SSCKC: Strategic framework for Semantics Driven Classification and Knowledge Based Enrichment of Crime Data using the Perspective of Current and Historical News

Anubrat Bora[1,†] and Gerard Deepak[2,*,†]

[1] *Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, India*
[2] *BMS Institute of Technology and Management, Yelahanka, Bengaluru, India*

Abstract

A strategic framework is essential for categorizing emerging and highly domain-specific datasets, such as those in criminology and news. Classification is crucial for enhancing the usefulness and relevance of these datasets. This research introduces a framework that systematically classifies and structures complex domain-specific data, improving its utility for expert systems. By integrating the LLaMA (Large Language Model Architecture) with a lightweight machine learning classifier, XGBoost, the framework enhances the categorization process and enriches datasets with valuable information. The approach also includes the amalgamation of standard news APIs and uses the Cricket Algorithm for optimization with a diversity index to derive optimal solutions, focusing on improving precision, accuracy and F1 score in crime classification.

Keywords

Knowledge-Based Classification, LLaMA, Standard News API

1. Introduction

Classification and knowledge enrichment are critical to the advancement of theoretical research and practical applications in the specialized field of criminology. The deep relationships between many entities associated with crime are typically difficult to capture by traditional methods of data classification, owing to the dynamic and multifaceted character of criminal activity. The lack of comprehensive and precise crime data makes this problem worse and makes it more difficult to create efficient classification schemes. Integrating the dynamic, real-time data offered by news APIs is essential to meeting this problem. These APIs provide an ongoing flow of up-to-date information that, when paired with past crime data, allows for a more thorough and sophisticated assessment of crime trends. Researchers and practitioners can obtain important insights into trends, correlations, and developing patterns—all crucial for efficient crime analysis and prevention—by examining both historical and contemporary data. By introducing an innovative categorization system that integrates the newest generative AI technologies and machine learning techniques, the suggested framework seeks to close the gap in existing methodology.

The application of LLaMA (Large Language Model Architecture), which offers a solid foundation for semantic comprehension and classification, is essential to this strategy. The ability of LLaMA to synthesize specific information from several data sources improves the framework's capacity to analyze and classify intricate crime data. Moreover, Web 3.0 technologies are employed by the framework to enhance the classification procedure. Through a Sparkle-based engine to query news APIs, the system combines historical and current data, allowing partial knowledge graphs and ontological representations to be created. A more intuitive comprehension of the links and trends within the crime data is made possible by these knowledge graphs, which provide a visual and

AISD-2024: Second International Workshop on Artificial Intelligence: Empowering Sustainable Development, October 2, 2024, co-located with the Second International Conference on Artificial Intelligence: Towards Sustainable Intelligence (AI4S-2024), Virtual Event, Lucknow, India.
* Corresponding author.
† These authors contributed equally.
anubrat.mitblr2022@learner.manipal.edu (A. Bora); gerard.deepak.christuni@gmail.com(G. Deepak)
0009-0005-8755-6558 (A. Bora); 0000-0003-0466-2143 (G. Deepak)

structured representation of the data. By utilizing the Cricket Algorithm for metaheuristic optimization and Shannon Entropy for feature selection, the framework guarantees the classification process's precision and effectiveness. This research enhances criminology and offers a useful tool for comprehending and preventing criminal activity by addressing the important demand for efficient classification and knowledge enrichment in the discipline.

1.1. Motivation

Because criminal activity is dynamic and complex, it is becoming more important than ever to classify things in specialized fields like criminology. If properly categorized, patterns in both historical and modern crimes might provide insightful information for future study and useful applications. Nevertheless, current models are unable to represent the complex interactions between things because they lack the semantic depth and enrichment that come from contemporary APIs. By creating a solid, semantic-driven framework that makes use of Web 3.0 technologies and cutting-edge classification concepts, this research aims to close this gap. The proposed framework aims to advance the field and meet a critical need by integrating a sophisticated system of semantic similarity mechanisms to provide a more accurate and comprehensive approach to categorizing criminal activities based on historical and real-time data.

1.2. Contributions

The proposed framework introduces several key contributions: it leverages LLaMA, a state-of-the-art AI-generated language model, to enhance semantic understanding and classification of datasets. This framework also incorporates a novel dataset perspective using XGBoost, a lightweight machine learning algorithm, to effectively synthesize and classify features. By integrating advanced semantic AI techniques with dynamic background knowledge from APIs, and applying Simpson's Diversity Index and Shannon Entropy, the framework ensures diverse and robust classification outcomes. Additionally, the use of the Cricket Algorithm for metaheuristic optimization transforms feasible solutions into optimal ones. Finally, the framework not only classifies datasets with precision but also synthesizes knowledge representations in the form of trees or partially linked graphs, adding a distinctive semantic perspective that enhances the understanding of classified data.

1.3. Organization

The paper's organization is as follows: Section 2. comprehensively lists out a set of works related to this paper. Section 3 provides an overview of the proposed system architecture. Section 4 enumerates the results yielded by the proposed model. Finally, Section 5 concludes the paper.

2. Related Works

2.1. Crime Classification and Prediction Techniques

A Bi-LSTM neural network was proposed by G. Deepak et al. [1] for classification purposes. The network was able to classify various forms of crime based on data gathered from Twitter and Google News. The initial labeling step and preprocessing of the data were done using the Fuzzy C-means algorithm. To extract features, GloVe word embeddings were used. To improve the quality of crime classification, a graph with weights was constructed from Google News and Social Web sites such as Twitter has been integrated into automatically created ontologies with minimum human supervision. After trials, their suggested approach has demonstrated that it can achieve assessment metrics more effectively than the current approaches. It was evaluated on four different datasets and contrasted with four alternative approaches; for the four noteworthy datasets, it demonstrated an increase in accuracy and a drop in FDR. Bappee et al. [2] have enhanced quantitative crime research by finding elements that contribute to crime by using better access to urban data, including infrastructure, criminal records, and demographics. But thorough data is typically lacking in small cities, which makes developing consistent forecasting frameworks more difficult. This study aims to improve Halifax, Nova Scotia's crime forecast through cross-domain knowledge transfer from Toronto and Vancouver. To create models, it used ensemble learning techniques and instance-based

transfer learning along with seasonal data fusion. The findings demonstrated that combining various data sources enhanced the accuracy of crime prediction. Sharma, Nishant, et al. [3] concentrated on identifying criminal patterns, and used machine learning techniques to identify criminal episodes by kind based on time and place. They used several supervised classification models, such as Decision Tree, Gaussian Naive Bayes, k-NN, Logistic Regression, AdaBoost, and Random Forest, and used San Francisco crime records from 2003 to 2015. They used under sampling strategies like Edited NN and Neighborhood Cleaning Rule and oversampling approaches like SMOTE to handle imbalanced crime categories. Their method classified crimes with an accuracy of about 81%.

2.2. Integration of Diverse Data Sources

Chhatwal et al. [4] have presented an architecture for the automated creation of knowledge bases from a variety of heterogeneous sources. With an emphasis on criminology, tools for predictive and preventive analysis by utilizing their vast domain knowledge were created. They used data from Google's news API, criminology publications, e-books, web portals, and publicly available datasets to augment a seed criminology ontology. Transformers are used for instance alignment, LSTM is used for classification, and description logics are used for reasoning in the system. The methodology outperformed prior methods in producing a vast, comprehensive knowledge base with an average accuracy of 95.35%. Yang Dingqi and colleagues [5] have addressed the difficult issue of crime prevention. Unlike other methods that rely solely on past crime data, CrimeTelescope combines crime data with urban and social media data to increase the accuracy of forecast. To pinpoint important characteristics and display hotspots on an interactive map, the platform continuously gathers and examines this data using linguistic and statistical techniques. This method enhances forecast accuracy by as much as 5.2% when compared to systems based solely on historical records, as shown using real-world data from New York City. The System usefulness Scale (SUS) survey was another tool used to verify the platform's usefulness.

2.3. Utilization of Artificial Intelligence and Machine Learning

SS Kshatri et al. [6] have demonstrated how artificial intelligence, data mining, and machine learning are being used increasingly in contemporary police investigations to prevent crime. Predicting and preventing future crimes required an understanding of criminal behavior from historical information. Researchers can group data and categorize crime patterns by analyzing larger datasets, which helps them pinpoint a city's most vulnerable areas. The thesis was concerned with techniques for crime prediction based on past data. Using three algorithms—SMO, Zero R, and J48 decision trees—it applies data mining techniques through WEKA to predict the risk of crime. The Naive algorithm accurately forecasted the frequency of crimes, according to a review of more than 10,000 records from the Indian police department. Junxiang Yin [7] has studied crime prediction, an important social computing system that predicts future crimes by gleaning information from criminal histories. This helps law enforcement and warns the public about dangers. Big data, IoT (Internet of Things), and AI developments are causing a rapid evolution in deep learning-based crime prediction models. Yin's survey examines the theories, practices, and widely used datasets of these models to classify them into deep learning and standard machine learning methods. To address issues such as limited data volume and ambiguous data kinds, the survey provides a machine learning-driven big data research model that provides guidance for more research.

2.4. Crime Data Analysis and Sentiment Detection

Shoeibi, et al. [8] have tackled the problem of preserving a robust online community while allowing individuals to express themselves freely. Their work created a framework for Twitter monitoring through the extraction of tweet data and the application of metadata and graph analysis to study user behavior. Interest-worthy profiles were selected by the platform for additional analysis. It analyzed information using natural language processing, and using an SVM + TF-IDF model, it was able to identify tweets related to crimes with 88.89% accuracy. DistilBERT + FFNN sentiment analysis achieved 80% accuracy in determining the sentiment of tweets pertaining to crimes. The platform sought to support law enforcement in suppressing hate speech and terrorist propaganda. Bogomolov et al. [9] presented a novel method for forecasting crime in specific geographic locations. In contrast to earlier techniques that depended on offender profiling or historical knowledge, their strategy

makes use of basic demographic data and human behavioral data from mobile network activities. Based on actual crime statistics from London, experimental results were able to forecast whether region s will become hotspots for crime with about 70% accuracy. The ramifications of these findings for data-driven crime investigation are also covered in the paper.

2.5. Challenges and Insights in Crime Data Analysis

Jeyaboopathiraja et al. [10] investigated the difficulties faced by police analysts in comprehending complex crime data. They emphasized how hard it is to deal with several crimes and how criminals' knowledge keeps changing. To solve crimes and prevent others from happening, the review stressed the value of crime analysis and prediction. The study investigated the ways in which artificial intelligence (AI) and deep learning approaches can improve the accuracy and applicability of crime rate projections. The review compared deep learning and machine learning methods to inform efficient crime rate prediction, assessing the benefits and drawbacks of various data approaches and model parameters.

3. Proposed Architecture

Figure 1. shows the proposed system architecture for the crime classification framework using a generated AI-driven semantic intelligence. The dataset of news and criminology is subject to extraction of terms and categories directly extracted from the keywords of a dataset and categories are extracted from the dataset categories. The terms and categories extracted from the dataset of news and criminology using the LLaMA model. Large language models (LLMs) such as LLaMA are used for the criminal classification framework because of their superior ability to comprehend and analyze complicated human language. LLMs are perfect for managing the huge and complex data found in criminology and the news because they can effectively extract meaningful terms and categories from large databases. To be more precise, the LLaMA (Language Learning and Machine Analysis) model was selected due to its high accuracy, scalability, adaptability, and sophisticated natural language processing capabilities. LLaMA's powerful algorithms guarantee accurate extraction of terms and categories from criminology and news datasets, which is essential to the framework's performance. Furthermore, LLaMA is engineered to manage massive datasets with efficiency, adapting to the vast and ever-expanding characteristics of criminology and news data without sacrificing functionality. Furthermore, the captions generated by the LLaMA model are subject to feature selection using the Shannon entropy, the step deviation of 0.12 is empirically set. This value, 0.12, is empirically set as the step deviance measure. The selection of features based on this step deviance is because the LLMs are already generated and are domain specific. Therefore, the entropy of Shannon's Entropy is relaxed accordingly.

The dataset of news and criminology is subjected to cloud-based classification using the features selected with XGBoost as the classifier of choice. The reason for using XGBoost, which is a lightweight machine learning classifier that runs a strong deep learning-based classification model, is the fact that it efficiently handles large datasets and provides high accuracy. The system is modeled using AgentSpeak for building the agent. The Spark human agent gathers the entities which are classified using the XGBoost classifier based on the class and the classified instances and appends them as a query using SPARQL. This query is then used to query the News API stack, namely the Newsomatic API, BBC Platform API, Zeit Online Content API, and Guardian Open Platform API. It queries each of these APIs independently and obtains the feasible solution phases. Since the query is based on the entities obtained from the domain, this can be called the feasible solution case.

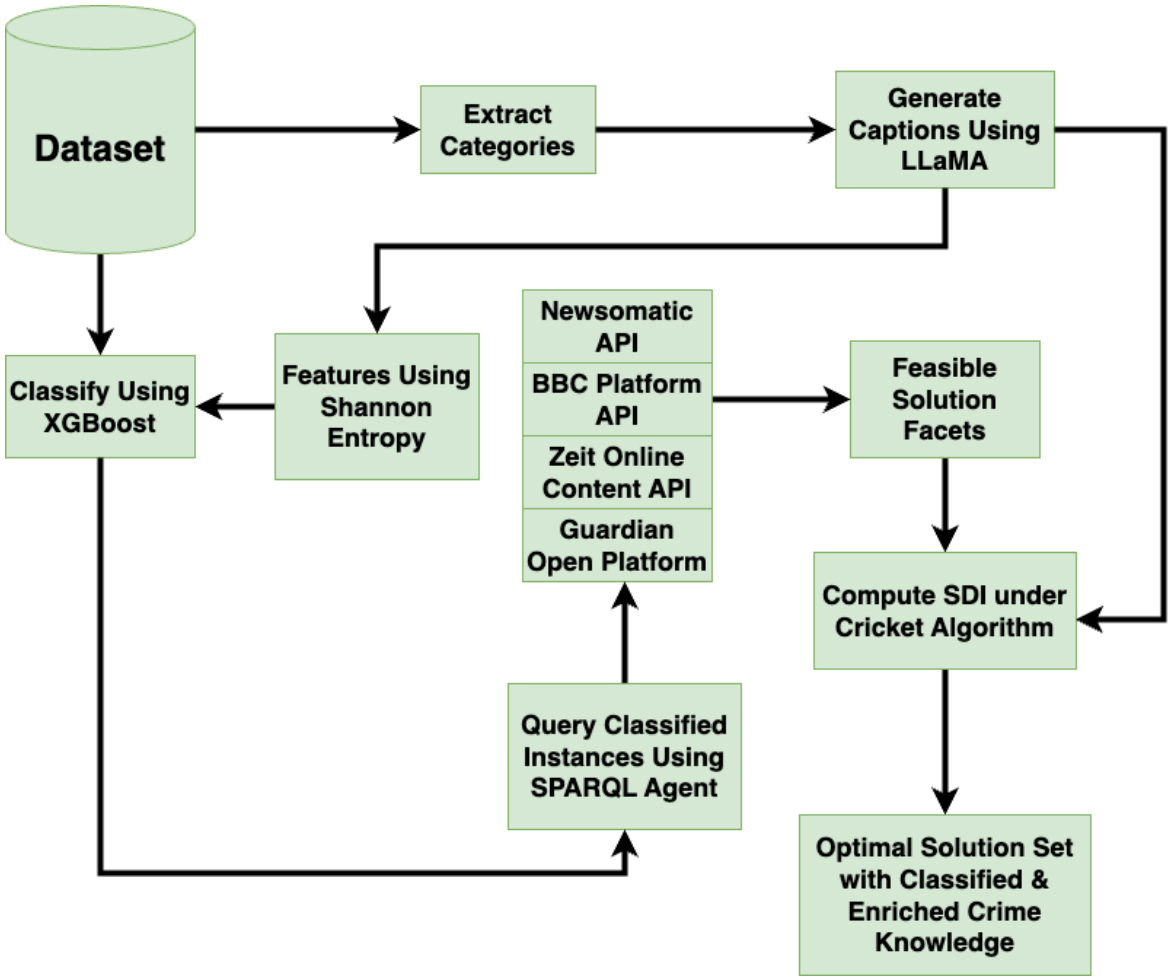

Figure 1: The proposed structure architecture of the SSCKC.

3.1. Newsomatic API

For getting current and pertinent news items, the Newsomatic API offers a comprehensive platform. With an emphasis on education, it provides a carefully chosen collection of news stories written in an approachable and captivating style for younger readers. Through the API, programmers can incorporate this content into apps to provide users with up-to-date news and instructional resources. The Newsomatic API has features like multimedia content, photos, and in-depth news stories that are all designed to promote learning and improve comprehension of current events. Applications can promote media literacy and awareness while providing richer news experiences by utilizing this API.

3.2. BBC Platform API

The British Broadcasting Corporation offers an abundance of news and multimedia content that may be accessed through the BBC API. Among the many services it provides are article retrieval, real-time news updates, and multimedia content including audio and video snippets. With the help of the API, developers can incorporate the BBC's comprehensive news coverage into their apps and provide customers reputable, high-quality news material. Access to breaking news, in-depth articles, and specialist information from several BBC news categories are among the features. Applications may provide their customers with dependable, varied news coverage by utilizing the BBC API, guaranteeing that users are informed about current events and stories throughout the world.

3.3. Guardian Open Platform API

One of the top news organizations in the world, The Guardian, provides comprehensive access to its news and media content through the Guardian Open Platform API. With this API, developers can include news stories, video files, and interactive elements among other types of content from The

Guardian's portfolio into their apps. Breaking news, in-depth articles, opinion pieces, and other editorial content can be retrieved via the API's endpoints. With features like subject classification, article metadata, and search capabilities, the Guardian Open Platform API enables applications to enhance user engagement with rich, carefully curated material from a reliable news source while delivering high-quality, current news content.

3.4. LLaMA

LLaMA, an innovative advancement in artificial intelligence and natural language processing, was developed to extend the capabilities of machine learning models. It is predicated on the ideas of transformer architectures, which serve as the basis for models like BERT and GPT-4. However, the unique architectural enhancements and optimizations that make LLaMA unique enable it to handle massive data volumes and perform a wide range of language-related activities with astounding precision and effectiveness. Because it can be used for a wide range of tasks, from basic ones like text generation and translation to more complex ones like summarization and question answering, LLaMA is a versatile tool in the AI (Artificial Intelligence) arsenal. One of the most important aspects of LLaMA is its scalability. The model can handle bigger datasets and more parameters without requiring a proportional increase in processing capacity because of its efficient scaling. Because of the progress made in data processing and model training, LLaMA can function at a prominent level even when task complexity increases since it is scalable.

Additionally, the architecture of LLaMA incorporates strategies to mitigate common issues associated with big language models, such as the difficulty of preserving coherence in lengthier text outputs and the tendency to provide convincing but inaccurate information (hallucinations). Because of these developments, LLaMA is now a key player in the ongoing development of AI-driven language models. There are many different industries and fields in which LLaMA is used. When it comes to content creation, LLaMA can assist marketers and writers by creating unique content that is contextually relevant. This can boost productivity and save time. It can make it possible for customer support chatbots to understand and respond to user requests with accuracy, improving the customer experience. Furthermore, LLaMA can assist with data analysis, literature reviews, and the generation of original ideas in academic and research settings by analyzing and synthesizing enormous amounts of data. As LLaMA advances, it will further integrate into and transform a multitude of aspects of everyday and professional life, showcasing the immense potential of state-of-the-art AI technology.

3.5. XGBoost

XGBoost is a technology developed by Tianqi Chen and colleagues based on the gradient boosting approach, which iteratively boosts prediction accuracy by combining the strengths of numerous weak learners, usually decision trees. Thanks to its novel approach to model training and optimization, XGBoost can handle large-scale data sets and difficult prediction tasks with remarkable speed and performance. Owing to its capacity to predict customer behavior and identify fraudulent transactions, XGBoost has gained popularity across a variety of industries. XGBoost stands out from other boosting algorithms in part due to its innovative system architecture, which blends algorithmic optimizations with hardware-aware enhancements. The model efficiently handles sparse features and missing data by employing a novel sparsity-aware technique that is frequently observed in real-world data sets.

Furthermore, XGBoost uses second-order gradient information (Hessian) to perform the gradient descent process more precisely, leading to faster convergence and more accurate models. XGBoost is incredibly useful in applications where time is of importance because it can train models significantly faster than traditional gradient boosting approaches because of these technological advancements and parallel processing capabilities. The extensive use of XGBoost in academia and industry is evidence of its massive impact on the machine learning sector. It has consistently outperformed in numerous machine learning competitions, such as Kaggle, where it has proven essential to winning numerous challenges. Regression, classification, and ranking issues are just a few of the many applications that XGBoost excels at solving thanks to its scalability and robustness. Its open-source architecture and comprehensive documentation have also facilitated its integration with other machine learning pipelines and frameworks. XGBoost will continue to be excellent in

modern machine learning approaches and become even more relevant and valuable as the demand for accurate and efficient prediction models grows.

3.6. Shannon Entropy

The Shannon entropy, first proposed by Claude Shannon in 1948, quantifies the average quantity of information generated by a stochastic source by gauging the degree of uncertainty or unpredictability in a group of occurrences. It determines the average bit count needed to encode a message from a source optimally; higher entropy levels denote more complicated or ambiguous information.

$$H(X) = -\sum_i p_i \log_2 p_i, \tag{1}$$

3.7. CoSimRank

CoSimRank is a version of the SimRank method which was created to overcome some of its drawbacks and increase the effectiveness of similarity calculations in huge graphs. As the graph size grows, SimRank may become computationally costly; in contrast, CoSimRank provides a more scalable alternative. CoSimRank does this by using an iterative calculation technique that is more effective and decreases the temporal complexity involved in determining node similarities. CoSimRank offers a useful method for calculating similarities in large-scale networks by streamlining the recursive similarity calculation process. This makes it appropriate for use in recommendation systems, social network analysis, and web page ranking, among other uses.

3.8. Simpson's Diversity Index

The Simpson's Diversity Index (SDI), which considers the total number of species present as well as their relative abundances, is a crucial indicator for evaluating biodiversity within a habitat. Ecologists and researchers can compare several habitats or monitor changes over time within a single habitat with the use of this index, which offers a quantifiable assessment of how diverse an ecosystem is. The Simpson's Diversity Index has a value between 0 and 1, where a number near 1 denotes minimal diversity, frequently because of a few dominant species, and a value of 0 suggests infinite diversity (with no single species dominating the community). Understanding the health of ecosystems, the state of conservation, and how environmental changes affect biodiversity all depend on this metric.

$$\mathrm{D} = 1 - \sum \frac{\mathrm{n_i}(\mathrm{n_i}-1)}{\mathrm{N}(\mathrm{N}-1)}, \tag{2}$$

3.9. Cricket Algorithm

The Cricket Algorithm is an optimization method inspired by nature that solves challenging optimization issues by mimicking the behavior of crickets, especially their mating sounds. The algorithm was created by S. S. N. Singh and associates, and it was inspired by the way crickets use sound to find partners and create a mating network. This strategy is based on the ideas of swarm intelligence, in which unique actions support the process of overall optimization. Every cricket in the Cricket Algorithm stands for a fix for an optimization issue. The "calls" that the crickets make are symbolic of the caliber of their solutions. A cricket's attractiveness to others is determined by the frequency and volume of its calls. A cricket's fitness level affects the volume and frequency of its calls, increased fitness results in louder calls. This procedure is like how crickets utilize their sounds to entice mates and create a network of fixes. The two primary functions of the algorithm are exploration and exploitation. Crickets wander around the solution space in search of new locations during exploration and concentrate on fine-tuning solutions in areas of potential interest during exploitation. The secret to the Cricket Algorithm's success is its ability to achieve equilibrium between exploration and exploitation, which allows it to bypass local optima and converge towards superior solutions. It has been used in many different domains, including data mining, scheduling, and engineering design. Through the utilization of adaptive mechanisms and the application of cricket behavior, the Cricket Algorithm offers a reliable and adaptable solution for intricate optimization issues. Because of its ability to replicate natural processes and adjust to various issue environments, it is a useful supplement to the family of swarm intelligence algorithms.

4. Performance Evaluation and Results

A single, sizable dataset created by combining many criminology and criminal news datasets was used for the experiments. Among other datasets, the combined dataset contains the DICE (2022) dataset, titled *DICE: A Dataset of Italian Crime Event News* [13], and the Work With Data (2024) dataset, focusing on *News about Grantsmanship for Criminal Justice and Criminology* [14]. Additionally, the dataset incorporates information from Work with Data (2024) on the *International Library of Criminology, Delinquency, and Deviant Social Behavior* [15]. Zee Solver (2023) contributed with the *International Crime* dataset [16], while PeakMetrics (2020) provided data on *News Articles Related to Mass Shootings* [17]. The dataset also includes the Work with Data (2024) dataset featuring *News of SpringerBriefs in Psychology, Behavioral Criminology* [18], and Government (2023) with the *Uniform Crime Reports* [19]. Standardized keywords, common categories, and annotations produced by an annotation tool were used to combine these disparate datasets. A modified web crawler was designed to formalize a similar dataset. This crawler carried out web crawling in accordance with the Web 3.0 structural metadata paradigm. It was created based on the given keywords and dataset annotations. To provide an extensive and rich dataset for analysis, the objective was to compile pertinent documents and data.

Table 1
Comparison of Performance of the proposed SSCKC with other approaches

Model	Average Precision %	Average Recall %	Average Accuracy %	Average F-Measure %	FDR
CCPS [11]	89.90	91.55	90.73	90.71	0.11
DLCCP [12]	90.75	92.83	91.79	91.80	0.10
CACP [13]	92.61	93.76	93.19	93.21	0.08
Proposed SSCKC	97.12	98.76	97.84	98.03	0.03

The proposed SSCKC (Semantic Strategic Crime Knowledge Classification) framework demonstrates high performance in classifying and enriching crime data using a combination of current and historical news scenarios. The framework achieves a precision of 97.12%, recall of 98.76%, accuracy of 97.84%, F-measure of 98.03%, and a false discovery rate (FDR) of 0.03. The approach hybridizes a Large Language Model (LLM) with a strategic machine learning classifier, XGBoost, to enhance feature selection and caption generation. The dataset, which contains historical news and criminology data, is novel in its composition. In addition, the SSCKC framework integrates various news APIs, such as Newsomatic, BBC Open API, Zeit API, and the Guardian Open Platform, creating a hybrid model that blends machine learning with community-verified, domain-expert-based news sources. These APIs provide authenticated and editorially referenced information, which strengthens the framework's knowledge base.

The framework also employs a semantic approach using Shannon entropy and the Simpson's Diversity Index, calculated via the cricket algorithm—a metaheuristic optimization technique. This ensures that the solutions generated are not only feasible but also optimal, contributing to the framework's robustness. The SSCKC is thus a comprehensive and strategically designed model that leverages lightweight machine learning, strong LLMs, and verified community news repositories to deliver a highly accurate and semantically enriched classification system. The CCPS (Conventional Crime Prediction System) model does not perform as expected primarily because it relies on individual machine learning models such as K-Means Nearest Neighbour, Decision Trees, Support Vector Machines, K-Nearest Neighbour, Extra Tree Classifier, and Artificial Neural Networks. These models are used in isolation rather than being hybridized in the framework. Without hybridization, these standalone models fail to deliver best-in-class results, as they lack the complementary strengths that a combined approach could offer. Moreover, the CCPS model does not incorporate Large Language Models (LLMs), which are crucial for generating contextual understanding and semantic enrichment of data. Additionally, the absence of community-verified facts, which serve as

background knowledge in the SSCKC framework, further weakens the CCPS model's ability to provide accurate and reliable predictions. As a result, the CCPS model underperforms when compared to the proposed SSCKC framework, which benefits from a more sophisticated integration of machine learning, LLMs, and verified knowledge sources. The DLCCP (Deep Learning Crime Classification and Prediction) framework does not perform as expected because it relies solely on deep learning architecture for crime classification and prediction. While deep learning classifiers can be powerful, the model lacks semantics-oriented reasoning and learning, particularly in terms of quantitative reasoning and semantic similarity computation. This absence of semantic analysis diminishes the model's ability to understand and process crime data in a nuanced and context-aware manner.

Furthermore, the DLCCP framework does not incorporate authentic auxiliary knowledge from the Web 3.0 paradigm, limiting the depth and breadth of knowledge available to the model. This minimalistic knowledge base hinders the model's overall performance. Additionally, the framework does not utilize optimization techniques or nature-based metaheuristic models, which are essential for refining solutions and achieving optimal performance. As a result, the DLCCP framework underperforms when compared to the proposed SSCKC framework, which benefits from a more comprehensive approach that includes semantic reasoning, Web 3.0 knowledge, and advanced optimization methods. The CACP (Crime Analysis and Classification Platform) framework underperforms because it relies solely on two classifiers, Random Forest, and Support Vector Machine (SVM), without any advanced hybridization or integration with other models. The dataset used is basic, which limits the model's ability to handle more complex crime data.

Although the original article reports high accuracy, this drops significantly when applied to the proposed dataset due to the framework's lack of hybridization, absence of semantics-oriented augmented knowledge, and failure to include Large Language Models (LLMs) for caption generation or agents for query formalization. Most critically, the CACP framework does not leverage community-contributed, verified knowledge, which is essential in the proposed framework. These limitations result in CACP's inability to perform at the level of the more sophisticated SSCKC framework.

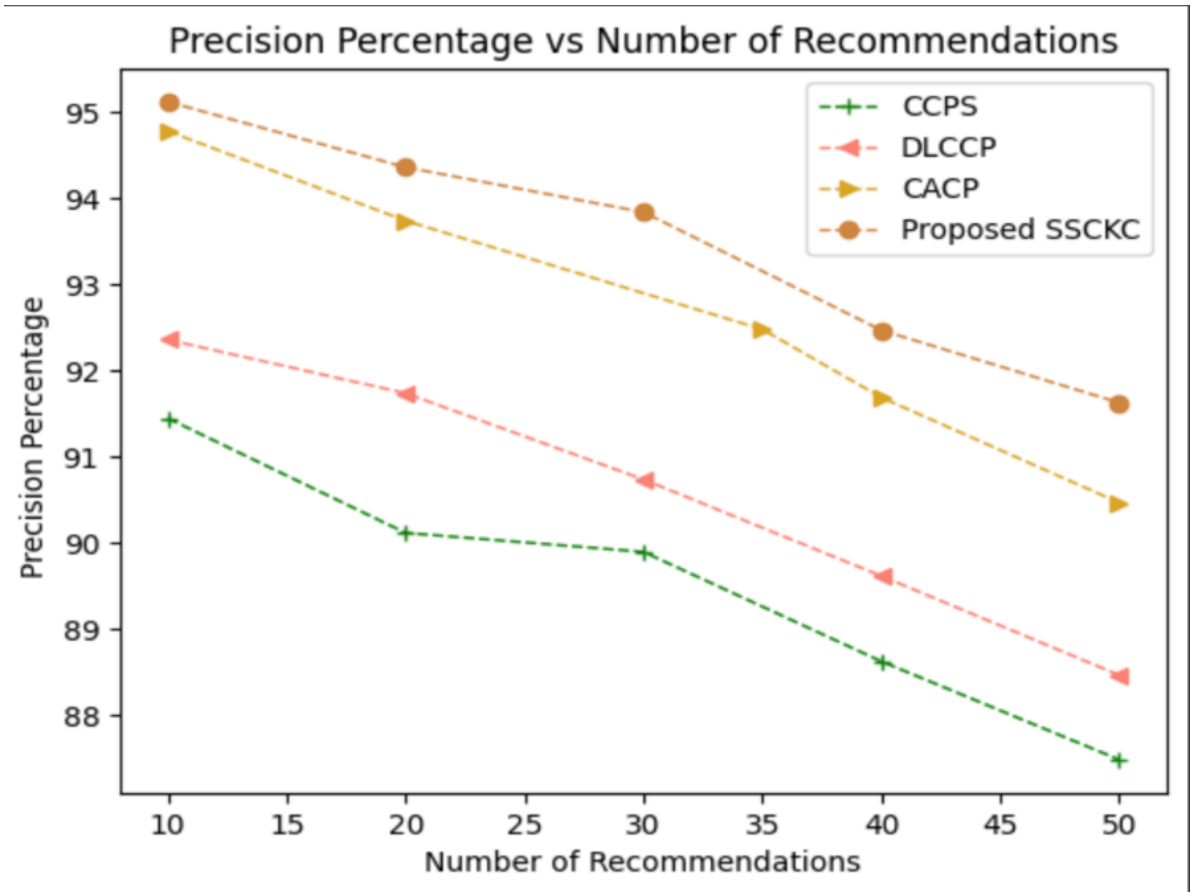

Figure 2: Model Comparision; Precession Percentage vs Number of Recommendations

For four models—CCPS, DLCCP, CACP, and SSCKC—the precision % is shown against the total number of recommendations in the graph. SSCKC has the highest precision at 95.12% of 10 recommendations, followed by CACP at 94.78%, DLCCP at 92.36%, and CCPS at 91.45%. SSCKC continues to be the best performer at 91.63% when recommendations reach 50, while CACP falls to 90.47%, DLCCP to 88.47%, and CCPS to 87.49%. This proves that SSCKC maintains the maximum precision at all recommended levels consistently.

The Proposed SSCKC, using a range of criteria such as false discovery rate (FDR), average precision, recall, accuracy, and F-measure. With an average precision of 97.12%, the Proposed SSCKC model outperforms the other models, which include CCPS at 89.90%, DLCCP at 90.75%, and CACP at 92.61%. SSCKC leads in average recall once more with 98.76%, followed by CCPS with 91.55%, DLCCP with 92.83%, and CACP with 93.76%. SSCKC has a higher average accuracy of 97.84% than CCPS (90.73%), DLCCP (91.79%), and CACP (93.19%). SSCKC scores a remarkable 98.03% on the F-measure, compared to scores of 90.71%, 91.80%, and 93.21% for CCPS, DLCCP, and CACP, respectively. Compared to CCPS (0.11), DLCCP (0.10), and CACP (0.08), SSCKC has the lowest false discovery rate (FDR), at 0.03. The Proposed SSCKC model is the most successful model in this comparison because it performs better overall in lowering false discovery rates and excels in precision, recall, accuracy, and F-measure.

5. Conclusion

This paper suggests a strategic framework for knowledge classification and generation specifically tailored for news and criminology. The framework uniquely combines the generative AI capabilities of LLaMA with the computational efficiency of the XGBoost machine learning algorithm, making it both cost-effective and practical. It integrates dynamic knowledge from current crime data and trends through news APIs, alongside historical data, to enhance its predictive capabilities. Utilizing a Sparkle-based engine for querying news APIs, Shannon Entropy for feature selection, and the Cricket Algorithm with Simpson's Diversity Index, the framework stands out for its ability to deliver high-quality, accurate solutions. This approach not only ensures precise crime classification but also enriches datasets with valuable knowledge, positioning the framework as a leading model in semantic crime and news classification with minimal false discovery rates.

References

[1] Deepak, Gerard, S. Rooban, and A. Santhanavijayan. "A knowledge centric hybridized approach for crime classification incorporating deep bi-LSTM neural network." *Multimedia Tools and Applications* 80.18 (2021): 28061-28085.

[2] Bappee, Fateha Khanam, et al. "Examining the impact of cross-domain learning on crime prediction." *Journal of big data* 8.1 (2021): 96. S. Cohen, W. Nutt, Y. Sagic, Deciding equivalances among conjunctive aggregate queries, J. ACM 54 (2007). doi:10.1145/1219092.1219093.

[3] Shama, Nishat. *A machine learning approach to predict crime using time and location data*. Diss. BRAC University, 2017.

[4] Chhatwal, Gurunameh Singh, and Gerard Deepak. "Integrative KnowGen: integrative knowledge base generation for criminology as a domain of choice." *International Conference on Digital Technologies and Applications*. Cham: Springer International Publishing, 2022.

[5] Yang, Dingqi, et al. "CrimeTelescope: crime hotspot prediction based on urban and social media data fusion." *World Wide Web* 21 (2018): 1323-1347. I. Editor (Ed.), The title of book two, The name of the series two, 2nd. ed., University of Chicago Press, Chicago, 2008. doi:10.1007/3-540-09237-4.

[6] Kshatri, Sapna Singh, and Bhawana Narain. "Analytical study of some selected classification algorithms and crime prediction." *Int J Eng Adv Technol* 9.6 (2020): 241-247.

[7] Yin, Junxiang. "Crime Prediction Methods Based on Machine Learning: A Survey." *Computers, Materials & Continua* 74.2 (2023).

[8] Shoeibi, Niloufar, et al. "Ai-crime hunter: An AI mixture of experts for crime discovery on twitter." *Electronics* 10.24 (2021): 3081. S. W. Smith, An experiment in bibliographic mark-up:

Parsing metadata for xml export, in: R. N. Smythe, A. Noble (Eds.), Proceedings of the 3rd. annual workshop on Librarians and Computers, volume 3 of LAC '10, Paparazzi Press, Milan Italy, 2010, pp. 422–431. doi:99.9999/woot07-S422.

[9] Bogomolov, Andrey, et al. "Once upon a crime: towards crime prediction from demographics and mobile data." *Proceedings of the 16th international conference on multimodal interaction*. 2014. D. Harel, LOGICS of Programs: AXIOMATICS and DESCRIPTIVE POWER, MIT Research Lab Technical Report TR-200, Massachusetts Institute of Technology, Cambridge, MA, 1978.

[10] Jeyaboopathiraja, J., and G. Maria Priscilla. "A Thorough Analysis of Machine Learning and Deep Learning Methods for Crime Data Analysis." *Inventive Computation and Information Technologies: Proceedings of ICICIT 2020* (2021): 795-812.

[11] Pratibha, A. Gahalot, Uprant, S. Dhiman and L. Chouhan, "Crime Prediction and Analysis," 2nd International Conference on Data, Engineering and Applications (IDEA), Bhopal, India, 2020, pp. 1-6, doi: 10.1109/IDEA49133.2020.9170731.

[12] Stalidis, P.; Semertzidis, T.; Daras, P. Examining Deep Learning Architectures for Crime Classification and Prediction. *Forecasting* **2021**, *3*, 741-762. https://doi.org/10.3390/forecast3040046.

[13] Zaidi, N.A.S., Mustapha, A., Mostafa, S.A., Razali, M.N. (2020). A Classification Approach for Crime Prediction. In: Khalaf, M., Al-Jumeily, D., Lisitsa, A. (eds) Applied Computing to Support Industry: Innovation and Technology. ACRIT 2019. Communications in Computer and Information Science, vol 1174. Springer, Cham. https://doi.org/10.1007/978-3-030-38752-5_6.

[14] DICE (2022). DICE: A Dataset of Italian Crime Event News [Dataset]. https://paperswithcode.com/dataset/italian-crime-news.

[15] Work With Data (2024). News about Grantsmanship for Criminal Justice and Criminology [Dataset]. https://www.workwithdata.com/datasets/news?f=1&fcol0=news_title_matched&fop0=includes&fval0=Grantsmanship+for+criminal+justice+and+criminology.

[16] Work With Data (2024). News about International Library of Criminology, Delinquency, and Deviant Social Behavior [Dataset]. https://www.workwithdata.com/datasets/news?f=1&fcol0=news_title_matched&fop0=includes&fval0=International+library+of+criminology%2C+delinquency+and+deviant+social+behavi.

[17] Zee Solver (2023). International Crime [Dataset]. https://www.kaggle.com/datasets/zeesolver/international-crime.

[18] PeakMetrics (2020). News Articles Related to Mass Shootings [Dataset]. https://www.kaggle.com/peakmetrics/news-articles-related-to-mass-shootings.

[19] Work With Data (2024). News of SpringerBriefs in Psychology, Behavioral Criminology [Dataset]. https://www.workwithdata.com/datasets/news?col=news_link%2Cnews_title_matched&f=1&fcol0=news_title_matched&fop0=includes&fval0=SpringerBriefs+in+psychology.+Behavioral+criminology.

[20] Government (2023). Uniform Crime Reports [Dataset]. https://data.world/government/uniform-crime-reports.