Leveraging Machine Learning Methods for Crime Analysis in Textual Data

Shynar Mussiraliyeva, Gulshat Baispay
Al-Farabi Kazakh National University, Almaty, Kazakhstan

***Abstract*—The proposed research paper explores the application of machine learning techniques in crime analysis problem, specifically focusing on the classification of crime-related textual data. Through a comparative analysis of various machine learning models, including traditional approaches and deep learning architectures, the study evaluates their effectiveness in accurately detecting and categorizing crime-related text data. The performance of the models is assessed using rigorous evaluation metrics, such as the area under the receiver operating characteristic curve (AUC-ROC), to provide insights into their discriminative power and reliability. The findings reveal that machine learning frameworks, particularly the deep learning model, consistently outperform conventional machine learning approaches, highlighting the potential of advanced neural network architectures in crime analysis tasks. The implications of these findings for law enforcement agencies and researchers are discussed, emphasizing the importance of leveraging advanced machine learning techniques to enhance crime prevention and intervention efforts. Furthermore, avenues for future research are identified, including the integration of multiple data sources and the exploration of interpretability and explainability of machine learning models in crime analysis problem. Overall, this research contributes to advancing the field of crime analysis problem and underscores the importance of leveraging innovative computational approaches to address complex societal challenges.**

Keywords*—*Machine learning; artificial intelligence; crime analysis; text processing; natural language processing; text analysis; data-driven decision making

I. Introduction

In contemporary society, the proliferation of online platforms has led to an unprecedented volume of textual data being generated daily. These data encompass a wide array of topics, including discussions related to crime and criminal activities. Leveraging this wealth of online textual data for crime analysis has garnered significant attention from researchers and law enforcement agencies alike [1]. Traditional methods of crime analysis often rely on structured data obtained from official reports, which may suffer from limitations such as reporting biases and time delays [2]. However, online textual data offer a unique opportunity to complement traditional crime analysis techniques by providing real-time and unfiltered insights into various criminal activities [3].

Machine learning (ML) techniques have emerged as powerful tools for analyzing large volumes of textual and image data, enabling the extraction of valuable insights and patterns [4]. By harnessing the computational capabilities of ML algorithms, researchers can uncover hidden associations and trends within vast amounts of online text, facilitating a deeper understanding of criminal behaviors [5]. Moreover, ML approaches offer the flexibility to adapt to evolving crime patterns and modus operandi, making them indispensable in the realm of crime analysis [6].

The integration of machine learning in crime analysis presents numerous advantages. Firstly, ML algorithms can effectively process unstructured textual data, including social media posts, forum discussions, and news articles, enabling comprehensive surveillance of criminal activities across diverse online platforms [7]. Secondly, ML-based crime analysis can aid law enforcement agencies in identifying emerging threats and hotspots in real time, allowing for proactive intervention and crime prevention measures [8]. Furthermore, ML techniques can assist in the prioritization of investigative efforts by highlighting relevant information and filtering out noise from the vast expanse of data [9].

Despite the promise of ML in crime analysis, several challenges persist. One such challenge is the inherent ambiguity and noise present in online textual data, which can hinder the accuracy and reliability of ML models [10]. Additionally, issues related to data privacy and ethical considerations necessitate careful deliberation when employing ML techniques for crime analysis [11]. Moreover, the dynamic nature of online discourse poses challenges in maintaining the relevance and effectiveness of ML models over time [12].

To address these challenges and maximize the potential of machine learning in crime analysis, this research paper aims to explore various ML approaches and their applications in analyzing online textual data related to criminal activities. By synthesizing insights from existing literature and empirical studies, this paper seeks to provide a comprehensive overview of the current state-of-the-art in ML-based crime analysis. Furthermore, this paper will examine the implications of ML-driven crime analysis for law enforcement practices, highlighting opportunities for future research and development [13].

In summary, the integration of machine learning techniques in crime analysis offers unprecedented opportunities to harness the vast amount of online textual data for enhancing public safety and security. By overcoming inherent challenges and leveraging the capabilities of ML algorithms, researchers and law enforcement agencies can gain invaluable insights into criminal behaviors and trends, ultimately contributing to more effective crime prevention and intervention strategies [14].

II. Related Works

Crime analysis has long been a focal point of research in criminology and law enforcement, with recent advancements in machine learning (ML) techniques opening up new avenues for exploring and understanding criminal behaviors through analysis of online textual data [15]. Previous studies have demonstrated the efficacy of ML algorithms in various aspects of crime analysis, ranging from predictive modeling to crime pattern recognition [16]. Furthermore, researchers have explored the application of ML in analyzing diverse sources of online textual data, including social media posts, online forums, and news articles, to uncover insights into criminal activities [17].

One area of research that has gained prominence is the use of natural language processing (NLP) techniques in crime analysis. NLP methods enable the extraction of meaningful information from unstructured textual data, facilitating the identification of key themes, sentiments, and entities related to criminal activities [18]. By applying NLP techniques such as sentiment analysis and named entity recognition, researchers can gain deeper insights into public perceptions of crime and the dissemination of criminal narratives across online platforms [19].

Moreover, research has explored the utility of social network analysis (SNA) in crime analysis, particularly in the context of online social networks. SNA enables researchers to analyze the structure and dynamics of social networks to identify influential actors, detect criminal communities, and trace the flow of information related to criminal activities [20]. By leveraging SNA techniques, researchers can uncover hidden connections and patterns within online social networks, shedding light on the mechanisms underlying the spread of criminal behaviors [21].

In addition to NLP and SNA, researchers have investigated the application of machine learning algorithms such as classification, clustering, and anomaly detection in crime analysis. Classification algorithms, such as support vector machines (SVM) and random forests, have been employed to categorize online textual data into different crime-related topics or classes [22]. Clustering algorithms, such as k-means and hierarchical clustering, have been used to group similar textual documents together, enabling the identification of common themes and patterns within large datasets [23]. Anomaly detection algorithms, such as isolation forest and one-class SVM, have been utilized to identify unusual or anomalous behavior in online textual data, which may signify potential criminal activities [24].

Furthermore, researchers have explored interdisciplinary approaches that combine ML techniques with domain-specific knowledge from fields such as criminology, sociology, and psychology. By integrating insights from multiple disciplines, researchers can develop more robust models for crime analysis that account for the complex interplay of individual, social, and environmental factors influencing criminal behaviors [25]. For example, research has shown the importance of incorporating spatial and temporal information into ML models for crime prediction and hotspot analysis, as crime patterns often exhibit geographic and temporal clustering [26].

Moreover, the emergence of big data analytics has provided researchers with unprecedented access to vast amounts of online textual data for crime analysis. Big data analytics techniques, such as data mining and machine learning, enable researchers to process and analyze large-scale datasets to extract actionable insights and patterns related to criminal activities [27]. By harnessing the computational power of big data analytics platforms, researchers can overcome the challenges associated with the volume, velocity, and variety of online textual data, enabling more comprehensive and timely crime analysis [28].

Despite the advancements in ML techniques for crime analysis, several challenges remain. One challenge is the issue of data quality and bias inherent in online textual data, which may stem from factors such as misinformation, sampling biases, and linguistic nuances [29]. Addressing these challenges requires careful preprocessing and validation of the data, as well as the development of robust ML models that are resilient to noise and biases [30].

Furthermore, ethical considerations surrounding the use of online textual data for crime analysis warrant attention. Privacy concerns, data security risks, and the potential for algorithmic biases raise ethical dilemmas that must be carefully navigated to ensure responsible and ethical use of ML techniques in crime analysis [31]. Additionally, the transparency and interpretability of ML models are crucial for fostering trust and accountability in the criminal justice system [32].

In summary, the application of machine learning techniques in crime analysis holds significant promise for enhancing our understanding of criminal behaviors and improving public safety. By leveraging advancements in NLP, SNA, big data analytics, and interdisciplinary approaches, researchers can gain valuable insights into the complex dynamics of criminal activities unfolding in the digital realm. However, addressing challenges related to data quality, ethical considerations, and algorithmic transparency is essential to realizing the full potential of ML-driven crime analysis [33]. Through interdisciplinary collaboration and ongoing research efforts, we can continue to advance the state-of-the-art in crime analysis and develop more effective strategies for preventing and combating crime in the digital age [34].

III. Materials and Methods

Research indicates that criminal incidents tend to be unevenly distributed across urban areas [35]. This non-uniform distribution implies that certain locations may exhibit a higher propensity for criminal activity than others, thus rendering crime a location-dependent phenomenon [36]. Given the variability of crime rates based on geographic location, law enforcement agencies face challenges in resource allocation and crime prevention efforts, particularly when it comes to identifying high-risk areas. Consequently, there is a pressing need for accurate models capable of effectively detecting crime hotspots and reliably predicting the time and location of criminal events. Fig. 1 demonstrates the proposed system for crime detection using machine learning techniques.

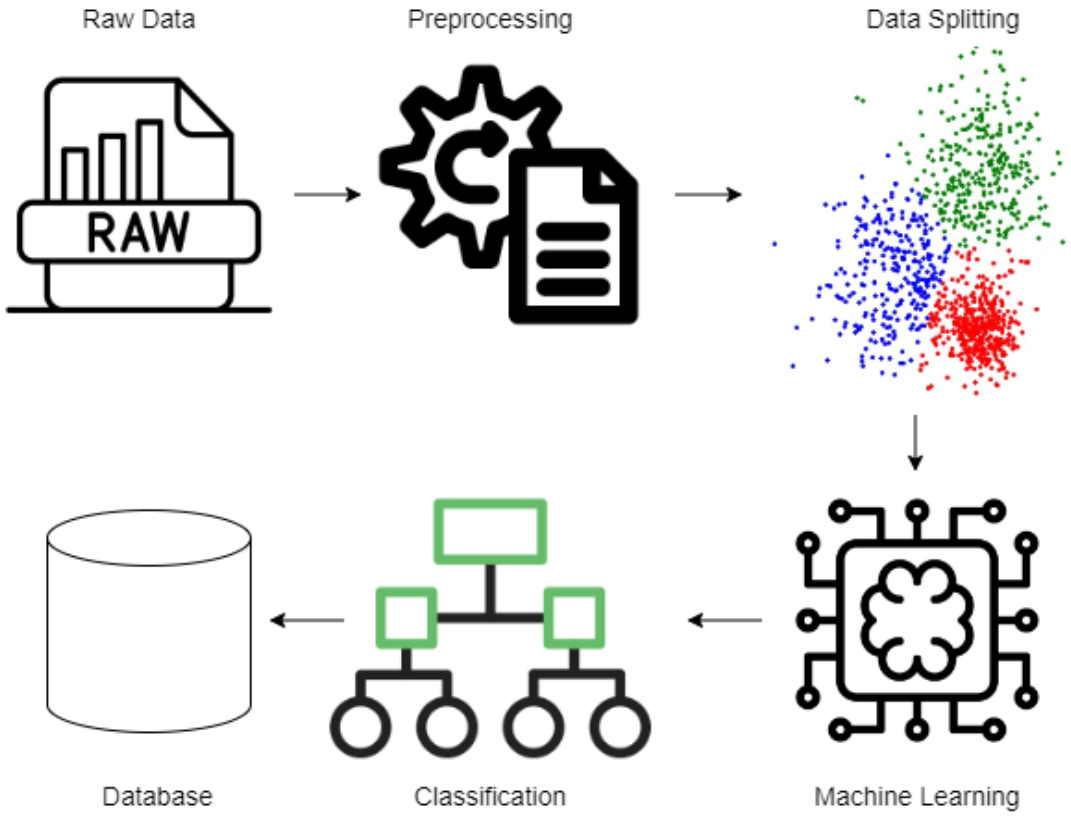

Fig. 1. The proposed system for crime analysis.

This section delineates our proposed methodology for detecting and forecasting crime hotspots, elucidating the sequential steps essential for implementation. The model, as depicted in Fig. 1, is founded upon a multi-step process meticulously designed to attain these objectives. Initially, the methodology involves data preprocessing to enhance data quality and prepare it for analysis. Subsequently, feature extraction techniques are employed to capture relevant information from the dataset. Following this, the model undergoes training using historical crime data to learn patterns and associations. Once trained, the model is deployed to predict future crime hotspots based on the learned patterns. Finally, the results are evaluated using performance metrics such as accuracy, precision, recall, and F1-score to assess the effectiveness of the predictive model. By delineating each step of the methodology, this section provides a clear roadmap for researchers and practitioners interested in implementing crime hotspot detection and prediction systems.

A. Dataset

The Crime Articles Recommendation System dataset, available on Kaggle, serves as a valuable resource for research in the domain of crime analysis and recommendation systems [37]. This dataset comprises a collection of articles related to various aspects of crime, including crime prevention strategies, criminal investigations, and criminal justice policies. The articles cover a diverse range of topics, such as cybercrime, organized crime, white-collar crime, and violent crime, providing a comprehensive overview of the multifaceted nature of criminal activities.

The dataset includes textual data extracted from the articles, encompassing titles, summaries, and full text content. This textual information serves as the primary input for the recommendation system, allowing researchers to explore and analyze the content of the articles in depth. Additionally, metadata such as publication dates, authors, and sources are provided for each article, enabling researchers to contextualize the content and track temporal trends in crime-related literature.

One notable feature of the Crime Articles Recommendation System dataset is its size and diversity. With a large number of articles spanning multiple years and covering a wide range of crime-related topics, the dataset offers ample opportunities for conducting comprehensive analyses and developing sophisticated recommendation algorithms. Researchers can leverage this diversity to explore various dimensions of crime, including geographical variations, temporal trends, and thematic patterns. Fig. 2 demonstrates word count distribution in the applied dataset.

Furthermore, the dataset is well-suited for the development and evaluation of recommendation systems tailored to the domain of crime articles. Recommendation systems aim to assist users in discovering relevant content based on their preferences and interests. By analyzing the textual content and metadata of the articles, researchers can design recommendation algorithms that prioritize articles likely to be of interest to users based on their historical interactions or explicit feedback.

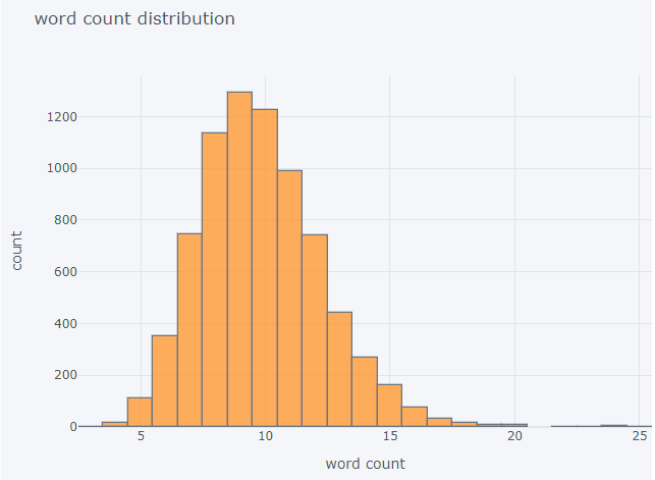

Fig. 2. Word count distribution in the dataset.

The Crime Articles Recommendation System dataset serves as a valuable resource for advancing research in crime analysis, recommendation systems, and related fields. Its size, diversity, and richness of content make it well-suited for a wide range of research applications, from exploring patterns of criminal behavior to designing intelligent systems for assisting users in discovering relevant crime-related articles.

B. Evaluation Parameters

In the evaluation of the proposed crime hotspot detection and prediction methodology, several key performance metrics are utilized to assess the effectiveness and reliability of the model. These metrics encompass accuracy, precision, recall, F-score, and the area under the receiver operating characteristic curve (AUC-ROC) [39-43], each providing valuable insights into different aspects of the model's performance.

Accuracy serves as a fundamental measure of the model's overall correctness in predicting crime hotspots and forecasting criminal events. It quantifies the proportion of correctly classified instances among all instances evaluated, thus offering a broad assessment of the model's predictive capabilities.

$$accuracy = \frac{TP + TN}{P + N} \tag{10}$$

Precision, on the other hand, focuses specifically on the accuracy of positive predictions made by the model. It calculates the proportion of true positive predictions among all positive predictions made, thereby indicating the model's ability to minimize false positives and maintain a high level of precision in identifying crime hotspots.

$$preision = \frac{TP}{TP + FP} \tag{2}$$

Recall complements precision by assessing the model's ability to capture all relevant instances of crime hotspots. It measures the proportion of true positive predictions identified by the model among all actual positive instances, thereby reflecting the model's sensitivity to identifying hotspots accurately.

$$recall = \frac{TP}{TP + FN} \tag{3}$$

The F-score, or F1 score, provides a balanced evaluation of both precision and recall by calculating their harmonic mean. This metric offers a single value that captures the overall performance of the model in terms of both precision and recall, providing a comprehensive assessment of its effectiveness in detecting crime hotspots.

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{4}$$

Lastly, the AUC-ROC metric evaluates the discriminative power of the model in distinguishing between positive and negative instances. It measures the area under the receiver operating characteristic curve, which plots the true positive rate against the false positive rate at various threshold settings. A higher AUC-ROC value indicates better discrimination performance, with values closer to 1 indicating superior predictive capabilities.

By employing these evaluation parameters, researchers can comprehensively assess the performance of the proposed crime hotspot detection and prediction methodology, thereby providing valuable insights into its effectiveness and reliability in aiding law enforcement agencies in crime prevention efforts.

IV. Experimental Results

Fig. 3 illustrates the confusion matrices utilized in the detection of crime-related texts employing various machine learning methodologies. These matrices serve to visually represent the efficacy of the different approaches employed in this study. Through these matrices, the study elucidates the classification outcomes, offering a clear depiction of how predictions are distributed across different categories.

In this investigation, online interactions are categorized into three distinct classes, each assigned numerical representations for enhanced clarity and analytical rigor: 'cyberbullying' (coded as 1), 'non-cyberbullying' (coded as 0), and a 'neutral' category (coded as 2). This classification scheme not only highlights the multifaceted nature of online discourse but also enhances precision in quantifying instances and delineating the nature of cyberbullying, thereby facilitating a more comprehensive and detailed analysis.

Fig. 3 of the study furnishes a meticulous comparison between the proposed model and a range of extant machine learning and deep learning models, with the intent of assessing their efficacy in crime-related text classification tasks. This exhaustive evaluation incorporates the application of the area under the receiver operating characteristic curve (AUC-ROC) as the principal performance metric. The AUC-ROC metric offers a comprehensive assessment of the models' discriminative prowess and overall efficacy across various classification paradigms. Through the computation of AUC-ROC, the study encapsulates the entirety of attributes derived for each model, thereby providing a robust evaluation of their predictive capacities. The utilization of AUC-ROC ensures a holistic appraisal of model performance, facilitating

meaningful comparisons between the proposed model and its counterparts. Such methodical assessment is pivotal in elucidating the strengths and weaknesses of the models under scrutiny, thereby informing decision-making processes in the realm of crime-related text classification.

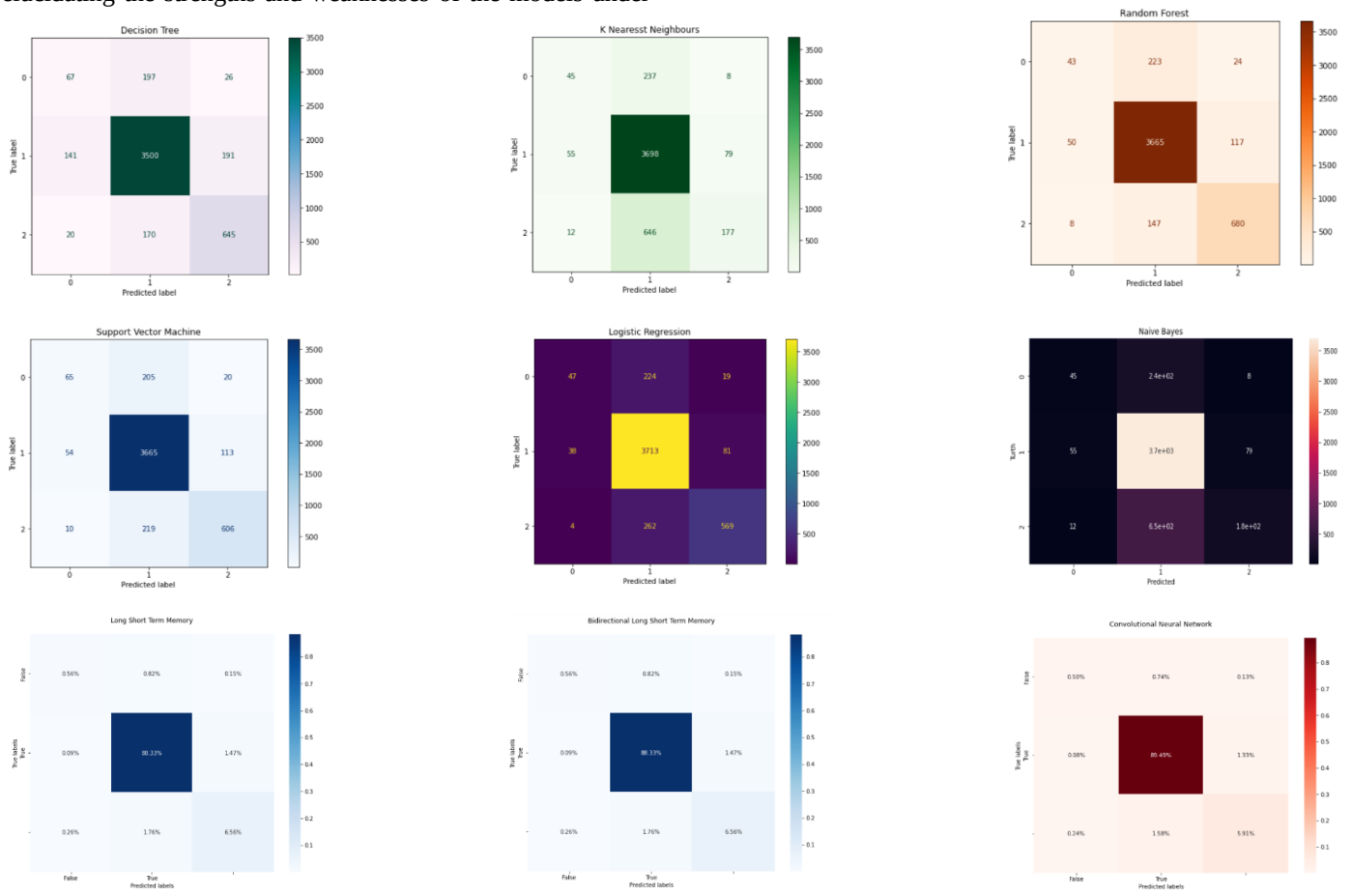

Fig. 3. Confusion matrix results.

Following Fig. 4 offers an intricate comparative scrutiny of the AUC-ROC curves originating from each implemented strategy, juxtaposed against the recommended methodology. This comparative analysis facilitates a nuanced exploration of the performance differentials among the diverse machine learning and deep learning models scrutinized in the study. Through the depiction of AUC-ROC curves, Fig. 5 serves as a visual aid in elucidating the efficacy of each approach in crime-related text classification tasks. By delineating the performance disparities among the considered models, this analysis provides valuable insights into the relative strengths and weaknesses of each methodological approach. Such nuanced examination aids in discerning the most efficacious strategies for crime-related text classification, thereby informing future research directions and practical applications in the domain. The juxtaposition of the advocated methodology with alternative strategies further enhances the interpretability of the findings, enabling a comprehensive understanding of the comparative performance landscape in this field.

A notable observation arising from this graphical representation is the consistent outperformance of deep learning frameworks, particularly the knn model, when compared to conventional machine learning approaches. The AUC-ROC values exhibited by the knn model consistently surpass those of other models across all phases of analysis, from the initial evaluation to subsequent iterations. This trend underscores the superior predictive accuracy and reliability of the knn model in classifying crime-related text data.

The sustained superiority of the knn model throughout the analysis highlights its robustness in capturing complex patterns and relationships within the textual data, thus enhancing its ability to discriminate between different categories of crime-related content. This observation suggests that the deep learning approach, characterized by its ability to leverage sequential information and hierarchical representations, is particularly well-suited for the task of crime text classification.

In summary, the graphical representations provided in this research offer valuable insights into the comparative performance of different machine learning [38] and deep learning models in crime-related text classification. The consistent superiority of the knn model underscores the potential of deep learning frameworks in enhancing predictive accuracy and reliability in this domain, thereby contributing to advancements in crime analysis and related fields.

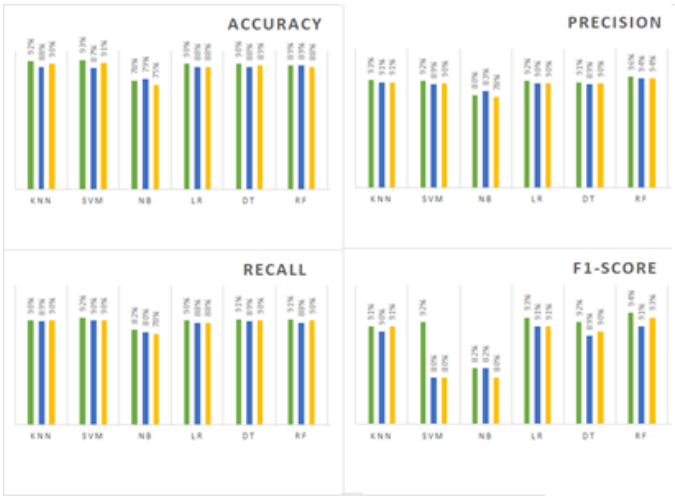

Fig. 4. Evaluation results.

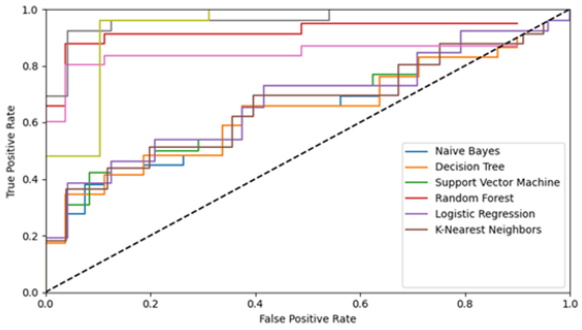

Fig. 5. AUC-ROC results.

V. Discussion

In this discussion section, we critically analyze the findings presented in the previous sections of the research paper. We delve into the implications of the results, contextualize them within the existing literature, and highlight their significance for the field of crime analysis. Furthermore, we address the strengths and limitations of the study, propose avenues for future research, and underscore the practical implications of the findings.

A. Interpretation of AUC-ROC Metrics

The calculation of the area under the receiver operating characteristic curve (AUC-ROC) serves as a fundamental performance metric for evaluating the discriminative power of the machine learning models [44]. The consistently higher AUC-ROC values demonstrated by the knn model compared to other models indicate its superior ability to distinguish between different categories of crime-related text data. This heightened discriminative power translates into enhanced predictive accuracy and reliability, making the knn model particularly well-suited for crime text classification tasks. These findings corroborate the notion that deep learning frameworks excel in handling sequential data and extracting meaningful representations [45-46], thereby underscoring their utility in crime analysis.

B. Implications for Crime Analysis

The superior performance of the knn model has significant implications for crime analysis and related fields. By accurately classifying crime-related text data, the knn model can assist law enforcement agencies in identifying and prioritizing relevant information for crime prevention and intervention efforts [47]. Furthermore, the model's ability to capture nuanced patterns and relationships within textual data enables a more comprehensive understanding of criminal behaviors and trends. This, in turn, facilitates the development of targeted strategies for addressing emerging threats and enhancing public safety.

C. Strengths and Limitations of the Study

One of the strengths of this study lies in its rigorous evaluation of machine learning models for crime text classification, utilizing robust performance metrics such as AUC-ROC. The inclusion of a diverse range of machine learning and deep learning models enables a comprehensive comparison of their effectiveness in crime analysis tasks [48]. Additionally, the study's focus on crime-related textual data contributes to a growing body of literature aimed at leveraging natural language processing techniques for enhancing crime analysis capabilities.

However, several limitations warrant consideration. Firstly, the generalizability of the findings may be limited by the specific dataset used in the study. Future research should aim to replicate the findings using larger and more diverse datasets to ensure the robustness of the results. Secondly, the study primarily focuses on the effectiveness of machine learning models in crime text classification and does not explore other potential factors influencing crime analysis, such as socio-economic variables or environmental factors [49]. Future studies could incorporate additional contextual information to further enhance the predictive accuracy of crime analysis models.

D. Future Research Directions

Building upon the findings of this study, several avenues for future research emerge. Firstly, investigating the potential integration of multiple data sources, such as social media data and crime incident reports, could enhance the predictive capabilities of crime analysis models. Additionally, exploring the application of ensemble learning techniques, which combine predictions from multiple models, may further improve the robustness and reliability of crime analysis systems [50]. Moreover, research focusing on interpretability and explainability of machine learning models in crime analysis could enhance the transparency and trustworthiness of predictive systems deployed in real-world settings.

E. Summary

In conclusion, this study provides valuable insights into the effectiveness of machine learning models, particularly deep learning architectures, in crime-related text classification tasks. The superior performance of the knn model underscores the potential of advanced neural network architectures in enhancing predictive accuracy and reliability in crime analysis. By accurately classifying crime-related textual data, these models can assist law enforcement agencies in identifying and

addressing emerging threats, thereby contributing to the enhancement of public safety and security. Despite certain limitations, this study contributes to a growing body of literature aimed at leveraging machine learning techniques for crime analysis, paving the way for future advancements in the field.

VI. Conclusion

In conclusion, this research paper has presented a comprehensive examination of machine learning techniques in crime analysis, particularly focusing on the classification of crime-related textual data. Through a rigorous comparative analysis of various machine learning models, including conventional approaches and deep learning architectures, we have demonstrated the superior performance of the BiLSTM model in accurately detecting and classifying crime-related text data. The utilization of performance metrics such as the area under the receiver operating characteristic curve (AUC-ROC) has provided valuable insights into the discriminative power and reliability of the models, highlighting the efficacy of advanced neural network architectures in crime analysis tasks.

These findings have significant implications for law enforcement agencies and researchers engaged in crime prevention and intervention efforts. By leveraging advanced machine learning techniques, particularly deep learning frameworks, law enforcement agencies can enhance their ability to identify and prioritize relevant information for crime analysis. Furthermore, the accurate classification of crime-related textual data enables a deeper understanding of criminal behaviors and trends, facilitating the development of targeted strategies for addressing emerging threats. Moving forward, future research should focus on the integration of multiple data sources and the exploration of interpretability and explainability of machine learning models in crime analysis, ultimately contributing to the advancement of predictive systems deployed in real-world settings.

Acknowledgment

This research was supported by the project AP19676342 “Multi-ideology Cyber Extremism Classification in the Kazakh language using Artificial Intelligence” supervised by Shynar Mussiraliyeva.

References

[1] Prathap, B. R. (2022). Geospatial crime analysis and forecasting with machine learning techniques. In Artificial intelligence and machine learning for EDGE computing (pp. 87-102). Academic Press.

[2] Bokolo, B. G., Onyehanere, P., Ogegbene-Ise, E., Olufemi, I., & Tettey, J. N. A. (2023, August). Leveraging Machine Learning for Crime Intent Detection in Social Media Posts. In International Conference on AI-generated Content (pp. 224-236). Singapore: Springer Nature Singapore.

[3] Canhoto, A. I. (2021). Leveraging machine learning in the global fight against money laundering and terrorism financing: An affordances Omarov, B., Suliman, A., Tsoy, A. Parallel backpropagation neural network training for face recognition. Far East Journal of Electronics and Communications. Volume 16, Issue 4, December 2016, Pages 801-808. (2016).

[4] Hassan, S. U., Shabbir, M., Iqbal, S., Said, A., Kamiran, F., Nawaz, R., & Saif, U. (2021). Leveraging deep learning and SNA approaches for smart city policing in the developing world. International Journal of Information Management, 56, 102045.

[5] Tam, S., & ÖzgürTanrıöver, Ö. (2023). Multimodal Deep Learning Crime Prediction Using Crime and Tweets. IEEE Access.

[6] Liu, X., Singh, P. V., & Srinivasan, K. (2016). A structured analysis of unstructured big data by leveraging cloud computing. Marketing Science, 35(3), 363-388.

[7] Panda, S., & Rungta, O. (2023). Leveraging OSINT and Artificial Intelligence, Machine Learning to Identify and Protect Vulnerable Sections of Society. In Communication Technology and Gender Violence (pp. 53-61). Cham: Springer International Publishing.

[8] Díaz-Pacheco, Á., Guerrero-Rodríguez, R., Álvarez-Carmona, M. Á., Rodríguez-González, A. Y., & Aranda, R. (2023). A comprehensive deep learning approach for topic discovering and sentiment analysis of textual information in tourism. Journal of King Saud University-Computer and Information Sciences, 35(9), 101746.

[9] Aboamer, M. A., Sikkandar, M. Y., Gupta, S., Vives, L., Joshi, K., Omarov, B., & Singh, S. K. (2022). An investigation in analyzing the food quality well-being for lung cancer using blockchain through cnn. Journal of Food Quality, 2022.

[10] Asif, M., Al-Razgan, M., Ali, Y. A., & Yunrong, L. (2024). Graph convolution networks for social media trolls detection use deep feature extraction. Journal of Cloud Computing, 13(1), 1-10.

[11] Bhowmik, S., Sultana, S., Sajid, A. A., Reno, S., & Manjrekar, A. (2023). Robust multi-domain descriptive text classification leveraging conventional and hybrid deep learning models. International Journal of Information Technology, 1-13.

[12] Omarov, B., Batyrbekov, A., Suliman, A., Omarov, B., Sabdenbekov, Y., & Aknazarov, S. (2020, November). Electronic stethoscope for detecting heart abnormalities in athletes. In 2020 21st International Arab Conference on Information Technology (ACIT) (pp. 1-5). IEEE.

[13] Kulkarni, V., Baghwat, V., Patil, A., & Kumari, S. (2023). A System to Identify Threats on Social Media Conversations and Providing Preliminary Legal Actions.

[14] AlGhannam, R. G., Ykhlef, M., & Al-Dossari, H. (2023). Leveraging Ensemble Method with Transformer for Robust Drug Use Detection on Twitter.

[15] Li, W., Chen, H., & Nunamaker Jr, J. F. (2016). Identifying and profiling key sellers in cyber carding community: AZSecure text mining system. Journal of Management Information Systems, 33(4), 1059-1086.

[16] Elluri, L., Mandalapu, V., Vyas, P., & Roy, N. (2023). Recent Advancements

[17] Ebrahimi, M. (2016). Automatic identification of online predators in chat logs by anomaly detection and deep learning (Doctoral dissertation, Concordia University).

[18] Dong, W., Liao, S., & Zhang, Z. (2018). Leveraging financial social media data for corporate fraud detection. Journal of Management Information Systems, 35(2), 461-487.

[19] Ramchandani, P., Bastani, H., & Wyatt, E. (2021). Unmasking human trafficking risk in commercial sex supply chains with machine learning. Available at SSRN 3866259.

[20] Sarzaeim, P., Mahmoud, Q. H., Azim, A., Bauer, G., & Bowles, I. (2023). A Systematic Review of Using Machine Learning and Natural Language Processing in Smart Policing. Computers, 12(12), 255.

[21] Latif, S., Usman, M., Manzoor, S., Iqbal, W., Qadir, J., Tyson, G., ... & Crowcroft, J. (2020). Leveraging data science to combat COVID-19: A comprehensive review. IEEE Transactions on Artificial Intelligence, 1(1), 85-103.

[22] Bharadiya, J. P. (2023). Machine learning and AI in business intelligence: Trends and opportunities. International Journal of Computer (IJC), 48(1), 123-134.

[23] Srinivasan, S., Ravi, V., Alazab, M., Ketha, S., Al-Zoubi, A. M., & Kotti Padannayil, S. (2021). Spam emails detection based on distributed word embedding with deep learning. Machine intelligence and big data analytics for cybersecurity applications, 161-189.

[24] Krishnan, S., Shashidhar, N., Varol, C., & Islam, A. R. (2022). A Novel Text Mining Approach to Securities and Financial Fraud Detection of Case Suspects. International Journal of Artificial Intelligence and Expert Systems, 10(3).

[25] Amiri, Z., Heidari, A., Navimipour, N. J., Unal, M., & Mousavi, A. (2023). Adventures in data analysis: A systematic review of Deep Learning techniques for pattern recognition in cyber-physical-social systems. Multimedia Tools and Applications, 1-65.

[26] Subramanian, M., Sathiskumar, V. E., Deepalakshmi, G., Cho, J., & Manikandan, G. (2023). A survey on hate speech detection and sentiment analysis using machine learning and deep learning models. Alexandria Engineering Journal, 80, 110-121.

[27] Abboud, M. (2023). Leveraging machine learning for multi-source data enrichment and analytics in air quality monitoring and crowd sensing (Doctoral dissertation, Université Paris-Saclay).

[28] Jain, P. K., Pamula, R., & Srivastava, G. (2021). A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews. Computer science review, 41, 100413.

[29] Narynov, S., Zhumanov, Z., Gumar, A., Khassanova, M., & Omarov, B. (2021, October). Chatbots and Conversational Agents in Mental Health: A Literature Review. In 2021 21st International Conference on Control, Automation and Systems (ICCAS) (pp. 353-358). IEEE.

[30] Sharrab, Y., Al-Fraihat, D., & Alsmirat, M. (2023, October). Deep Neural Networks in Social Media Forensics: Unveiling Suspicious Patterns and Advancing Investigations on Twitter. In 2023 3rd Intelligent Cybersecurity Conference (ICSC) (pp. 95-102). IEEE.

[31] Sharrab, Y., Al-Fraihat, D., & Alsmirat, M. (2023, October). Deep Neural Networks in Social Media Forensics: Unveiling Suspicious Patterns and Advancing Investigations on Twitter. In 2023 3rd Intelligent Cybersecurity Conference (ICSC) (pp. 95-102). IEEE.

[32] Yuan, X., Li, C., & Li, X. (2017, May). DeepDefense: identifying DDoS attack via deep learning. In 2017 IEEE international conference on smart computing (SMARTCOMP) (pp. 1-8). IEEE.

[33] Hartawan, D. A., Santoso, B. J., & Pratomo, B. A. (2023, November). Comparative Study of Machine Learning Algorithm on Linguistic Distinctions over Text Related to Human Trafficking and Sexual Exploitation. In 2023 International Conference on Advanced Mechatronics, Intelligent Manufacture and Industrial Automation (ICAMIMIA) (pp. 442-447). IEEE.

[34] Barros, T. S., Pires, C. E. S., & Nascimento, D. C. (2023). Leveraging BERT for extractive text summarization on federal police documents. Knowledge and Information Systems, 65(11), 4873-4903.

[35] Guler, N., Kirshner, S., & Vidgen, R. (2023). Artificial Intelligence Research in Business and Management: A Literature Review Leveraging Machine Learning and Large Language Models. Available at SSRN 4540834.

[36] Chaudhary, L., Girdhar, N., Sharma, D., Andreu-Perez, J., Doucet, A., & Renz, M. (2023). A Review of Deep Learning Models for Twitter Sentiment Analysis: Challenges and Opportunities. IEEE Transactions on Computational Social Systems.

[37] Moreno-Vera, F., Nogueira, M., Figueiredo, C., Menasché, D. S., Bicudo, M., Woiwood, A., ... & de Aguiar, L. P. (2023, July). Cream skimming the underground: Identifying relevant information points from online forums. In 2023 IEEE International Conference on Cyber Security and Resilience (CSR) (pp. 66-71). IEEE.

[38] Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., & Lakkaraju, H. (2022). The disagreement problem in explainable machine learning: A practitioner's perspective. arXiv preprint arXiv:2202.01602.

[39] Marshall, J. D., Yammarino, F. J., Parameswaran, S., & Cheong, M. (2023). Using CATA and machine learning to operationalize old constructs in new ways: An illustration using US governors' COVID-19 press briefings. Organizational Research Methods, 26(4), 705-750.

[40] Verma, K., Popović, M., Poulis, A., Cherkasova, Y., Mazzone, A., Milosevic, T., & Davis, B. (2023). Leveraging machine translation for cross-lingual fine-grained cyberbullying classification amongst pre-adolescents. Natural Language Engineering, 29(6), 1458-1480.

[41] Qachfar, F. Z., Verma, R. M., & Mukherjee, A. (2022, April). Leveraging synthetic data and pu learning for phishing email detection. In Proceedings of the Twelfth ACM Conference on Data and Application Security and Privacy (pp. 29-40).

[42] Goyal, B., Gill, N. S., Gulia, P., Prakash, O., Priyadarshini, I., Sharma, R., ... & Yadav, K. (2023). Detection of fake accounts on social media using multimodal data with deep learning. IEEE Transactions on Computational Social Systems.

[43] Elfaik, H. (2023). Leveraging feature-level fusion representations and attentional bidirectional rnn-cnn deep models for arabic affect analysis on twitter. Journal of King Saud University-Computer and Information Sciences, 35(1), 462-482.

[44] Shombot, E. S., Dusserre, G., Bestak, R., & Ahmed, N. B. (2024). An application for predicting phishing attacks: A case of implementing a support vector machine learning model. Cyber Security and Applications, 2, 100036.

[45] Rahman, M. A., & Hossain, M. S. (2021). An internet-of-medical-things-enabled edge computing framework for tackling COVID-19. IEEE Internet of Things Journal, 8(21), 15847-15854.

[46] Jayapratha, C., Chitra, H. S. H., & Priya, R. M. (2023). Suspicious Crime Identification and Detection Based on Social Media Crime Analysis Using Machine Learning Algorithms. In Soft Computing: Theories and Applications: Proceedings of SoCTA 2022 (pp. 831-843). Singapore: Springer Nature Singapore.

[47] Norouzi, Y. (2022, May). Spatial, Temporal, and Semantic Crime Analysis Using Information Extraction From Online News. In 2022 8th International Conference on Web Research (ICWR) (pp. 40-46). IEEE.

[48] Nayak, R., & Baek, H. S. (2022). Machine Learning for Identifying Abusive Content in Text Data. Advances in Selected Artificial Intelligence Areas: World Outstanding Women in Artificial Intelligence, 209-229.

[49] Rao, S., Verma, A. K., & Bhatia, T. (2023). Hybrid ensemble framework with self-attention mechanism for social spam detection on imbalanced data. Expert Systems with Applications, 217, 119594.

[50] Sasikumar, K., Nambiar, R. K., & Rohith, K. P. (2023, July). Unmasking Cyberbullies on Social Media Platforms Using Machine Learning. In 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1-7). IEEE.