Role of Text mining in Digital Forensic Investigation- A Review

Muhammad Umar Rasheed[1], Muhammad Junaid Anjum[2], Fatima Tariq[3]

[1]University of West of the Scotland, London, B01795957@studentmail.uws.ac.uk
[2]Computer Science Department, COMSATS University Islamabad Lahore Campus, Lahore, Punjab, Pakistan
[3]Computer Science Department, Lahore College Women University, Lahore, Punjab, Pakistan

***Abstract*-** There is a noticeable increase in the ratio of crimes that are done by using computing technologies. The investigation of such crimes involves the investigation of digital evidence, data and communication that is done by the suspect's computer (Digital Forensic Investigation DFI). With the increasing storage of computer devices, the amount of data in a device has also increased [4] and it takes time to investigate every data of the computer. Using data mining techniques, the DFI can be simple and less time-consuming. Data from a suspect's computer or any digital evidence that is found at a crime scene can be grouped into the form of clusters using some keywords to classify data into the clusters. The role of this study is to present a detailed survey on the advancement of text mining for the purpose of Digital Forensic Investigation.

Introduction

Nowadays, in the world of fast computing, all things are being digitized. More things are being sold online now. The media has transformed into social media from television etc. Criminals have also used computing as a platform for committing crimes such as hacking, unauthorized access, child pornography etc. in these changing dynamics of digital crimes there is a need of some digital forensic tool for identifying and analyzing these crimes. That is why digital forensics is getting popular day by day in the field of crime investigation. But when digital devices like cell phones/smartphones, notebooks, and laptops are found at the crime scenes then the digital investigation must go through the analysis of documents stored in the suspect's devices but as storage increases, the number of documents stored too. Going through each and every document for an investigation purpose is a very time-consuming job. So, using data mining techniques for Digital Forensic Investigation has become popular area of research.

Clustering is a technique of dividing data into groups of similar objects. Clustering is used in many fields, i.e. mathematics, machine learning, and data mining and hence it's a topic of research in different fields [1]. Clustering is one of the fundamental operations of Data Mining. The most appropriate and efficient algorithm for implementing clustering is k-means [2]. As K-means algorithm is effective and efficient when the data is large and its performance increases as the number of clusters increases [3].

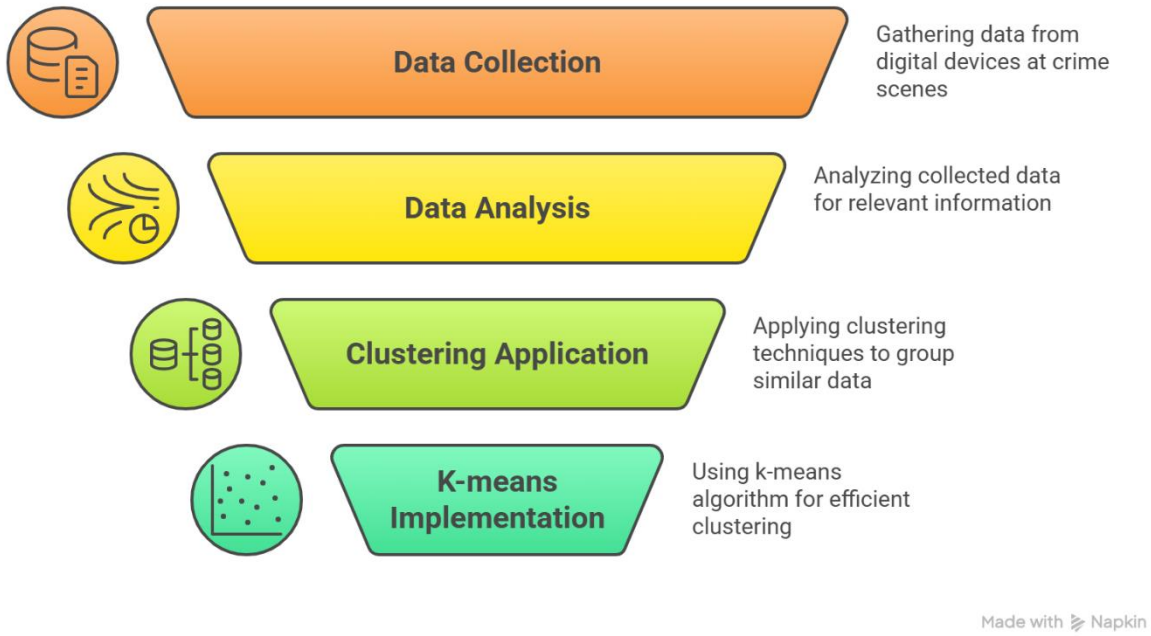

Figure 1 Streamlining Digital Forensic Investigation

The goal of this study is to provide a detailed survey on the advancements of **text mining** techniques, a sub-field of data mining, as they are applied for the purpose of Digital Forensic Investigation. Text mining, particularly through clustering, offers a promising path for efficiently analyzing the vast textual data encountered in modern digital forensic cases.

Literature Review

Tools to commit crimes and other unethical things have been spreading day by day which increased the need of Digital Investigation techniques to be more efficient [5]. This is why the review of this area has wide range of research. Selected works are described, in detail, in this article on the basis of sub-categories of this vast area.

Role of Data Mining in Crime Investigation

These days, crimes are increasing widely due to which there is need of enhanced security. In the police department, the analyzed and clustered data of criminals was stored in criminal database. In [6] author Sukanya determines the criminal tasks hotspot and find the criminals with the help of algorithms. Investigating and arresting guilty is the most challenging task because it requires comprehensive knowledge of crimes and criminals. Hotspot detection helps police to shun these activities in future. The classification is performed on the basis of crime type. Hotspot can be viewed by GIS. The application that is based on intranet should be facilitated with high security and not be accessible to unauthorized people. For protecting people effectively from crimes, criminal mapping is used which indicates where crime occurred. A Digital map helps to check the crime scene quickly. Crimes are classified on the basis of three attributes such as 1) Classification on the basis of place, 2) Classification on the basis of crime type, 3) Classification on the basis of

crime time. To place similar instances into sets the structured crime classification algorithm is used. For this purpose, data clustering algorithms can also be used. The authors claimed that the techniques presented in this paper will help in the identification of hotspot which will led in decrement of crimes.

Chen chung et al described the efficient and error free technique of mining crime data to be useful for investigation by saving the investigator time so that he can invest [7]. They gave examples of different crimes such as traffic violation, sex crimes, theft, fraud cybercrime etc. they defined data mining as powerful and useful tool of investigation as it saves cost of time and hiring personnel. Some techniques of data mining are traditional like clustering, prediction, classification, sequential pattern mining or entity extraction. The authors found the relationship between data mining techniques that applied for crime investigation of above stated crimes in a graph. On the basis of this kind of literature and Tucson police department, they presented a general framework to analyze the crime data through data mining. Their framework determines the relation between different techniques of data mining that are applied (such as traffic violation, sex crimes, theft, fraud cybercrime etc.) and intelligence and criminal analysis. They focused on entity extraction, prediction, pattern visualization and association techniques. For association and prediction clustering can be effectively used while for pattern visualization and social networks analysis can be effectively done. They implemented their framework in coplink case study to demonstrate its application. They used the coplink data for research and used clustering and association technique to reveal the identities of cybercrime.

Richard Adderley inspected the contribution of data mining in crime scene investigator performance [8]. In this paper a new technique was presented to determine the performance of crime scene investigators which used computer based unsupervised learning algorithms. Data mining offers a wide range of techniques in various areas like data visualization, neural networks and statistic etc. cross industry standard process is a cyclic technique of data mining that was used in police north. The research was started by using insightful miner which is a data mining software. Nodes were placed on the worksheet that was developed on the needed process in order to achieve results. This research recommends CSI performance modeling as a practical selection for managers.
S. V Nath discussed the usage of K-means clustering algorithm that will help in detection and identification of patterns of crimes [9]. An attribute-based weighting scheme was also developed that helps others in enhancing low enforcement officers. After implementing K-means clustering algorithm the data was being prepared for investigation. The data is then transformed into denormalized form by using extraction. Then checks are executed to check the quality of data. tectives have a look at clusters and offers their recommendations. The only limitation of this system is that it only helps detectives but can't replace crime patterns.

The author Nicole Lang Beebe narrated that the digital forensic search aims to hunt each byte of evidence to trace desired string of text [10]. Due to large amount of data, the investigator drowns into it and spends his time on irrelevant searches. For this problem there are two possible solutions available which are 1) Decrease the number of irrelevant search triumphs, 2) Present search hits in a manner which helps the investigator to determine related hits quickly. First approach was confusing for several investigators whereas the other approach was observed attractive. The present approach is based on the second solution. If some text mining activities are applied on forensic text string searches, then it will fit in with the first solution. Those activities 1) Information

Extraction: Detects conceptual associations by using syntactic outlines, 2) Content Summarization: shrinks the contents of documents, 3) Information Visualizations: graphically represents textual data, 4) Concept Linkage: recognize associations among documents.

Clustering Techniques for Forensic Investigation

Bide et al. argued that, due to remarkable escalation in documents day by day, it is very necessary to isolate the documents of suspects' computers in appropriate clusters [11]. Quicker categorization of documents is essential in forensic exploration, but examination of these documents is very challenging. The authors presented an algorithm in which it is necessary to define number of clusters, and its output is totally dependent on the provided input. The algorithm takes keywords as input and divides the document into small groups using divide and conquer technique. To resolve the issues of typical K-means algorithm, current algorithm was proposed. It uses divide and conquer, and merge sort techniques on the documents. In the preprocessing uninformative words are removed. Preprocessing is performed before applying Vector Space Model which consists of 5 steps. These steps are Filtering, Tokenization, stop word removal, Stemming and Pruning. Filtering eliminates punctuation marks from documents. Tokenization divides sentence into words. Stop Word Removal eliminates words which have no meaning. Stemming shrinks the words to its origin and Pruning eliminates the words with low frequency. VSM calculates the term frequency and number of words in a document. The input is distributed into small documents using divide and conquer to assemble parts of documents. The algorithm was tested on 20 news groups. A complete system was designed using all the models and the exactitude was measured according to F1 score. This algorithm takes less time than before. The conclusion of the experiment reveals that the F1 score is higher than existing algorithm and the time for clustering is also low.

Table 1 Algorithms and their primary applications in DFI

Technique/Algorithm	**Primary Application in DFI/Crime Investigation**	**Key Benefit**
Clustering (General)	Dividing documents into groups of similar objects to speed up DFI.	Simplifies DFI and makes it less time-consuming.
K-means	Efficiently implementing clustering, especially with large data sets.	Effective and efficient when data is large; helps in crime pattern detection.
Entity Extraction	Analyzing crime data; finding the relationship between data mining techniques and criminal analysis.	Saves investigator time and effort.
Information Extraction	Detecting conceptual associations in forensic text string searches.	Helps fit text mining into solutions for irrelevant searches.
Subject-based Semantic Clustering (SVSM)	Grouping documents into overlapping clusters based on investigator-chosen subjects.	Quicker categorization of documents and targeted investigation.
LDA + K-means	Clustering digital forensic string search output.	Enhances the precision rate of search hits (up to 4.24%).
Average Link & Complete Link	Document clustering for forensic analysis.	Proved to be the most appropriate clustering algorithms for this domain.

Nassif et al. proposed a method in which document clustering algorithm to forensic analysis was implemented [12]. The proposed method was explained by 6 famous clustering algorithms i.e. (K-means, Average Link, SingleLink, Complete Link, K-medoids, and CSPA) which were implemented to 5 databases. In the paper an algorithm was proposed which uses two indexes for validity which estimates about the total number of clusters. These experiment reveals that Average and Complete Link algorithm are most appropriate for this domain. Whenever any related document is found, the forensic examiner performs the analysis on the documents that belong to the area of interest. The algorithms mentioned above were executed by using the combination of its parameters. The execution of algorithm results in 16 vibrant algorithmic instances. The algorithms which were based on the SOM, clusters files on the basis of keyword search. Relevant application domain groups emails by using structural, lexical and semantics. Some preprocessing steps were taken before the execution of clustering algorithms. In present model every document is denoted by a vector which contains a frequency of word occurrences whose quantity was from 4 to 25. To determine the gap among documents they use cosine-based distance and Shtein distance. To predict the total number of clusters, an approach was used which takes partitions of dataset along with vibrant clusters after that. The required partitions were selected. The K-mediods algorithm is same as K-means. The only difference is that it computes mediods but not centroids. Hierarchal algorithms were executed and assessed each portion from dendrograms by silhouette. If the selected. If the selected portion is singleton object, then the process of clustering repeated recursively. The only limitation of this system can be its scalability. Silhouette was proved to be more precise than its modest versions. Our consequences are that using the file names with the document content information may be beneficial for cluster communal algorithms.

Sergio Decherchi et al. presented a methodology for text mining which was then applied via experiments [13]. The consequent investigation analysis was often done by time effort expensive human-based analysis. The examiner did weighted examinations on contents gathered from forensic acquisition. In this task the text data established one of the central data cradles which may involve related information. Text extraction was the procedure which produces a pool of raw files that contains text. This procedure depends on powerful tool which handles huge data. The outcome of text mining helps in discovering patterns, tracking, classifying and extracting by using vibrant algorithms like detection and tracking etc. Extraction of textual data in majority relies on two techniques i.e. Meta data based and application-based approach. In the first technique the detected files are recovered. Application based approach is when the metadata is smeared. Pieces of data are investigated with initials of header and footer. After that header and footer are investigated using same approach as data. At the end the files that are not populated with text only were investigated, text is extracted and then that text was processed by using text mining tool. After performing text mining, the documents were divided into clusters on the basis of similarity.

D. B. Skillcorn proposed ATHENS system design. ATHENS was general purpose system that elaborates huge information [14]. The ATHENS system finds terms from existing knowledge in massive data. ATHENS system starts with some terms that show user's current knowledge. By using the background knowledge of user novel contextualized queries are formed. The outcomes of these queries are clustered. After that the entire procedure is repeated. The major steps of ATHENS systems in this study are 1) Closure (It specifies the content of keyword's list. Here is the set of nouns which are rated by their significance. It also enables the users to be casual. 2) Probe (It performs the task of seeking novel information from closure. 3) Cluster (It organizes the resultant data of probe queries. The pages that don't fit in any cluster are discarded.) 4) Iterate.

The above-mentioned phases are repeated many times. The validation of their system was problematic because the people who use ATHENS are ontological, and the people suppose discrete pages in a cluster to be similar only in a way that people would ponder similar. Their proposed system was designed to overcome both issues of piggybacking. The efficiency and effectiveness of this system are shown by the corresponding results of their study.

Text and Documentation Clustering

Dagher and Fung introduced a subject-based semantic document clustering algorithm [5]. That will support the forensic investigation of the provincial police of SQ (Canada). In this digital forensic process, the document and communication of a suspect's computers are gathered in overlapping clusters on the basis of some subjects chosen by the investigator. The subject will be based on the area of interest of investigation, or which can identify any criminal activity. Their clustering solution used SVSM (Subject Vector Space Model). SVSM is used to represent documents, subjects or terms as vectors and demonstrate the relation between them. For building the definition of subject they proposed they proposed another scalable algorithm. The definition of subject will be based on initial definition in this algorithm. The initial definition is given as input from the investigator; each subject belongs to an area of crime investigation.

Thilagavathi also introduced a subject based semantic document clustering algorithm with the bisecting K-Means which can be used by investigator to group the documents into a cluster on the basis of a specific subject and a cluster of document which do not belong to any specific generic cluster [15]. They tried to increase the accuracy of this clustering technique, which has been improved by using K-Means. Bisecting K-Means is a combination of hierarchal clustering and K-means, used for the generic cluster. They also used SVSM for comparing terms with ESL Lookup, Word Net Synonym, and top frequent terms. They used Word Net to extract term synonyms and word sense disambiguation to determine the sense of words. Then they used subject semantics clustering algorithm along with bisecting K means. The authors also used F- measure, precision and recall improving the performance of clustering. The authors claim that they have made their presented technique unsupervised that there is no need for user to predetermine the subject and the clusters will made on the basis of subject.

Mascarnes et al also focused on the document clustering through which the investigator can cluster the document stored in a device found on crime scene systematically and it can provide subject suggestion for searching [16]. Their proposed DFI system can select subjects both ways either by investigation or through suggestion given by system (via subject suggestion module). Subject-suggestion Module suggests the most frequent keywords for investigators to help his further investigation so that the investigator can select subjects through suggested list or at his own. They implemented their framework using Java and NetBeans. According to the authors, subject suggestions can help investigators by selecting the appropriate query for search investigators at the same time.

Nicole L. Beebe used physical level text string search output relies on more than 2 million search triumphs that were found in about 50,000 allocated files [17]. Their study shows that LDA+ K-means uses centroid based user navigation process for best result. An experiment was conducted that used M57 patents datasets. The datasets involve daily images. Police seizure was specifically

used. A query based on 36 terms was searched over police seizure. Four clustering algorithms LDA, SOM, K-means and Kohonen, for the experiment it is necessary to sustain the quantity of clusters constantly in order to enhance cross algorithm evaluation. Average information retrieval (IR) in both cases relevant and irrelevant triumphs was presented to users. The IR becomes an important deliberation in noisy search. User navigation was simulated by cluster navigation algorithm. Cluster navigation algorithms can be a part of 2 classes which are 1) Clusters are selected randomly but documents are selected in each cluster with a sequence that starts from centroid. 2) Clusters and documents are selected randomly. They claimed that the above-mentioned four algorithms enhance the precision rate of search hits up to 4.24%.

In another study, Nicole L Beebe enhanced text mining and information fetching research to digital forensic text string procedure [18]. They used self-organizing neural networks for clustering search triumphs. As clustering is an approach that can enhance the effectiveness of IR. Algorithm of clustering were applied to textual IR as procedure of pre-retrieval to decrease search space and torecover vocabulary difference issue. Clustering text serves some unsupervised functions such as categorization of text, determining the semantic relationships and visualization. Clustering methods also play a role in communal visualization systems as human decision making is assisted by visual stimuli and humans can efficiently practice more concurrent information visually than through other senses. Clustering investigation concentrated on web field and user-level Internet surfing applications. Digital forensic text string search method contains four key steps; 1) Searching--The search repossesses all objects of the search strings irrespective of context, allocation status and data type, 2) Cluster pre-processing--The intention of Cluster Pre-Processing is to produce document routes for those documents that hold search hits from Step One. 3) Clustering--It practices a clustering algorithm to evaluate the mathematical connection of the document vectors and develop abstract clusters of documents. 4) Search hit analysis--The search hits are clustered by thematic cluster. The examiner chose a cluster and investigated the search triumphs within that cluster in the manner in which they are offered. They have shown that the examiner can rapidly analyze and estimate clusters for similarity to examine the intentions. They instantiated novel text string search procedure into a functioning search engine prototype, named dfGrouper, and matched the outcomes with two viable digital forensic search engines: EnCase and FTK. EnCase by using a string inspecting algorithm to trace all objects of the query terms.

Discussion

DFI is a complex process that cannot be simplified into a single approach. The literature presented herein confirms that data mining is being extensively applied in DFI in order to achieve effective and efficient investigations and ultimately reduce the workload of law enforcers. Data mining techniques carry out a variety of critical functions in multiple fields [20] including the field of digital forensics:

- **Crime Pattern Detection**: Data mining techniques such as clustering and classification are effectively used for identifying and detecting crime patterns. This involves identifying criminal task hotspots, which enables the police to prevent future activities.
- **Information Finding for Intelligence and Counterterrorism**: Systems such as the ATHENS system are developed to discover new and significant information from large datasets for intelligence and counterterrorism [21].

- **Text Clustering and String Searching**: The core of the majority of DFI problems is the discovery of significant information from large volumes of text documents. Text mining applications like Information Extraction, Content Summarization, Information Visualization, and Concept Linkage can be utilized in forensic text string searching to limit irrelevant discoveries and present hits in a more meaningful manner. Clustering is specifically highlighted as a means to enhance the effectiveness of information retrieval (IR) by decreasing the search space and mitigating vocabulary differences.
- **Performance Modeling**: Data mining has been applied in the performance modeling of crime scene investigators (CSI) using computer-based unsupervised learning algorithms.

Among the prevailing topics in the literature is the use of document clustering in the classification of the files on the suspect's computer. The vast number of documents demands a quicker classification method.

- **Clustering Algorithms**: Several algorithms have been tried and tested in the area of forensic analysis, including K-means, Average Link, Single Link, Complete Link, K-medoids, and CSPA. Research reveals that Average and Complete Link algorithms are best applied in this domain. Other studies have also proposed modified algorithms, such as one that uses divide and conquer and merge sort to address issues with the regular K-means with higher accuracy and less clustering time. The use of LDA + K-means with a centroid-based user navigation process has also been shown to yield good performance with the potential to enhance the precision rate of search hits.
- **Subject-Based Clustering**: A highly promising approach is subject-based semantic document clustering, where documents are grouped into overlapping clusters based on subjects or keywords defined by the researcher. The Subject Vector Space Model (SVSM) is key to this approach, as it is utilized in mapping documents and subjects onto vectors.
- **The Subject Selection Challenge**: Among the initial challenges of subject-based clustering is the manual effort required for the researcher to pre-define the subjects. This has led to the development of systems that make subject suggestions (via a Subject-suggestion Module) based on the most frequent keywords, helping the researcher select appropriate queries and save time.

Although data mining holds significant advantages, there are certain limitations in existing systems: Firstly, such systems pose incapacity to replace human judgment: One system limitation recognized is that it assists detectives but cannot fully replace human-based crime pattern analysis. For certain algorithms, such as K-medoids and hierarchical methods, scalability can be a limitation with constantly growing data volumes. Additionally, generic cluster issues can occur. Methods that group documents based on specific subjects have a tendency to create a "generic cluster" for the other documents. The generic cluster could, however, contain sensitive words or evidence that are potentially crucial for subsequent in-depth examination.

The current research reiterates the importance of text mining and clustering towards enhancing DFI through tackling the overwhelming volume of digital evidence. The literature has helped to develop the research issue for this research: the need for a more sophisticated subject-based semantic document clustering approach. In the best way forward, this ought to incorporate subject suggestions to help the researcher and, notably, a method of re-clustering the general cluster so that no potentially sensitive or key evidence is not analyzed in the in-depth investigation phase.

Conclusion

DFI is a complex task, and it is not limited to just one technique. Data mining is vastly used in DFI. It uses a number of different techniques for an effective and efficient investigation to reduce police work as described in [7]. The techniques of data mining are efficient for crime pattern detection[9][6][19], information finding for counter terrorism and intelligence [14], text clustering and string searching for finding important information hidden in the text documents [17][18][10][12] [4] [13] and forming the clusters on specific subjects for investigation [5][16][15]. The main focus of this study is about clustering of documents, found on a device of suspect, over a specific subject. This technique has been used in [5] but in this the clustering is done over the subjects given by investigator manually which is a daunting process. The problem of choosing subject by the suggestions is presented in [16] and it makes a generic cluster of all other documents. But this generic cluster can have some sensitive terms, which can be useful for further investigation. The study of literature has helped this research in refining the problem. And the direction of this research moves toward the subject-based semantic document clustering with the subject suggestion and re-grouping the generic cluster for a detailed investigation approach.

References:

[1] P. Berkhin, "Survey of Clustering Data Mining Techniques," pp. 1–56.

[2] Z. Huang, "A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining 1 Introduction 2 Categorical Data."

[3] A. Joshi, "A Review : Comparative Study of Various Clustering Techniques in Data Mining," vol. 3, no. 3, pp. 55–57, 2013.

[4] N. Beebe and G. Dietrich, "Chapter 12 A N E W P R O C E S S MODEL F O R TEXT STRING SEARCHING," vol. 242, pp. 179–191, 2007.

[5] G. G. Dagher and B. C. M. Fung, "Subject-based semantic document clustering for digital forensic investigations," *Data Knowl. Eng.*, vol. 86, pp. 224–241, 2013.

[6] S. M. Road, "Criminals and crime hotspot detection using data mining algorithms : clustering and classification," vol. 1, no. 10, pp. 225–227, 2012.

[7] J. Jie and M. Chau, "Crime Data Mining : A General Framework," *IEEE Comput. Soc.*, 2004.

[8] R. Adderley, M. Townsley, and J. Bond, "Use of data mining techniques to model crime scene investigator performance," *Knowledge-Based Syst.*, vol. 20, no. 2, pp. 170–176, 2007.

[9] S. Nath, "Crime pattern detection using data mining," *Web Intell. Intell. Agent Technol. ...*, vol. 1, no. 954, p. 4, 2006.

[10] N. L. Beebe and J. G. Clark, "Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results," *Digit. Investig.*, vol. 4, no. SUPPL., pp. 49–54, 2007.

[11] P. Bide, N. Mumbai, R. Shedge, and N. Mumbai, "Improved Document Clustering using K-means Algorithm," no. 978, pp. 0–4.

[12] L. F. Da Cruz Nassif and E. R. Hruschka, "Document clustering for forensic analysis: An approach

for improving computer inspection," *IEEE Trans. Inf. Forensics Secur.*, vol. 8, no. 1, pp. 46–54, 2013.

[13] S. Decherchi, S. Tacconi, J. Redi, A. Leoncini, F. Sangiacomo, and R. Zunino, "Text clustering for digital forensics analysis," *Adv. Intell. Soft Comput.*, vol. 63 AISC, pp. 29–36, 2009.

[14] D. B. Skillicorn and N. Vats, "Novel information discovery for intelligence and counterterrorism," *Decis. Support Syst.*, vol. 43, no. 4, pp. 1375–1382, 2007.

[15] G. Thilagavathi and J. Anitha, "Document Clustering in Forensic Investigation by Hybrid Approach," vol. 91, no. 3, pp. 14–19, 2014.

[16] S. Mascarnes and J. Gomes, "Subject based Clustering for Digital Forensic Investigation with Subject Suggestion," *{International J. Comput. Appl.*, vol. 102, no. 11, p. 8887, 2014.

[17] N. L. Beebe and L. Liu, "Clustering digital forensic string search output," *Digit. Investig.*, vol. 11, no. 4, pp. 314–322, 2014.

[18] N. L. Beebe, J. G. Clark, G. B. Dietrich, M. S. Ko, and D. Ko, "Post-retrieval search hit clustering to improve information retrieval effectiveness: Two digital forensics case studies," *Decis. Support Syst.*, vol. 51, no. 4, pp. 732–744, 2011.

[19] G. Oatley, B. Ewart, and J. Zeleznikow, "Decision support systems for police: Lessons from the application of data mining techniques to 'soft' forensic evidence," *Artif. Intell. Law*, vol. 14, no. 1–2, pp. 35–100, 2006.

[20] Shaheen, M., Awan, S. M., Hussain, N., & Gondal, Z. A. (2019). Sentiment analysis on mobile phone reviews using supervised learning techniques. *International Journal of Modern Education and Computer Science*, *10*(7), 32.

[21] Farooq MS, Gilani SMA, Manzoor MF, Shaheen M. Detecting Fake News in Urdu Language Using Machine Learning, Deep Learning, and Large Language Model-Based Approaches. *Information*. 2025; 16(7):595. https://doi.org/10.3390/info16070595