Journal of Computer Science

Research Article

A Hybridized BERT-Based Approach for Crime News Collection and Classification from Online Newspapers

Ashour Ali, Shahrul Azman Mohd Noah, Lailatul Qadri Zakaria and Saeed Amer Al Ameri

Centre for Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), Bangi, Selangor, Malaysia

Article history
Received: 30-08-2024
Revised: 13-01-2025
Accepted: 04-02-2025

Corresponding Author:
Shahrul Azman Mohd Noah
Centre for Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), Bangi, Selangor, Malaysia
Email: shahrul@ukm.edu.my

Abstract: Crime news analysis is crucial for understanding criminal activity, enhancing public safety, and informing policy decisions. The exponential growth and unstructured nature of online news articles, however, present significant challenges for efficient and accurate information extraction. This study aims to enhance the efficiency and accuracy of crime news data collection and classification through advanced Natural Language Processing (NLP) techniques and pre-trained language models. We propose a hybridized approach that combines topic modelling, an external knowledge base, and a BERT-based pre-trained model fine-tuned specifically for crime-related content. Our comprehensive experiments demonstrate that this method significantly outperforms existing models, achieving a new state-of-the-art result with a 0.58% increase in accuracy for crime news classification. These findings underscore the practical applicability of our approach in real-world scenarios for improving public safety and crime awareness.

Keywords: BERT, Crime News Classification, Natural Language Processing, Web Scraping, Topic Modeling, Knowledge Bases, Deep Learning, Text Classification, Data Filtering, Online News

Introduction

The proliferation of online news platforms has led to an unprecedented surge in readily available crime-related data. This constantly updated and easily accessible information has become an invaluable resource for analysis (Alameri & Mohd, 2021; Nafea, *et al.*, 2024a; Nafea, *et al.*, 2024b; Prieto Curiel *et al.*, 2020; Rahem & Omar, 2014; Reyes-Ortiz, 2019). Reputable online news outlets offer comprehensive coverage of criminal incidents, drawing from diverse sources such as police reports, court documents, and eyewitness accounts. This wealth of freely available crime-related articles provides researchers with a rich dataset for exploration (Rollo & Po, 2020). In Malaysia, notable news agencies contributing to this information landscape include BERNAMA (Malaysian News Agency) and NST (New Strait Times).

The study of crime patterns has emerged as a critical field within criminology. Practical analysis allows for the identification of recurring trends and the strategic allocation of law enforcement resources. This optimization is crucial not only within individual states but also nationwide, fostering collaboration between various police organizations. Moreover, automated crime data collection and classification facilitates the timely identification and prevention of criminal activity, ensuring responsiveness to the evolving needs of society.

However, the ever-growing volume and complexity of crime-related data present significant challenges. The vast spatial disparities and intricate relationships within this data necessitate advanced collection and classification methods. Historically, information sharing between law enforcement agencies has been inefficient, hindering timely access to crucial data (Valasik, 2024). In recent years, the analysis of crime-related news data has gained importance, with automated collection and classification of online news articles becoming an essential means of shedding light on criminal activities, enhancing public safety, and reducing fear of crime (Ali *et al.*, 2012; Roche *et al.*, 2016).

Despite advancements in data collection and classification methods, managing various crime types remains challenging. This underscores the importance of understanding the relationships between different crimes and how they influence one another (Shu *et al.*, 2017). The primary challenge lies in the unstructured nature of textual data. While rule-based methods and machine-learning approaches have been used for text

classification, the latter demands substantial training data and effort, making crime news classification a relatively underexplored domain (Gasparetto *et al.*, 2022; Hissah & Al-Saif, 2018; Hsu, 2020). Additionally, the complexity of the task extends beyond mere event counting, requiring a thorough examination of various criminogenic factors influencing criminal behavior (Ali *et al.*, 2022; 2012; Mohd *et al.*, 2012; Mohd & Mohamad Ali, 2011). An innovative model known as Bidirectional Encoder Representation from Transformers (BERT) has shown promise by pre-training comprehensive bidirectional representations from unlabeled text (Devlin *et al.*, 2019; González-Carvajal & Garrido-Merchán, 2020). BERT emerged as an alternative to existing language models like BiLM (ELMo) (Peters *et al.*, 2018) and Fine-tuned Transformer LM (Radford *et al.*, 2018), both of which utilized bidirectional models for acquiring general text representations. BERT undergoes training on plain text to execute two tasks: predicting masked words and foreseeing the following sentence. While training BERT from the ground up demands a substantial dataset and significant training time (Nugroho *et al.*, 2021), applying BERT can potentially address the challenges associated with crime news classification.

Most existing work uses unsupervised and supervised approaches to address the problem, incorporating text classification algorithms. However, these techniques often fall short in the specific domain of crime news classification due to their generic approach, which lacks the incorporation of crime-specific attributes necessary for enhanced precision and systematic accuracy, particularly with the increasing volume of data. Moreover, there is currently a scarcity of comprehensive knowledge bases on crime, especially those that provide a detailed understanding of crime documents. The lack of sufficient knowledge impedes the ability of many end-users to prevent and predict crime effectively. There is a pressing need for an advanced classification model that can efficiently categorize crime data. This model should utilize topic modeling, external knowledge base representations, and fine-tuned BERT applied to current and archived data from online newspapers, ensuring accurate affirmation and prediction of crime-related news.

This research makes several significant contributions to the field of crime news analysis and classification. First, unlike other works, we introduce a hybrid approach that leverages the BERT model's capabilities alongside external knowledge bases to create a more robust system for crime news processing. Our approach effectively addresses key challenges in collecting and classifying various types of crime-related content from online Malaysian news sources, specifically BERNAMA and NST. The research advances the field through its innovative integration of multiple technologies. By combining natural language processing techniques, external knowledge bases (DBpedia and Wikidata), and semantic similarity analysis, our approach reduces cognitive learning requirements while enhancing classification accuracy. The system demonstrates notable versatility, being adaptable to various datasets beyond crime-related content when appropriate pre-processing is applied. Importantly, this research establishes a framework that can be extended and adapted for similar classification tasks in other domains, making it a valuable contribution to the broader field of text classification and analysis.

Related Work

Text collection and classification has a rich history with diverse methodologies, yet research specifically focused on the classification of crime-related news remains relatively limited. This section provides a comprehensive overview of recent and prevalent approaches employed for collecting and classifying crime-related news data.

Collection Methods

Despite the extensive development of text collection methods, approaches explicitly tailored for gathering crime news are scarce. Several strategies have been adopted, each with its own set of advantages and limitations.

Web Scraping: Researchers frequently utilize web scraping tools, both generic and specialized (Christian *et al.*, 2022; Maybir & Chapman, 2021), alongside Natural Language Processing (NLP) libraries such as Beautiful Soup, Scrapy, and Selenium (Arumi & Sukmasetya, 2020; Kynabay *et al.*, 2021). Machine learning techniques have also been incorporated to extract relevant information from scraped data (Gopal *et al.*, 2020; Srinivasa *et al.*, 2019).

Keyword-Based Scraping: Approaches like those employed by Thielmann *et al.* (2023) rely on specific keywords to identify and scrape relevant documents. While this method offers simplicity, it can be constrained by the availability and diversity of relevant keywords, potentially leading to incomplete or biased data, especially when dealing with complex topics.

Topic Modeling-Assisted Scraping: Recent work has explored the integration of topic modelling to refine web scraping processes. Srinivasa *et al.* (2019) demonstrated the use of topic modelling to improve the identification of relevant articles and filter out unrelated content. However, challenges remain in effectively excluding false negatives, highlighting the need for further refinement and the potential incorporation of external knowledge bases.

Classification Methods

Recent research in crime news classification has witnessed significant advancements through deep

learning and hybrid approaches, yielding promising results across multiple languages and methodologies.

Transformer-based approaches have shown particular promise in this domain. Hossain *et al.* (2023) developed an innovative transformer network-based deep learning system for analyzing Bengali crime news articles. Their methodology employed BERT's zero-shot classification capabilities for both binary and multiclass classification, complemented by Named Entity Recognition (NER) for geographic location identification in drug-related content. Their system achieved remarkable results, with binary classification reaching 96% precision, 92% recall, and 94% F1-scores, while multiclass classification performed even better at 95% precision, 98% recall, and 97% F1-scores. Though highly effective, their approach was constrained by limited labeled data availability.

LSTM-based architectures have also demonstrated strong performance in crime news classification. Vidal *et al.* (2020) conducted a comparative analysis of Bi-LSTM, Attention Bi-LSTM, and BERT models for Spanish crime news classification, with BERT achieving superior accuracy at 99.18%. Building on this foundation, Deepak *et al.* (2021) and Vidal *et al.* (2021) further validated the effectiveness of Bi-LSTM networks for multi-label crime classification, achieving accuracies of 96.55% and 98.87% respectively.

Hybrid approaches combining multiple neural network architectures have emerged as another effective strategy. Saha *et al.* (2020) and Wang *et al.* (2020) developed systems integrating Convolutional Neural Networks (CNNs) with Bi-directional Long Short-Term Memory networks. While Saha *et al.* focused on crime pattern recognition and statistical analysis, Wang *et al.* enhanced their Chinese crime news classification system with an attention mechanism.

Additional methodological innovations include Singh Bhati *et al.* (2019) application of CNNs for crime analysis and prediction, and Sundhara Kumar and Bhalaji (2020) integration of Recurrent Neural Networks with Extreme Learning Machines, achieving 92.30% accuracy in crime hotspot classification.

Despite these advances, significant challenges remain in the field. Current research often focuses on limited category sets, with attempts to expand coverage by Deepak *et al.* (2021) and Saha *et al.* (2020) revealing accuracy trade-offs as category scope increases. Additionally, while transformer-based models like BERT have shown exceptional results in specific languages, as demonstrated by Vidal *et al.* (2021), their effectiveness across different languages and multilingual contexts requires further investigation.

Methodology

This research follows our previous work (Ali *et al.*, 2023) published as a preprint and employs a two-phased methodology for the comprehensive analysis of online crime news articles. The initial phase focuses on data acquisition and refinement, introducing an enhanced approach for the collection and filtering of crime news articles from diverse online sources. Subsequently, the second phase delves into the fine tuning and evaluation of BERT model, presenting a robust method for fine-tuning and testing the BERT model specifically designed for crime news classification. Following successful model fine-tuning, the entirety of the collected data from the first phase undergoes classification using the fine-tuned BERT model. This sequential framework ensures a rigorous and systematic approach to crime news analysis, facilitating both the creation of a high-quality dataset and the development of a reliable classification model.

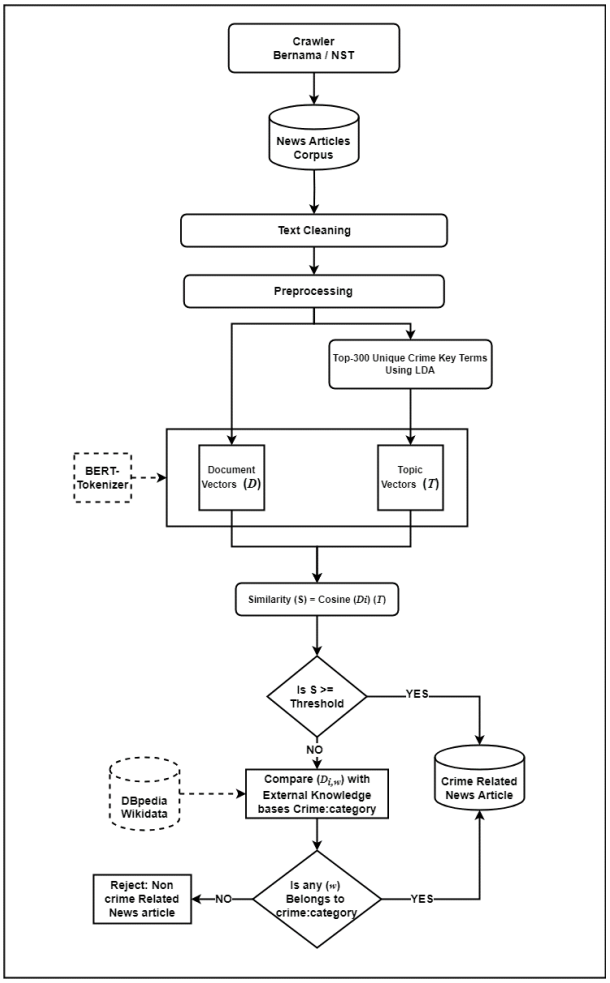

Fig. 1: Collection and filtering of news articles workflow

Multi-Stage Approach to Collection and Filtering of Crime News Articles

In this process, the approach introduces a multi-stage method for collecting and refining online crime news articles. Drawing inspiration from the work of Srinivasa *et al.* (2019), the approach integrates web crawler, text pre-processing, document and topic representation, similarity measures, and external knowledge base verification to ensure a comprehensive and accurate

dataset. The process is shown in Fig. 1, and each stage is discussed in detail below.

Web Crawler

The first stage involves the development of a web crawler specifically designed to extract crime-related news articles from online sources. The crawler will leverage optimizations inspired by previous studies (Boldi *et al.*, 2004; Heydon & Najork, 1999; Srinivasa *et al.*, 2019) to minimize the load on the target websites while ensuring effective data collection. The data collection process encompasses the following steps:

- Specifying Target URLs: The crawler is configured to target specific URLs from relevant news sources, the BERNAMA and New Straits Times, for a defined timeframe (from February 2017 to October 2023).
- Headline-based Sub-URL Selection: The crawler identifies and collects sub-URLs of potential news articles based on the presence of crime-related keywords within the headlines.
- Full-Text Scraping: Upon identifying a relevant sub-URL, the crawler extracts the complete text content of the corresponding news article.
- Extracted Text Storage: The collected text data is systematically stored for subsequent processing and analysis.

Text Pre-Processing

The Natural Language Toolkit (NLTK) will serve as the primary platform for this task. NLTK offers a comprehensive suite of tools for working with natural human language data within the Python programming environment. The raw collected text data undergoes a series of pre-processing steps using NLTK to prepare it for subsequent analysis. These steps encompass text cleaning, which involves the removal of irrelevant information such as HTML tags, special characters, and punctuation; tokenization, which consists of the segmentation of text into individual words or sub-words; normalization, which involves the conversion of text to a standardized format (e.g., lowercase); duplicate removal: which is used for eliminating of redundant data points; and lemmatization, which is responsible for reducing words to their base or root form.

Document Vector Representation

The document vector representation approach implements a document vectorization technique using the BERT Tokenizer model, as illustrated in Fig. 2. The process begins with a carefully cleaned and pre-processed dataset of 22,006 news articles.

The articles are individually broken down into their constituent words or tokens, creating a list of words for each document. Each document (D1, D2, D3, ..., Dn) is then represented as an array of words (w1, w2, ..., wn). The BERT Tokenizer maps these words to a fixed-size vector of 300 dimensions, creating a dense numerical representation for each document. This representation captures the semantic meaning and importance of words in their context. To ensure consistent vector sizes, documents with fewer than 300 words have their remaining entries filled with zeros, while those exceeding 300 words are truncated. The resulting uniform vector size allows for efficient comparison and analysis of documents based on their content.

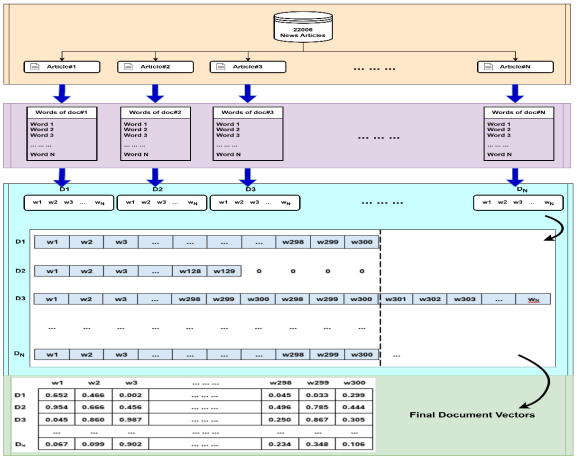

Fig. 2: Document Vectorization Process for News Articles

Unified Crime Key-Terms Vector Representation

This process aims to discover underlying topics within the crime news corpus and extract crime-related key terms using Latent Dirichlet Allocation (LDA), introduced by Blei *et al.* (2003). The method creates a unified vector representation across all topics, ensuring uniqueness and representativeness of crime key terms. As shown in Fig. 3, the process involves the steps below.

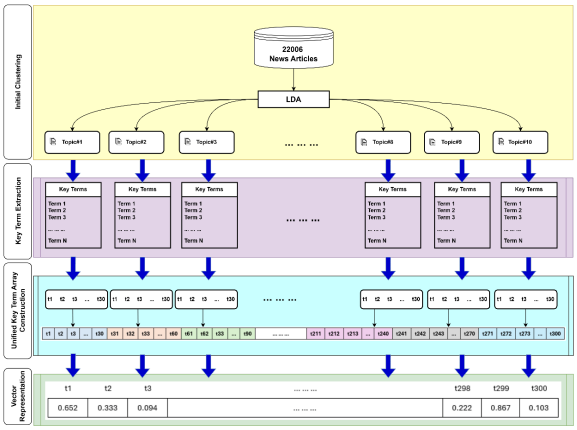

Fig. 3: Unified Crime Key-Terms Vector Representation

Step 1 (Initial Clustering): All 22k documents have been clustered into 10 predefined topics using LDA. Each topic will contain a different number of documents based on the clustering results.

Step 2 (Key Term Extraction): For each topic, key terms are extracted from the documents classified under

it. This results in topic-specific key term sets, which may overlap across topics.

Step 3 (Unified Key Term Array Construction): A unified array of 300 unique key terms has been built using the following procedure:

1. Start with Topic 10: Extract the top 30 key terms and add them to the array
2. For Topics 9 to 1 (in descending order to ensure that topics with a higher number of key terms contribute more significantly to the final representation, while still allowing each topic to add unique terms)
3. Extract the top N key terms (where N is sufficient to find 30 unique terms)
4. Add unique terms to the array, avoiding duplicates (redundant)
5. Continue until 30 new unique terms are added or all top N terms are processed
6. Repeat this process for each topic until the array contains 300 unique key terms

Step 4 (Vector Representation): The final 300 unique key terms in the array are used to create a vector representation with a length of 300 using BERT tokenizer.

Following these steps ensures a comprehensive and unique vector representation by giving priority to topics with a greater number of key terms, while also preserving contributions from all topics. This strategy maximizes the ratio between the concentration on certain topics and cross-topic uniqueness in the final representation. It prioritizes topics with a larger number of key terms while ensuring each topic contributes to the final representation

Similarity Measures

The cosine similarity between each document vectors ($\vec{D}_i$) and crime key-terms vector ($\vec{\mathrm{T}}$) as shown in Fig. 4 will be computed.

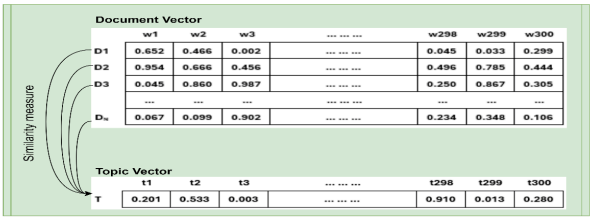

Fig. 4: Similarity between document vector and topic vector

Cosine similarity, as shown in Equation (1), measures the directional similarity between two vectors. Cosine similarity is optimal for comparing document and crime key-term vectors, as it measures semantic similarity while normalizing for length and works well in high-dimensional spaces.

$$\text{cosine similarity}\left(\vec{D}_i, \vec{T}\right) = \frac{\vec{D}_i \cdot \vec{T}}{\|\vec{D}_i\| \, \|\vec{T}\|} \quad (1)$$

where ($\vec{D}_i$) is the vector representation of the news article, ($\vec{\mathrm{T}}$) is the vector of the crime key-terms, ($\cdot\cdot$) represents the dot product, and ($\|\vec{D}_i\| \, \|\vec{T}\|$) denotes the norm of vectors $\vec{D}_i$ and $\vec{T}$.

This calculation yields a single scalar value representing the overall semantic similarity of the news article to the crime-related benchmarked key-terms. Higher values indicate stronger relevance to crime-related content. The cosine similarity score for each article will then be expressed as a percentage. Articles exceeding a predefined threshold (50%) will be classified as crime related.

Integrating DBpedia and Wikidata

The previous stages of the methodology effectively capture a significant portion of crime-related news articles. However, this approach may miss relevant articles with low similarity scores to the key-term vector. To address this limitation and enhance dataset comprehensiveness, the proposed method incorporates external knowledge bases, specifically DBpedia (Morsey *et al.*, 2012) and Wikidata (Vrandecic & Krotzsch, 2014). These knowledge bases are leveraged due to their extensive collections of structured information encompassing a vast array of entities and concepts. They provide rich semantic connections and categorizations that can be utilized to validate the thematic relevance of news articles to the crime domain. This additional layer of analysis offers a complementary perspective to the initial vector similarity approach.

The integration process entails querying DBpedia and Wikidata using the words represented in articles that fell below the keyword-topic similarity threshold in the previous stage. The system analyses these keywords thematic classifications and relationships within the knowledge bases to determine their association with crime-related concepts. Articles containing keywords demonstrably classified within crime categories in either knowledge base are subsequently reclassified as relevant to the crime domain. This additional step helps to capture crime-related articles that may use unconventional terminology or discuss emerging crime trends not yet reflected in the main key-term vector. It also helps to mitigate potential biases in the initial vector representation, ensuring a more comprehensive and nuanced crime news dataset.

BERT-Based Crime News Classification

As discussed early in (Ali *et al.*, 2023) after successfully collecting and filtering the online news articles, we propose a robust method to classify them according to the crime events reported in them. Due to the nature and complexity of online data on crime news articles, it is evident that the proposed model must have different aspects to predict and classify different crime events precisely and accurately. Fig. 5 illustrates the end-

to-end process used in this study for crime news classification using a BERT-based model.

The core of this research involves utilizing a BERT-based model to classify crime news articles. BERT, or Bidirectional Encoder Representations from Transformers, excels at natural language processing NLP tasks due to its ability to capture contextual information from text data. The proposed approach takes advantage of transfer learning by fine-tuning a pre-trained BERT model on a specific dataset of crime news articles. This allows the model to learn the nuanced language and terminology associated with different crime types and effectively categorize unseen news articles.

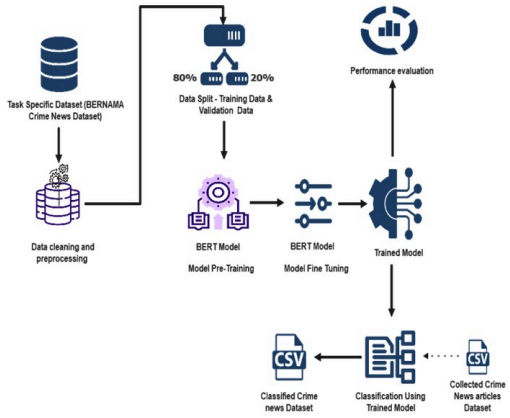

Fig. 5: Training and Fine-Tuning workflow for the BERT Model

The BERT model employed in this research is a deep neural network architecture built upon the transformer encoder. It consists of 12 transformer blocks, each with 12 self-attention heads and a hidden size of 768. The model accepts input sequences of up to 512 tokens and generates representations for each token.

The classification process involves adding special tokens ([CLS] and [SEP]) to the input sequence, tokenizing the text, and generating word embeddings. The pre-trained BERT model then processes the sequence and produces an output layer at the token level, including a class label at the [CLS] position. For multi-class classification, the final hidden state of the [CLS] token is used as a representation of the entire sequence, and a softmax classifier predicts the probability of each label.

There are two key steps in utilizing BERT for classification: pretraining and fine-tuning. The model is initially pre-trained on a massive unlabeled corpus, followed by fine-tuning on the labeled CriNED dataset to adapt it to the specific task of crime news classification

Experiments Setup

We evaluate BERT on CriNED dataset and compare it with existing crime news classification models.

CriNED: Crime News Event Dataset Description

The research utilizes a Crime News Event Dataset (CriNED) sourced from the Malaysian News Agency (BERNAMA) containing 2304 crime news articles. Each article is manually annotated with one of 7 distinct crime categories: Violence Crime, Financial Crime, Illegal Activity, Social or Political Crime, Property Crime, Drug Offense, and Cyber Crime. Then each crime is annotated with its specific sub-category. Fig. 6 provides a visual representation of the class distribution within the dataset. As observed, the dataset exhibits some degree of class imbalance, with certain crime categories being more prevalent than others. To address this issue, a class weighting method was implemented during model fine tuning, assigning higher weights to minority classes to prevent the model from being biased towards majority classes.

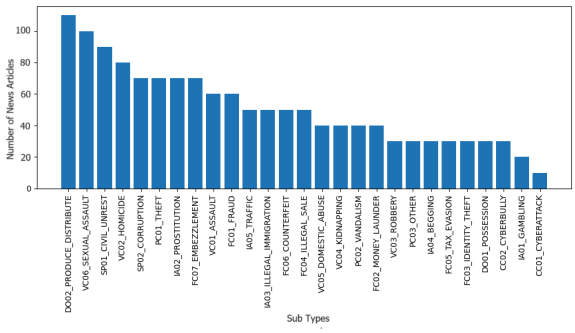

Fig. 6: Distribution of Crime Sub-Categories in CriNED dataset

The CriNED dataset was preprocessed by removing line breaks, consolidating text files, and creating a single CSV file with relevant attributes (ID, Title, Text, Types, Sub-Types).

Implementation of Collection Stage

Implementation of Web Crawler

The initial stage involves the development of a web crawler designed to extract crime-related news articles from online sources efficiently. The process involves specifying target URLs (BERNAMA, New Straits Times) ranging from February 2017 to October 2023, selecting article sub-URLs based on headlines, scraping the full text of articles, and storing the extracted text for further processing.

The web crawler is implemented by utilizing the Python packages BeautifulSoup4 and Selenium. BeautifulSoup4 is a powerful tool for parsing HTML and XML documents, even those with errors, making it ideal for extracting data from websites. Selenium complements this by automating web browsers, enabling the crawler to interact with dynamic websites and gather information that would otherwise be inaccessible through static parsing alone.

The website for both newspapers is crawled and the total number of collected news articles is 22,006 news articles from Bernama and NST published over the past seven years mainly in crime-related categories. These articles' publication dates range from February 2017 to October 2023. Fig. 7 presents the distributionof articles over time.

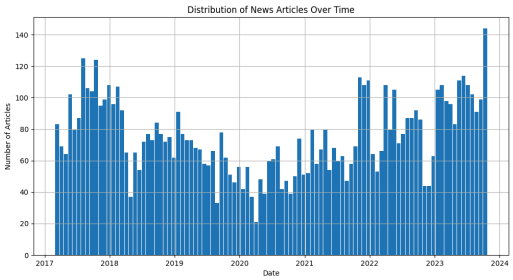

Fig. 7: Distribution of Collected News Articles Over Time

Text Pre-processing

The initial step involves preprocessing the content of the "article_text" column to enhance their suitability for analysis and to yield reliable results. This is achieved by applying a regular expression and using NLTK tool kit to remove punctuation and by converting all text to lowercase. Subsequently, each sentence is tokenized into a list of words, with punctuation and extraneous characters being removed.

To capture meaningful multi-word expressions, bigrams and trigrams are identified. Bigrams are pairs of words that frequently occur together, while trigrams consist of three words. Examples include terms like "Seri Kembangan", "Tan Sri", and "Dato' Seri". Gensim's Phrases model facilitates the creation and implementation of bigrams, trigrams, quad grams, and more. The key parameters for this model are min_count = 5 and threshold = 100, which help in refining the phrase detection process.

With the phrase models prepared, functions are defined to remove stopwords, generate trigrams, and perform lemmatization. These functions are executed sequentially to further clean and standardize the text data.

Unified Crime Key-Terms Vector Representation

This study employed the Latent Dirichlet Allocation (LDA) for topic discovery within a document collection. The experiment was setup according to the Algorithm 1.

The primary inputs to the LDA topic model are the dictionary (id2word) and the corpus. The dictionary serves the crucial function of mapping each unique word to a unique ID, ensuring that every term in the dataset has a distinct identifier. The corpus, on the other hand, represents the documents in a bag-of-words format, listing word IDs alongside their respective frequencies. Together, these components form the foundation for the topic modelling process, enabling the analysis and extraction of thematic structures from the textual data.

Algorithm 1: Extraction of Top Crime-Related Key Terms Using LDA

Input: List of news articles

Output: Top crime key terms

Procedure: ExtractLDATopics(texts **D**, num_topics **T**, num_words **W**)

1. **Create** dictionary and corpus from **D**
2. **Preprocess** texts (tokenize, remove stopwords, lemmatize/stem, optional bigrams)
3. **Train** LDA model with **T** topics on the corpus
4. **Return** top **W** words for each topic from the trained LDA model
5. **Post-process** topic words (filter non-crime terms, merge duplicates, rank by relevance)

End Procedure

Table 1: LDA Key Hyperparameters

Hyperparameters	Description
Number of Topics	Determines the number of distinct topics (K) the model will identify.
Alpha (α)	A Dirichlet hyperparameter that affects document-topic density
Beta (β)	A Dirichlet hyperparameter that influences word-topic density.

Training the baseline LDA model involves using the prepared corpus and dictionary. Table 1 details the key hyperparameters crucial to the configuration of the model. For the base model, the Alpha (α) and Beta (β) are set to 1.0/num_topics, reflecting their default values. The chunksize parameter, which affects the number of documents processed at one time, is configured to optimize training speed while considering memory constraints. Additionally, the passes parameter, set to 10, dictates the number of times the model iterates over the entire corpus, while the iterations parameter controls the number of iterations over each document within each pass.

The performance of the LDA model is evaluated by training it with an initial configuration of 10 topics. Each topic is represented by a combination of keywords, where each keyword has a specific weight contributing to the topic. To assess the model's efficacy, metrics such as perplexity and coherence score are computed. These metrics provide insights into the model's interpretability and the quality of the topics generated, with coherence score particularly emphasizing the comprehensibility of the topics. To optimize the LDA model, sensitivity tests are conducted on key hyperparameters: Number of Topics (K), Alpha (α), and Beta (β). Each parameter undergoes individual testing while the others remain constant, allowing for a focused analysis of its impact on model performance. These tests are performed on two

different validation corpus sets (100 and 75%), with the coherence score serving as the performance metric. The top 10 coherence score values shown in Table 2 guide the final model configuration, leading to the selection of hyperparameters that yield the highest coherence score for the trained model.

Table 2: LDA Model Top 10 Coherence Score

Validation_Set	Topics	Alpha	Beta	Coherence
100% Corpus	10	0.001	0.98	0.68
75% Corpus	8	0.61	0.91	0.65
75% Corpus	8	symmetric	0.61	0.65
75% Corpus	8	asymmetric	0.61	0.65
100% Corpus	5	0.01	0.91	0.65
100% Corpus	9	symmetric	0.91	0.64
100% Corpus	8	0.01	0.91	0.64
75% Corpus	9	symmetric	0.91	0.64
100% Corpus	5	0.01	0.61	0.64
75% Corpus	7	symmetric	0.31	0.64

The final LDA model is fine-tuned with the following parameters shown in Table 3.

Table 3: Final LDA Model Training Parameters

Parameter	Values
Number of Topics (k)	10
Random State	100
Chunk size	100
Passes	10
Alpha (α)	0.001
Beta (β)	1.0

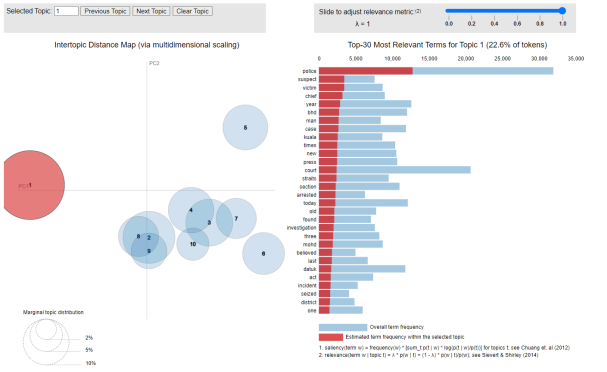

Fig. 8: Intertopic distance map and the top 30 most relevant terms for topic 1

The fine-tuned model is implemented using Gensim's LdaMulticore function. This comprehensive approach ensures that the model is robust and optimized for extracting meaningful topics from crime news articles. With the number of topics (k) determined as 10, the analysis focuses on the top 30 most relevant terms per topic, as illustrated in Fig. 8 This results in a collection of 300 (30 terms/topic * 10 topics) highly relevant crime terms across all topics, which are utilized as a benchmark for the subsequent analysis.

Fig. 8 presents a topic model analysis of 22,006 crime news articles. The Intertopic Distance Map on the left uses multidimensional scaling to display topic relationships, with Topic 1 selected and emphasized. Circle sizes represent topic prevalence. The right side features a bar chart of the 30 most relevant terms for Topic 1, including common crime reporting words like "police," "report," and "victim." Each term displays its overall corpus frequency alongside its estimated frequency within Topic 1. The interface allows for topic selection and adjustment, enabling interactive exploration of thematic content throughout the crime news corpus.

Similarity Measures

As detailed in the Methodology section, this study proposes a semantic-based method to assess the relevance of news articles to the crime domain. The approach utilizes a comparative analysis of semantic similarity between the document vector representing the news article and the unified crime key-terms identified through the LDA model. Cosine similarity, a measure of relatedness between words based on shared meaning and context, is leveraged to quantify the degree of alignment between article content and crime-related themes.

Algorithm 2: Vector Representation of the Documents

Input: List of words for each document (**D**)

Output: Document vector

Procedure: DocumentVectorization(D)

1. **Initialize** BERT tokenizer and model
2. **For each** word **w** in **D**, call **GetWordVector(w)**
3. **Aggregate** all word vectors (e.g., average or weighted mean)
4. **Return** the resulting document vector

FUNCTION: GetWordVector(word **w**)

1. **Tokenizew** with BERT tokenizer
2. **Feed** tokenized input to BERT model
3. **Extract** embeddings from the final hidden layer
4. **Return** the averaged embedding vector for **w**

End Function

End Procedure

The implementation of this approach is delineated in Algorithms 2, 3 and 4 which detail the specific steps for vector representation of documents, vector representation of LDA-extracted key-terms, and the calculation of cosine similarity, respectively. These similarity measures are designed to assess the relatedness between article words and the top crime key terms identified using LDA.

Table 4 presents the similarity scores (%) between document vector and top crime key terms extracted by LDA for ten different articles. The similarity scores range from 50.05 to 75.34%. These percentages indicate varying degrees of alignment between the documents and the LDA-extracted crime key terms across the different

articles, with most articles showing moderate to high levels of similarity.

Algorithm 3: Representation of Top Crime Key Terms Extracted by LDA

Input: List of news articles (**D**), Number of topics **T** (10 in this case), Number of unique terms required (**U = 300**)

Output: Unified vector representation of length 300

Procedure: ExtractAndRepresentCrimeTerms(D, T, U)

1. **Create** dictionary and corpus from **D**
2. **Train** LDA model with **T** topics
3. **Cluster** documents into **T** topics
4. **Initialize unified_terms = []**
5. **FOR** each topic in range(**T**, 0, -1):
   a. **topic_terms** = ExtractTopKeyTerms(topic, **N**) // **N** is sufficiently large
   b. **FOR** each term in **topic_terms**:
      i. **IF** term not in **unified_terms**:
         - **Add** term to **unified_terms**
      ii. **IF** length of **unified_terms == U**:
         - **Break**
6. **Return** VectorizeTerms(**unified_terms**)

End Procedure

Algorithm 4: Calculating Cosine Similarity

Input: Articles - List of news articles

Output: RelatedArticles - List of articles classified as crime-related

Procedure: CalculateCosineSimilarity(Document_vectors D, Topic_vector T)

1. **Initialize** similarity_scores **S = []**
2. **FOR** each $\mathbf{D_i}$ in **Document_vectors**:
   a. **Compute** similarity = ($\mathbf{D_i}$ â‹… **T**) / (||$\mathbf{D_i}$|| * ||**T**||)
   b. **Add** similarity to **S**
3. **Return** S

Procedure: CalculateOverallSimilarity(S, threshold Ï„)

4. **Compute** is_crime_related = (S â‰¥ Ï„)
5. **Return** overall_similarity, is_crime_related

End Procedure

Table 4: Semantic Similarity Scores: Crime news articles and Crime-Related Terms Extracted using LDA

Article ID	Overall Similarity %
216035	75.34
216041	59.14
216054	62.30
216062	50.05
216066	66.45
216068	54.77
216119	69.59
216137	69.24
216141	52.72
216149	65.35

Baseline Models Implementation

We benchmark the performance of the recent crime news classification models on the CriNED dataset, including Bidirectional LSTM and Attention Bi-LSTM models proposed by Vidal *et al.* (2021), and Attention-based CNN-Bi-LSTM model introduced by Saha *et al.* (2020) and Wang *et al.* (2020).

Implementation of Bi-LSTM Model

This model is also employed, particularly to leverage its capability to capture long-range dependencies in text data. This model is especially useful in maintaining the contextual information of both past and future tokens, which is pivotal for accurately classifying lengthy or complex article structures. The training of Bi-LSTM involved multiple epochs with an emphasis on balancing the training loss across different crime categories to prevent biases. The Bi-LSTM model was implemented using Keras with TensorFlow and trained to minimize categorical cross-entropy loss using the ADAM optimizer. The training process employed a batch size of 64 over 50 epochs. To determine the optimal number of units for the hidden layers and the appropriate dropout rate, we performed a Random Search. The architecture of Bi-LSTM model as follows:

LSTM and Fully Connected Layers: The network configuration includes 190 units for both the LSTM and fully connected layers.

Regularization and Dropout: The L2 regularization applied is on the order of 10−3, and the dropout rate was set to 0.15.

Embedding Layer: For the embedding layer, we utilized 300-dimensional word vectors.

Implementation of Attention Bidirectional LSTM

This network is optimized to minimize binary cross-entropy loss using the ADAM optimizer with a batch size of 32 over 60 epochs. We followed the authors and conducted a Random Search to determine the optimal number of units for the LSTM hidden layers and the dropout rate. As a result, we configured the LSTM with 180 units and set the dropout rate to 0.25. The L2 regularization applied is on the order of 10−3. For the embedding layer, we also follow the authors and utilize 300-dimensional word vectors, which are also used in the bidirectional LSTM.

Implementation of Attention-Based CNN-Bi-LSTM Model

The implementation of the Attention-based CNN-Bi-LSTM model in this study follows the original work steps, which involve encoding each word in sentences as 100-dimensional vectors, which are first processed by a CNN layer to extract local n-gram features. These features are fed into a Bidirectional Long Short-Term

Memory (Bi-LSTM) layer, which reads sequences forward and backwards to capture contextual dependencies. An attention-pooling layer follows, enabling the model to focus on critical linguistic elements without increasing computational complexity. As per Saha *et al.* (2020) This architecture allows for effectively classifying crime indicators by focusing on relevant words in context. The final classification is achieved through a fully connected layer with a sigmoid activation function. Extracted indicators are resolved to unique identifiers for indexing and linking documents, forming dynamic threads based on overlapping crime indicators.

Fine Tuning the BERT Model

The pre-trained BERT model was fine-tuned on the CriNED dataset with specific hyperparameters aas shown in Table 5, including the number of epochs, batch size, optimizer, learning rate, loss function, and maximum sequence length. All fine-tuning and evaluation hyperparameters performed for this model used the architecture of the TF Hub BERT collection.

Table 5: Hyperparameters of the Fine-Tuned BERT Model

Hyperparameter	Values
No of Epochs	25
Batch Size	16
Optimizer	Adam
Learning Rate	1e-5
Optimizer decay	1e-6
Loss Function	Categorical Cross Entropy
Maximum Sequence Length	256

The BERT model's fine-tuning process begins with the integration of [CLS] tokens at the start of each input sequence, crucial for classification tasks. Texts are tokenized, and embeddings are formed by aggregating token and segment embeddings. During fine-tuning, we adjust both the BERT model parameters, and a softmax classifier layer is added on top, which computes the probability distribution over the class labels. The fine-tuning employs a low learning rate (1e-5) to adapt the pre-trained model to our specific task without overfitting.

To simplify, an epoch comprises one or more batches. In our model, we partitioned our dataset into 80% training data and 20% test data. The training dataset was then segmented into a batch size of 16 batches, with each batch passing 144 articles to the model.

The Optimizer is a pivotal parameter responsible for determining the adjustments to weight and learning rate to minimize model losses. In the context of multi-class categorization, Adam stands out as one of several variants of the Adam Optimizer, utilizing a parameter known as optimizer decay.

The Learning Rate governs how swiftly the model learns in response to errors. A higher learning rate may transition the model from underfitting to overfitting. In our model, we set the learning rate to 1e-5 to strike a balance and prevent both underfitting and overfitting.

The Loss Function is another parameter influencing the type of function applied to calculate the model's loss during training and validation.

Lastly, the Maximum Sequence Length parameter defines the maximum size of the token sequence accepted as input by the model.

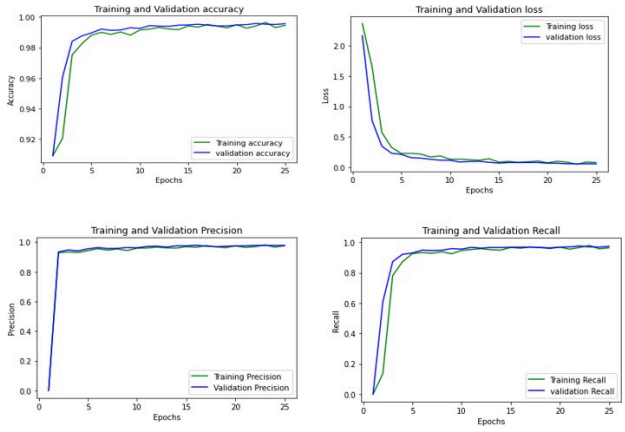

Fig. 9: BERT Model Learning Curves for the Training and Validation Accuracy, Loss, Precision, and Recall

Performance Analysis: Learning curves were generated to analyze the training and validation performance of the models. Fig. 9 specifically presents the learning curves for the BERT model, illustrating its accuracy, loss, recall, and precision over epochs. Each learning curve plot includes one curve representing the model's performance on the training dataset and another for the validation dataset. Notably, the training loss consistently remained lower than the validation loss throughout training, indicative of a generalization gap between the two. This gap signifies the model's ability to generalize beyond the training data. Importantly, the curves demonstrate that our model exhibits a favorable fitting pattern, as the training loss steadily decreases to a stable point, and the validation loss follows suit, maintaining a small gap with the training loss.

Results and Discussion

Collection and Filtering of Crime News Articles

The initial analysis of the 22,006 news articles collected revealed an average similarity score of 61.58% between documents and the top crime key terms identified using LDA for each news article. When implementing a similarity threshold, the dataset is refined and focused on the most unique and informative articles. Setting a similarity threshold of 50% reduced the number of articles to 19,545, ensuring only those with a substantial degree of uniqueness were included. The percentage of the rejected articles in this stage is 11.18%, representing 2461 news articles and highlighting the importance of establishing an appropriate threshold

to maintain a balance between data comprehensiveness and relevance.

To further enhance the accuracy of crime-related article selection, the proposed method integrated external knowledge sources (DBpedia and Wikidata) by querying these knowledge bases using all words from articles that fell below the document-topic similarity threshold (i.e. the 2461 news articles). While traditional methods relying solely on keyword identification often led to the inclusion of irrelevant articles (false positives), the combined approach offers a more nuanced understanding of context. Similarity scores help decipher the thematic context of crime-related terms, while external knowledge sources provide additional information to differentiate between truly relevant and irrelevant articles. This ultimately leads to a more precise and reliable dataset for analysis.

Table 6 demonstrates the effectiveness of this multi-faceted approach by analysing examples of six articles and their similarity scores. The selected articles represent three distinct cases:

- Two articles were accepted as crime-related news due to high overall similarity scores.
- Two articles with overall similarity were below the threshold but were accepted due to supporting information from external knowledge bases.
- Two articles fell below the threshold and lacked supporting information from external bases.

Table 6: Evaluation of the Results of Crime News Articles Collection and Filtering

ID	Source	Title	URL	Average Similarity %	Identification in DBpedia and Wikidata	Status
333472	NST	EXCLUSIVE: VASANTHAPIRIYA SUICIDE CASE: INVESTIGATION PAPER INCOMPLETE	https://www.nst.com.my/news/crime-courts/2018/02/333472/exclusive-vasanthapiriya-suicide-case-investigation-paper	84.06	-	Accepted
216054	NST	LAWAS FAMILY ARRESTED FOR SUSPECTED DRUG TRAFFICKING	https://www.nst.com.my/news/2017/02/216054/lawas-family-arrested-suspected-drug-trafficking	64.53	-	Accepted
500976	NST	TRIO FREED OF KIDNAPPING FEMALE DOCTOR IN MUAR	https://www.nst.com.my/news/crime-courts/2019/07/500976/trio-freed-kidnapping-female-doctor-muar	43.15	https://commons.wikimedia.org/wiki/Category:Kidnapping https://dbpedia.org/page/Kidnapping	Accepted
227418	NST	HIGH COURT UPHOLDS DECISION TO JAIL, FINE YOUTH IN LOW YAT SMARTPHONE THEFT CASE	https://www.nst.com.my/news/2017/04/227418/high-court-upholds-decision-jail-fine-youth-low-yat-smartphone-theft-case	39.99	https://commons.wikimedia.org/wiki/Category:Theft https://dbpedia.org/page/Category:Theft	Accepted
307472	NST	24 POLICEMEN, OFFICERS WATCHING OVER 12 PPR	https://www.nst.com.my/news/crime-courts/2017/11/307472/24-policemen-officers-watching-over-12-ppr	21.61	No crime entry	Rejected
222946	NST	TWELVE, INCLUDING DATUK SERI, CHARGED WITH RIOTING IN PENANG NIGHTCLUB ROW	https://www.nst.com.my/news/2017/03/222946/twelve-including-datuk-seri-charged-rioting-penang-nightclub-row	23.02	No crime entry	Rejected

The table illustrates how external knowledge sources can validate the relevance of articles even when their similarity scores fall below the established threshold. For example, articles with scores of 43.15% and 39.99% were ultimately deemed relevant due to corroborating information found in the external knowledge bases. This approach enhances the system's ability to accurately identify crime-related news articles by considering multiple factors beyond simple similarity scores.

Finally, 50.06% of the rejected articles representing 1232 articles are considered as crime related news articles and 1229 news articles (49.94%) are the final rejected articles. This demonstrates the crucial role of external knowledge sources in preventing the exclusion of potentially valuable articles based solely on similarity scores. However, the absence of supporting information in these sources also helps filter out irrelevant articles, as seen in the example of an article rejected due to its low similarity score and lack of corresponding information in DBpedia and/or Wikidata. Interestingly, all discarded articles originated from the NST newspaper, suggesting a potential discrepancy in the reliability of different news sources. This emphasizes the need to consider the source of information when evaluating the overall quality and relevance of collected data. The analysis underscores the importance of a multifaceted crime news collection and filtering approach. A reliable dataset for further analysis and understanding of the crime domain was obtained by combining similarity thresholds with topic modelling and external knowledge sources.

The final number of crime related news articles filtered using this approach is 20,777. Figs. 10 and 11 show a detailed statistic of this dataset. Fig. 7 shows the number of articles published each year. The data indicates trends in article publication over time, helping us understand the temporal distribution of articles within the dataset. We can see from the graph that the number of reported crime news articles sharply decreased during the years 2020 and 2021 due to COVID-19 MCO (Movement Control Order).

Fig. 10 shows the distribution of the number of sentences per article, which indicates that most articles have between 5 and 20 sentences, with a peak of around 10 to 15 sentences. However, fewer articles have a very high or very low sentence count.

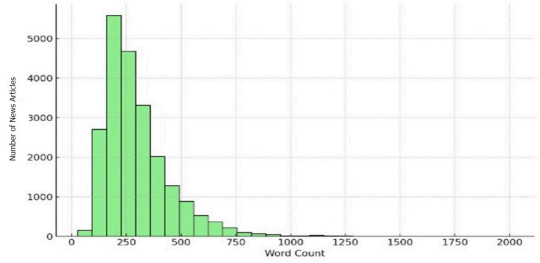

Fig. 10: Distribution of Sentence Count

Fig. 11 displays the distribution of the number of words per article indicating that the majority of articles contain between 100 and 500 words, with a noticeable peak around 200-300 words. However, there are fewer articles with extremely high or low word counts.

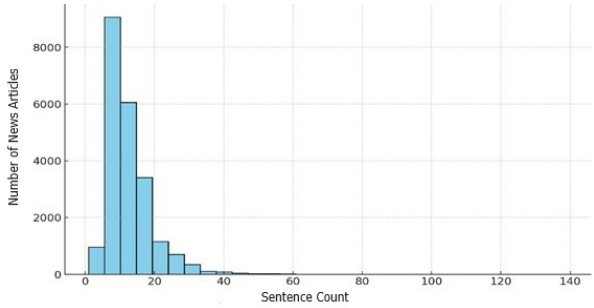

Fig. 11: Distribution of Word Count

Classification of Crime News Articles

Focusing on crime news classification, this section analyzes the effectiveness of the fin-tuned BERT-based model compared to other neural network architectures. Table 7 provides performance metrics (Accuracy, Precision, Recall, and F1-score) for four different models: Bi-LSTM, Attention-Based Bi-LSTM, Attention-Based CNN-Bi-LSTM, and a Fine-Tuned BERT model.

The Bi-LSTM model demonstrates strong performance with high accuracy, precision, recall, and F1-score. The precision and recall values are quite close, indicating a balanced performance in correctly identifying positive cases and minimizing false positives. However, there is a slight edge towards precision over recall, which means it slightly favors correctly identifying true positives over false positives.

The Attention-Based Bi-LSTM model shows a significant improvement over the plain Bi-LSTM model. The accuracy is almost 2% higher, and both precision and recall are in the high 97-98 range. The attention mechanism helps the model focus on the most relevant parts of the input sequences, improving its ability to correctly classify them. This results in the highest F1-score among the models discussed, showing an excellent balance between precision and recall.

Table 7: Comparison of the Baseline and Proposed Model Using Accuracy, Precision, Recall, and F1-Score

Model	Accuracy	Precision	Recall	F1-score
Bi-LSTM	96.70	93.29	92.05	92.67
Attention Based Bi-LSTM	98.87	97.86	98.4	98.2
Attention-based CNN-Bi-LSTM	92.3	94.4	89.47	91.87
Fine Tuned BERT (Ours)	99.45	97.38	96.92	97.15

The Attention-Based CNN-Bi-LSTM model has the lowest accuracy among the models listed, at 92.3%. However, it still maintains a high precision of 94.4%, indicating that when it predicts a positive case, it is likely correct. The recall is somewhat lower at 89.47%, suggesting that the model misses some positive cases. The combination of CNN and Bi-LSTM, along with the attention mechanism, provides useful features but seems to struggle slightly compared to other models in this table. This might be due to the complexity and the way this model handles long-term dependencies.

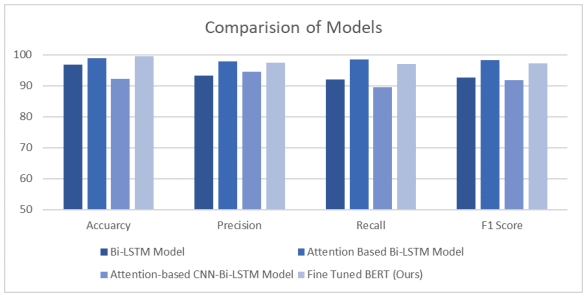

Fig. 12: Comparison of Models performance

The Fine-Tuned BERT model achieves the highest accuracy at 99.45%, indicating that it correctly classifies almost all instances. The precision and recall are also very high, at 97.38 and 96.92% respectively, leading to a high F1-score of 97.15. BERT's transformer-based architecture, which excels at capturing contextual relationships within text, contributes to its superior performance. Fine-tuning BERT on the specific data*set al.*ows it to leverage its pre-trained knowledge and adapt to the nuances of the task, resulting in outstanding overall performance.

As shown in the bar chart depicted in Fig. 12, we can notice both attention-based models outperform the standard Bi-LSTM, suggesting that attention mechanisms enhance the model's ability to focus on relevant parts of the input text for improved classification. The superior performance of the fine-tuned BERT model underscores its effectiveness in capturing the nuances of language and context within crime news articles. However, the relatively lower performance of this model suggests that the CNN layer might not be suitable for this specific task or requires

further optimization. Analyzing the errors and misclassifications could provide valuable insights.

Final Classification Results Using Fine Tuned BERT

In this task of our work, we feed the fine-tuned BERT model with the collected data gathered and filtered in the collection and filtering stage. A total of 20,777 crime news-related articles will be classified into their appropriate main category as well as the sub-categories.

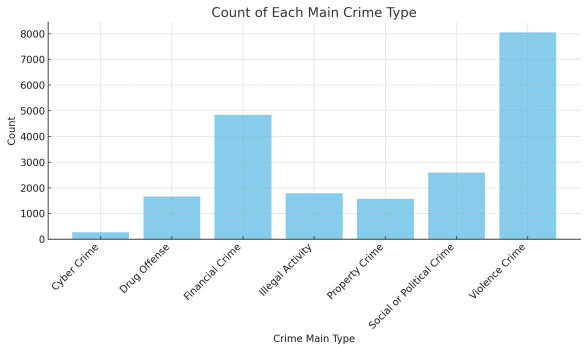

Fig. 13: Total number of classified Crime news articles of each main Crime Type

Fig. 13 illustrates the distribution of different main crime types. Violence crime has the highest count, indicating it is the most prevalent main crime type. The number of incidents surpasses 8000, representing 38.80% of the crime news articles. In contrast, cybercrimes are the least frequent among the listed crime types, with a relatively low count of only 1.3%. The second most common type is financial crime, with around 4000 news articles, 23.3% reflecting a significant issue in financial-related offences. Social or political crimes also have a substantial count of 12.5%, though less than financial crime, indicating a notable amount of unspecified social or political activities. Drug offences, property crimes, and illegal activity have a moderate count of 8.0, 7.5 and 8.6%, respectively, suggesting they are less frequent but still notable.

From our point of view and according to the classification report, the high incidence of violent crimes suggests that law enforcement agencies might need to allocate more resources and strategies to address and mitigate violent crime. Also, significant financial crimes indicate potential weaknesses in financial regulations and monitoring, necessitating stricter controls and compliance measures. Consequently, crimes like drug offences and property crimes, while less frequent, still require public awareness campaigns and policy interventions to reduce their occurrence.

Fig. 14 shows the details of the distribution of crime subtypes within the broader categories. Like the main crime distribution graph, we can monitor the variation of reported crime news articles with specific and broader types. Homicide, which is a violent crime type is the most frequent subtype, highlighting a critical area for law enforcement and public safety efforts. Both corruption and assault are highly prevalent, indicating issues with public trust and safety. Fraud, this subtype is notably high, reflecting significant challenges in financial security and fraud prevention. Money laundering and robbery also have high counts, suggesting ongoing issues with financial crimes and violent theft. Civil unrest is another notable subtype, potentially indicating social or political instability. Finally, the graph shows a wide variety of crime subtypes, each with different frequencies, illustrating the complexity of crime patterns.

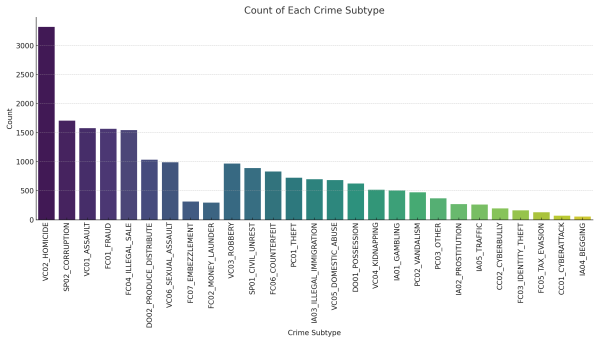

Fig. 14: Total number of classified Crime news articles of each Crime sub-Type

To sum up, the fine-tuned BERT classification results provide a valuable starting point for understanding crime trends reflected in news media. By considering potential biases and exploring the data further, we can better understand crime patterns and their implications.

Conclusion

This research investigated the potential of utilizing BERT and external knowledge bases for crime news classification. The proposed multi-stage framework successfully collected and filtered a large dataset of crime news articles, demonstrating the effectiveness of combining similarity measures with topic modeling and knowledge base verification. Furthermore, the fine-tuned BERT model achieved superior performance compared to baseline models, highlighting its ability to capture the nuances of language and context within crime news articles. This study emphasizes the value of integrating advanced NLP techniques with external knowledge sources for enhanced crime news analysis, ultimately contributing to a better understanding of crime trends and facilitating effective responses.

Future Work

This research opens several promising avenues for future investigation and research. A primary focus would be expanding the model's capabilities to handle multilingual crime news analysis, particularly incorporating local Malaysian languages, while implementing real-time news streaming functionality for continuous monitoring. The integration of advanced

transformer architectures beyond BERT, such as GPT-4 or LLaMA and using different approaches, such as zero-shot, Few-shot and Chain of Thoughts (CoT), could potentially enhance classification accuracy and robustness we are currently working on. The knowledge base framework could be strengthened by expanding beyond DBpedia and Wikidata, developing domain-specific crime event ontology, and implementing automatic population mechanisms we are also currently working on. Finally, developing user-friendly interfaces and APIs would facilitate practical adoption by law enforcement agencies, bridging the gap between academic research and real-world applications in crime prevention and analysis.

Acknowledgment

The authors would like to express their sincere gratitude to the Faculty of Information Science and Technology (FTSM), Universiti Kebangsaan Malaysia (UKM), for their continuous support and provision of research facilities. Special thanks are also extended to all staff members whose assistance and encouragement contributed to the completion of this work.

Funding Information

This research received no external funding.

Author's Contributions

Ashour Ali: Conceived the study, developed the methodology, conducted data analysis, implemented the model and code, and prepared the initial manuscript draft.

Shahrul Azman Mohd Noah: Provided supervision, critically reviewed the methodology and results, and contributed to manuscript editing.

Lailatul Qadri Zakaria: Critical review, and manuscript editing.

Saeed Amer Al Ameri: Contributed to experimental design, validation, and interpretation of results, and revised the manuscript for technical accuracy.

All authors read and approved the final manuscript.

Ethics

This article is an original work and includes previously unpublished material.

Conflicts of Interest

The authors have no conflicts of interests to declare.

References

Alameri, S. A., & Mohd, M. (2021). Comparison of Fake News Detection using Machine Learning and Deep Learning Techniques. *2021 3rd International Cyber Resilience Conference (CRC)*, 1–6. https://doi.org/10.1109/crc50527.2021.9392458

Ali, A., Noah, S. A. M., & Zakaria, L. Q. (2022). Representation of Event-Based Ontology Models: A Comparative Study. *International Journal of Computer Science & Network Security*, *22*(7), 147–156.

Ali, A., Noah, S. A. M., & Zakaria, L. Q. (2023). A BERT-Based model: Improving Crime News Documents Classification through Adopting Pre-trained Language Models. *Research Square*. https://doi.org/10.21203/rs.3.rs-2582775/v1

Ali, N. M., Mohd, M., Yacob, N. F., Saad, S., Omar, N., Aziz, M. J. A., & Noah, S. A. M. (2012). Investigating User Perception in a Development of Crime News Retrieval System. *International Journal of Digital Content Technology and Its Applications*, *6*(10), 118–126. https://doi.org/10.4156/jdcta.vol6.issue10.14

Al-Saif, H., & Al-Dossari, H. (2018). Detecting and Classifying Crimes from Arabic Twitter Posts using Text Mining Techniques. *International Journal of Advanced Computer Science and Applications*, *9*(10), 377–387. https://doi.org/10.14569/ijacsa.2018.091046

Arumi, E. R., & Sukmasetya, P. (2020). Exploiting Web Scraping for Education News Analysis Using Depth-First Search Algorithm. *Jurnal Online Informatika*, *5*(1), 19–26. https://doi.org/10.15575/join.v5i1.548

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichl*et al.*ocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Boldi, P., Codenotti, B., Santini, M., & Vigna, S. (2004). UbiCrawler: a scalable fully distributed Web crawler. *Software: Practice and Experience*, *34*(8), 711–726. https://doi.org/10.1002/spe.587

Christian, J., Valiveti, S., & Jain, S. (2022). Profiling Cyber Crimes from News Portals Using Web Scraping. *Futuristic Trends in Networks and Computing Technologies*, *936*, 1007–1016. https://doi.org/10.1007/978-981-19-5037-7_72

Deepak, G., Rooban, S., & Santhanavijayan, A. (2021). A knowledge centric hybridized approach for crime classification incorporating deep bi-LSTM neural network. *Multimedia Tools and Applications*, *80*(18), 28061–28085. https://doi.org/10.1007/s11042-021-11050-4

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

Gasparetto, A., Marcuzzo, M., Zangari, A., & Albarelli, A. (2022). A Survey on Text Classification Algorithms: From Text to Predictions. *Information*, *13*(2), 83. https://doi.org/10.3390/info13020083

González-Carvajal, S., & Garrido-Merchán, E. C. (2020). Comparing BERT against traditional machine learning text classification. *Computation and Language*. https://doi.org/10.47852/bonviewJCCE3202838

Gopal, L. S., Prabha, R., Pullarkatt, D., & Ramesh, M. V. (2020). Machine Learning based Classification of Online News Data for Disaster Management. *2020 IEEE Global Humanitarian Technology Conference (GHTC)*, 1–8. https://doi.org/10.1109/ghtc46280.2020.9342921

Heydon, A., & Najork, M. (1999). Mercator: A scalable, extensible Web crawler. *World Wide Web*, *2*(4), 219–229. https://doi.org/10.1023/a:1019213109274

Hossain, Md. M., Chowdhury, Z. R., Rezwanul Haque Akib, S. M., Sabbir Ahmed, Md., Hossain, Md. Moazzem., & Miah, A. S. M. (2023). Crime Text Classification and Drug Modeling from Bengali News Articles : A Transformer Network-Based Deep Learning Approach. *2023 26th International Conference on Computer and Information Technology (ICCIT)*, 1–6. https://doi.org/10.1109/iccit60459.2023.10441195

Hsu, B.-M. (2020). Comparison of Supervised Classification Models on Textual Data. *Mathematics*, *8*(5), 851. https://doi.org/10.3390/math8050851

Srinivasa, K., & Thilagam, P. S. (2019). Crime base: Towards building a knowledge base for crime entities and their relationships from online news papers. *Information Processing & Management*, *56*(6), 102059. https://doi.org/10.1016/j.ipm.2019.102059

Kynabay, B., Aldabergen, A., & Zhamanov, A. (2021). Automatic Summarizing the News from Inform.kz by Using Natural Language Processing Tools. *2021 IEEE International Conference on Smart Information Systems and Technologies (SIST)*, 1–4. https://doi.org/10.1109/sist50301.2021.9465885

Maybir, J., & Chapman, B. (2021). Web scraping of ecstasy user reports as a novel tool for detecting drug market trends. *Forensic Science International: Digital Investigation*, *37*, 301172. https://doi.org/10.1016/j.fsidi.2021.301172

Mohd, M., Bsoul, Q. W., Ali, N. M., Noah, S. A. M., Saad, S., Omar, N., & Aziz, M. J. A. (2012). Optimal initial centroid in k-means for crime topic. *Journal of Theoretical & Applied Information Technology*, *45*(1).

Mohd, M., & Mohamad Ali, N. (2011). An Interactive Malaysia Crime News Retrieval System. *IEEE*, 220–223. https://doi.org/10.1109/STAIR.2011.5995792

Morsey, M., Lehmann, J., Auer, S., Stadler, C., & Hellmann, S. (2012). DBpedia and the live extraction of structured data from Wikipedia. *Program*, *46*(2), 157–181. https://doi.org/10.1108/00330331211221828

Nafea, A. A., Majeed, R. R., Ali, A., Yas, A. J., Alameri, S. A., & AL-Ani, M. M. (2024a). A Brief Review of Big Data in Healthcare: Challenges and Issues, Recent Developments, and Future Directions. *Babylonian Journal of Internet of Things*, *2024*, 10–15. https://doi.org/10.58496/bjiot/2024/002

Nafea, A. A., Muayad, M. S., Majeed, R. R., Ali, A., Bashaddadh, O. M., Khalaf, M. A., Sami, A. B. N., & Steiti, A. (2024b). A Brief Review on Preprocessing Text in Arabic Language Dataset: Techniques and Challenges. *Babylonian Journal of Artificial Intelligence*, *2024*, 46–53. https://doi.org/10.58496/bjai/2024/007

Nugroho, K. S., Sukmadewa, A. Y., & Yudistira, N. (2021). Large-Scale News Classification using BERT Language Model: Spark NLP Approach. *6th International Conference on Sustainable Information Engineering and Technology 2021*, 240–246. https://doi.org/10.1145/3479645.3479658

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237. https://doi.org/10.18653/v1/n18-1202

Prieto Curiel, R., Cresci, S., Muntean, C. I., & Bishop, S. R. (2020). Crime and its fear in social media. *Palgrave Communications*, *6*(1), 57. https://doi.org/10.1057/s41599-020-0430-7

Radford, A. (2018). Improving language understanding with unsupervised learning. *Ntegration of CiNii Dissertations and CiNii Books into CiNii Research*.

Rahem, K. R., & Omar, N. (2014). Drug-related crime information extraction and analysis. *Proceedings of the 6th International Conference on Information Technology and Multimedia*, 250–254. https://doi.org/10.1109/icimu.2014.7066639

Reyes-Ortiz, J. A. (2019). Criminal Event Ontology Population and Enrichment using Patterns Recognition from Text. *International Journal of Pattern Recognition and Artificial Intelligence*, *33*(11), 1940014. https://doi.org/10.1142/s0218001419400147

Roche, S. P., Pickett, J. T., & Gertz, M. (2016). The Scary World of Online News? Internet News Exposure and Public Attitudes Toward Crime and Justice. *Journal of Quantitative Criminology*, *32*(2), 215–236. https://doi.org/10.1007/s10940-015-9261-x

Rollo, F., & Po, L. (2020). Crime Event Localization and Deduplication. *The Semantic Web – ISWC 2020. ISWC 2020.*, *12507*, 361–377. https://doi.org/10.1007/978-3-030-62466-8_23

Saha, R., Naskar, A., Dasgupta, T., & Dey, L. (2020). A System for Analysis, Visualization and Retrieval of Crime Documents. *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, 317–321. https://doi.org/10.1145/3371158.3371405

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22–36. https://doi.org/10.1145/3137597.3137600

Singh Bhati, V., Tiwari, S., & Mandloi, J. (2019). Machine Learning and Deep Learning Integrated Model to Predict, Classify and Analyze Crime in Indore City. *Proceedings of Recent Advances in Interdisciplinary Trends in Engineering & Applications (RAITEA) 2019*. https://doi.org/10.2139/ssrn.3364984

Sundhara Kumar, K. B., & Bhalaji, N. (2020). A Novel Hybrid RNN-ELM Architecture for Crime Classification. *Second International Conference on Computer Networks and Communication Technologies Conference Paper*, *44*, 876–882. https://doi.org/10.1007/978-3-030-37051-0_98

Thielmann, A., Weisser, C., Krenz, A., & S fken, B. (2023). Unsupervised document classification integrating web scraping, one-class SVM and LDA topic modelling. *Journal of Applied Statistics*, *50*(3), 574–591. https://doi.org/10.1080/02664763.2021.1919063

Valasik, M. (2024). *Crime Mapping and Spatial Analysis*. https://doi.org/10.1093/acrefore/9780190264079.013.821

Vidal, M. T., Rodr guez, E. S., & Reyes-Ortiz, J. A. (2020). Classification of Criminal News Over Time Using Bidirectional LSTM. *Pattern Recognition and Artificial Intelligence*, *12068*, 702–713. https://doi.org/10.1007/978-3-030-59830-3_61

Vidal, M. T., Rodr guez, E. S., & Reyes-Ortiz, J. A. (2021). Classification of Spanish Criminal News Using Neural Networks. *Advances in Pattern Recognition and Artificial Intelligence*, 193–211. https://doi.org/10.1142/9789811239014_0012

Vrandecic, D., & Kr tzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, *57*(10), 78–85. https://doi.org/10.1145/2629489

Wang, M., Cai, Q., Wang, L., Li, J., & Wang, X. (2020). Chinese news text classification based on attention-based CNN-BiLSTM. *MIPPR 2019: Pattern Recognition and Computer Vision*. MIPPR 2019: Pattern Recognition and Computer Vision, Wuhan, China. https://doi.org/10.1117/12.2538132