Neural Computing and Applications (2025) 37:7183–7205
https://doi.org/10.1007/s00521-024-10833-8

ORIGINAL ARTICLE

Beyond administrative reports: a deep learning framework for classifying and monitoring crime and accidents leveraging large-scale online news

Suppawong Tuarob[1] · Phonarnun Tatiyamaneekul[1,2] · Siripen Pongpaichet[1] · Tanisa Tawichsri[3] · Thanapon Noraset[1]

Received: 28 March 2024 / Accepted: 26 November 2024 / Published online: 15 February 2025

Abstract
The escalating prevalence of violent crimes and accidents underscores the urgent need for efficient and timely monitoring systems. Traditional methods reliant on administrative reports often suffer from significant delays. This paper proposes CRIMSON, a novel framework that leverages large-scale online news to provide real-time insights into crime and accident trends. CRIMSON utilizes a multi-label classification technique that leverages a fine-tuned, pre-trained, cross-lingual language model to accurately categorize news articles. Our experimental results, conducted on a substantial dataset of Thai news articles, demonstrate superior performance, achieving an average F1 score of 86%. Beyond classification, CRIMSON aggregates categorized news into real-time statistics, revealing strong correlations between news-reported incidents and official crime data. This study pioneers online news as a reliable and timely crime and accident monitoring source, offering valuable insights for law enforcement, policymakers, and researchers.

Keywords Multi-label crime/Accident classification · Cross-lingual language models · Online news articles · Deep learning

1 Introduction

The occurrence of crimes and accidents can have significant economic consequences, affecting not only individuals, but also entire communities and nations [31, 88]. In addition, a recent study has demonstrated that increased crime rates might have a detrimental effect on the mental well-being of the local community [69]. Despite reports of declining crime rates in many developed nations [34, 35], serious crimes and accidents remain pervasive in developing countries [23, 66]. In addition to crimes, such developing nations also suffer from prevalent traffic accidents that result in both personal injuries and fatalities [86], and impeding economic growth [92]. Although the distinction between crimes and accidents often hinges on the actors' intent, timely monitoring of the development of crimes and accidents can prove helpful for law enforcement, policymakers, and relevant stakeholders in proactively and sustainably combating these incidents. For example, if policymakers discover an uptick in theft in a region that relies on tourism, they can investigate the root causes of the problem. If the cause is poverty, they can implement incentives to create tourism-related jobs for the poor, which would reduce theft, attract more travelers, and promote safety and prosperity. Furthermore, this information can also assist citizens in preparing for potential crimes and accidents in the area. As another example, if there is a spike in traffic accidents during a festival period, local police could investigate the causes and allocate appropriate healthcare resources to the affected area. If drunken driving is a significant cause, police could dispatch

Most of the research was done while P. Tatiyamaneekul was with the Faculty of Information and Communication Technology, Mahidol University.

✉ Thanapon Noraset
thanapon.nor@mahidol.edu

[1] Faculty of Information and Communication Technology, Mahidol University, Nakhon Pathom, Thailand

[2] Department of Computer Science, University College London, London, UK

[3] Puey Ungphakorn Institute for Economic Research, Bank of Thailand, Bangkok, Thailand

checkpoints or patrols around alcohol-selling places, and policymakers could impose heavier fines, restrict alcohol-selling periods, and create jobs in the area by providing low-cost pick-up services to reduce the incentive to drink and drive.

Literature has proposed using social networks such as Twitter and Facebook as alternative sources to mitigate issues posed by administrative data for monitoring crimes and accidents [76]. These techniques identify crime/accident-related messages and use their geo-tagged locations for real-time visualization on the map. Although social media data can provide real-time information that is easily accessible, its colloquial nature and high volume of communication make it challenging to develop automated systems that accurately filter incident-related messages while remaining resilient against false information [42].

To mitigate these colloquial and misinformative issues in social media data, recent literature on crime and accident analytics has explored large-scale online news articles as a source of reliable and factual information [96]. Reputable news publishers must report timely and verified content to the public. Furthermore, news articles generally use a more standardized language with fewer writing style variations compared to social networks, which can result in better predictive accuracy. Indeed, recent studies have investigated the ability to extract crime and accident information from news articles, focusing on classifying news reports into finer-grained crime/accident types or extracting relevant crime metadata from news articles [96]. However, such previous work did not verify if the information extracted from news articles could reflect real-world statistics, which is crucial for policymakers to understand the overall situation at the national level accurately.

Therefore, our research aims to establish large-scale online news articles as reliable real-world crime and accident information sources. We also present a cross-validation study investigating the correlation between crime/accident information extracted from online news articles and the relevant real-world statistics provided by administrative authorities. Specifically, we propose the intelligent framework *CRIMSON* for *CRIM*e and accident *S*urveillance from *O*nline *N*ews articles. Our framework includes a multi-label crime/accident news classification task that can categorize news articles into multiple crime/accident types. To address the issue of low-resource languages, we experiment with multi-lingual and cross-lingual language models in addition to language-specific and conventional bag-of-words models. Furthermore, we perform cross-validation studies to verify the ability of the extracted crime/accident statistics from news articles to represent real-world trends. This validation compares the statistics extracted from news articles with the corresponding real-world statistics provided by administrative authorities.

We conducted a case study on the efficacy of the news classification models using two reputable online news sources and seven prevalent types of crimes/accidents in Thailand. Our results indicated that the extensive version of the XLMR (XLMR-Large) [26] model performed best, achieving a macro-average F1 of 86%. In addition, our correlation analysis reveals a high correlation between the collective crimes/accidents extracted from online news sources and battery/assault crimes and a moderate correlation with sexual abuse crimes, theft/burglary crimes, deaths from traffic accidents, and injuries from traffic accidents. These findings demonstrate the potential of using large-scale online news articles as a reliable source of information for crime and accident monitoring applications.

To summarize, this paper makes the following key contributions:

- **Dataset creation**: We contributed a novel labeled dataset for multi-label crime and accident classification in Thai online news articles. The dataset comprises 8,567 news articles composed in Thai, multi-labeled into corresponding crime types. This dataset addresses the scarcity of labeled data for Thai text classification tasks.
- **Evaluation of pre-trained language models on a low-resource task**: We comprehensively evaluated various neural network architectures for multi-label classification on Thai news (a low-resource language). This includes benchmarking traditional machine learning methods against deep learning approaches, comparing the performance of a BiLSTM with Thai2Vec embeddings to pre-trained models, such as RoBERTa (fine-tuned on Thai corpora) and multi-lingual BERT, and evaluating the effectiveness of RoBERTa-based cross-lingual models for Thai text classification.
- **Rigorous evaluation methodology**: We ensured the robustness of our findings by employing a standard information retrieval protocol for news classification evaluation. Comprehensive parameter sensitivity analyses were conducted to investigate the impact of various hyperparameters on model performance.
- **Cross-validation with real-world data**: We performed a cross-validation study to assess the real-world applicability of our model's predictions. This involves comparing news-reported crime/accident statistics extracted from our system with official data from the Royal Thai Police and Thailand's Road Accident Victims Protection Company Limited (ThaiRSC). We employ correlation analysis to quantify the relationship between predicted and actual crime/accident data.

The source code of our framework is available on GitHub.[1] The rest of this paper is organized as follows. In Sect. 2, we review relevant research on crime and accident analysis using both historical statistics and online media. Section 3 presents the details of the proposed CRIMSON framework. In Sect. 4, we describe the datasets and protocols used to evaluate different modules in the framework. Section 5 provides further discussions on the relevant experiment results and limitations of the proposed framework. Section 6 briefly examines possible ethical issues and implications for society. Finally, Sect. 7 concludes the paper.

2 Related work

Crimes and accidents in many developing countries are challenging to control for various reasons and have become prevalent public concerns [3, 19, 75]. The capacity to track the progression of crimes and accidents in real time could prove valuable for law enforcement and policymakers in managing public safety and formulating policies to tackle their root causes [47]. The advent of artificial intelligence technologies as well as the accumulation of historical data has given rise to many predictive policing and crime/accident analytics studies and applications [81]. Most prior research has used historical data, such as case reports and crime/accident statistics, to categorize, predict, and depict crime and accident occurrences [84]. Therefore, this section first discusses previous research that used historical administrative data for analytical purposes. In addition, since using online news to monitor crimes and accidents is a novel aspect of our study, we also review literature that uses online media, such as online social networks and news articles, for crime and accident analytics.

2.1 Crime and accident analysis using historical reports and statistics

Police departments or related organizations in many countries keep track of reported crime and violence events, along with their metadata, such as criminals' and victims' information, date/time, locations, and fine-grained crime types [16, 29, 59]. These longitudinal historical records have been utilized for crime and accident data mining research, such as fine-grained classification of crimes, predicting the number of crime cases, and forecasting criminality trends [101].

Administrative crime and accident reports are often structured in a way where conventional machine learning algorithms can be directly used to perform analyses. A straightforward, primary task would be classifying each record into a finer-grained crime/accident type. Tayal et al. [94] were among the first who analyzed crime statistics from the National Crime Records Bureau in India by clustering crime incidents and proposed to use k-nearest neighbor (kNN) to find similar past crimes. Their framework also aided visualization by mapping crime cases onto Google Maps. Similarly, ToppiReddy et al. [97] used the historical data from the UK police department to visualize both the 2D (top-down) map and the street view. They also experimented with kNN and Naive Bayes to classify crimes into one of the five crime types, including anti-social behaviors, drug, theft, robbery, and vehicle crime. Kumar and Nagpal [54] addressed the crime pattern analysis by classifying crime and violent incidents into one of the five crime types, namely disorder, property crime, vehicle theft, traffic, and drug, using the date/time and location as input features and NaiveBayes as the classifier. Alsaqabi et al. [7] retrieved 45,620 violent events in Saudi Arabia, as well as their metadata, from the GDELT project [56] from January 2018 to September 2018. They then studied factors that affect these crimes using PCA (principle component analysis) and FAMD (factor analysis of mixed data) using the event code as the target attribute. They also experimented with crime classification using many conventional machine learning algorithms and found Naive Bayes with FAMD features to yield the best accuracy. Qazi and Wong [78] studied burglary crimes in the UK using 1.6 million crime reports along with associated offenders' and victims' information from the UK Law Enforcement Agency. They proposed a human-centered, data-oriented approach to facilitate users' navigation through these reports and crime-related entities. Recently, Kshatri et al. [53] collected 60,000 crime cases from India's National Crime Record Bureau between 2001 and 2015, comprising 11 types of crimes in 28 states of India. They proposed an ensemble-stacking-based crime prediction method (SBCPM) to classify these crime records into one of the 11 types. A total of 36 features are extracted from each crime record and used to train different conventional machine learning classifiers that are ensembled by an SVM classifier in a stacking manner.

In addition to incident classification tasks, previous studies have investigated the possibility of predicting and forecasting the actual numbers of crime/accident cases from historical records. Ingilevich and Ivanov [46] proposed to predict the numbers of banditry, massacre, and robbery cases in each region in Saint Petersburg, Russia, during 2014–2017 by framing this problem into a regression task where linear regression, logistic regression, and gradient boosting algorithms were validated. Rummens et al. [83] conducted a study to identify whether the ambient or residential populations should be used for normalizing crime rates. Their study discovered that using

[1] https://github.com/Zenonist/Crimson.

the ambient population, where the physical locations are determined by mobile phone usage, to analyze crime rates was more appropriate because such a population is dynamic, real time, and can prevent data loss in certain unpopulated areas. In an attempt to forecast future crime trends, Catlett et al. [17, 18] proposed to divide a city into crime-dense regions using DBSCAN. For each region, ARIMA was used to train a regressor to predict future crimes. A case study of Chicago was used to validate their proposed method. Hu et al. [93] investigated using the spatiotemporal Bayesian model, widely used in epidemiology, to predict potential burglary hotspots and identify developing trends in Wuhan city, China. Their proposed model was generated from diverse data sources, including reported cases, administrative boundaries, population, and points of interest (POIs). Feng et al. [36] proposed forecasting crime rates in major cities such as San Francisco, Chicago, and Philadelphia. They experimented with various conventional and deep learning sequence models such as Phophet [95], LSTM, and neural networks, and varying configurations such as lags, change points, and other model hyperparameters. Ahmed et al. [2] focused on human and drug trafficking crime and pointed out that finding sufficient data to perform relevant analyses was challenging. They then proposed consolidating incident reports, crime reports, and court records in Kentucky, USA, to construct a comprehensive dataset. They also framed the crime forecasting problem as a classification task where each 7-day sliding window is classified as whether there would be a human trafficking incident or not. In terms of accidents, Sunny et al. [91] proposed to use Holt-Winters and ARIMA methods to forecast the aggregate number of road accidents in Kerala, India, using the historical statistics from 1999–2016 as a case study. Lee et al. [55] proposed to predict traffic accident severity in Seoul, Republic of Korea. In addition to the historical accident statistics, they also incorporated the road geometry and weather conditions when training the random forest, artificial neural network, and decision tree models to predict the accident level in a given period. Recently, Jomnonkwao et al. [48] highlighted that traffic accidents are of significant concern in Thailand and proposed to remedy such an issue with the ability to forecast road traffic deaths. Various socioeconomic variables were investigated for viable indicators of the aggregate numbers of traffic deaths, including Thailand's GDP, the number of registered vehicles, and the energy consumption of the transportation sector. These pieces of sequential information were used to train a multivariate time series forecasting model using multiple linear regression, where the best performance was reported with 6.4% MAPE.

The above-mentioned approaches rely on historical crime/accident reports or statistics collected and made available by relevant police departments or responsible organizations to perform analyses. Furthermore, the findings and evaluations were performed retrospectively on the available data. However, in some countries, especially developing ones, such administrative data are delayed, infrequent, or inaccessible. For example, in Thailand, at the time of writing in March 2023, aggregate crime statistics were reported and made available annually, where the most recent record was dated September 2020.[2] Such delay and low-frequency characterizing the official data hinder the ability to timely and dynamically monitor the development of crime and accident incidents. Mitigating such drawbacks posed by administrative data, recent studies have investigated using user-generated online media such as social networks, online forums, and online news as alternative sources for monitoring real-world phenomena, as next reviewed in the following subsection.

2.2 Crime and accident analysis using online media

Many real-world incident monitoring applications have investigated utilizing ubiquitous online media to address the delay, infrequency, and inaccessibility problems characterizing administrative data. For example, Twitter data have been used to monitor epidemics [25], real-time traffic [28], disastrous events [90], and suicide attempts [68]. Furthermore, large-scale online news sources have been used to enhance the accuracy in predicting stock movements [60], monitoring COVID-19 incidents [52], and detecting business events [79].

For crime and accident analysis purposes, online social networks have been established as viable real-time sources for monitoring both aggregate level and fine-grain incidents. Gerber [38] was among the first to explore the use of online social media for crime prediction. The author collected geo-tagged tweets in Chicago, USA, and used them to generate a topic model using latent Dirichlet allocation (LDA). Then, the city area was divided into small grids, where a regression model was trained with the topic distribution of the tweets mapped to each grid to predict the number of incidents for each of the 25 crime types. The findings showed that incorporating the topic distribution from tweets improved the prediction accuracy for 19 out of the 25 crime types. Later, Chen et al. [22] also investigated the crime prediction problem in Chicago. Arguing that weather conditions could impact criminals' decisions to commit crimes, they proposed to use both sentiments extracted from tweets and weather information to enhance crime prediction. Their experimental results, however, only

[2] http://edw-opendata.moi.go.th/dataset/page/5e9fb64e35a3945ea521caba5cc1e2e915ed575168900.

showed a slight improvement over the baseline using the standard hot-spot model on the theft incident prediction task. For accident analysis, Ali et al. [4] proposed a framework for extracting road accident metadata from social messages. Their framework first identifies traffic-related messages using the LDA topic distribution. Then, polarity and location are extracted from each traffic-related message. Finally, FastText and Word2Vec were used to generate word embedding where BiLSTM was used to classify each tweet, whether it discusses a traffic or car accident. Similarly, Azhar et al. [9] proposed a framework to detect accident-related tweets. The task was framed as a binary classification problem where sentiment, emotions, weather, location, and timestamp were extracted from tweets where GRU, RNN, and LSTM were validated for classifiers. Their experiment results reported a classification accuracy of 94%.

While social networks offer timely and accessible information that could be valuable for real-time monitoring systems of crime and accident activities, extracting crucial knowledge from such user-generated, colloquial sources of information could be challenging. First, automatic classification algorithms for social media messages are still far from perfect. Due to the massive amount of social media data generated each day, it is vital that accurate classification algorithms are deployed to find the needles in this haystack. However, especially in developing countries where low-resource and non-standard languages are used, accurate natural language processing models for social media texts are challenging to develop [68]. These classification errors could then propagate to downstream knowledge extraction tasks that result in amplifying inaccuracies. Second, information available on social networks has been scrutinized for lack of veracity (i.e., inaccurate or fake information), mainly due to inadequate moderation and verification [5, 24, 71]. While it is essential for law enforcement and policy-related personnel to receive as accurate and meaningful information as possible, the decision support system must rely on not only timely, but also trustworthy information sources [14].

Therefore, reputable online news outlets have been established as alternative timely sources of real-world information that are easier to process and more trustworthy than user-generated social media data [67]. Since standard language is used to compose news articles with minor variations in linguistic styles, most modern machine learning classification models often yield acceptable accuracy. Furthermore, since news publishers' businesses are built upon factual and informative content, the information in news articles is often validated before publishing, bypassing the credibility issues posed by social media information.

Indeed, since crimes and accidents are among the topics that most news outlets target to report, recent work has focused on developing intelligent techniques to extract meaningful insights from crime/accident reporting news articles. Alruily et al. [6] proposed a linguistic-based algorithm for extracting crime metadata such as crime types, location, and nationality in Arabic news articles. In addition, Srinivasa and Thilagam [89] extracted crime metadata from online news articles in India using a set of language-specific rules. However, these algorithms rely on local grammar, which prevents them from generalizing to other languages.

Another body of research attempts to identify crime-related articles from online news articles and often frames the problem as a binary (crime vs. non-crime) or multiclass classification task. Kalmegh [49] evaluated REPTree, Simple Cart, and RandomTree for the classification of Indian news articles into one of the seven categories, where each article is represented as a bag of words. Similarly, Ghankutkar et al. [39] addressed the crime news identification problem as a binary classification task where TF-IDF term weights were used to represent each document, and SVM, Naive Bayes, and Random Forest were validated for classifiers. Furthermore, Magnusson et al. [63] proposed utilizing a simple conjugate Bayesian model to identify news leads that would be of interest to local journalists. A case study of reported offense news was used to validate the efficacy of their proposed model.

Since many respectable news sources already have distinct sections for crimes and accidents, the added value of using crime/accident news filtering algorithms is minimal. Thus, another body of research on mining news articles for crimes and accidents focuses on news stories already pre-categorized as crime/accident reports by the publishers. The majority of these studies define the issue as a multi-class classification task to categorize a crime/accident news article into one of the finer-grained types for further fine-grained analyses. Rajapakshe et al. [80] collected crime/accident news articles in Sri Lanka in 2018 and validated decision tree, random forest, and SVM for their ability to classify these articles into one of the nine crime types: murder, kidnapping, robbery, drug dealing, accident, rape, assault, and burglary. Umair et al. [99] obtained 900 crime/accident news articles from eight popular news outlets in Pakistan, dating from 2011 to 2019, and classified them into eight crime/accident types, including robbery, accident, blast, kidnapping, murder, shot, suicide, and arrest. Representing each document with n-grams, they experimented with kNN and random forest for the multiclass classification task. Furthermore, each document was geocoded by GeoPy[3] for visualization on the map.

Besides using traditional n-gram-based and conventional classification algorithms, the advent of deep learning

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

has stepped into crime analyses using news articles. Rollo et al. [82] represented an Italian news article with Word2Vec embeddings where various traditional machine learning classification algorithms were validated for categorizing these documents into one of the 13 crime categories: theft, drug dealing, illegal sale, robbery, aggression, scam, murder, kidnapping, mistreatment, evasion, sexual violence, money laundering, and fraud. Furthermore, the data imbalance issues were also explicitly studied and mitigated using SMOTE [20]. Deepak et al. [30] addressed the fine-grained crime classification in Google News, where fuzzy c-means clustering was first used to guide data labeling. Then, GloVe was used to extract word embeddings for training a BiLSTM classifier to classify a news article into one of the 14 crime types. Khan et al. [51] validated the use of a Bangla-BERT-based model against LSTM, BiLSTM, and other traditional machine learning models for their ability to classify a Bangla crime news headline into one of the six crime types, including terrorism, murder, corruption, harassment, drug, and robbery. A case study of 7,897 Bangla news headlines was manually labeled and showed that the Bangla-BERT-Base model performed the best. Closest to our work was the study by Thaipisutikul et al. [96], who proposed a multiclass classification algorithm for categorizing news articles in Thailand, which were pre-categorized into the "Crime" category, into five finer-grained classes, including burglary, accident, corruption, drug, and murder. They represented each news article with TF-IDF features and trained standard machine learning classification algorithms such as multinomial Naive Bayes, gradient boosting machine, random forest, kNN, multinomial logistic regression, and support vector machine.

In our research, we propose a classification scheme of eight classes, including gambling, murder, sexual abuse, theft/burglary, drug, battery/assault, accident, and non-crime/accident. Note that these crime/accident types were defined due to their prevalence in Thailand. However, our proposed framework is independent of the classification scheme and can be easily generalized to other classification schemes tailored specifically for different local regions. Contrary to the aforementioned work, we notice that a crime/accident incident could potentially fall into multiple types and accordingly frame the problem as a multi-label classification task where a news article can belong to more than one class. Furthermore, since our case study comprises news article composed in a low-resource language (i.e., Thai), to allow the framework to generalize to other human-centered applications tailored for different cultures and languages [27], our framework proposes to comparatively experiment with a cutting-edge RoBERTa-based model pre-trained on Thai corpus (i.e., WangchanBERTa [62]), multi-lingual BERT-base model (i.e., MBERT [32]), and cross-lingual RoBERTa-based model (i.e., XLMR [26]) and validated their performance on extensive online news datasets, comparing against the traditional BiLSTM classifier with Thai2Vec embedding and conventional machine learning classification methods such as Naive Bayes, SVM, and extreme gradient boosting with TF-IDF representation. Finally, most of the above-mentioned work on analyzing crimes/accidents from news articles assumed that incidents reported in the news are representative of real-world ones and did not cross-validate with real-world statistics. In our framework, the aggregate crime statistics from online news articles are cross-validated with real-world reported crime/accident statistics to determine their ability to represent the landscape of crimes and accidents at the national level. Regardless, the ability to accurately classify incidents reported in news articles into fine-grained crime/accident types could prove useful for further downstream tasks in criminology that focus on individual incidents rather than at the aggregate level, such as automatic crime metadata extraction [15], incident profiling [37], and extracting linked information for predictive policing applications [12, 45].

3 Methodology

This paper presents a framework for utilizing news as an alternative source of crime and accident incidents. The framework first crawls news articles from reputable online publishers and classifies news articles into crime/accident categories. Finally, statistics of each category can be used for monitoring. This section describes the methods and techniques used to create and evaluate the proposed framework. The overall process of the methodology is illustrated in Fig. 1.

3.1 Data collection

This study used two main types of raw data: online news articles and the historical statistics of crimes and accidents. The online news articles are the main raw data of the framework. We selected well-established national news publishers and crawled publicly accessible news articles from their websites. Although online news publishers typically provide a coarse-grained classification of the crime/accident news articles, finer-grained classification is needed for monitoring applications that target specific crime/accident types. Furthermore, we noticed that some crime/accident articles were categorized into other non-crime/accident categories to better reach their target audiences—collecting articles from the pre-categorized "crime" category would miss these relevant articles. In this work, we collected all news articles regardless of the

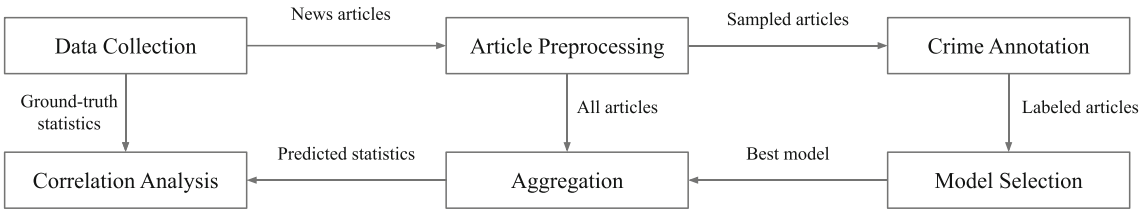

Fig. 1 A high-level diagram of the methodology used to study the proposed *CRIMSON* framework

publishers' classifications. All news articles were used to further process and evaluate the proposed framework.

Ultimately, we would like the crime and accident statistics extracted from the news articles to correlate with those happening in the real world. Such correlation would establish news-reported statistics as being representative of the real-world phenomenon that inherently benefits the downstream monitoring applications at the national level. Thus, in addition to the news articles, the historical statistics of crimes and accidents collected from official authorities were used as the ground-truth data for evaluating the framework. Two categories of ground-truth statistics were collected for this research. First, the annual statistics of the crimes are broken down into different subtype of crimes, namely batter/assault, murder, rape, and theft/burglary. While the news articles can provide statistics at a higher frequency than annually, these are the only available crime statistics publicly published in Thailand. Therefore, we also collected traffic accident statistics aggregated into monthly data for higher-frequency correlation evaluation.

3.2 Data preprocessing

The news articles crawled from the online news publishers were in the HTML format containing irrelevant information, such as HTML tags, advertisements, navigation menus, and links to other articles. We would like to use only textual news data to determine its categories. News articles are often written in a simple structure of title, introduction, and full description. The title is short but catchy. The introduction is around three to four sentences, providing an overview of the reported incident. The description can be of arbitrary length with detailed information. For all of our news articles, we removed special characters and HTML elements and kept only the published date, title, introduction, and description of each news article.

3.3 Data annotation

An essential component of the framework is crime news classification models. We require a labeled dataset not only for training the models, but also for testing their classification performance, while most news outlets already have delicate categories for crimes and/or accidents. These are coarse-grained categories that do not facilitate analyses into specific crime types. Furthermore, some crime and accident reports could be categorized as local news, while some non-crime/accident articles may also appear as ones.

First, we defined a classification scheme comprising eight categories, including gambling, murder, sexual abuse, theft/burglary, drug, battery/assault, accident, and non-crime/accident. Such a scheme was primarily designed to capture violent incidents prevalently reported in Thailand's news and closely match the categories used in the official crime statistics. Note that the non-crime/accident class also includes news articles that report other crimes not primarily listed above, such as financial and digital crimes, as these crimes typically do not involve violence and are not frequently reported in Thailand's news articles. However, future work could easily adjust the classification scheme to suit the specific information needs. To obtain a labeled dataset, we randomly sampled news articles from all sources, focusing on the crime and local news categories where most crime and accident news events were reported and asked volunteers to annotate them. The volunteers studied the categories with example news articles and were asked to provide at least one category for each news article.

3.4 Crime news classification

This step involves building classification models and selecting the best one for further analysis. Since a single news article could belong to more than one category, we approached the crime news classification as a multi-label text classification task. In other words, given a news article, $\boldsymbol{x}$, there is a label vector $\boldsymbol{y} = [y^{(1)}, ..., y^{(K)}]$, where $y^{(k)} \in \{0, 1\}$. The text classification consists of two components: feature extraction and classification. Specifically, the features of each input article, $\boldsymbol{x}$, are first extracted into a representation vector, $\boldsymbol{h} = f(\boldsymbol{x}; \theta)$, where θ is the parameters of the feature extractor. Then, a classification model predicts the probability of each label $p(y^{(k)} = 1|\boldsymbol{x}) = g_k(\boldsymbol{h}; \phi)$, where ϕ is the parameter of the classification model. Finally, a threshold was applied to make a category prediction $\hat{y}^{(k)} = \mathbb{1}[p(y^{(k)} = 1|\boldsymbol{x}) > \tau^{(k)}]$,

where $\tau^{(k)}$ is the threshold for the category k. While recent work in text classification focuses on using pre-trained language models, classical methods such as bag-of-words models or recurrent neural network models could provide higher accuracy in some settings. In this work, we investigated a wide range of feature extraction methods ($f(\cdot)$) and classification learning algorithms ($g(\cdot)$).

3.4.1 Pre-trained language models

A deep learning text classification method typically uses a pre-trained feature extractor, $f(\cdot;\theta_p)$, and then fine-tunes it along with the classification model using domain-specific datasets. The pre-trained feature extractors are usually transformer-based language models such as BERT [32] and RoBERTa [61]. Given the set of labeled news articles, $D = \{(\boldsymbol{x_1}, \boldsymbol{y_1}), ..., (\boldsymbol{x_N}, \boldsymbol{y_N})\}$, and $f(\cdot;\theta_p)$, we trained the crime classification model using a gradient-based algorithm to minimize the binary cross-entropy loss:

$$L(\theta, \phi; D) = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_i^{(k)} \cdot \log(p(y_i^{(k)} = 1|x_i)) + (1 - y_i)^{(k)} \cdot \log(1 - p(y_i^{(k)} = 1|x_i)), \quad (1)$$

$$p(y_i^{(k)} = 1|\boldsymbol{x}_i) = g_k(f(\boldsymbol{x}_i;\theta);\phi) \quad (2)$$

where $g_k(\cdot)$ was the k^{th} output of a linear classifier having the sigmoid activation function. Since the primary data studied in this work was in the Thai language, we selected a Thai pre-trained language model, WangchanBERTa [62], and other multi-lingual and cross-lingual pre-trained language models, such as multi-lingual BERT (MBERT) [32] and XLMR [26], for comparison.

These transformer-based models are fine-tuned for four epochs, as reported optimal for text classification [32]. Other hyperparameters include the AdamW optimizer with epsilon = 1e-6, the learning rate of 2e-5, and a weight decay of 0.01.

3.4.2 Word embedding

The original datasets used to pre-train the language models might not be suitable for the crime news classification, potentially causing the classification models to perform poorly. To check this mismatch, we included a standard deep learning model for comparison. Specifically, the configuration and the loss function were similar to the contextual embedding models. But, we replaced the transformer-based feature extractor with the bi-directional long short-term memory network (BiLSTM) [41] using Thai pre-trained word embedding from Thai2Vec [74].

3.4.3 Bag-of-words approaches

To establish a robust baseline for our crime news classification task, we adopted a traditional machine learning approach inspired by Thaipisutikul et al. [96], who pioneered the use of online news for crime analysis in Thailand. Their work demonstrated the feasibility of classifying crime news into fine-grained categories using SVM with bag-of-words features.

Building upon this foundation, we employed a standard text representation technique, TF-IDF [85], as the feature extractor ($f(\cdot)$). In this setting, we treated a news article as a document and computed IDF as the parameter of the feature extractor from the training samples. To classify an article, we trained three widely used machine learning classifiers ($g(\cdot)$): Naive Bayes (NB) [102], support vector machines (SVM) with the linear kernel [73], and XGBoost with logistic loss function for binary classification (`binary:logistic`) [21]. These models, trained independently for each crime category, served as our benchmark for evaluating the performance of more sophisticated neural network models. This traditional approach, while effective for certain tasks, often falls short when dealing with the complexities of natural language, motivating our exploration of deep learning alternatives.

3.5 Model selection

In the context of machine learning experimentation, model selection refers to the process of identifying the most suitable algorithm or model from a pool of candidates to categorize documents into their respective classes effectively. The downstream tasks in the crime monitoring applications from online news rely on accurately classifying articles into fine-grained types. Therefore, the ability to precisely distinguish non-crime articles is still vital. The crime news classification models described in Sect. 3.4 can have varied performance due to different configurations of the input articles ($\boldsymbol{x}$), the feature extractors ($f(\cdot)$), the classification models ($g(\cdot)$), and the thresholds (τ). Selecting the best configuration for further analyses is an important step to reduce the time complexity as we have to apply the classification model to millions of news articles. Various standard evaluation metrics for text classification were reported, including precision, recall, F1, accuracy, Matthews correlation coefficient (MCC), and AUC-ROC. The F1 of the positive class was used as the main evaluation criterion to select the best models. Furthermore, the stratified tenfold cross-validation protocol was adopted to reduce any bias from the test data and to produce more reliable performance results. Algorithm 1 summarizes the above classification model selection process.

Algorithm 1 Classification model selection

Input: $\langle \mathbb{X}, \mathbb{Y} \rangle$: News articles and corresponding labels; $f(\cdot)$: Feature extractor with parameters θ; $\mathbb{G}$: Classification models
Output: g^*: The best classification model
$\mathbb{T} = \oslash$
for $x \in \mathbb{X}$ **do**
  $x' \leftarrow Preprocess x$ $h \leftarrow f(x', \theta)$
  $y \leftarrow Y(x)$
  Add $\langle h, y \rangle$ to $\mathbb{T}$;
end
$\mathbb{T} \leftarrow$ Stratified 10-fold split $\mathbb{T}$
for $g \in \mathbb{G}$ **do**
  10-fold cross-validate g with dataset $\mathbb{T}$
end
$g^* \leftarrow g \in \mathbb{G}$ with the highest F1
return g^*;

3.6 Crime activity monitoring

The traditional crime monitoring methods heavily rely on delayed and often incomplete official administrative reports, limiting their effectiveness in providing timely insights. This reliance on bureaucratic processes hinders law enforcement agencies, policymakers, and the public from accessing up-to-date information on crime trends and patterns. In contrast, our framework introduces a novel approach that leverages the power of online news to provide a real time and transparent overview of crime activity at the regional level. By utilizing advanced natural language processing techniques, we are able to extract critical information from news articles, offering a more comprehensive and timely picture of the crime landscape. This innovative method not only complements, but also surpasses traditional monitoring methods by providing a more dynamic and informative perspective on crime trends.

Given a crime news classification model, we can classify news articles into our pre-defined categories ($\hat{y}$). The total number of articles in a predicted category is a proxy for the crime statistics of the corresponding category. Specifically, given a collection of news articles over a period of time t, D_t, we applied the best model to obtain the predicted categories ($\hat{y}_t$). Then, we computed the proxy statistics as $\hat{z}_t = \sum_{x \in D_t} y_t$, where $\hat{z}_t$ is a vector whose k^{th} element was a statistic of a crime category k in the period t. In addition, we normalized these statistics, and a relevant discussion was given within the period. The news-reported statistics were aggregated using the same frequencies as the corresponding ground-truth data (i.e., annually for crimes and monthly for accidents).

To validate the crime monitoring task, Pearson correlation analysis was used to find the statistical relationship between the numbers of crime/accident news articles and actual administrative reported crimes for each crime/accident type. Formally, given two series of ground-truth statistics $z^{(k)}$ and $\hat{z}^{(k)}$, we compute the correlation $r^{(k)}$ as follows:

$$r^{(k)} = \frac{\sum_t (z_t^{(k)} - z_\mu^{(k)})(\hat{z}_t^{(k)} - \hat{z}_\mu^{(k)})}{\sqrt{\sum_t (z_t^{(k)} - z_\mu^{(k)})^2} \sqrt{\sum_t (\hat{z}_t^{(k)} - \hat{z}_\mu^{(k)})^2}} \tag{3}$$

where $z_\mu^{(k)}$ and $\hat{z}_\mu^{(k)}$ are the mean of the ground-truth cases and news-reported cases, respectively.

4 Experiment, results, and discussion

This section details the news datasets, annotation, and ground-truth statistical data used for validating our proposed framework. Then, the experiment results are reported with relevant discussion.

4.1 Datasets

Two types of data were used in the experiments: news articles and ground-truth historical statistics of crimes and accidents. The news articles were used to validate the efficacy of the proposed classification algorithms, while the historical statistics were used to verify if crimes and accidents extracted from news articles could collectively represent those in the real world. These data and statistics pertain to Thailand; however, the proposed framework could easily be generalized to other languages, owing to the capability of multi-lingual and cross-lingual language models.

Thai news articles were collected from two reputable online news publishers in Thailand, anonymized as S1 and S2, respectively. Both news sources were founded in 1950, starting with paper-based news publications. They started to expand their publications in the online realm around 2009. Therefore, we started collecting publicly accessible news articles from both sources from 2009 to 2021, totaling around 1.5 million articles, comprising 730,996 and 790,622 articles, respectively. Note that we collected all the news articles, including all categories of articles, not just the crime and accident categories. The reason was that we noticed that some crime/accident articles were sometimes categorized as *local* news. Furthermore, some articles, such as crime-related TV drama snippets, can mimic crime-reporting articles. Therefore, we would like to validate the classifiers' ability to detect these potential false positives as well. Table 1 shows the distribution of news articles collected from both publishers each year. Figure 2 compares news articles collected from both the selected news sources during 2009–2021.

Table 1 Statistics of the collected online news articles

Source	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	Total
S1	25,636	33,476	4,555	60,465	80,492	85,286	79,269	73,375	68,292	60,907	62,011	64,535	32,697	730,996
S2	1	37,450	87,557	85,764	73,846	58,752	42,312	84,508	75,820	63,118	57,561	62,784	61,149	790,622
Total	25,637	70,926	92,112	146,229	154,338	144,038	121,581	157,883	144,112	124,025	119,572	127,319	93,846	1,521,618
Percentage	1.68%	4.66%	6.05%	9.61%	10.14%	9.47%	7.99%	10.38%	9.47%	8.15%	7.86%	8.37%	6.17%	100.00%

Articles from the crime, accident, and other categories were randomly selected for annotation into finer-grained crime/accident and non-crime/accident types as set forth in Sect. 3.3. Three independent, well-educated annotators were asked to label each article into at least one of the predefined types. The final labels were resolved with majority votes. A total of 8,567 articles were labeled as illustrated in Table 2. Note that since the annotation scheme was multi-labeling (tagging), one article can belong to many crime/accident types; therefore, the sum of articles in Table 2 is greater than the actual number of annotated articles.

4.2 Crime news classification performance

In the experiments to select the best crime news classification model, we used the labeled datasets and ran tenfold cross-validation experiments. For each fold, 80%, 10%, and 10% of the data were allocated in a stratified manner for training, validation, and testing sets. The experiments were conducted on a Linux machine with 20 CPU Threads, 128 GB of RAM, and an RTX 3090 GPU.

Tables 3 and 4 highlight classification results in terms of F1 and AUC-ROC, respectively, for each class. Furthermore, the macro-average F1 and AUC-ROC of the positive classes (crimes and accidents) are also reported to quantify how well each classifier distinguishes different crime/accident types on average.

Easing analyses, Fig. 3 visualizes the comparison of macro-average precision, recall, F1, accuracy, MCC, and AUC-ROC of all the classification methods. It is a consensus that XLMR-Large has the best performance in terms of an average F1 of 0.86. Such a model also has the best F1 in all classes except gambling, where XLMR-base has a better performance. Though TF-IDF-based models such as NB, SVM, and XGB are not the best-performing algorithms, their performance is not as plummeting as one would expect when compared with state-of-the-art deep learning methods. Specifically, the best TF-IDF-based model, SVM, yields 16.4% better performance than BiLSTM and only 3.6% worse than XLMR-Large in terms of average F1. This could mean that the deep learning technologies for computing low-level semantic representation for the Thai language are still not mature enough that the sole traditional term-weight features could yield a slightly inferior performance. These findings could also shed light on the need to improve the semantic interpretation of deep learning methods, especially transformer-based language models for low-resource languages.

It is interesting to note that though WanchanBERTa was pre-trained specifically on large copula of Thai documents and was reported the best performance in many downstream Thai document classification tasks [44, 62], its performance is still inferior to the cross-lingual XLMR-

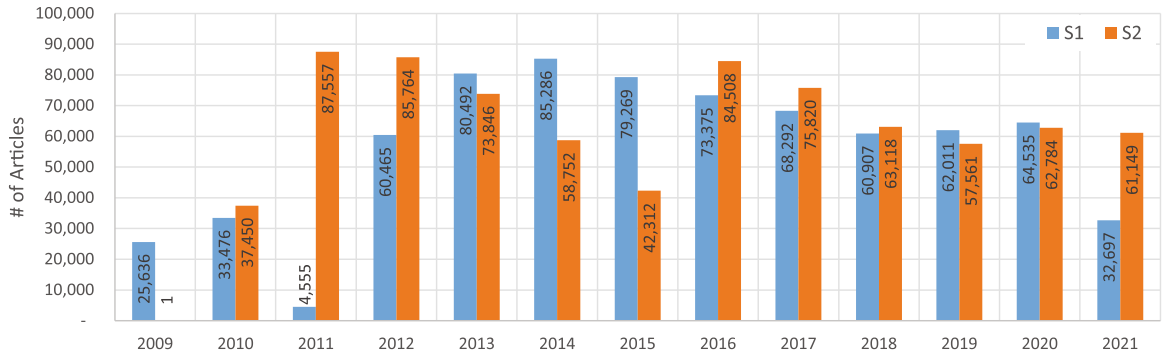

Fig. 2 Comparison of the number of news articles from both the news sources (i.e., S1 and S2) collected from 2009–2021

Table 2 Statistics of the annotated news articles for model evaluation

Class	# Samples	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/Accident	1,406	16.41%

Base and XLMR-Large models by 0.2% and 2.0%, respectively, in terms of average F1. Though WanchanBERTa and XLMR are based on RoBERTa, such performance differences could be due to the cross-lingual pre-training task of the XLMR models, where semantics from different languages could be transferred among each other, resulting in better overall performance, compared to WangchanBERTa, which was only pre-trained on Thai documents.

It is also worth noting that MBERT performs much worse than one would expect from a multi-lingual model, especially for the class gambling, where MBERT's F1 is only 0.008. Such a trend also appears in NB's performance, where the gambling class has the lowest performance compared to others. It is our conjecture that such low performance could have been caused by the low amount of gambling articles in the dataset (i.e., only 2.91%), resulting in biased performance or insufficient priors to learn gambling-specific patterns. As a result, MBERT yields the worst performance among the deep learning models, underperforming XLMR-Large by 23.14%.

Considering the class-wise performance of XLMR-Large in Table 3, the F1 scores do not differ much across all the classes (i.e., 0.75–0.92). However, some classes perform better than others. For example, classes murder, sexual abuse, and drug yield an F1 above 0.9, while the battery/assault class has an F1 of 0.753. After investigation, we found that murder, sexual abuse, and drug news articles could be easily spotted by the presence of indicating keywords such as *kill*, *die*, *rape*, and/or illegal *drugs' names*. On the contrary, the language used in battery/assault news reports can appear similar to murder articles, with an

Table 3 Performance comparison, in terms of F1, of different classification algorithms for each class

Classifier	Gambling	Murder	Sexual abuse	Theft/ Burglary	Drug	Battery/ assault	Accident	Non-crime/ Accident	Avg. crime/ Accident	Avg. all classes
NB	0.335	0.760	0.744	0.598	0.690	0.611	0.682	0.733	0.631	0.644
SVM	0.883	0.886	0.864	0.776	0.872	0.716	0.816	0.816	0.831	0.829
XGB	0.875	0.887	0.877	0.756	0.867	0.689	0.781	0.803	0.819	0.817
BiLSTM	0.707	0.833	0.711	0.611	0.750	0.606	0.724	0.755	0.706	0.712
WangchanBERTa	0.888	0.905	0.889	0.778	0.907	0.720	0.845	0.809	0.848	0.843
MBERT	0.008	0.854	0.753	0.658	0.849	0.641	0.733	0.789	0.642	0.661
XLMR-Base	***0.903***	0.907	0.889	0.784	0.903	0.727	0.824	0.823	0.848	0.845
XLMR-Large	0.887	***0.917***	***0.904***	***0.818***	***0.916***	***0.753***	***0.846***	***0.839***	***0.863***	***0.860***

The bold-italic figures indicate the highest performance in their respective categories

Table 4 Performance comparison, in terms of ROC, of different classification algorithms for each class

Classifier	Gambling	Murder	Sexual abuse	Theft/ Burglary	Drug	Battery/ assault	Accident	Non-crime / Accident	Avg. crime / Accident	Avg. all classes
NB	0.628	0.847	0.829	0.747	0.801	0.780	0.799	0.841	0.776	0.784
SVM	0.908	0.916	0.904	0.837	0.902	0.803	0.866	0.869	0.877	0.876
XGB	0.925	0.917	0.917	0.831	0.916	0.783	0.840	0.855	0.876	0.873
BiLSTM	0.823	0.879	0.827	0.759	0.836	0.743	0.830	0.841	0.814	0.817
WangchanBERTa	0.927	0.934	0.936	0.857	0.945	0.807	0.910	0.869	0.902	0.898
MBERT	0.502	0.887	0.827	0.775	0.911	0.755	0.834	0.864	0.785	0.794
XLMR-Base	***0.948***	0.936	0.945	0.867	0.947	0.818	0.900	0.875	0.909	0.905
XLMR-Large	0.944	***0.942***	***0.948***	***0.891***	***0.952***	***0.839***	***0.917***	***0.894***	***0.919***	***0.916***

The bold-italic figures indicate the highest performance in their respective categories

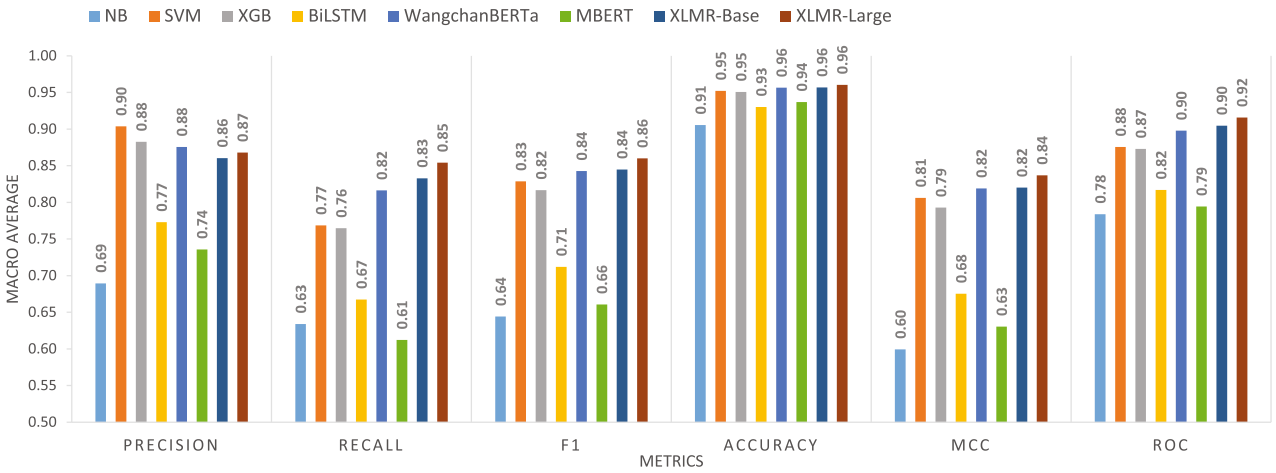

Fig. 3 Comparison of the classification performance (macro-averages of all classes), in terms of precision, recall, F1, accuracy, MCC, and AUC-ROC

identical set of violence-indicating keywords. Note that the main difference between murder and battery/assault articles is that the latter incidents do not result in the victim's death. However, the ways and the choices of words that the reporters use for narration could be similar.

4.3 Correlation analysis with real-world statistics

The previous sections discussed the performance of the news article classification models, where we determined that XLMR-Large was the best model for classifying news articles in our datasets. However, to establish news sources as proxies for real-world aggregate phenomena, they must be cross-validated with real-world ground-truth statistics. Therefore, in this section, correlation analysis was performed between the numbers of news articles categorized as a particular crime/accident type with its corresponding aggregate real-world crime/accident statistics. The ground-truth battery/assault, murder, rape, and theft/burglary statistics during 2016–2020 were collected from Thailand's Ministry of Interior,[4] where the statistics were available at the annual frequency. The accident cases in 2020 were collected from Thailand's Road Accident Victims Protection Company Limited (ThaiRSC).[5] Though the accident data were available daily, we aggregated them into monthly statistics to reduce possible sensitive bias from delayed reports.

Table 5 reports the correlation between the numbers of selected crime/accident types computed from news articles from both the news sources (anonymized as S1 and S2) and their corresponding administrative statistics collected from official sources. The correlation analysis was performed on both the actual and normalized numbers of articles. The normalization was performed by simply computing the fraction of the target number of articles with respect to all the articles in a given time period. Note that the accident

[4] http://edw-opendata.moi.go.th/dataset/page/5e9fb64e35a3945ea521caba5cc1e2e915ed575168900.

[5] https://www.thairsc.com/.

Table 5 Pearson correlation coefficients between real-world cases of battery/assault, murder, sexual abuse, and accident, and the numbers of news articles classified into their corresponding crime/accident categories, both with and without normalization

Normalization	Crime/Accident type	Ground-truth statistics	Period	# of Articles	
				S1	S2
None	Battery/Assault	Battery/Assault	2016–2020 (Yearly)	0.981***	0.858**
	Murder	Murder	2016–2020 (Yearly)	0.367	0.025
	Sexual abuse	Rape	2016–2020 (Yearly)	0.570*	0.564*
	Theft/Burglary	Theft/Burglary	2016–2020 (Yearly)	0.217	0.001
	Accident	Accident-Death	Jan–Dec 2020 (Monthly)	0.173	0.035
	Accident	Accident-Injure	Jan–Dec 2020 (Monthly)	0.360	0.153
	Accident	Accident-Total	Jan–Dec 2020 (Monthly)	0.358	0.151
	Accident	Accident-Death	Jan–Oct 2020 (Monthly)	0.623*	0.463*
	Accident	Accident-Injure	Jan–Oct 2020 (Monthly)	0.620*	0.401*
	Accident	Accident-Total	Jan–Oct 2020 (Monthly)	0.621*	0.402*
Normalized	Battery/Assault	Battery/Assault	2016–2020 (Yearly)	−0.518	0.569*
	Murder	Murder	2016–2020 (Yearly)	0.419*	−0.137
	Sexual abuse	Sexual abuse	2016–2020 (Yearly)	−0.642	−0.641
	Theft/Burglary	Theft/Burglary	2016–2020 (Yearly)	0.313	0.509*
	Accident	Accident-Death	Jan–Dec 2020 (Monthly)	−0.061	−0.146
	Accident	Accident-Injure	Jan–Dec 2020 (Monthly)	−0.020	−0.051
	Accident	Accident-Total	Jan–Dec 2020 (Monthly)	−0.020	−0.053
	Accident	Accident-Death	Jan–Oct 2020 (Monthly)	0.246	0.223
	Accident	Accident-Injure	Jan–Oct 2020 (Monthly)	0.159	0.181
	Accident	Accident-Total	Jan–Oct 2020 (Monthly)	0.161	0.182

$***$, $**$, and $*$ denote very strong (0.9−1.0), strong (0.7−0.89), and moderate (0.4−0.69) correlation levels, respectively, according to Schober et al. [72]'s criteria

cases have three sub-categories: those resulting in death (Accident-Death), injuries (Accident-Injure), and the combined cases (Accident-Total). Furthermore, we noticed that there was a declining trend of news-reporting accident incidents toward the end of the year from both news sources, despite the fact that accident rates should rise due to holiday travel. Therefore, we performed another set of correlation analyses with the accident cases during January–October 2020 until it became better to understand why such a disagreement between news-reported and actual accident trends was observed around the last two months of the year.

Considering the correlation with the non-normalized news articles, S1 gives a very strong correlation with the aggregate battery/assault cases ($r = 0.981$) and a moderate correlation with the rape cases ($r = 0.570$) and 10-month accident cases (January–October 2020), with $r = 0.623$, 0.620, and 0.621 for accident-death, accident-injure, and accident-total, respectively. While the actual news-reported cases have a good correlation with some of the crime/accident types, the normalized version does not exhibit such strong signals, except for murder and theft/burglary, where the normalized numbers of murder (S1) and theft/burglary (S2) news reports have a moderate correlation of 0.419 and 0.509, respectively. An explanation for these high correlations between news-based statistics and the above crime/accident ground-truth statistics could be the fact that these are serious violent incidents that news reporters aim to disseminate. Similarly, these violent crimes/accidents are likely reported to the police since they involve victims, as reflected by the ground-truth statistics.

Figure 4 (left) and 4 (right) illustrates the actual numbers of battery/assault and rape cases (bars) in comparison with the numbers of battery/assault and sexual abuse articles extracted from S1 and S2 during 2016–2020. Visually, a high correlation is observed between the news-reported statistics (especially from S1) and the corresponding ground-truth statistics, as evidenced in Table 5.

Figure 5 visualizes similar information for the total accident cases (both resulting in death and injuries). Visually, a high correlation is observed with the news-reported statistics from both S1 and S2 until October 2020, when the news-reported accidents started declining while the actual accidents continued to rise. It is our conjecture that while each news publication has a certain capacity to report events, other stories might be of interest to the

Fig. 4 Comparison of the real-world reported battery/assault (left) and sexual abuse (right) cases with the corresponding numbers of news articles, categorized as battery/assault and sexual abuse, respectively, from the two selected news sources (i.e., S1 and S2)

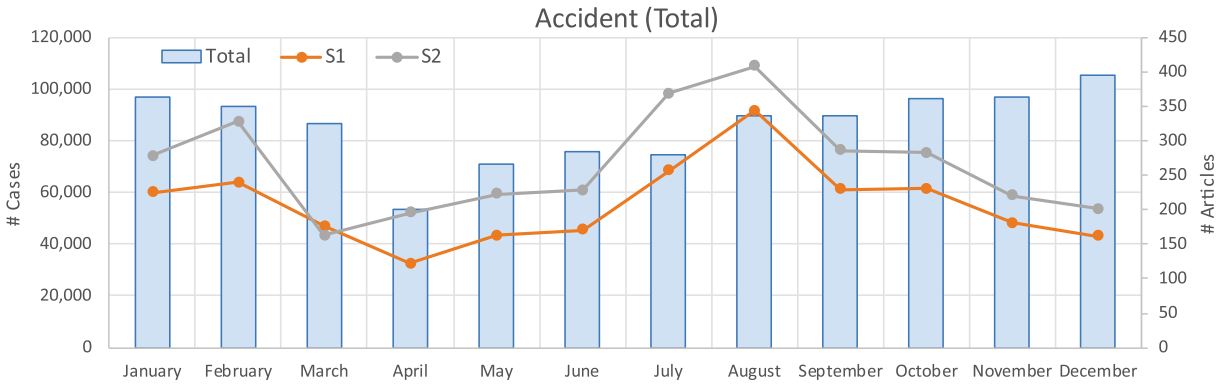

Fig. 5 Comparison of the real-world reported total accident cases with the corresponding numbers of news articles, categorized as accident, from both news sources

audience than repeated similar accidents. However, such a conjecture requires further scientific and statistical validation to conclude.

It is worth observing that correlations vary across different types of crimes. For example, lower correlations are observed for the murder and theft/burglary cases compared to battery/assault and rape cases. Explaining such a phenomenon could lead to representation bias in news media [100]. Therefore, de-biasing and normalizing reported incidents could be the next step of our research to establish online news as a potential source of information for accurate monitoring of real-world crimes and accidents.

In summary, the traditional crime monitoring methods heavily rely on delayed and often incomplete official administrative reports, limiting their effectiveness in providing timely insights. This reliance on bureaucratic processes hinders law enforcement agencies, policymakers, and the public from accessing up-to-date information on crime trends and patterns. In contrast, we proposed a novel approach that leverages the power of online news to provide a real time and transparent overview of crime activity at the regional level. By utilizing advanced natural language processing techniques, critical information could be extracted from news articles, offering a more comprehensive and timely picture of the crime landscape. This innovative method complements and surpasses traditional monitoring methods by providing a more dynamic and informative perspective on crime trends. We have demonstrated the feasibility of this approach by correlating crime statistics extracted from online news with official records. Our findings reveal a strong correlation for specific crime categories such as battery/assault, sexual abuse, and accidents, highlighting the potential of online news as a valuable complement to traditional crime monitoring methods.

5 Discussion

The correlation results in Table 5 presented a novel finding that aggregated predictions from online news articles could be a timely source of some types of crime/accident statistics. In addition, these predictions were obtained from a fine-tuned language model using a small amount of training data. These findings confirmed previous work that studied similar combinations of text classification and online text,

such as social media and mental health [68], online news articles and stock prices [98], and online news sentiments and oil prices [58]. In addition, the results in Fig. 3 showed that a multi-lingual model outperformed a local language model. This provided additional support in pre-training a multi-lingual language model [26]. The rest of this section provides additional discussions and limitations of this work.

5.1 Impact of classification models

The CRIMSON framework relies on the accurate classification of crime news articles. The performance metrics (F1 score and AUC-ROC) for various classification models in Tables 3 and 4 provide insight into the strengths and weaknesses of traditional machine learning and neural network approaches for this specific task.

This research employed both traditional machine learning classifiers, including NB, SVM, and XGB, and deep learning-based models, such as BiLSTM, WangchanBERTa, MBERT, and XLMR. In general, traditional machine learning classifiers represent a document with bag-of-words features where each word is represented with a numerical weight. These models generally exhibit strong performance on well-defined, structured data, and are often interpretable, making it easier to understand the factors influencing classification decisions. SVM, in particular, excels at handling high-dimensional data and finding optimal decision boundaries, yielding the average F1 of 82.9%. However, they often struggle to capture complex patterns and relationships within text data. This could be attributed to their limited ability to handle the inherent ambiguity and nuances of natural language. As observed in the table, traditional methods, especially NB, show relatively lower performance compared to neural network models, especially in terms of average F1 score and performance on specific crime categories.

On the contrary, neural networks, especially deep learning models, are adept at capturing intricate patterns and relationships within text data. They can learn rich representations of words and sentences, leading to improved performance on complex tasks like text classification that requires semantic understanding. For example, BiLSTM can capture sequential dependencies in text, while pre-trained language models like WangchanBERTa, MBERT, and XLMR benefit from extensive pre-training on large datasets, enabling them to generalize better to new tasks. For example, XLMR-Large highlights the benefit of larger models and more extensive pre-training with copula from diverse languages. Both XLMR-Base and XLMR-Large models outperform other models, demonstrating the effectiveness of cross-lingual language models for this task. However, these neural networks can be computationally expensive to train and require large amounts of data. They often lack interpretability, making it challenging to understand the decision-making process.

Extending the prior study [96] that found SVM to perform best in a similar task, our research showed that neural networks, especially pre-trained language models, significantly outperform traditional machine learning methods for Thai crime news classification. Furthermore, language-specific pre-training is crucial for achieving optimal performance in low-resource languages like Thai. Finally, larger and more complex models generally lead to better performance.

5.2 Impact of different levels of news information

A typical news article is not simply monotonic but structured into different levels to provide relevant information to the right information needs [64]. For generalization, this research assumes that each news article comprises a title, an introduction, and a description. While it is legitimate that combining all these pieces of content would provide full information to machine learning algorithms, it is also fair to raise questions about how each of these different article zones would impact the classification performance. The answers to this question can guide the implementation of this framework when computational resources or accessible news information is limited. For example, one may want to expand the data scope to cover multiple news sources with a constant computing resource. For example, certain news publishers may only allow public access to the articles' titles or short snippets.

The best TF-IDF-based classifier (i.e., SVM) and representative deep learning methods (i.e., WangchanBERTa, MBERT, XLMR-Base, and XLMR-Large) were used to investigate the impact of different levels of news information on the classification efficacy. Each algorithm was trained with different parts of news articles, namely titles, introductions, descriptions, and combined information, where tenfold cross-validation was applied. The comparative experimental results regarding the precision, recall, and *F1* are enumerated in Table 6, whose average *F1* scores are depicted in Fig. 6 for ease of analysis.

The results from SVM, WangchanBERTa, XLMR-Base, and XLMR-Large agree that the classification efficacy depends on the length of the input text. Comparing using titles, introduction, and descriptions for training, models trained with only titles yield the lowest performance, while training the models with descriptions gives the highest classification efficacy. The models trained with introductions yield performance somewhere in between. Combining the title, introduction, and description into a single

Table 6 Performance comparison, in terms of macro-averages, of different classification algorithms trained with only the title, introduction, description, and combined parts of news articles

Classifier	Title			Introduction			Description			Combined		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
SVM	0.84	0.68	0.75	0.87	0.70	0.77	0.88	0.75	0.81	0.90	0.77	0.83
WangchanBERTa	0.85	0.72	0.78	0.86	0.74	0.79	0.86	0.82	0.84	0.88	0.82	0.84
MBERT	0.80	0.59	0.68	0.77	0.60	0.67	0.69	0.58	0.63	0.74	0.61	0.66
XLMR-Base	0.82	0.73	0.77	0.85	0.77	0.80	0.85	0.81	0.83	0.86	0.83	0.84
XLMR-Large	0.82	0.73	0.77	0.85	0.77	0.80	0.85	0.81	0.83	0.87	0.85	0.86

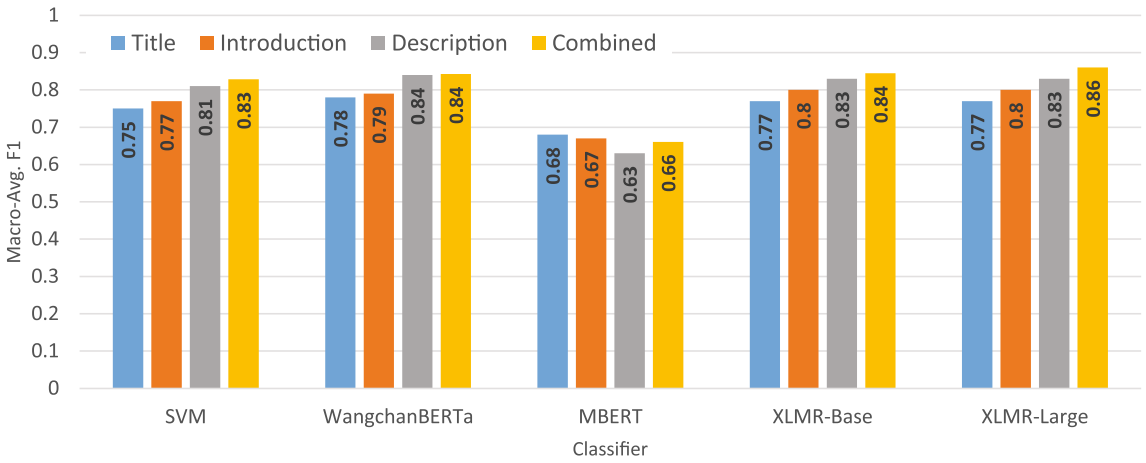

Fig. 6 Comparison of macro-average F1 of different classification algorithms trained with different levels of news articles' information

document yields the best results. It is worth noting that the performance from MBERT does not follow the above analysis, where training MBERT with only titles yields the best results, while the descriptions yield the lowest performance. This could be due to MBERT's deteriorating ability to handle longer texts as reported in certain low-resource languages [70].

Focusing on the classification performance by the best classifier, i.e., XLMR-Large, training the model with titles and introductions alone yields an average *F1* of 0.77 and 0.8, lower than that of the combined information (*F1* = 0.86) by only 10.47% and 6.98%, respectively. Since titles and introductions are relatively shorter than the full description when computation resources or available news information is restricted, these models still yield relatively acceptable performance.

5.3 Impact of fine-grained threshold tuning

The classification of crime news articles was framed as a multi-label classification task, and the one-versus-rest protocol was used. Therefore, each crime/accident type is treated as a binary classification task. Typically, a binary classifier outputs a probability in the range of [0,1], where the default cut-off threshold of 0.5 is used to squash the decision. However, research has shown that such a default value may not be optimal for document classification and should be tuned in a similar manner as other model hyperparameters [57].

This section, therefore, investigates this matter. The experiments were conducted with WangchanBERTa and XLMR-Large models, where the probability thresholds were tuned to maximize the *F1* of the positive class using the validation data. Table 7 compares the precision, recall, *F1*, accuracy, and MCC between using the default threshold of 0.5 and the tuned threshold for each classifier on each class. $\Delta F1(\%)$ denotes the relative performance (*F1*) change with respect to the default threshold version of the same classifier.

On average, tuning the thresholds slightly worsens the performance. Specifically, the F1 scores of WanchanBERTa and XLMR-Large drop by 0.19% and 0.47%, respectively. Inspecting each class, only the accident and non-crime/accident classes benefit from threshold tuning. An explanation for why threshold tuning does not help much in our experiment with WanchanBERTa and XLMR-Large could be that these models were trained with sufficient training data and epochs that they confidently output the probabilities close to 0 or 1; therefore, moving the threshold a little away from 0.5 would not affect the outcomes much.

Table 7 Comparison of the classification performance before (threshold = 0.50) and after tuning the probability threshold for each class, using WanchanBERTa and XLMR-Large as the classifiers

Class	Classifier	Threshold	P	R	F1	ΔF1 (%)	Accuracy	MCC
Gambling	WangchanBERTa	0.5	0.934	0.855	0.888	–	0.994	0.888
	XLMR-Large	0.5	0.887	0.892	0.887	–	0.994	0.885
	WangchanBERTa (Tuned)	0.26	0.848	0.908	0.873	−1.73	0.994	0.871
	XLMR-Large (Tuned)	0.44	0.870	0.904	0.882	−0.58	0.994	0.881
Murder	WangchanBERTa	0.5	0.902	0.909	0.905	–	0.946	0.865
	XLMR-Large	0.5	0.911	0.923	0.917	–	0.945	0.882
	WangchanBERTa (Tuned)	0.52	0.907	0.906	0.906	0.06	0.951	0.866
	XLMR-Large (Tuned)	0.6	0.913	0.913	0.913	−0.49	0.937	0.876
Sexual abuse	WangchanBERTa	0.5	0.902	0.880	0.889	–	0.986	0.881
	XLMR-Large	0.5	0.908	0.903	0.904	–	0.987	0.897
	WangchanBERTa (Tuned)	0.45	0.896	0.890	0.891	0.21	0.986	0.883
	XLMR-Large (Tuned)	0.38	0.890	0.909	0.897	−0.73	0.987	0.890
Theft/ Burglary	WangchanBERTa	0.5	0.838	0.728	0.778	–	0.961	0.760
	XLMR-Large	0.5	0.844	0.796	0.818	–	0.972	0.802
	WangchanBERTa (Tuned)	0.48	0.828	0.731	0.775	−0.41	0.961	0.757
	XLMR-Large (Tuned)	0.46	0.834	0.779	0.803	−1.84	0.972	0.787
Drug	WangchanBERTa	0.5	0.914	0.901	0.907	–	0.980	0.895
	XLMR-Large	0.5	0.916	0.916	0.916	–	0.977	0.904
	WangchanBERTa (Tuned)	0.46	0.910	0.900	0.904	−0.39	0.975	0.891
	XLMR-Large (Tuned)	0.47	0.905	0.911	0.908	−0.91	0.977	0.895
Battery/ Assault	WangchanBERTa	0.5	0.789	0.664	0.720	–	0.870	0.654
	XLMR-Large	0.5	0.765	0.742	0.753	–	0.879	0.685
	WangchanBERTa (Tuned)	0.29	0.682	0.793	0.731	1.44	0.865	0.652
	XLMR-Large (Tuned)	0.33	0.727	0.784	0.752	−0.09	0.869	0.681
Accident	WangchanBERTa	0.5	0.860	0.832	0.845	–	0.971	0.832
	XLMR-Large	0.5	0.846	0.849	0.846	–	0.972	0.833
	WangchanBERTa (Tuned)	0.45	0.845	0.846	0.845	−0.04	0.972	0.831
	XLMR-Large (Tuned)	0.61	0.864	0.839	0.850	0.47	0.972	0.838
Non-Crime/ Accident	WangchanBERTa	0.5	0.866	0.761	0.809	–	0.943	0.777
	XLMR-Large	0.5	0.868	0.812	0.839	–	0.958	0.809
	WangchanBERTa (Tuned)	0.52	0.868	0.759	0.806	−0.30	0.943	0.776
	XLMR-Large (Tuned)	0.39	0.863	0.827	0.844	0.57	0.959	0.815
Macro Average	WangchanBERTa	–	0.876	0.816	0.843	–	0.957	0.819
	XLMR-Large	–	0.868	0.854	0.860	–	0.960	0.837
	WangchanBERTa (Tuned)	–	0.848	0.842	0.841	−0.186	0.956	0.816
	XLMR-Large (Tuned)	–	0.858	0.858	0.856	−0.469	0.958	0.833

5.4 Challenges in dataset selection

While the proposed framework can easily be generalized to different countries and languages, this research focused on crime statistics in Thailand and utilized a Thai-annotated dataset as a case study for the following reasons:

– **Limited Applicability of Existing Crime News Datasets**: Our research focused on a fine-grained classification scheme specific to the Thai context. Existing annotated crime news datasets in high-resource languages (e.g., Italian [13], Chinese [8], English [65, 87]) often employ different classification schemes with varying granularities. For instance, the Italian dataset features 13 categories, but only four map directly to ours. Furthermore, the Chinese dataset has five classes, including theft, intentional injury, dangerous driving, fraud, and traffic accident, which overlap

with only three of our classes (i.e., theft/burglary, battery/assault, and accident). This discrepancy reflects potential variations in crime prevalence across different countries, as evidenced by the significant imbalance favoring theft in the Italian data (73.37%) compared to ours (9.03%). While some English datasets exist, they lack the required fine-grained crime-type classification.

- **Data Size Considerations**: While exploring high-resource datasets, we found that their size is comparable to or smaller than ours. For example, despite having around 10,395 articles, the Italian dataset suffers from significant class imbalance. Removing the dominant "Theft" class reduces the usable data to approximately 2768 articles. This size falls short of our own dataset, which contains over 8567 labeled articles.
- **Uniqueness of Thai Language and Crime Reporting**: We focused on Thai to address a specific need: developing a crime monitoring system tailored for Thailand. This system would benefit law enforcement, travelers, and investors seeking to gauge crime risks in specific locations. Additionally, as the authors' primary language, Thai facilitated efficient annotation by subject matter experts, who could categorize news articles into fine-grained crime types. Finally, the accessibility of ground-truth crime statistics in Thailand allowed us to explore the correlation between news-based and actual crime data (as reported in Sect. 4.3). This aspect would be challenging to replicate without readily available ground-truth crime statistics in other languages/countries.

The points mentioned above highlight the need to create labeled data in a local language. This is consistent with research on classifying crime news in languages with limited resources. In those cases, creating annotated datasets in the local language rather than relying on labeled data from languages with more resources has been the approach. Although we were unable to find suitable annotated crime news datasets for cross-resource learning, we discovered that cross-lingual models pre-trained with large-scale texts from multiple languages (such as XLMR) performed best in our tasks. Specifically, they outperformed RoBERTa-based models trained specifically with only Thai text (WangchanBERTa). As research on crime analysis from online news continues to progress, there is expected to eventually be enough annotated datasets in various languages to train cross-lingual models, reducing the need to create additional datasets. We believe our focus on Thai and Thai-annotated data allows for a more nuanced understanding of crime news classification in a low-resource language context. However, we recognize the potential of cross-lingual approaches as the availability of diverse language datasets grows.

5.5 Vision and future directions

This paper represents a foundational step toward a comprehensive, human-centered intelligent system for real-time crime and accident monitoring. By establishing online news as a reliable data source and demonstrating the efficacy of our multi-label classification approach, we have laid the groundwork for a more sophisticated ecosystem.

To fully realize the potential of this research, several avenues for future exploration emerge. First, the integration of diverse data sources, including social media, weather data, and socioeconomic indicators, will enrich the system's ability to detect emerging crime patterns and inform preventative measures. Second, the application of advanced deep learning techniques, especially those pre-trained on multi-lingual crime textual copula, is crucial for enhancing the system's capabilities.

Specifically, large language models (LLMs) can be leveraged to improve text understanding and information extraction from news articles, enabling more accurate and nuanced crime classification. Active learning strategies can be employed to prioritize the labeling of the most informative data points, thereby reducing annotation costs and improving model performance. Additionally, contrastive learning can be explored to learn robust representations of crime-related entities and events, facilitating improved similarity search and anomaly detection.

The integration of these techniques could also enable the construction of a more sophisticated knowledge graph, capturing intricate relationships between entities and events. This knowledge graph can serve as the foundation for advanced analytics, including graph neural networks for link prediction and anomaly detection. Moreover, the development of explainable AI models is crucial for building trust and transparency in the system's decision-making processes.

Ultimately, our vision is to create a dynamic, human-centered system that empowers law enforcement, policymakers, and the public to proactively address crime and safety challenges. By combining cutting-edge AI with human expertise, we can develop a system that not only monitors crime and accidents but also predicts, prevents, and responds to these incidents effectively.

5.6 Limitations

While the experiment results of both the crime/accident news article classification task and the correlation analyses with real-world administrative statistics are encouraging, we have encountered challenges that could potentially give rise to future novel research problems.

First, after conducting the error analysis from the news classification results, major causes of error were identified. These include false positive identification of original crime/accident incidents, most of which are follow-up reports and crime-mimicking articles. In Thailand, some major crime and accident incidents require time to solve, wherein news reporters keep their audience informed by publishing follow-up stories. However, these follow-ups do not contribute to new incidents, but some were false-positively identified as ones. Furthermore, some of Thailand's newspapers publish TV drama snippets that sometimes narrate crime or accident scenes. We found some of these made-up stories appeared as part of the false positives as well. Therefore, a future direction could improve the classification performance by investigating more into these misclassified samples and identifying features that characterize follow-up and other crime/accident-mimicking articles.

Furthermore, the proposed framework only performs the correlation analyses between the news-reported events and the ground-truth crime/accident statistics. Even with certain incident types that correlate highly with news-reported statistics, the system would only be useful for investigating the current trend of crimes and accidents. However, it would be more useful to policymakers if the system could also forecast the actual numbers of crime/accident cases in the future. Such ability requires developing forecasting models that learn from historical statistics, online media, and other socioeconomic variables. However, current ground-truth crime statistics in our case study (Thailand) are only available at an annual frequency, resulting in insufficient data points to train learning-based forecasting models. Research into the direction of spatial transfer learning [1], which transfers the models from similar geographical data-rich regions to data-sparse ones, could be investigated as our next steps.

6 Ethical issues and societal implications

Freely available online data power the proposed framework. Ethical concerns and social ramifications could arise if adopted and implemented for real-world applications. Some of these issues are addressed here.

6.1 Media bias

Studies showed that news media could be biased [11, 43]. Though extensive research has investigated the news biases in political [50], race [77], and gender [10] domains, studying the bias in crimes/accidents reported in news media is limited [33]. The business model of newspapers is driven by reporting information that readers want to know [40], which could potentially derail journalists from reporting true samples that reflect the actual distribution of the crimes and accidents but only focus on those that could make catchy headlines. While trustworthy news sources are valuable for fact extraction, the selection and representation biases in crime/accident news could lead to partial information that bends the policymakers' attention to solving *popular* crimes rather than those that truly incur the major cost to society.

6.2 Exploitation

Although we hope that the proposed framework will be employed for the greater good of society, the ability to investigate real-world crimes and accidents, the majority of which might inflame public opinion may prove to be a double-edged sword. The government may adopt the framework for beneficial reasons, such as enabling policymakers to monitor the changes in crime and accident patterns in response to certain policies or to act promptly when certain crimes begin to develop. On the other hand, one might use the information for one's own personal gain. Seeing, for instance, an expanding trend of public dread of an uncontrolled murder, one may benefit egotistically by selling fraudulent surveillance systems or raising the prices of such equipment. Also, competing political parties may use the public's worry to attack the government's inability to handle related situations.

The difficulties associated with exploitation have the potential to reach a global scale. The fact that the news data employed in the proposed framework is accessible to the public implies that anybody in the world may learn about a nation's crime/accident landscape. After finding that a developing nation is afflicted with drug crimes, for instance, illicit drug dealers may use this information to change their unlawful selling locations or devise means to evade capture.

6.3 Burden to the people

Implementing systems that continuously gather and process large-scale data comes with a cost. Gaining access to even publicly accessible data is not always free. These expenses include programmers who create and manage the data gathering and processing pipeline, storage, cybersecurity enforcement, and data access fees dictated by the terms and conditions of various news media services. If the framework is to be implemented by governmental authorities, the government must invest funds to acquire access to and collect data from news outlets. Whatever the extent of the project, the cash for this budget will almost definitely come from tax revenues. If the project is significant and requires a large quantity of news data from many platforms, the

economic burden will almost certainly be passed on to the general people through increased taxes. Although it is not our goal for the people to shoulder this burden, the agency in charge of materializing the proposed framework would have to assess the benefits that the people would receive against the expenses involved with implementing the system.

7 Conclusions and future directions

This paper proposed *CRIMSON*, an intelligent framework for collecting online news articles, classifying them into fine-grained crime and accident types, and cross-validating the news-reported statistics with the official ground-truth data. The news classification problem was framed as a multi-label classification task, where multi-lingual, cross-lingual, and traditional text classification models were investigated for their ability to categorize news articles written in low-resource languages. Parameter sensitivity was analyzed to empirically show how training the models with different parts of a news article and tuning probability cut-offs would impact the classification efficacy. A case study of seven crime/accident types prevalent in Thailand and two local trustworthy online news sources was used to validate the news classification models. Furthermore, the six ground-truth official statistics, including reported cases of battery/assault, murder, rape, theft/burglary, death from accidents, and injuries from accidents, at the national level, were used in the correlation analysis with the news-reported incidents. The overarching objective of our research is to implement a system that local police officers and national level policymakers can use to monitor real-time trends of crimes and accidents. This research represents only the initial step toward reaching that goal. The next steps involve studying de-biasing and normalizing news-reported statistics, expanding the analysis to cover emerging crime types such as digital and financial crimes, and developing deep learning models to parse news articles and automatically extract crime/accident metadata.

Acknowledgements This research project was supported by Mahidol University (MU-MiniRC02/2564). The authors have no financial or proprietary interests in any material discussed in this article.

Author contributions Suppawong Tuarob involved in conceptualization, methodology, software, validation, formal analysis, investigation, writing—original draft, visualization, and funding acquisition; Phonarnun Tatiyamaneekul involved in resources and software; Siripen Pongpaichet involved in resources and software; Tanisa Tawichsri involved in conceptualization, resources, and writing—review and editing; Thanapon Noraset involved in conceptualization, methodology, writing—review and editing, supervision, and project administration.

Funding Open access funding provided by Mahidol University. Mahidol University (MU-MiniRC02/2564), Grantee: Suppawong Tuarob.

Data availability The source code associated with the proposed method and experiments is available at https://github.com/Zenonist/Crimson.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Abubakr M, Akoush B, Khalil A, Hassan MA (2022) Unleashing deep neural network full potential for solar radiation forecasting in a new geographic location with historical data scarcity: a transfer learning approach. Eur Phys J Plus 137(4):474
2. Ahmed S, Gentili M, Sierra-Sosa D, Elmaghraby AS (2022) Multi-layer data integration technique for combining heterogeneous crime data. Inf Process Manag 59(3):102879
3. Ajide FM (2020) Criminal activities and road accidents in Nigerian transport industry. Transp Dev Econ 6:1–10
4. Ali F, Ali A, Imran M, Naqvi RA, Siddiqi MH, Kwak K-S (2021) Traffic accident detection and condition analysis based on social networking data. Accid Anal Prev 151:105973
5. Alkhamees M, Alsaleem S, Al-Qurishi M, Al-Rubaian M, Hussain A (2021) User trustworthiness in online social networks: a systematic review. Appl Soft Comput 103:107159
6. Alruily M, Ayesh A, Zedan H (2014) Crime profiling for the Arabic language using computational linguistic techniques. Inf Process Manag 50(2):315–341
7. Alsaqabi A, Aldhubayi F, Albahli S (2019) Using machine learning for prediction of factors affecting crimes in Saudi Arabia. In: Proceedings of the 2019 International Conference on Big Data Engineering, p 57–62
8. Amanda-WangXiao. Bert-based-crime-news-classification. URL https://github.com/Amanda-WangXiao/BERT-based-crime-news-classification
9. Azhar A, Rubab S, Khan MM, Bangash YA, Alshehri MD, Illahi F, Bashir AK (2022) Detection and prediction of traffic accidents using deep learning techniques. Clust Comput 26(1):1–17
10. Bauer NM (2022) Who covers the qualifications of female candidates? examining gender bias in news coverage across national and local newspapers. Journal Mass Commun Q 101(3):10776990221100514

11. Beelen K, Lawrence J, Wilson Daniel CS, Beavan D (2022) Bias and representativeness in digitized newspaper collections: introducing the environmental scan. Digit Scholarsh Humanit 38(1):fqac037
12. Berk RA (2021) Artificial intelligence, predictive policing, and risk assessment for law enforcement. Annu Rev Criminol 4:209–237
13. Bonisoli G, Di Buono MP, Po L, Rollo F (2023) Dice: A dataset of italian crime event news. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, p 2985–2995
14. Carden F (2009) Knowledge to policy: making the most of development research. IDRC
15. Castano S, Ferrara A, Falduti M, Montanelli S (2019) Crime knowledge extraction: an ontology-driven approach for detecting abstract terms in case law decisions. In Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, p 179–183
16. Castro M, Tirso C (2023) The impacts of the age of majority on the exposure to violent crimes. Empir Econ 64(2):983–1023
17. Catlett C, Cesario E, Talia D, Vinci A (2018) A data-driven approach for spatio-temporal crime predictions in smart cities. In: 2018 IEEE International Conference on Smart Computing (SMARTCOMP), p 17–24. IEEE
18. Catlett C, Cesario E, Talia D, Vinci A (2019) Spatio-temporal crime predictions in smart cities: a data-driven approach and experiments. Pervasive Mob Comput 53:62–74
19. Chanci L, Kumbhakar SC, Sandoval L (2023) Crime underreporting in bogotá: a spatial panel model with fixed effects. Empir Econ 66(5):1–32
20. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
21. Chen Tianqi, He Tong, Benesty Michael, Khotilovich Vadim, Tang Yuan, Cho Hyunsu, Chen Kailong, Mitchell Rory, Cano Ignacio, Zhou Tianyi et al (2015a) Xgboost: extreme gradient boosting. R Package Version 0.4-2 1(4):1–4
22. Chen X, Cho Y, Jang SY (2015b) Crime prediction using twitter sentiment and weather. In: 2015 systems and information engineering design symposium, p 63–68. IEEE
23. Chokprajakchat S, Techagaisiyavanit W, Mulaphong D, Iyavarakul T, Kuanliang A, Laosunthorn C (2023) Tracking violence in Thailand: the making of violent crime index. Secur J 37(1):1–20
24. Collins B, Hoang DT, Nguyen NT, Hwang D (2021) Trends in combating fake news on social media-a survey. J Inf Telecommun 5(2):247–266
25. Comito Carmela (2021) How covid-19 information spread in us? The role of twitter as early indicator of epidemics. IEEE Trans Ser Comput 15(3):1193–1205
26. Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave E, Ott M, Zettlemoyer L, Stoyanov V (2020) Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, p 8440–8451, Online, Association for Computationa Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747
27. Dan W, Fan S, Yao S, Shuang X (2023) An exploration of ethnic minorities' needs for multilingual information access of public digital cultural services. J Doc 79(1):1–20
28. D'Andrea E, Ducange P, Lazzerini B, Marcelloni F (2015) Real-time detection of traffic from twitter stream analysis. IEEE Trans Intell Transp Syst 16(4):2269–2283
29. Das P, Das AK (2019) Graph-based clustering of extracted paraphrases for labelling crime reports. Knowl-Based Syst 179:55–76
30. Deepak G, Rooban S, Santhanavijayan A (2021) A knowledge centric hybridized approach for crime classification incorporating deep bi-lstm neural network. Multimed Tools Appl 80(18):28061–28085
31. Detotto C, Otranto E (2010) Does crime affect economic growth? Kyklos 63(3):330–345
32. Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. p 4171–4186. URL https://doi.org/10.18653/v1/n19-1423
33. Ditton J, Duffy J (1983) Bias in the newspaper reporting of crime news. Brit J Criminol 23:159
34. Elonheimo H (2014) Evidence for the crime drop: survey findings from two finnish cities between 1992 and 2013. J Scand Stud Criminol Crime Prev 15(2):209–217
35. Farrell G, Tseloni A, Mailley J, Tilley N (2011) The crime drop and the security hypothesis. J Res Crime Delinq 48(2):147–175
36. Feng M, Zheng J, Ren J, Hussain A, Li X, Xi Y, Liu Q (2019) Big data analytics and mining for effective visualization and trends forecasting of crime data. IEEE Access 7:106111–106123
37. Francese S (2019) Criminal profiling through maldi ms based technologies-breaking barriers towards border-free forensic science. Aust J Forensic Sci 51(6):623–635
38. Gerber Matthew S (2014) Predicting crime using twitter and kernel density estimation. Decis Support Syst 61:115–125
39. Ghankutkar S, Sarkar N, Gajbhiye P, Yadav S, Kalbande D, Bakereywala N (2019) Modelling machine learning for analysing crime news. In: 2019 International Conference on Advances in Computing, Communication and Control (ICAC3), p 1–5. IEEE
40. Grant PH, Otto PI (2008) The mass media and victims of rape. In Controversies in victimology, p 49–71. Routledge
41. Graves A, Schmidhuber J (2005) Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Netw 18(5–6):602–610
42. Guo B, Ding Y, Yao L, Liang Y, Zhiwen Y (2020) The future of false information detection on social media: new perspectives and trends. ACM Comput Surv (CSUR) 53(4):1–36
43. Hamborg F, Donnay K, Gipp B (2019) Automated identification of media bias in news articles: an interdisciplinary literature review. Int J Digit Libr 20(4):391–415
44. Harnmetta P, Samanchuen T (2022) Sentiment analysis of thai stock reviews using transformer models. In: 2022 19th International Joint Conference on Computer Science and Software Engineering (JCSSE), p 1–6. IEEE
45. Hayward KJ, Maas MM (2021) Artificial intelligence and crime: a primer for criminologists. Crime Media Cult 17(2):209–233
46. Ingilevich V, Ivanov S (2018) Crime rate prediction in the urban environment using social factors. Procedia Comput Sci 136:472–478
47. Jefferson BJ (2018) Predictable policing: predictive crime mapping and geographies of policing and race. Annals Am Assoc Geogr 108(1):1–16
48. Jomnonkwao S, Uttra S, Ratanavaraha V (2020) Forecasting road traffic deaths in Thailand: applications of time-series, curve estimation, multiple linear regression, and path analysis models. Sustainability 12(1):395
49. Kalmegh S (2015) Analysis of weka data mining algorithm reptree, simple cart and randomtree for classification of indian news. Int J Innov Sci Eng Technol 2(2):438–446
50. Kang H, Yang J (2022) Quantifying perceived political bias of newspapers through a document classification technique. J Quant Linguist 29(2):127–150
51. Khan N, Islam Md S, Chowdhury F, Siham AS, Sakib N (2022) Bengali crime news classification based on newspaper headlines

using nlp. In: 2022 25th International Conference on Computer and Information Technology (ICCIT), p 194–199. IEEE
52. Khotimah PH, Arisal A, Rozie AF, Nugraheni E, Riswantini D, Suwarningsih W, Munandar D, Purwarianti A (2023) Monitoring Indonesian online news for covid-19 event detection using deep learning. Int J Electr Comput Eng (2088-8708) 13(1)
53. Kshatri SS, Singh D, Narain B, Bhatia S, Quasim MT, Sinha GR (2021) An empirical analysis of machine learning algorithms for crime prediction using stacked generalization: an ensemble approach. IEEE Access 9:67488–67500
54. Kumar R, Nagpal B (2019) Analysis and prediction of crime patterns using big data. Int J Inf Technol 11:799–805
55. Lee J, Yoon T, Kwon S, Lee J (2019) Model evaluation for forecasting traffic accident severity in rainy seasons using machine learning algorithms: Seoul city study. Appl Sci 10(1):129
56. Leetaru K, Schrodt PA (2013) Gdelt: global data on events, location, and tone, 1979–2012. In: ISA annual convention, vol 2, p 1–49. Citeseer
57. Lewis DD (1995) Evaluating and optimizing autonomous text classification systems. In: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, p 246–254
58. Li J, Zhenjing X, Huijuan X, Tang L, Lean Y (2017) Forecasting oil price trends with sentiment of online news articles. Asia-Pacific J Oper Res 34(02):1740019
59. Li Q, Long W (2018) Do parole abolition and truth-in-sentencing deter violent crimes in virginia? Empir Econ 55:2027–2045
60. Li Q, Tan J, Wang J, Chen H (2020) A multimodal event-driven lstm model for stock prediction using online news. IEEE Trans Knowl Data Eng 33(10):3323–3337
61. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: a robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692
62. Lowphansirikul L, Polpanumas C, Jantrakulchai N, Nutanong S (2021) Wangchanberta: pretraining transformer-based thai language models. arXiv preprint arXiv:2101.09635
63. Magnusson M, Finnäs J, Wallentin L (2016) Finding the news lead in the data haystack: automated local data journalism using crime data. In: Computation+ Journalism Symposium
64. Meier B, Stadelmann T, Stampfli J, Arnold M, Cieliebak M (2017) Fully convolutional neural networks for newspaper article segmentation. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol 1, p 414–419. IEEE
65. Misra R (2022) News category dataset. URL https://www.kaggle.com/datasets/rmisra/news-category-dataset
66. Natarajan M (2016) Crime in developing countries: the contribution of crime science
67. Newman N (2011) Mainstream media and the distribution of news in the age of social media
68. Noraset T, Chatrinan K, Tawichsri T, Thaipisutikul T, Tuarob S (2022) Language-agnostic deep learning framework for automatic monitoring of population-level mental health from social networks. J Biomed Inform 133:104145
69. Pak A, Gannon B (2023) The effect of neighbourhood and spatial crime rates on mental wellbeing. Empir Econ 64(1):99–134
70. Panchenko D, Maksymenko D, Turuta O, Luzan M, Tytarenko S, Turuta O (2022) Ukrainian news corpus as text classification benchmark. In: ICTERI 2021 Workshops: ITER, MROL, RMSEBT, TheRMIT, UNLP 2021, Kherson, Ukraine, September 28–October 2, 2021, Proceedings, Springer, p 550–559
71. Papadopoulos S, Bontcheva K, Jaho E, Lupu M, Castillo C (2016) Overview of the special issue on trust and veracity of information in social media. ACM Trans Inf Syst (TOIS) 34(3):1–5
72. Patrick S, Christa B, Lothar AS (2018) Correlation coefficients: appropriate use and interpretation. Anesth Analg 126(5):1763–1768
73. Pisner DA, Schnyer DM (2020) Support vector machine. In: Machine learning, p 101–121. Elsevier
74. Polpanumas C, Phatthiyaphaibun W (2021) thai2fit: Thai language implementation of ulmfit. URL https://doi.org/10.5281/zenodo.4429691
75. Prateeppornnarong D, Young R (2019) A critique of the internal complaints system of the thai police. Polic Soc 29(1):18–35
76. Prathap BR (2022) Geospatial crime analysis and forecasting with machine learning techniques. In: Artificial intelligence and machine learning for EDGE computing, p 87–102. Elsevier
77. Principe F, van Ours JC (2022) Racial bias in newspaper ratings of professional football players. Eur Econ Rev 141:103980
78. Qazi N, Wong BLW (2019) An interactive human centered data science approach towards crime pattern analysis. Inf Process Manag 56(6):102066
79. Qian Y, Deng X, Ye Q, Ma B, Yuan H (2019) On detecting business event from the headlines and leads of massive online news articles. Inf Process Manag 56(6):102086
80. Rajapakshe C, Balasooriya S, Dayarathna H, Ranaweera N, Walgampaya N, Pemadasa N (2019) Using cnns rnns and machine learning algorithms for real-time crime prediction. In: 2019 International Conference on Advancements in Computing (ICAC), p 310–316. IEEE
81. Rigano C (2019) Using artificial intelligence to address criminal justice needs. Natl Instit Justice J 280(1–10):17
82. Rollo F, Bonisoli G, Po L (2021) Supervised and unsupervised categorization of an imbalanced italian crime news dataset. In: Information Technology for Management: Business and Social Issues: 16th Conference, ISM 2021, and FedCSIS-AIST 2021 Track, Held as Part of FedCSIS 2021, Virtual Event, September 2–5, Extended and Revised Selected Papers, p 117–139. Springer, 2022
83. Rummens A, Snaphaan T, Van de Weghe N, Van den Poel D, Pauwels JRL, Hardyns W (2021) Do mobile phone data provide a better denominator in crime rates and improve spatiotemporal predictions of crime? ISPRS Int J Geo-Inf 10(6):369
84. Saravanan P, Selvaprabu J, Raj LA, Azeez KAA, Sathick KJ (2021) Survey on crime analysis and prediction using data mining and machine learning techniques. In *Advances in Smart Grid Technology: Select Proceedings of PECCON 2019-Volume II*, p 435–448. Springer
85. Schütze H, Manning CD, Raghavan P (2008) Introduction to information retrieval, vol 39. Cambridge University Press Cambridge
86. Seresirikachorn K, Singhanetr P, Soonthornworasiri N, Amornpetchsathaporn A, Theeramunkong T (2022) Characteristics of road traffic mortality and distribution of healthcare resources in Thailand. Sci Rep 12(1):20255
87. Serreli L, Marche C, Nitti M (2023) Global news 60k. https://doi.org/10.21227/vek7-e690
88. Sharkey P, Torrats-Espinosa G (2017) The effect of violent crime on economic mobility. J Urban Econ 102:22–33
89. Srinivasa K, Santhi Thilagam P (2019) Crime base: towards building a knowledge base for crime entities and their relationships from online news papers. Inf Process Manag 56(6):102059
90. Sufi Fahim K, Khalil I (2022) Automated disaster monitoring from social media posts using ai-based location intelligence and

sentiment analysis. IEEE Transactions on Computational Social Systems
91. Sunny Christine M, Nithya S, Sinshi KS, Vinodini V, Aiswaria Lakshmi KG, Anjana S, Manojkumar TK (2018) Forecasting of road accident in Kerala: a case study. In: 2018 International Conference on Data Science and Engineering (ICDSE), p 1–5. IEEE
92. Suphanchaimat R, Sornsrivichai V, Limwattananon S, Thammawijaya P (2019) Economic development and road traffic injuries and fatalities in Thailand: an application of spatial panel data analysis, 2012–2016. BMC Public Health 19(1):1–15
93. Tao H, Zhu X, Duan L, Guo W (2018) Urban crime prediction based on spatio-temporal bayesian model. PloS one 13(10):e0206215
94. Tayal DK, Jain A, Arora S, Agarwal S, Gupta T, Tyagi N (2015) Crime detection and criminal identification in India using data mining techniques. AI Soc 30:117–127
95. Taylor Sean J, Letham B (2018) Forecasting at scale. Am Stat 72(1):37–45
96. Thaipisutikul T, Tuarob S, Pongpaichet S, Amornvatcharapong A, Shih Timothy K (2021) Automated classification of criminal and violent activities in thailand from online news articles. In *2021 13th International Conference on Knowledge and Smart Technology (KST)*, p 170–175. IEEE
97. ToppiReddy HK, Saini B, Mahajan G (2018) Crime prediction & monitoring framework based on spatial analysis. Procedia Comput Sci 132:696–705
98. Tuarob S, Wettayakorn P, Phetchai P, Traivijitkhun S, Lim S, Noraset T, Thaipisutikul T (2021) Davis: a unified solution for data collection, analyzation, and visualization in real-time stock market prediction. Financ Innov 7:1–32
99. Umair A, Sarfraz MS, Ahmad M, Habib U, Ullah MH, Mazzara M (2020) Spatiotemporal analysis of web news archives for crime prediction. Appl Sci 10(22):8220
100. van der Meer Toni GLA, Kroon Anne C, Rens V (2022) Do news media kill? How a biased news reality can overshadow real societal risks, the case of aviation and road traffic accidents. Soc Forces 101(1):506–530
101. Wang Q, Jin G, Zhao X, Feng Y, Huang J (2020) Csan: a neural network benchmark model for crime forecasting in spatio-temporal scale. Knowl-Based Syst 189:105120
102. Zhang H (2004) The optimality of naive bayes. In: Valerie Barr and Zdravko Markov, editors, Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, Miami Beach, Florida, USA, p 562–567. AAAI Press. URL http://www.aaai.org/Library/FLAIRS/2004/flairs04-097.php