Digital Forensics Analysis from Social Media Data

MICT-2000

Project Group Members:
Md. Sakib Muhtadee(2254991030)
BUM Ashif Rabbani(2254991034)
Jannatul Ferdous(2254991006)

Supervisor:
Dr. Abu Sayed Md. Mostafizur Rahaman
Professor,
Department of Computer Science and Engineering,
Jahangirnagar University.

1 Abstract

Digital forensics is defined as the process of preserving, identifying, extracting and recording computer evidence that can be used in court.It is the science of finding evidence from digital media such as computers, mobile phones, servers or networks.It provides forensic teams with the best techniques and tools to solve complex digital cases.

Block-chain is a shared, immutable ledger that facilitates the process of recording transactions and tracking assets within a business network.Assets can be tangible (house, car, money, land) or intangible (intellectual property, patents, copyrights, trademarks).Virtually anything of value can be tracked and traded on the block-chain network, reducing risk and costs for everyone involved.

NLP combines computational linguistics (rule-based human language modeling) with statistical models, machine learning, and deep learning.Together, these technologies allow computers to process human language as text or speech data and understand its full meaning, as well as the intention and feeling of the speaker or writer.

In our project, we shall collect data from the social media and analyze the data using NLP. Then we shall store the digital evidences using block chain technology. With this technology we can achieve immutable data source that can be used in the court as evidence.

2 Introduction

In the realm of digital forensics, a new avenue has emerged: harnessing social media evidence. This untapped resource holds immense potential for bolstering investigations into a variety of offenses. Yet, extracting

meaningful proof from social media is no easy feat. This venture integrates cutting-edge natural language processing (NLP) techniques and a robust blockchain framework.

NLP is central to this innovation, aiding in tasks like data collection, representation, vectorization, and classifier evaluation. This systematic approach refines data, revealing the intricate details within the social media landscape.

Complementing NLP, the system employs blockchain to fortify security. This safeguards against hacking and network breaches, ensuring the reliability and confidentiality of the data.

To showcase its prowess, the system's capabilities are demonstrated using real-world data. This synergy of NLP and blockchain illustrates its potential in processing and analyzing social media evidence.

In summary, the fusion of NLP and blockchain opens doors to effective social media investigation in digital forensics. This breakthrough promises to reshape the field, amplifying investigative precision in the digital age.

3 Project Objectives

Main objective of our project is :

- To build an NLP based digital forensics analysis tool which will work on efficient evidence collection and fake news detection from social media data.
- To enhance our tool by implementing block-chain technology to ensure data authenticity and data security.

4 Literature Review

Some existing research are based on relationship graph for individuals relationship in a social media.Where we pick the most closest a associates of a known suspect[1, 2].Such approaches ignores the content of the conversation.This approach can be highly case dependent, have lower accuracy and may not serve the purposes of digital investigation.

There are two challenges we shall face during implementing digital forensics analysis with NLP and securing forensics data with block chain. One challenge is to obtain forensics data from social media and another is digital forensics challenges in block-chain.

Dongming et. al introduces an NLP-based digital investigation platform for cyber-related cases, showcasing its superior performance over existing approaches with an F1-score of 0.65 and 80.83 precision compared to 0.51 F1-score and 0.49 precision for LogAnalysis and 0.59 F1-score and 0.58 precision for SIIMCO in a real world dataset [3].

Zeinab et. al emphasizes the use of NLP techniques and blockchain to securely analyze and process social media data, demonstrating its potential with a real-world dataset [4].

5 Methodology

The integration of NLP techniques with blockchain for digital forensic analysis process is described below[4]:

5.1 Processing Layer

Processing layer performs the following processes:

- Responsible for identifying and acquiring system input.

- Inputs sourced from social networks and incident notifications.
- Identification requires incident specification and incident boundary specification.
- Includes forensic data identification, extraction, data collection, parsing, and preservation.

5.2 Data Layer

- Forensic data becomes the input for this layer.
- Utilizes hybrid data mapping for global and local ontology.
- Stores data for further processing.

5.3 Analysis Layer

Analysis operators include:

- Correlation events and document analysis.
- Location analysis.
- Assessment of relatedness of findings.
- Frequency analysis.
- Relationship analysis.

5.4 Interface Layer

Contains various components for presenting analysis results:

- Timelines.
- Tweet cloud.

- Temporal graph.
- Interaction graph.
- Frequency chart.
- Location chart.

5.5 Knowledge Layer

Focuses on processing the relationship between the extracted dataset and its links.

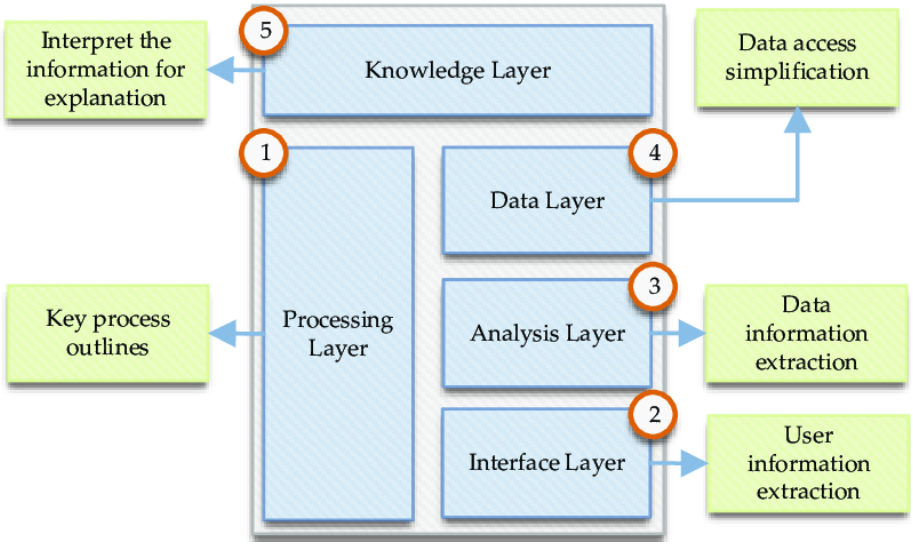

Figure 1: Multilayered Data Processing[4]

5.6 Block-chain Integration

- After NLP steps and data processing, the analyzed dataset is saved into a block-chain framework.
- Block-chain is chosen for security reasons, limiting access to prevent hacking or attacks.

- The blockchain framework encompasses data collection, investigation, and verification processes.
- Provides verified data for use in court proceedings, both for defense and prosecution.

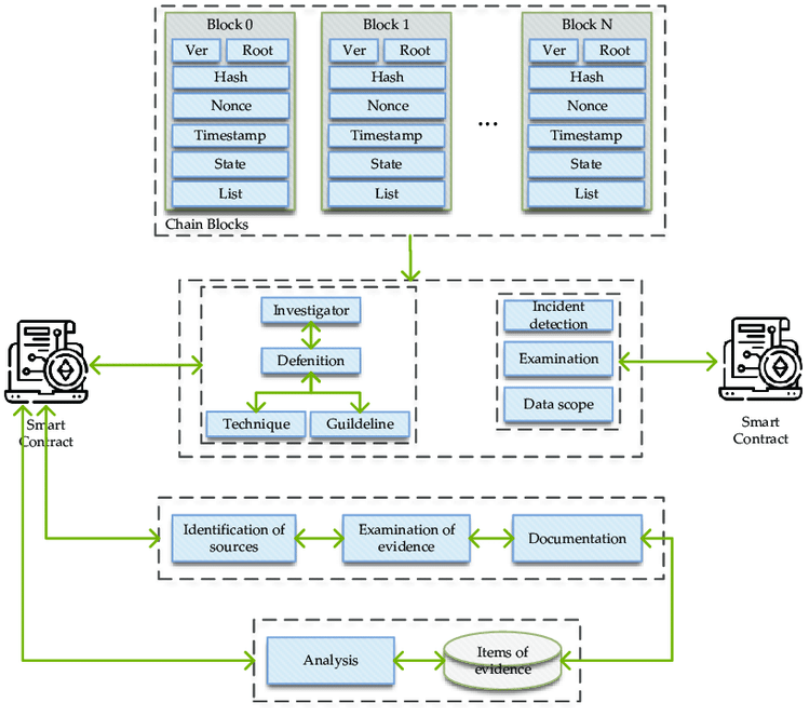

Figure 2: Block-chain Based Data Storage[4]

6 Expected Result

We are expecting rise in precision F1- score in the real world dataset.

7 Conclusion

There is a high prospect of NLP and Block Chain Technology to improve the security and performance of online digital forensics.Usually state of art language model like GPT 3.5/ 4.0 can enhance the performance of analysing forensics data.Whereas,The utilization of a blockchain framework presents a valuable avenue for securely storing and safeguarding the outcomes of digital forensic procedures, along with their associated details. In the context of future research, this system can be extended to address contemporary challenges in the realm of cybercriminal activities and fraud, offering effective solutions to these pressing issues.

References

[1] E. Lee, J. Woo, H. Kim, and H. K. Kim, "No silk road for online gamers!: Using social network analysis to unveil black markets in online games," 2018.

[2] W. Liu, D. Gong, M. Tan, J. Q. Shi, Y. Yang, and A. G. Hauptmann, "Learning distilled graph for large-scale social network data clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 7, pp. 1393–1404, 2020.

[3] D. Sun, X. Zhang, K.-K. R. Choo, L. Hu, and F. Wang, "Nlp-based digital forensic investigation platform for online communications," *Computers Security*, vol. 104, p. 102210, 01 2021.

[4] Z. hahbazi and Y.-C. Byun, "Nlp-based digital forensic analysis for online social network based on system security." *International journal of environmental research and public health*, vol. 19, p. 127027, 06 2022.