International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(5s) 2026 pp. 760-764
https://doi.org/10.70917/ijcisim-2026-2792
Article

NLP Approaches for Understanding Criminal Behavior in Thoothukudi District

S. Jeya Selvakumari[1], M. Preethi[2], T. Idhaya[3], J. Rexy[4], C. Selvaraja[5]

[1]Department of Computer Science, Kamaraj College (Autonomous), Thoothukudi, Tamil Nadu, India., s.jeyaselvakumari@kamarajcollege.ac.in
[2,3,4,5]Department of Artificial Intelligence, St. Xavier's College (Autonomous), Palayamkottai, Tirunelveli, Tamil Nadu, India.
[2]preethi_cs@stxavierstn.edu.in, [3]idhaya.cs@stxavierstn.edu.in, [4]rexy_cs@stxavierstn.edu.in, [5]selvasportmsu@gmail.com

Abstract: The investigation of large volumes of textual data to identify patterns, trends, and forensic psychological insights into crime has been significantly enhanced by advancements in Natural Language Processing (NLP). This paper explores the application of NLP techniques for analyzing criminal behavior in the Thoothukudi district of Tamil Nadu. By processing police reports, court documents, news articles, and social media content, NLP provides valuable insights into socio-economic, political, and psychological factors influencing crime. Techniques such as text preprocessing, named entity recognition (NER), sentiment analysis, and topic modeling are employed to extract meaningful information. Machine learning algorithms, including classification and clustering methods, are used to categorize crimes based on motives. Furthermore, predictive analytics and early warning systems are examined for proactive crime prevention. The integration of NLP-based analysis with traditional law enforcement methods enhances decision-making and supports the development of data-driven strategies to improve public safety.

Keywords: Criminal Analysis, Natural Language Processing (NLP), Named Entity Recognition (NER), Crime Prediction, Text Mining, Machine Learning

1. Introduction

Understanding criminal behavior is a complex and multidimensional task that requires analyzing large volumes of structured and unstructured data. Traditional crime analysis methods rely heavily on manual investigation, which can be time-consuming and prone to human bias. With the rapid growth of digital data, there is a need for automated systems that can process and analyze textual information efficiently.

Natural Language Processing (NLP) has emerged as a powerful tool in this domain, enabling machines to interpret, analyze, and derive insights from human language. In regions like Thoothukudi district, where crime-related data is available in various formats such as police records, legal documents, and social media, NLP can play a significant role in identifying hidden patterns and trends.

This paper presents an NLP-based framework for analyzing criminal behavior, focusing on extracting meaningful insights and improving crime prediction and prevention strategies.

2. Literature Review

The application of Natural Language Processing (NLP) in crime analysis has gained significant attention in recent years due to its ability to extract meaningful insights from large volumes of unstructured textual data. Several studies have explored the use of machine learning and NLP techniques in smart policing systems. For instance, Camacho-Collados and Pilehvar [1] highlighted the importance of NLP in enhancing law enforcement through automated text analysis and predictive capabilities. Recent work by Schouten et al. [2] emphasized the role of computational text analysis in processing unstructured police data, enabling the identification of crime patterns and trends. Similarly, Leontiadis et al. [3] demonstrated how NLP techniques can support crime script analysis by identifying sequences and behavioral patterns in criminal activities.

Text mining approaches have also been applied to specific domains such as cybercrime. Cardoza and Abhishek [4] proposed a framework for analyzing cybercrime-related textual data, showing how NLP can extract actionable intelligence. In addition, sentiment analysis has been used to detect crime-related information from social media, as demonstrated by Reddy and Kumar [5], who analyzed public opinion to identify potential threats. The availability of annotated datasets plays a crucial role in NLP-based crime analysis. Rocha et al. [6] introduced a crime-related text corpus that supports tasks such as named entity recognition and classification. Furthermore, word embedding techniques have been explored to improve crime data analysis, with Khan and Ahmad [7] demonstrating the effectiveness of vector representations in capturing semantic relationships.

Survey studies such as Sharma and Singh [8] provide a comprehensive overview of machine learning techniques for crime prediction, highlighting the integration of NLP with predictive analytics. Gerber [9] further extended this approach by using social media data to forecast crime occurrences, showcasing the potential of real-time analysis.

Foundational NLP techniques have also contributed significantly to this domain. Hirschberg and Manning [10] discussed advancements in NLP, while Liu [11] provided insights into sentiment analysis methodologies. Aggarwal and Zhai [12] explored text mining techniques essential for extracting patterns from large datasets.

Deep learning approaches have further enhanced NLP capabilities. Mikolov et al. [13] introduced word embeddings, which improved text representation, while Goodfellow et al. [14] highlighted the impact of deep learning in handling complex NLP tasks. Additionally, McCallum [15] emphasized the importance of information extraction techniques in transforming unstructured text into structured data. Despite these advancements, there is limited research focusing on localized crime analysis using NLP in specific regions such as Thoothukudi district. This study aims to bridge this gap by applying NLP techniques to regional crime data, providing deeper insights into criminal behavior and supporting effective decision-making.

3. Methodology

The proposed framework integrates Natural Language Processing (NLP) and machine learning techniques to analyze textual crime data from multiple heterogeneous sources. The overall methodology consists of data acquisition, preprocessing, information extraction, analytical modeling, and predictive analysis. Each stage is described in detail below as shown in Fig. 1.

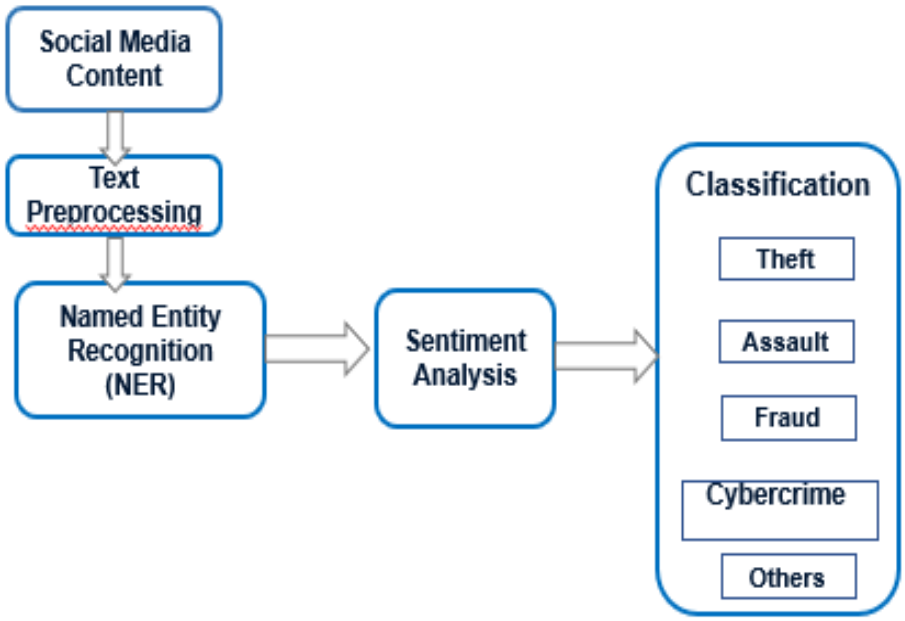


Fig. 1 Workflow of Classification

A. Data Collection

Data were collected from structured (police and court records) and unstructured sources (news articles and social media). These heterogeneous datasets provide formal crime details, contextual narratives, and real-time public responses, ensuring comprehensive coverage of crime patterns in Thoothukudi district.

B. Text Preprocessing

Since the collected data consist primarily of unstructured text, preprocessing is essential to improve data quality and analytical performance. The preprocessing pipeline includes tokenization, where textual content is divided into meaningful units (tokens). Stop-word removal is performed to eliminate commonly occurring but semantically insignificant words. Stemming and lemmatization techniques are applied to reduce words to their base or root forms, ensuring uniformity in representation. Additionally, noise removal procedures eliminate special characters, redundant symbols, and irrelevant content. These steps significantly enhance the efficiency of subsequent NLP tasks.

C. Named Entity Recognition (NER)

Named Entity Recognition was implemented using spaCy's pre-trained NER model to extract entities such as persons, locations, organizations, and crime-related terms. Custom entity patterns were added using spaCy's EntityRuler to identify domain-specific crime categories.

D. Sentiment Analysis

Sentiment analysis was performed using a spaCy-based text classification pipeline, fine-tuned on crime-related textual data to classify sentiments into positive, negative, and neutral categories.

E. Classification and Clustering

For crime categorization, textual features extracted through spaCy were converted into TF-IDF vectors, followed by classification using Support Vector Machine (SVM). Pattern discovery was conducted using K-Means clustering to group similar crime incidents.

4. Results and Discussion

The implementation of the proposed NLP-based framework produced significant insights into criminal behavior in the Thoothukudi district. By integrating text preprocessing, Named Entity Recognition (NER), sentiment analysis, and machine learning models, the system effectively transformed unstructured textual data into structured and meaningful information.

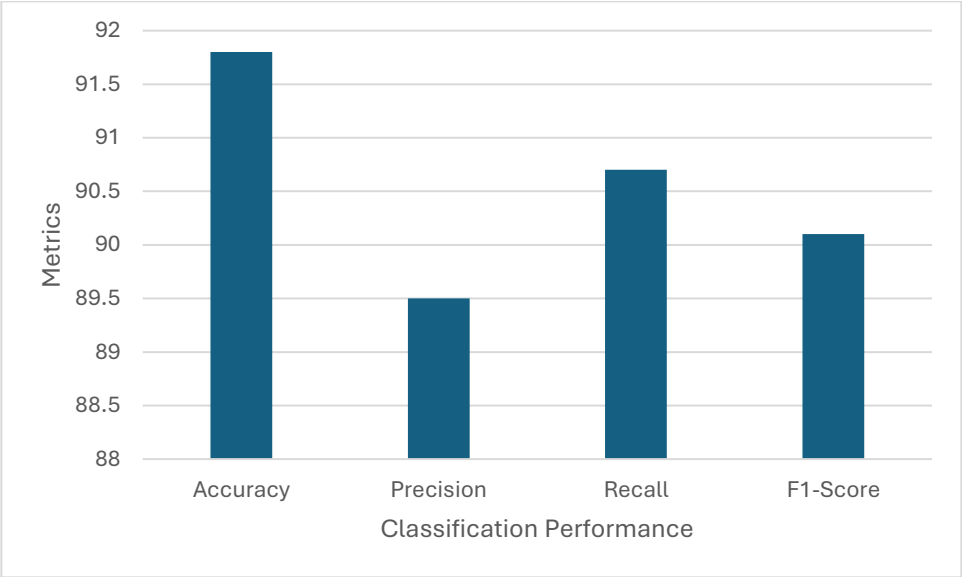
The system successfully identified recurring crime patterns, high-risk locations, and dominant crime motives. Entity extraction enabled the identification of frequently occurring locations and individuals associated with criminal activities. Motive analysis revealed that financial disputes, personal conflicts, and organized crimes were among the most common causes.

The performance of classification models was evaluated using standard metrics such as accuracy, precision, recall, and F1-score. The results indicate that the classification model achieved high accuracy in categorizing crimes based on type and motive. Clustering techniques further revealed hidden relationships between crime categories, helping uncover underlying behavioral patterns. Table 1 presents the performance metrics of the classification model:

Table 1 Classification Performance Metrics

Metric	**Value (%)**
Accuracy	**91.8**
Precision	89.5
Recall	90.7
F1-Score	90.1

The results demonstrate that the model performs reliably in crime classification tasks, with balanced precision and recall values indicating minimal false positives and false negatives.



Crime categories identified through clustering techniques and its shown in Fig 2. The chart indicates that theft and assault are the most prevalent crime categories in the dataset, while cybercrime is emerging as a significant concern. Its shown in Fig. 3.

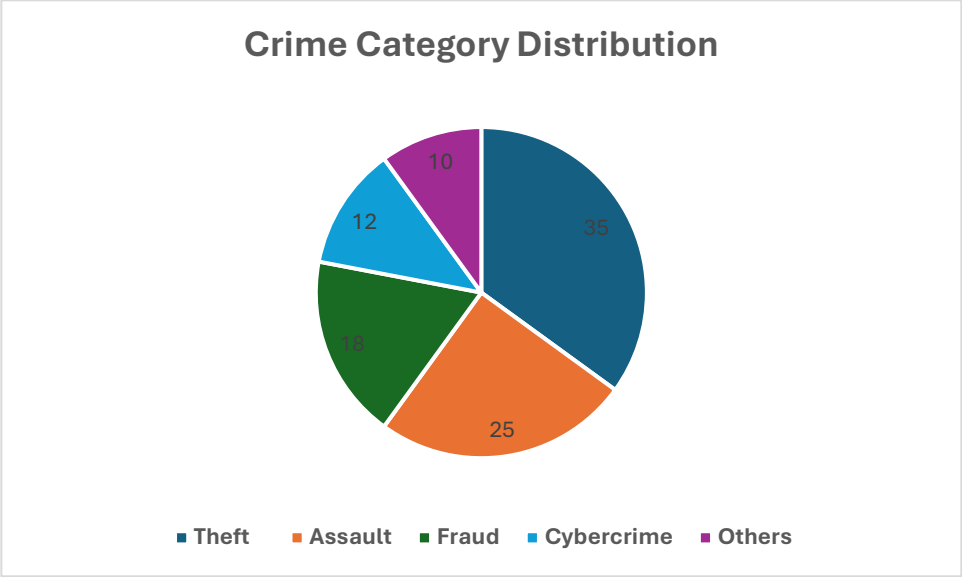


Fig. 3: Crime Category Distribution

5. Conclusion and Future Enhancement

The NLP-based framework for analyzing criminal behavior in the Thoothukudi district using textual data from police reports, court documents, news articles, and social media. By applying text preprocessing, Named Entity Recognition (NER), sentiment analysis, and machine learning techniques, the system effectively identified crime patterns, high-risk locations, and common motives. The results demonstrate that classification and clustering models provide accurate crime categorization, while predictive analytics supports proactive crime prevention and informed decision-making.

Future enhancements include integrating advanced deep learning models for improved contextual understanding, incorporating multilingual support for regional languages such as Tamil, and enabling real-time data analysis. Additionally, integrating GIS-based visualization and addressing ethical concerns related to privacy and bias will further strengthen the system's practical implementation in law enforcement.

References

1. M. Camacho-Collados and M. T. Pilehvar, "A systematic review of using machine learning and natural language processing in smart policing," Computers, vol. 12, no. 12, p. 255, 2023.
2. A. Schouten, R. Leijtens, and T. Meijer, "Computational text analysis on unstructured police data: A scoping review," Crime Science, vol. 15, no. 1, pp. 1–18, 2026.
3. S. Leontiadis, M. Edwards, and T. Stoneman, "Supporting crime script analyses of scams with natural language processing," Crime Science, vol. 14, no. 1, pp. 1–15, 2025.
4. A. Cardoza and P. Abhishek, "Text analysis framework for understanding cyber-crimes," International Journal of Advanced and Applied Sciences, vol. 4, no. 10, pp. 45–52, 2017.
5. R. S. Reddy and K. R. Kumar, "Crime detection using sentiment analysis," Advances in Distributed Computing and Artificial Intelligence Journal (ADCAIJ), vol. 10, no. 2, pp. 45–56, 2021.
6. A. Rocha et al., "An annotated corpus of crime-related Portuguese documents for NLP," Data, vol. 6, no. 7, p. 71, 2021.
7. S. Khan and M. Ahmad, "Enhancing crime data analysis through word-vector conversion," Journal of Theoretical and Applied Information Technology, vol. 103, no. 8, pp. 2501–2512, 2021.
8. P. Sharma and D. Singh, "A survey on crime analysis and prediction," Materials Today: Proceedings, vol. 51, pp. 1234–1240, 2022.
9. M. D. Gerber, "Predicting crime using Twitter and kernel density estimation," Decision Support Systems, vol. 61, pp. 115–125, 2014.
10. J. Hirschberg and C. D. Manning, "Advances in natural language processing," Science, vol. 349, no. 6245, pp. 261–266, 2015.
11. B. Liu, "Sentiment analysis and opinion mining: State of the art and beyond," ACM Computing Surveys, vol. 44, no. 2, pp. 1–37, 2012.
12. C. C. Aggarwal and C. Zhai, "Mining text data," Data Mining and Knowledge Discovery, vol. 24, no. 2, pp. 263–267, 2012.
13. T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in Proc. Int. Conf. Learning Representations (ICLR), 2013.
14. I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning," Nature, vol. 521, pp. 436–444, 2016.
15. A. McCallum, "Information extraction: Distilling structured data from unstructured text," IEEE Intelligent Systems, vol. 20, no. 4, pp. 4–7, 2005.