# A REVIEW: WEB SCRAPING AND SUMMARIZATION OF CRIME RELATED NEWS ARTICLES

Basit Ansari
*Department of Computer Engineering*
K.J Somaiya Institute of Technology
Mumbai, India

Parth Bhambure
*Department of Computer Engineering*
K.J Somaiya Institute of Technology
Mumbai, India

Ms. Pallavi Patil
*Department of Computer Engineering*
K.J Somaiya Institute of Technology
Mumbai, India

Radhey Borse
*Department of Computer Engineering*
K.J Somaiya Institute of Technology
Mumbai, India

***Abstract*—An active field of research that has received a lot of attention recently is the web scraping and summary of news stories on crimes. It has become difficult to weed out pertinent information and keep up with the most recent crime-related news due to the growing volume of data that is available online. In this review paper, we present a thorough analysis of the body of work on online scraping and summarizing news items about crimes. We start by talking about the different methods and equipment utilized for web scraping and data extraction. The approaches and summarizing algorithms used to create informative summaries of news items on crimes are then discussed. We also provide a critical assessment of the current methodologies, highlighting their advantages and disadvantages. Overall, this review paper offers a thorough summary of the challenges involved in online scraping and summarizing news items about crimes, and it might be an invaluable tool for researchers and practitioners in this subject.**

***Keywords*** *— Summarization, Crime Report, NLP, Text Extraction, Location-based*

## I. INTRODUCTION

With the increasing volume of text data generated from various sources, including news articles and social media, text mining and data mining techniques have gained attention in recent years for analyzing and extracting meaningful information related to crime. Crime-related news articles can provide valuable insights into the areas with high crime rates and assist law enforcement authorities in their efforts to curb criminal activities. Real-time crime statistics and safety scores can also help citizens make informed decisions regarding their safety.

Various studies have been conducted on the use of text mining and data mining techniques for crime-related information extraction and real-time crime statistics provision. These studies have proposed automated systems that crawl online news documents and mine data to provide an overview of crime in a specific location. The systems process data from various sources, including news articles and social media, using data mining techniques to extract and analyze crime-related information, such as the type and location of crimes.

This paper provides a review of various studies on the use of text mining and data mining techniques for crime-related information extraction and real-time crime statistics provision. It also discusses the use of Conditional Random Field (CRF) techniques with SMOTE enhancement for crime news information extraction from Indonesian news articles. Additionally, in the paper, there is a comprehensive overview of the different methods and strategies utilized in text summarization studies. These techniques comprise extractive and abstractive summarization, and the paper also covers the evaluation metrics used to assess the effectiveness of these methods. Overall, this paper highlights the potential benefits of text mining and data mining techniques in improving public safety and aiding law enforcement efforts.

## II. Literature review

The paper “Analyzing Large News Corpus Using Text Mining Techniques for Recognizing High Crime Prone Areas” [1] presents a study on the use of text mining techniques to analyze large news corpus for recognizing areas with a high propensity for crime. Since not all criminal activities are reported to law enforcement, and public access to crime records is limited, it can be challenging to obtain a complete picture of criminal activity in a given area, scanning internet news texts for crime-related information can be an effective method to extract such information. The researchers have proposed an automated system capable of crawling through online news articles and extracting data to generate an overview of criminal activity within a particular geographic area. The study demonstrates the effectiveness of this approach by processing a large dataset of news articles and providing insights into the areas with high crime rates. The proposed system can aid in curbing criminal activities and assist law enforcement authorities in their efforts. The paper provides valuable insights into the use of text mining techniques for analyzing large news corpus to extract meaningful information related to crime.

The paper “Rule-Based Crime Information Extraction on Indonesian Digital News” [2] presents a rule-based crime information extraction system that extracts crime-related information from Indonesian digital news. To enable the automated system to identify crime-related data from online news articles, the authors put forth a set of guidelines that encompass the type of crime, location, and perpetrator involved in the criminal activity. The system utilizes Natural Language Processing (NLP) techniques to parse the news articles and extract the relevant information. The extracted data is then stored in a structured format in a database, which can be used for further analysis and visualization. The authors evaluate the proposed system on a dataset of Indonesian news articles and report a high accuracy rate in extracting crime-related information. They also compare the performance of their rule-based approach with other existing methods and demonstrate its effectiveness in identifying crime-related information in Indonesian news articles.

The automated system proposed by the authors has the potential to be utilized in a variety of fields, including law enforcement, journalism, and public safety. The system can aid law enforcement agencies in their efforts to track and prevent crime by providing them with relevant crime-related information. It can also assist journalists in reporting crime-related news more efficiently and accurately. Additionally, the system can be used to provide citizens with real-time crime information and alerts, thereby improving public safety. Overall, the paper provides valuable insights into the use of rule-based approaches for crime information extraction from digital news articles and highlights the potential benefits of such systems in improving public safety and aiding law enforcement efforts.

Veronika C. M. and D. S. Naga presented a research paper on "Conditional Random Field for Crime News Information Extraction with Enhancement of SMOTE" [3]. The objective of the authors was to extract crime-related information from news articles in Indonesia, which they achieved by utilizing Conditional Random Field (CRF) techniques, along with SMOTE (Synthetic Minority Over-sampling Technique) enhancement. The CRF model was designed to classify the text into predefined categories such as time, location, perpetrator, and victim. The SMOTE technique was applied to overcome the imbalanced data problem in the dataset.

The authors tested the CRF model with SMOTE enhancement on a corpus of 100 news articles related to crimes. The model showed an average precision of 87.67%, an average recall of 88.29%, and an average F1-score of 87.97%. The findings of the study indicated that the incorporation of SMOTE enhancement in the CRF model resulted in a significant improvement in performance, as compared to the baseline CRF model. The study is relevant to the field of crime analytics and can be useful for law enforcement agencies to extract useful insights from crime-related news articles. The authors plan to extend their work by improving the accuracy of the CRF model and testing it on a larger corpus of crime-related news articles.

In their paper "Techniques and Research in Text Summarization - A Survey," [4] M. Shinde, D. Mhatre, and G. Marwal provide an overview of various techniques and approaches used in text summarization research. The authors discuss the importance of text summarization in today's information age, where we are inundated with large volumes of text data. In the paper, the authors explore different approaches to text summarization, which include extractive and abstractive summarization methods. Extractive summarization aims to extract and merge the most crucial phrases or sentences from the source text to create a summary, while abstractive summarization generates novel sentences that encapsulate the essential information from the original text. The authors also explore various evaluation metrics used to assess the quality of summarization, such as ROUGE, BLEU, and F-measure. They discuss the limitations and challenges associated with current summarization techniques, including the difficulty in capturing the context and the need for further improvement in summarization accuracy. Overall, the paper provides a comprehensive survey of current research in text summarization and highlights the importance of summarization in processing and understanding large volumes of text data. It also provides insights into the strengths and weaknesses of current techniques and the potential for future developments in this field

In the field of crime analysis, it is essential to have access to reliable and timely information. However, police departments often face limitations in their data collection and analysis capabilities. To tackle this problem, Y. Norouzi presented a new approach for carrying out spatial, temporal, and semantic analysis of criminal activities by extracting information from online news sources. The proposed method includes four main stages: data collection, preprocessing, information extraction, and analysis. The collected data is preprocessed to remove irrelevant information and enhance the quality of the remaining data. Following that, natural language processing techniques are employed during the information extraction phase to extract pertinent details, such as the kind of crime committed, the location of the incident, and the time it occurred. Subsequently, the extracted data is scrutinized to determine crime patterns and trends.

The proposed method was evaluated using a dataset of crime-related news articles from different sources. The results showed that the method was effective in extracting relevant information and identifying crime patterns. The proposed approach offers an efficient and scalable solution for crime analysis, particularly in areas where police departments face data limitations. It has the potential to aid law enforcement agencies in their efforts to identify and prevent crime, ultimately leading to safer communities

In the paper "Automated Text Summarization and Topic Detection on News Aggregation System Using BART and SVM", [8] F. Octavianus, A. Wihardi, M. K. Ario and D. Suhartono, the authors suggest an automated system for text summarization and topic detection, which can be implemented on a news aggregation platform. The system leverages two natural language processing techniques, BART and SVM, to extract significant information from news articles and produce a summary. To implement the system, the authors collected a dataset of news articles from various sources and preprocessed the data to remove noise and irrelevant information. They then used BART to generate a summary of each article and SVM to detect the topic of the article. To assess the system's efficacy, the authors employed several performance metrics, such as precision, recall, and F1-score. The outcome of the evaluation demonstrated that the proposed system could create precise summaries and detect topics with a high degree of accuracy. The authors

conclude that their system has the potential to improve the efficiency and effectiveness of news aggregation platforms, as it can quickly summarize and categorize large volumes of news articles. They suggest that future research can explore the use of other NLP techniques and datasets to further improve the system's performance.

In their paper "Recent Progress on Text Summarization," [7] S. Alhojely and J. Kalita discuss recent advancements in text summarization, focusing on the use of deep learning models for extractive and abstractive summarization. The authors highlight the importance of summarization in dealing with large volumes of text data and the need for accurate and efficient summarization techniques. The paper delivers a comprehensive review of a variety of deep learning models employed in text summarization, which includes convolutional neural networks (CNN), recurrent neural networks (RNN), and transformer models. The authors discuss the strengths and weaknesses of these models and provide examples of their application in summarization tasks.

Additionally, the paper examines the evaluation metrics employed to evaluate the quality of summarization and the difficulties encountered in generating accurate and coherent summaries. The authors emphasize the necessity of enhancing the contextual comprehension of text and the capability to produce summaries that resemble those created by humans. Overall, the paper provides valuable insights into the recent advancements in text summarization using deep learning models. It highlights the potential for further developments in this field and the importance of accurate and efficient summarization techniques in processing and understanding large volumes of text data.

The article titled "Natural Language Processing based on Semantic Inferentialism for Extracting Crime Information from Text" [6]. The paper's authors, V. Pinheiro, T. Furtado, T. Pequeno, and D. Nogueira, introduced a new natural language processing technique that enables the extraction of crime-related information from text. The proposed method leverages a semantic inferentialism approach to scrutinize the relationships between words and phrases in a given text. This approach enables the identification of significant entities and concepts that relate to crimes, such as the perpetrator, victim, location, and type of crime. Furthermore, the authors employed a statistical analysis technique to assess the relevance of each identified entity to the crime discussed in the text.

The authors tested their approach using a dataset of news articles and achieved promising results with a high degree of accuracy in identifying crime-related entities and concepts. The authors believe that their approach has the potential to enhance the capabilities of law enforcement agencies and assist in the analysis of crime-related data. In conclusion, the authors introduced a novel natural language processing technique that relies on semantic inferentialism to extract crime-related information from text. Their approach was tested on a dataset of news articles and showed promising results in identifying important entities and concepts related to crimes. The authors believe that their approach can be useful in enhancing the capabilities of law enforcement agencies in analyzing crime-related data.

The paper titled "Spatial, Temporal, and Semantic Crime Analysis Using Information Extraction From Online News" [5]. The author of the paper, Y. Norouzi, presented a method for analyzing crime-related data using information extracted from online news articles. The proposed method utilizes natural language processing techniques to extract various crime-related information from news articles, including but not limited to location, time, type of crime, and involved parties. The extracted information is then used to perform spatial, temporal, and semantic analysis of the crime data. The author also used a map visualization technique to display the crime data on a geographical map.

The author tested their approach using a dataset of online news articles related to crimes in Tehran, Iran. The results of the analysis showed that the proposed method can effectively extract crime-related information from online news articles and can be used to perform spatial, temporal, and semantic analysis of the data. In conclusion, the author presented a method for analyzing crime-related data using information extracted from online news articles. The proposed method utilizes natural language processing techniques and map visualization to perform spatial, temporal, and semantic analysis of the data. The results of the analysis on a dataset of online news articles related to crimes in Tehran, Iran showed the effectiveness of the proposed method in extracting crime-related information and analyzing the data.

## III. REVIEW ANALYSIS

| Sr. No. | Paper Name | Takeaways |
|---|---|---|
| 1 | “Analyzing Large News Corpus Using Text Mining Techniques for Recognizing High Crime Prone Areas” [1] | ● The news corpus is used in this work to detect information on crimes using a variety of natural language processing techniques, including topic modeling and named entity recognition.<br>● It uses a dataset of crime news articles from the Times of India newspaper for the period of 2010-2019. |
| 2 | “Rule-Based Crime Information Extraction on | ● The rule-based technique entails creating a set of |

| | | |
|---|---|---|
| | Indonesian Digital News" [2] | guidelines for extracting crime-related data from news items, such as incident categories, locations, and suspects.<br>● The paper applies various natural language processing techniques, such as part-of-speech tagging and named entity recognition, to the news article dataset for identifying crime-related information |
| 3 | "Conditional Random Field for Crime News Information Extraction with Enhancement of SMOTE" [3] | ● The approach involves using Conditional Random Field (CRF) to extract crime-related information, such as crime types, locations, and suspects, from news articles, and using Synthetic Minority Over-sampling Technique (SMOTE) to improve the performance of the classification.<br>● The paper applies various natural language processing techniques, such as tokenization, part-of-speech tagging, and named entity recognition, to the news article dataset for identifying crime-related information. |
| 4 | "Techniques and Research in Text Summarization - A Survey" [4] | ● The paper provides a detailed description of various techniques used in extractive summarization, including statistical methods, graph-based methods, and machine learning-based methods.<br>● The paper then explains the various methods used in abstractive summarization, including natural language generation and neural network-based models. |
| 5 | "Spatial, Temporal, and Semantic Crime Analysis Using Information Extraction From Online News" [5] | ● The paper proposes a framework for spatial, temporal, and semantic crime analysis using information extraction from online news.<br>● The framework involves using natural language processing techniques, such as named entity recognition, relation extraction, and event extraction, to extract crime-related information, such as crime types, locations, and time, from online news articles. |
| 6 | "Natural Language Processing based on Semantic inferentialism for extracting crime information from text" [6] | ● The paper proposes a natural language processing system based on semantic inferentialism for extracting crime-related information from text.<br>● The paper describes the three main components of the system: a lexical analyzer, a semantic analyzer, and an inference engine. |
| 7 | "Recent Progress on Text Summarization" [7] | ● Extractive methods involve selecting important sentences or phrases from the original text and presenting them in a summary. Abstractive methods involve generating a summary using natural language generation techniques.<br>● Evaluation metrics for text summarization include ROUGE, BLEU, and METEOR. |
| 8 | "Automated Text Summarization and Topic Detection on News Aggregation System Using BART and SVM" [8] | ● The system uses BART (Bidirectional and Auto-Regressive Transformers) for summarization and SVM (Support Vector Machines) for topic detection<br>● The BART model is fine-tuned using a combination of supervised and unsupervised learning. The supervised learning involves training the model on a dataset of news article summaries, while the unsupervised learning involves training the model on the entire dataset of news articles. |

## IV. CONCLUSION

This review paper provides a comprehensive analysis of the field of web scraping and summarization of crime-related news articles. It discusses the various challenges involved in extracting pertinent information from the vast amount of data available online and presents several techniques and algorithms proposed in the literature for addressing these

challenges. This paper suggests the need to design a system that can gather crime-related news based on the user's location and present it in a summarized and easily accessible format, which holds great promise in enhancing crime prevention efforts and increasing public awareness. Overall, web scraping and summarization of crime-related news articles have significant implications for law enforcement agencies, journalists, and the general public. This review paper aims to serve as a useful reference for researchers and practitioners in the field and encourages further research and development in this important area.

## V. FUTURE WORK

In the coming years, there is significant potential for the creation of a location-based system that collects and condenses crime-related news articles, which could significantly improve public awareness and efforts in crime prevention. It could be integrated with social media platforms and mobile applications to create a more individualized and comprehensive experience for users. One possible direction for future research is the development of more sophisticated algorithms that could enhance the precision and pertinence of the information extracted. This could involve utilizing advanced machine learning techniques and natural language processing tools to better recognize and extract pertinent information.

## • ACKNOWLEDGMENT

Basit Ansari, Parth Bhambure, Radhey Borse contributed equally for this paper and were guided by Prof. Pallavi Patil, K.J. Somaiya Institute of Technology.

## • REFERENCES

[1] S. Mukherjee and K. Sarkar, "Analyzing Large News Corpus Using Text Mining Techniques for Recognizing High Crime Prone Areas," 2020 IEEE Calcutta Conference (CALCON), Kolkata, India, 2020, pp. 444-450, doi: 10.1109/CALCON49167.2020.9106554.

[2] F. Rahma and A. Romadhony, "Rule-Based Crime Information Extraction on Indonesian Digital News," 2021 International Conference on Data Science and Its Applications (ICoDSA), Bandung, Indonesia, 2021, pp. 10-15, doi: 10.1109/ICoDSA53588.2021.9617509.

[3] V. C. M., Veronika and D. S. Naga, "Conditional Random Field for Crime News Information Extraction with Enhancement of SMOTE," 2022 Seventh International Conference on Informatics and Computing (ICIC), Denpasar, Bali, Indonesia, 2022, pp. 1-6, doi: 10.1109/ICIC56845.2022.10007023.

[4] M. Shinde, D. Mhatre and G. Marwal, "Techniques and Research in Text Summarization - A Survey," 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2021, pp. 260-263, doi: 10.1109/ICACITE51222.2021.9404670.

[5] Y. Norouzi, "Spatial, Temporal, and Semantic Crime Analysis Using Information Extraction From Online News," 2022 8th International Conference on Web Research (ICWR), Tehran, Iran, Islamic Republic of, 2022, pp. 40-46, doi: 10.1109/ICWR54782.2022.9786256.

[6] V. Pinheiro, V. Furtado, T. Pequeno and D. Nogueira, "Natural Language Processing based on Semantic inferentialism for extracting crime information from text," 2010 IEEE International Conference on Intelligence and Security Informatics, Vancouver, BC, Canada, 2010, pp. 19-24, doi: 10.1109/ISI.2010.5484783.

[7] S. Alhojely and J. Kalita, "Recent Progress on Text Summarization," 2020 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2020, pp. 1503-1509, doi: 10.1109/CSCI51800.2020.00278.

[8] F. Octavianus, A. Wihardi, M. K. Ario and D. Suhartono, "Automated Text Summarization and Topic Detection on News Aggregation System Using BART and SVM," 2022 International Symposium on Information Technology and Digital Innovation (ISITDI), Padang, Indonesia, 2022, pp. 108-113, doi: 10.1109/ISITDI55734.2022.9944521.