criminal hotspots: An analysis of total cognizable crimes in the state of Haryana," in *Proc. 2nd IEEE Int. Conf. Recent Trends Electron., Inf. Commun. Technol. (RTEICT)*, May 2017, pp. 1060–1066.

[46] G. Garcia-Zanabria, E. Gomez-Nieto, J. Silveira, J. Poco, M. Nery, S. Adorno, and L. G. Nonato, "Mirante: A visualization tool for analyzing urban crimes," in *Proc. 33rd SIBGRAPI Conf. Graph., Patterns Images (SIBGRAPI)*, Nov. 2020, pp. 148–155.

[47] A. I. Sirokhin and V. N. Shikhanov, "Structure of psychic attitude of a person to careless traffic violations and potential ways of accident prevention," *J. Siberian Federal Univ.—Humanities Social Sci.*, vol. 4, no. 1, pp. 474-480, 2008.

[48] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Netw.*, vol. 18, nos. 5–6, pp. 602–610, 2005.

[49] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, 2019, pp. 4171–4186, doi: 10.18653/v1/n19-1423.

[50] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2020, pp. 8440–8451. [Online]. Available: https://aclanthology.org/2020.acl-main.747

[51] L. Lowphansirikul, C. Polpanumas, N. Jantrakulchai, and S. Nutanong, "WangchanBERTa: Pretraining transformer-based Thai language models," 2021, *arXiv:2101.09635*.

[52] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[53] C. Polpanumas and W. Phatthiyaphaibun, "Thai2fit: Thai language implementation of ulmfit," Tech. Rep., Jan. 2021, doi: 10.5281/zenodo.4429691.

[54] S. Ramadani, E. Danil, F. Sabri, and A. Zurnetti, "Criminal law politics on regulation of criminal actions in Indonesia," *Linguistics Culture Rev.*, vol. 5, no. S1, pp. 1373–1380, Nov. 2021.

[55] J. C. Cochran, M. J. Lynch, E. L. Toman, and R. T. Shields, "Court sentencing patterns for environmental crimes: Is there a 'green' gap in punishment?" *J. Quant. Criminol.*, vol. 34, pp. 37–66, Sep. 2018.

[56] J. D. Unnever, F. T. Cullen, and J. D. Jones, "Public support for attacking the 'root causes' of crime: The impact of egalitarian and racial beliefs," *Sociol. Focus*, vol. 41, no. 1, pp. 1–33, 2008.

[57] U. Haggård-Grann, J. Hallqvist, N. Långström, and J. Möller, "The role of alcohol and drugs in triggering criminal violence: A case-crossover study," *Addiction*, vol. 101, no. 1, pp. 100–108, 2006.

[58] J. Gallaugher and S. C. Ramanathan, "Choosing a client/server architecture: A comparison of two-and three-tier systems," *Inf. Syst. Manag.*, vol. 13, no. 2, pp. 7–13, 1996.
[59] Office of the Council of State. (2017). *Code of Laws*. [Online]. Available: https://www.krisdika.go.th/law?lawId=4
[60] P. B. Cote. (2022). *Mapping With Aggregated Statistics*. GISManual.com. [Online]. Available: https://www.pbcgis.com/normalize/
[61] D. Wu, S. Fan, S. Yao, and S. Xu, "An exploration of ethnic minorities' needs for multilingual information access of public digital cultural services," *J. Documentation*, vol. 79, no. 9, pp. 1–20, 2022.
[62] C. Kruengkrai, T. H. Nguyen, S. M. Aljunied, and L. Bing, "Improving low-resource named entity recognition using joint sentence and token labeling," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 5898–5905.
[63] X. Hu, H. Zhang, and S. Hu, "Chinese named entity recognition based on BERTbased-BiLSTM-CRF model," in *Proc. IEEE/ACIS 22nd Int. Conf. Comput. Inf. Sci. (ICIS)*, Jun. 2022, pp. 100–104.
[64] H. M. Yohannes and T. Amagasa, "Named-entity recognition for a low-resource language using pre-trained language model," in *Proc. 37th ACM/SIGAPP Symp. Appl. Comput.*, Apr. 2022, pp. 837–844.
[65] F. Rollo, L. Po, and G. Bonisoli, "Online news event extraction for crime analysis," in *Proc. 30th Italian Symp. Adv. Database Syst. (SEBD)*, 2022, pp. 19–22.
[66] J. Fior, T. Favale, L. Cagliero, D. Giordano, M. Mellia, E. Baralis, S. Ronchiadin, P. Baracco, and D. Moncalvo, "Legal entity disambiguation for financial crime detection," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2022, pp. 6639–6641.
[67] Y. Norouzi, "Spatial, temporal, and semantic crime analysis using information extraction from online news," in *Proc. 8th Int. Conf. Web Res. (ICWR)*, May 2022, pp. 40–46.
[68] J. Miller, "The new news media: Democratic implications of undergraduate education and news consumption over social and traditional media," Simon Fraser Univ., Burnaby, BC, Canada, Tech. Rep., 2013.

SIRIPEN PONGPAICHET received the Ph.D. degree in computer science from the University of California at Irvine, Irvine, USA. Currently, she is a Lecturer with the Faculty of Information and Communication Technology, Mahidol University, Thailand. Her research interests include machine learning applications in education and multimedia information systems.

BOONYAPAT SUKOSIT received the bachelor's degree from the Faculty of Information and Communication Technology, Mahidol University, Thailand. His research interest includes machine learning applications in natural language processing tasks.

CHITCHAYA DUANGTANAWAT received the bachelor's degree from the Faculty of Information and Communication Technology, Mahidol University, Thailand. Her research interests include text mining and metadata extraction.

JIRAMED JAMJONGDAMRONGKIT received the bachelor's degree from the Faculty of Information and Communication Technology, Mahidol University, Thailand. His research interests include data mining and analytics.

CHANCHEEP MAHACHAROENSUK received the bachelor's degree from the Faculty of Information and Communication Technology, Mahidol University, Thailand. He specializes in gathering user requirements, creating functional and non-functional requirements, and developing diagrams to facilitate the software development process, with a focus on making software development more efficient.

KANTAPONG MATANGKARAT received the bachelor's degree from the Faculty of Information and Communication Technology, Mahidol University, Thailand. His research interest includes the visualization of intelligently processed information.

PATTADON SINGHAJAN received the bachelor's degree from the Faculty of Information and Communication Technology, Mahidol University, Thailand. His specializations include data engineering and operationalizing machine learning models.

THANAPON NORASET received the bachelor's degree from the Faculty of Information and Communication Technology, in 2007, and the Ph.D. degree in computer science from Northwestern University, USA, in 2017. He is currently a Faculty Member with the Faculty of Information and Communication Technology, Mahidol University, Thailand. His research work and interests include natural language processing and machine learning.

SUPPAWONG TUAROB (Member, IEEE) received the B.S.E. and M.S.E. degrees in computer science and engineering from the University of Michigan, Ann Arbor, and the Ph.D. degree in computer science and engineering and the M.S. degree in industrial engineering from The Pennsylvania State University. Currently, he is an Associate Professor in computer science with the Faculty of Information and Communication Technology, Mahidol University, Thailand. His research interests include data mining in large-scale scholarly, social media, and healthcare domains, and intelligent technologies in social sciences.