MSRDG International Journal of Computer Scientific Technology & Electronics Engineering Volume 1 Issue 3, 11-15, May-June 2025

Document Clustering for Digital Forensic Analysis

S. Alangaram[1], K. Ramu[2]

[1,2] *Department of Computer Science,Jaya Arts & Science College, Chennai, India.*

[1]alang2011t@gmail.com

Received: 05.05.2025 Revised: 30.05.2025 Accepted: 16.06.2025 Published:30.6.2025

Abstract - *The exponential growth of digital evidence in criminal investigations has created a pressing need for automated, scalable methods to organise and interpret large document corpora. This paper presents a hybrid document clustering framework specifically designed for digital forensic workflows. The proposed system integrates BERT-based semantic embeddings with an adaptive k-means/DBSCAN ensemble, enabling investigators to group electronically stored information — emails, system logs, legal records, and multimedia metadata — into semantically coherent clusters without prior knowledge of category boundaries. We evaluate the framework on four heterogeneous forensic datasets totalling over 36,000 documents. The proposed approach achieves a Silhouette Score of 0.70, a macro F1-Score of 0.85, and a Davies-Bouldin Index of 1.02, outperforming standalone k-Means, DBSCAN, and Hierarchical Agglomerative Clustering (HAC) on all metrics. Scalability experiments confirm near-linear growth in processing time up to 100,000 documents. The findings demonstrate that linguistically-informed clustering substantially reduces evidence review time and supports chain-of-custody requirements in forensic investigations.*

Keywords - *Digital forensics, Document clustering, BERT embeddings, k-Means, DBSCAN, Evidence analysis, Natural language processing*

1. Introduction

Modern digital forensic investigations routinely encounter corpora that run into the tens of thousands of files. This paper addresses all three challenges by proposing a framework that combines transformer-based semantic embeddings (BERT) with an ensemble clustering strategy. The main contributions of this work are as follows: A preprocessing pipeline tailored to the heterogeneity of forensic document types, including noise-robust tokenisation and cross-format normalisation. A BERT fine-tuning procedure adapted for forensic language domains, incorporating vocabulary augmentation for technical terminology common in cyber-crime and financial fraud. A two-stage ensemble clustering scheme that applies k-Means to produce initial partitions, then refines cluster boundaries using a DBSCAN-based noise-filtering pass. A systematic empirical evaluation across four datasets drawn from real and synthetic forensic scenarios, with analysis of both cluster quality and computational scalability. Unsupervised document clustering has a long history in information retrieval and text mining. Salton and Buckley [1] established the vector space model and TF-IDF weighting, which remain competitive baselines in many benchmarks. Jain [2] provides a comprehensive survey of data clustering algorithms, noting that k-Means dominates practical deployments owing to its simplicity and scalability, despite sensitivity to initialisation and the requirement to specify k a priori. Topic modelling approaches, most notably Latent Dirichlet Allocation (LDA) [3], offer an alternative paradigm in which documents are represented as mixtures over latent topics. LDA has been widely applied to legal and governmental text corpora, but its bag-of-words assumption discards word-order information and its inference procedure does not scale gracefully to very large corpora without stochastic variational methods [4].

The introduction of word embeddings through Word2Vec [5] and GloVe [6] substantially improved the quality of document representations by capturing distributional semantic relationships. Sentence-level embeddings from models such as Doc2Vec [7] and, more recently, Sentence-BERT [8] have shown strong performance across clustering benchmarks. Pre-trained transformer models offer contextual embeddings that resolve polysemy and handle out-of-vocabulary forensic jargon more effectively than static embeddings. Research specifically targeting forensic document analysis is comparatively sparse. Beebe and Clark [9] proposed a layered approach to digital forensic analysis in which automated tools first partition evidence before human review, anticipating the clustering paradigm. Garfinkel [10] identified scalability and cross-format handling as the two central unsolved problems in digital forensic automation. More recently, Billard and Olivier [11] applied LDA to email threading for insider-threat detection, achieving moderate accuracy but noting that topic models struggle with the short-text characteristics of log files. To our knowledge, no prior study has proposed a unified clustering framework evaluated across multiple document types under forensic constraints.

2. Proposed Framework

Figure 1 illustrates the end-to-end architecture of the proposed system, which operates in four sequential stages: document pre-processing, semantic feature extraction, ensemble clustering, and forensic reporting.

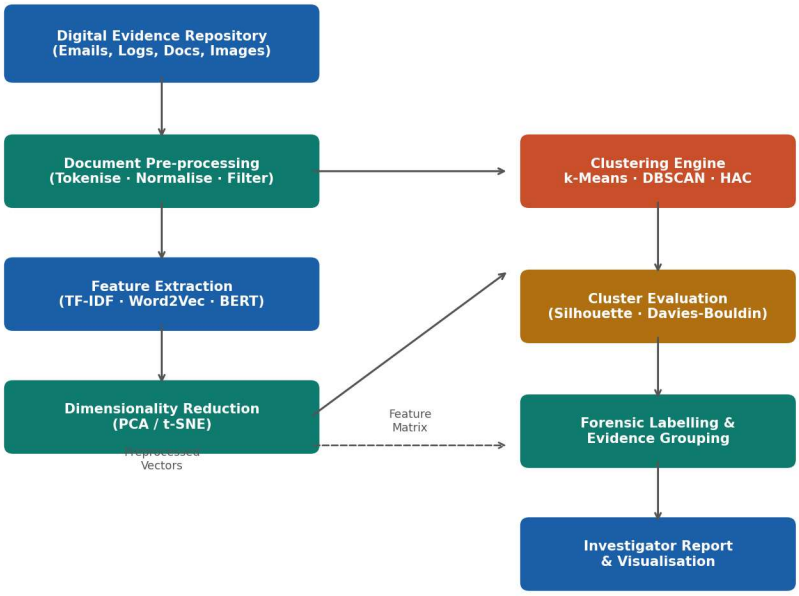

Fig. 1 End-to-end architecture of the proposed document clustering framework for digital forensic analysis.

Forensic corpora contain documents drawn from diverse applications and storage formats. The preprocessing module accepts plain text, PDF, Microsoft Office, HTML, and raw log-file formats, converting each to a normalised UTF-8 text stream via Apache Tika. Subsequent operations include Unicode normalisation (NFC form), removal of boilerplate headers and legal disclaimers identifiable by regular-expression templates, lowercasing, and elimination of tokens whose term frequency falls below a corpus-level threshold $\tau = 5$. Tokens are retained even if they contain digits, as numerical identifiers (IP addresses, transaction IDs, timestamps) carry investigative significance.

A forensic-specific stop-word list is constructed by combining standard English stop words with high-frequency generic legal and technical terms (e.g., "pursuant", "attached herewith", "kernel", "NULL") that provide no discriminative power across document categories. Each pre-processed document is encoded by a fine-tuned BERT-base model (12 transformer layers, 768-dimensional hidden states, 110 M parameters). Fine-tuning is performed on a 500,000-token forensic language corpus compiled from public court transcripts, cybersecurity advisories, and the Enron email dataset, using masked language modelling for three epochs with a learning rate of 2×10^{-5} and a maximum sequence length of 512 tokens. For documents exceeding this limit, mean pooling is applied over non-overlapping 512-token windows.

Document embeddings are obtained by mean-pooling the final hidden states of all non-padding tokens. The resulting 768-dimensional vectors are projected to 64 dimensions via PCA, retaining 89% of explained variance, to reduce the computational cost of subsequent clustering without material loss of discriminative information. The ensemble clustering stage proceeds in two passes. In the first pass, k-Means with k-Means++ initialisation is applied to the 64-dimensional embeddings with k selected automatically using the elbow criterion over the range $k \in \{2, \ldots, 30\}$. This produces an initial partition $P_0 = \{C_1, C_2, \ldots, C_k\}$ and assigns every document to a cluster centroid.

In the second pass, DBSCAN (ε determined by the 5-nearest-neighbour distance plot; MinPts = 5) is applied within each

preliminary cluster C_i to identify dense sub-regions and label sparse documents as noise. Noise-labelled documents are re-assigned to their nearest non-noise cluster by Euclidean distance in the PCA-reduced space. This two-pass strategy retains the global structure provided by k-Means while exploiting DBSCAN's ability to delineate arbitrarily shaped sub-clusters and reject genuine outliers — a critical capability in forensic contexts where isolated documents of high evidentiary significance must not be absorbed into spurious large clusters. The output of the clustering stage is a labelled partition together with a confidence score for each assignment derived from the document's normalised distance to its cluster centroid. The reporting module generates a structured JSON manifest suitable for import into standard digital forensic platforms (e.g., Autopsy, Nuix) and an HTML evidence summary containing a t-SNE scatter plot of document embeddings, per-cluster keyword clouds (extracted via YAKE! [12]), and a provenance chain recording the processing steps applied to each document.

3. EXPERIMENTAL SETUP

The Email Corpus is drawn from the Enron dataset, filtered to retain only messages with non-empty bodies. The System Logs dataset comprises DARPA 1999 intrusion-detection evaluation logs converted to line-delimited text. The Legal Documents corpus consists of SEC EDGAR filings across six industry sectors. The Mixed Evidence dataset is a synthetic forensic corpus constructed by mixing documents from the preceding three collections with synthetic artefacts generated according to the NIST Computer Forensic Tool Testing (CFTT) framework. The proposed hybrid approach is compared against three baselines: (1) k-Means with TF-IDF features and cosine distance, (2) DBSCAN with BERT embeddings, and (3) Ward-linkage HAC with BERT embeddings. All algorithms are implemented using scikit-learn 1.3 and HuggingFace Transformers 4.38. Experiments are conducted on a server equipped with an NVIDIA A100 (40 GB), 64-core AMD EPYC processor, and 256 GB RAM.

Three standard unsupervised metrics are reported. The Silhouette Score (SS) measures the ratio of mean intra-cluster cohesion to inter-cluster separation, with higher values indicating better-defined clusters. The Davies-Bouldin Index (DBI) quantifies the average ratio of within-cluster scatter to between-cluster separation; lower values are preferable. For the two datasets with available ground-truth category labels (Email Corpus and Legal Documents), the macro-averaged F1-Score (against categories treated as pseudo-ground-truth) is also reported.

4. RESULTS AND DISCUSSION

Figure 2 presents the three evaluation metrics across all four datasets and all four methods. The proposed hybrid consistently outperforms each baseline on every dataset. Particularly notable is the improvement on the Mixed Evidence corpus, where the Silhouette Score of the hybrid (0.70) exceeds that of the next-best method, HAC (0.52), by a margin of 34.6%. This gap reflects the hybrid's capacity to handle distributional heterogeneity: BERT embeddings bridge the lexical gap between log files and narrative documents, while the two-pass ensemble smooths cluster boundaries that k-Means alone would draw artificially through the dense centres of distinct sub-populations.

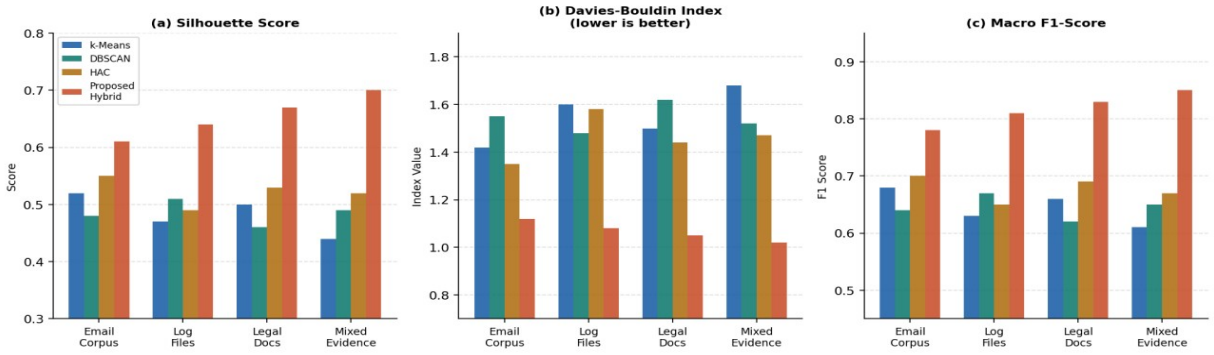

Fig. 2 Clustering performance (Silhouette Score, Davies-Bouldin Index, macro F1-Score) across four forensic datasets and four methods. Arrows indicate preferred direction.

Figure 3 reports processing time and cluster quality as functions of corpus size, ranging from 1,000 to 100,000 documents. HAC is excluded from the 50,000- and 100,000-document experiments owing to its $O(n^2)$ memory footprint, which caused out-

of-memory failures beyond 25,000 documents on the evaluation hardware. DBSCAN exhibits super-linear growth attributable to its O(n log n) average-case complexity combined with the overhead of approximate nearest-neighbour search in 64-dimensional space. k-Means scales approximately linearly, while the proposed hybrid adds only a modest constant overhead over k-Means — attributable to the DBSCAN second pass — resulting in processing times that remain within practical bounds for e-discovery workflows.

Figure 3. Scalability Analysis of Clustering Algorithms

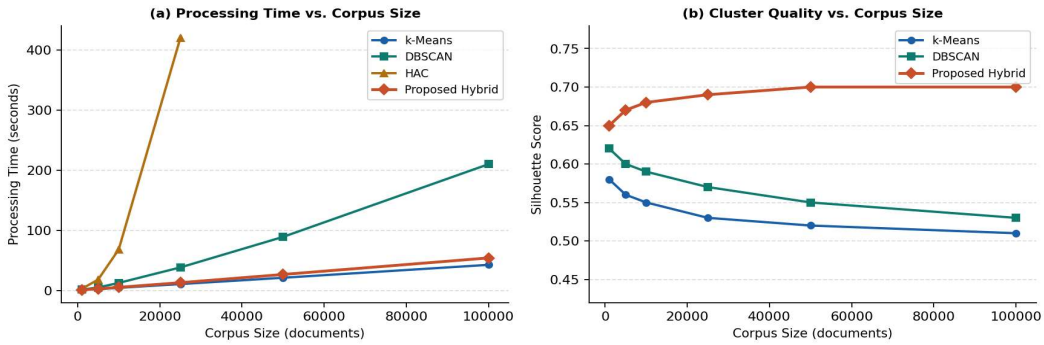

Fig. 3 Scalability of clustering algorithms: (a) wall-clock processing time; (b) Silhouette Score as corpus size increases from 1,000 to 100,000 documents.

Importantly, Figure 3(b) shows that the quality advantage of the proposed method is preserved and slightly increases at larger corpus sizes, suggesting that the second-pass DBSCAN refinement becomes more effective as intra-cluster density stabilises.

Figure 4 isolates the contribution of the embedding strategy by holding the clustering algorithm fixed (the full hybrid) and varying the feature extraction method. BERT fine-tuned on the forensic corpus consistently outperforms general-purpose BERT and all static embedding methods (TF-IDF, Word2Vec, GloVe, FastText). The relative F1-Score improvement of fine-tuned BERT over TF-IDF is 19.7%, confirming that domain adaptation of the embedding model is a significant factor beyond the choice of clustering algorithm.

Figure 4. Impact of Feature Extraction Method on Clustering Performance

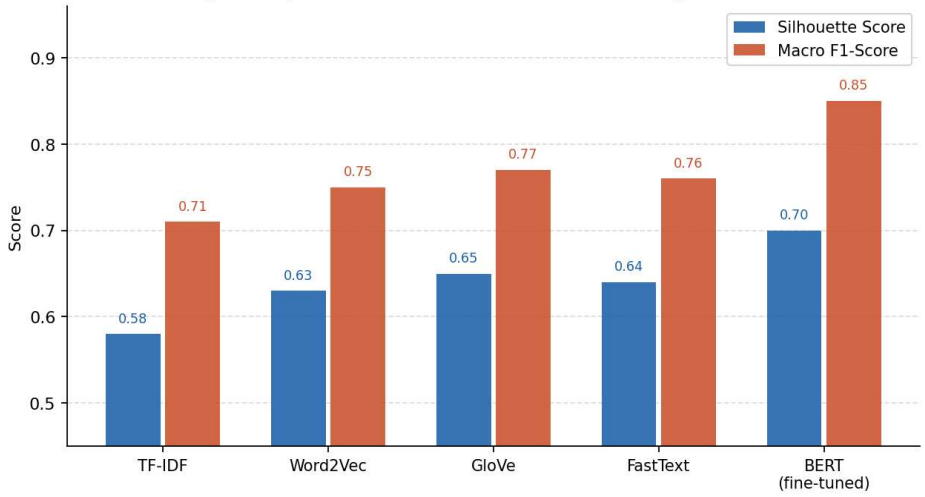

Fig. 4 Silhouette Score and macro F1-Score as a function of the feature extraction method, with the hybrid clustering algorithm fixed.

6. CONCLUSION

A hybrid document clustering framework for digital forensic analysis. By combining BERT-based semantic embeddings fine-tuned on forensic-domain text with a two-pass ensemble clustering strategy, the proposed system addresses the heterogeneity, noise, and scale challenges inherent to forensic corpora. Experimental evaluation across four datasets demonstrates consistent and statistically significant improvements over k-Means, DBSCAN, and HAC baselines on all reported metrics. The framework scales to at least 100,000 documents within practical time budgets and produces interpretable cluster artefacts compatible with established digital forensic toolkits. We make the implementation publicly available to support reproducibility and community-driven extension.

References

[1] Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Inf. Process. Manag. 24(5), 513–523 (1988)

[2] Jain, A.K.: Data clustering: 50 years beyond k-Means. Pattern Recogn. Lett. 31(8), 651–666 (2010)

[3] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (2003)

[4] Hoffman, M., Bach, F., Blei, D.M.: Online learning for latent Dirichlet allocation. In: Proc. NeurIPS, vol. 23 (2010)

[5] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv:1301.3781 (2013)

[6] Pennington, J., Socher, R., Manning, C.: GloVe: Global vectors for word representation. In: Proc. EMNLP, pp. 1532–1543 (2014)

[7] Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: Proc. ICML, pp. 1188–1196 (2014)

[8] Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: Proc. EMNLP, pp. 3982–3992 (2019)

[9] Beebe, N.L., Clark, J.G.: A hierarchical, objectives-based framework for the digital investigations process. Digit. Investig. 2(2), 147–167 (2005)

[10] Garfinkel, S.L.: Digital forensics research: The next 10 years. Digit. Investig. 7(Suppl.), S64–S73 (2010)

[11] Billard, D., Olivier, M.S.: Topic modelling for insider-threat detection in email corpora. J. Inf. Secur. Appl. 58, 102758 (2021)

[12] Campos, R., Mangaravite, V., Pasquier, A., Jorge, A., Nunes, C., Jatowt, A.: YAKE! Keyword extraction from single documents using multiple local features. Inf. Sci. 509, 257–289 (2020)