Westcliff International Journal of Applied Research
WIJAR

Westcliff International Journal of Applied Research

Vol. 10, No. 2. Summer 2026

https://doi.org/10.47670/wuwijar20262ADTDP

Smarter Search for Law Enforcement: Evaluating Context-Aware and Traditional Approaches

Archan Dutta
Westcliff University

Thaís D. Poi
San Joaquin Delta College

Abstract

Despite widespread and rapid advancements in natural language processing, most law enforcement case and report management systems still rely on keyword-based search engines. Information retrieval within law enforcement agencies depends on obtaining vast amounts of unstructured text data, including case reports, incident reports, and written statements. Investigators must sift through and analyze this data to find and discover meaningful leads. Keyword-based search engines face challenges with synonyms, equivalent terms, abbreviations, slang, paraphrases, and narrative fluctuation often found in police reports and case files. As a result, investigators may overlook potentially relevant cases unless they manually craft multiple keyword searches. This quantitative study evaluated, using a synthetic dataset generated for this study, whether context-aware models (also known as semantic models or embedding models) reach higher retrieval accuracy than traditional keyword-based search models on law enforcement reports. Results of the study showed that semantic models have significantly better performance than traditional keyword-based search, an improvement of 56 percentage points. Besides investigators, inspectors, and analysts, smarter and enhanced search may benefit other key parties like patrol officers, crime analysts, and prosecutors by allowing faster discovery of related incidents, improved case connection, and richer text analysis. These results highlight the potential to improve operational efficiency, reduce administrative and management burden, and strengthen public safety outcomes.

Keywords: semantic search, information retrieval, law enforcement, natural language processing, embedding models, crime reports, retrieval accuracy

Introduction

Law enforcement departments produce and manage massive volumes of unstructured text-based data every day, from incident reports and witness statements to investigation notes. In the majority of record-management and case processing systems, traditional keyword search continues to be the dominant search technology and may not consistently retrieve all relevant documents. Some examples of record-management systems are municipal file management systems, state repositories, and federal databases such as the National Incident-Based Reporting System and the National Crime Information Center. Traditional keyword search has limitations in terms of accurate retrieval. These limitations arise from the characteristics of

natural language, such as English, which has many narrative variations, paraphrases, interpretations, acronyms, synonyms, polysemy, slang, and other informal expressions. Reports describing the same event may vary widely in vocabulary and phrasing. For example, “car stolen from home garage,” “vehicle taken last night,” or “motor vehicle theft.”

The traditional keyword-based search models like Okapi BM25 and Term Frequency–Inverse Document Frequency (TF-IDF) are not “context-sensitive” and are based on literal and explicit word overlap between a query and a document (Robertson & Zaragoza, 2009). Keyword search models treat these as unrelated, missing vital connections between similar cases. These limitations impede investigators, supervisors, attorneys, and intelligence units from generating accurate insights from historical text data. Current advances in natural-language processing (NLP) provide an innovative alternative. Semantic search methods represent text as numerical vectors (also known as embeddings) that capture the meaning of the sentences. By measuring the similarity between these vectors, systems can retrieve conceptually related reports even when they share a small number of or no common terms. Among these models, the Sentence-BERT (SBERT) class of models has shown strong retrieval performance in the open domain, indicating potential for specialized fields such as law enforcement.

We benchmarked traditional keyword-based retrieval models (BM25, TF-IDF) against two SBERT versions (MPNet, MiniLM) on a document dataset of law enforcement reports, primarily focusing on retrieval accuracy. Through observation-based analysis of semantic and traditional search side by side, this research revealed the possibility of improving investigative workflows, intelligence analysis, and content access across all tiers of policing using context-sensitive Artificial Intelligence (AI) models. By improving search retrieval, smarter retrieval tools have the potential to shorten investigative timelines and support more effective and transparent public-safety operations.

Background

This section establishes the crucial role of information retrieval (IR) in the high-risk area of law enforcement, where efficiency and accuracy immediately impact investigations and community safety. We examined the inherent difficulties in traditional search methods, specifically dealing with the lexical gap problem. Subsequently, we detailed the transformative development of semantic search, resolving these limitations by using vector embeddings to capture the contextual meaning and interpretation of texts and queries. This provides the necessary background needed to understand traditional, keyword-based techniques and state-of-the-art, context-aware semantic methods.

Information Retrieval in Law Enforcement

The complexities of criminal analysis necessitate leveraging the intricacies found exclusively in descriptive, narrative-based data (Lovell et al., 2022; Parker, 2025). These free-text narratives consistently contain critical, subtle contextual details such as inferred officer attitudes concerning victim credibility or nuanced descriptions of perpetrator acts, which influence case results (Lovell et al., 2022; Parker, 2025). As a result, law enforcement and judicial systems increasingly depend on efficient and accurate access to immense amounts of unstructured textual data (Xing & Chen, 2024). If the retrieval/search mechanism is inaccurate, then critical qualitative information remains unseen and inaccessible. Therefore, improving retrieval performance is more than just a gradual efficiency gain but a main factor for helping analysis (Xing & Chen, 2024).

Lexical Gap Problem

Because the traditional retrieval methods are based on the exact word overlap between a query and a document, these systems are fundamentally limited by the vocabulary mismatch problem, also known as the lexical gap (Al-Haddad et al., 2025). This gap occurs when the query and the relevant document describe the same concept using different terminology, such as synonyms, paraphrasing, or misspellings (Al-Haddad et al., 2025). Table 1

contains examples that show the limitations of keyword-based models, highlighting that even though the word overlap between the police report and the query is low, they still have similar semantic meaning. The low word overlap means that keyword-based models will struggle to retrieve these documents, despite having similar semantic meaning.

Table 1

Examples of reports and queries showing limitations of a keyword-based search

Police Report	Query	Word Overlap	Semantic Relationship
The suspect broke into a parked vehicle and stole a laptop from the back seat.	Who took the computer from my car?	"stole" ≠ "took"; "laptop" ≠ "computer"; "vehicle" ≠ "car"	Both are about theft
A woman reported receiving threatening calls from an unknown number.	Has anyone complained about harassment over the phone?	"threatening calls" ≠ "harassment"	Both about phone harassment

Emergence of Semantic Search via Embeddings

In a semantic search model, text is mapped into a continuous, higher-dimensional vector space, known as embeddings (Opitz et al., 2025; Schutz & Hupfeld, 2021). In this space, similarity is measured based on semantic meaning rather than exact word form, typically using cosine similarity (Reimers & Gurevych, 2019). The cosine similarity measures semantic similarity; how similar the meanings of two sentences are. A higher cosine similarity score (closer to 1) means that the sentences have a more similar meaning, while a lower score (closer to 0) means they are semantically different than each other. This approach is widely used in natural language processing applications to identify semantic relationships between words, phrases, or entire texts (Reimers & Gurevych, 2019; Zhelezniak et al., 2019; Steck et al., 2024). Some examples of pairs of text (a report and a query) and their corresponding cosine similarity are shown in Table 2. The first example from Table 2 demonstrates that embedding models can identify the semantic similarity between the report and the query (as shown by the high cosine value) even when the word overlap is low.

Table 2

Examples of Reports and Queries, and their Cosine Similarity

Example	Police Report	Query	Cosine Similarity
1	The suspect broke into a parked vehicle and stole a laptop from the back seat.	Who stole the laptop from the car?	High 0.83
2	A woman reported receiving multiple harassing phone calls from an unknown number.	Any cases of identity theft reported this week?	Medium 0.52
3	A pedestrian was struck by a speeding motorcycle near Main Street.	How many missing persons cases are open right now?	Low 0.28

Understanding Keyword-based Models and Embedding Models

The two most popular keyword-based models are TF-IDF and BM25. This model captures TF (how frequently words occur in a document) and IDF (how unique they are across the corpus). TF-IDF is purely statistical and does not capture semantics (meaning) or word order. BM25, developed in 1994, is an improvement over TF-IDF because BM25 has an additional probabilistic model of relevance. BM25 is widely used in search engines and information retrieval systems to rank search results. MiniLM and MPNet are built on BERT to improve semantic and contextual understanding.

Methods and Materials

Dataset Synthesis

Access to comprehensive, real-world law enforcement data presents significant challenges due to data sensitivity, data privacy regulations, and restricted sharing policies (Güss et al., 2020). Redaction is mandatory for personally identifiable information (PII), such as full names, social security numbers, and specific locations, to comply with regulations like the GDPR and local laws (Böhlin, 2024). Due to those challenges, we synthesized two datasets specifically for this study (the data is not derived from real-world law enforcement records). However, the LLM was prompted with actual law enforcement records to ground the synthetic data in real-world law enforcement records. The first dataset consists of 50 crime reports. The crime reports mirror the structure and linguistic content of the National Incident-Based Reporting System (NIBRS). The synthetic corpus was generated using a large language model (LLM), specifically GPT-4o. The second dataset consists of 50 queries. For each query, a ground truth was established. Ground truth represents the "true" or correct answer, obtained through direct observation or measurement, which is used to evaluate the output of a model. Considering the labor intensity of human annotation, the study leveraged a large language model (LLM) for obtaining the ground truth. Table 3 presents sample queries from the synthesized dataset, and Table 4 presents sample crime reports.

Table 3

Samples from the Synthesized Query Datasets

ID	Queries	Expected Report ID
0	Report concerning the removal of electronic components following a forced entry into a business.	0
1	Driver impairment suspected after a crash where a subject failed to obey traffic lights.	1
2	Domestic violence charges filed after a physical altercation over shared finances.	2
3	Damage to an historic municipal monument requiring expensive specialty restoration.	3
4	Seniors tricked by someone posing as a government official demanding money via gift vouchers.	4

Table 4

Samples from the Synthesized Crime Report Dataset

ID	Report
0	On Tuesday, at 02:30 AM, a commercial entryway breach was reported at the 'Digital Depot' electronics store. Two individuals forcibly gained access via a back portal. They quickly targeted the secure storage unit, absconding with high-value computing hardware before leaving the location in a dark-colored utility vehicle. The estimated value of the removed property is significant.
1	A severe multi-car event occurred at the intersection of Highway 101 and Oak Avenue at 5:45 PM on Friday. One operator, showing clear signs of impaired faculties, failed to yield to a signal and struck a large passenger vehicle. The impaired subject resisted sobriety measures and was detained on site.
2	Officers responded to a disturbance involving a heated cohabitation conflict at apartment complex unit 4B. The male subject was preparing to exit the dwelling after physically assaulting the female victim over a debate concerning domestic finances. The suspect was taken into custody and faces charges related to domestic battery.
3	Extensive architectural defacement was discovered early Sunday morning at Central Park’s main pavilion. Large segments of the stone facade were covered in unauthorized painted messages with derogatory content. Restoration costs are projected to be high due to the required specialist cleaning.
4	The cybercrimes unit identified a complex impersonation scheme targeting elderly citizens. The perpetrators pretended to be revenue agents via voice calls, demanding immediate remittance through prepaid cards or crypto assets to avoid legal consequences. The public is urged to verify callers requesting funds.

Evaluation

The performance of both approaches is measured using a metric called Hit@3, which means retrieval accuracy in the top 3 reports. This metric measures the percentage of queries for which the relevant crime report is present in the top 3 results. The experiment step-by-step is recorded below.

- Generated a synthetic dataset of crime reports and queries.
- Obtained the ground truth for every query.
- For each model type, converted the reports and queries into the specific model type format suitable for retrieval.
- For every combination of query, report, and model.
    - Ran the retrieval for each query using cosine similarity.
    - Picked the top three reports based on the highest cosine similarity score.
    - If the ground truth existed, in the top three reports, marked it as a successful hit; otherwise, marked it as missed.
- Aggregated the hits and misses and calculated the retrieval accuracy in the top 3 as *#Hits / (#Hits + #Misses)*.

Results

Table 5 shows the retrieval results for a test query. The embedding-based models (MPNet-Base and MiniLM-L6) ranked the correct document (ID = 0) in the top position. In contrast, the keyword-based models (TF-IDF and BM25) did not retrieve the correct document in their top three result.

Table 5

Sample Retrieval Results for Queries

Queries	Expected Report ID	Model	Cosine Similarity of Expected Report ID	Top 3 Reports
Report concerning the removal of electronic components following a forced entry into a business.	0	MiniLM-L6	0.5613	[0, 8, 26]
Report concerning the removal of electronic components following a forced entry into a business.	0	MPNet-Base	0.6229	[0, 17, 20]
Report concerning the removal of electronic components following a forced entry into a business.	0	TF-IDF	0.0000	[26, 9, 42]
Report concerning the removal of electronic components following a forced entry into a business.	0	BM25	0.0000	[26, 2, 34]

Table 6 summarizes the overall retrieval performance across all queries using the Hit@3 metric. The MPNet-Base model achieved a Hit@3 score of 100% (on the synthesized dataset), while the MiniLM-L6 model achieved 98%. In comparison, the keyword-based models showed lower performance, with BM25 achieving 44% and TF-IDF achieving 28%.

Table 6

Retrieval Accuracy in Top 3 (Hit@3) by Model Type

Model	Type	Retrieval Accuracy in Top 3 (Hit@3)
MPNet-Base	Embedding	100%
MiniLM-L6	Embedding	98%
BM25	Keyword Based	44%
TF-IDF	Keyword Based	28%

Discussion/Implications

Table 5 shows the retrieval results for a test query. The embedding-based models (MPNet-Base and MiniLM-L6) successfully ranked the correct document (ID=0) in the top position. Their success is quantitatively supported by the high cosine similarity scores (0.6229 and 0.5613, respectively), measuring the semantic overlap between the report and the query. These scores indicate that the models correctly identified the meaning of "removal of electronic components" as being highly similar to "computing hardware," even when the exact words did not match. In contrast, the keyword-based models (TF-IDF and BM25) had a much lower retrieval accuracy, sometimes returning irrelevant documents in their top three ranks. Because keyword-based models only look for word overlap and disregard the underlying meaning, they effectively treated the query and the report as completely unrelated documents. Table 6 shows the Hit@3 for the 4 models, where the embedding-based models demonstrated notably better performance. The MPNet-Base model achieved a Hit@3 score of 100% (on the synthesized dataset), meaning every query successfully retrieved its corresponding crime report within the top three positions. The MiniLM-L6 model also performed exceptionally well, scoring 98%. In contrast, the traditional keyword-based models exhibited poor performance, BM25 achieved a Hit@3 of 44%, while TF-IDF was the lowest-performing model at 28%.

The results suggest the value of leveraging semantic search capabilities for information retrieval within complex, narrative-driven domains such as criminal reports. There is a 52-percentage point difference between MPNet-Base (100%) and BM25 (44%), a vast performance disparity between the embedding model with the highest Hit@3 and the keyword-based model with the highest Hit@3. TF-IDF was less accurate with a Hit@3 of 28%. Because TF-IDF and BM25 rely purely on the frequency and exact matching of terms between the query and the report, these models were unable to recognize the semantic equivalence of the text, resulting in their low retrieval scores. Table 6 highlights the low retrieval accuracy of traditional keyword-based models. The better performance of embedding models, MPNet-Base (100% Hit@3) and MiniLM-L6 (98% Hit@3), can be attributed to their underlying transformer architectures. These models encode the full contextual meaning of the query and the report into a high-dimensional vector space, where similarity was calculated based on semantic proximity measures such as cosine similarity, allowing the models to correctly identify the relevant reports, even when the query used an entirely different terminology.

Academic and Practical Implications

From an academic perspective, this study contributes to several important gaps in the literature on IR, NLP, and policing. First, much of the IR literature has focused on keyword-based and embedding models in general domains (Boubekeur & Azzoug, 2013). Our work extends that line of inquiry from the general domain into the specific domain of law enforcement. Second, our work adds to the literature on the usefulness of machine learning/NLP in law enforcement. NLP and ML tools in policing are growing but still face challenges in adoption, domain-specific evaluation, and bias/ethics (Sarzaeim et al., 2023).

The findings have direct and practical implications for law enforcement agencies, including investigators, patrol officers, crime analysts, and prosecutors in leveraging smarter search capabilities to improve operations, efficiency, and public-safety outcomes. The superior performance of semantic models in our study implies that law enforcement agencies may deploy these systems to retrieve relevant reports even when lexical overlap is low. The use of semantic models may enable faster discovery of related crimes, better case linkage, and earlier identification of patterns. Patrol officers may retrieve relevant past incidents or intelligence while on shift. Similarly, crime analysts may be able to surface latent patterns that keyword systems may miss.

Deployment Considerations

There are several implementation considerations that law enforcement agencies need to consider when adopting these smarter search capabilities. Evaluating their existing case/report management systems to assess how often keyword searches fail because of vocabulary variation. Additionally, the agency may benefit from training employees on embedding models and how they differ from keyword-based models. Finally, a direct comparison of computational cost and latency is needed before deployment. The[16] [17] higher retrieval accuracy of MPNet may come at the cost of high latency, making MiniLM-L6 a more efficient choice for real-time systems.

Ethical Implications of NLP in Law Enforcement

All NLP/ML is based on historical data. If the training data has biases, it might display as inaccurate results after the deployment. Historical law enforcement narrative descriptions often reflect systemic biases related to sensitive features such as race, ethnicity, and socio-economic status (Shrestha & Yang, 2019). NLP/ML in policing raises concerns about fairness and bias in the sources and systems (Dixon & Birks, 2021). Agencies must audit retrieval results and monitor for bias. Additionally, deployment must adhere strictly to legal and regulatory frameworks governing data privacy, including PII and protected health information (Böhlin, 2024). There is a direct tension between maximizing retrieval utility and ensuring legal compliance. The required extensive redaction to remove PII, a legal necessity, inadvertently removes contextual richness from the police narratives.

Limitations

The datasets of crime reports and queries had 50 samples each, explicitly engineered to create a wide lexical-semantic gap. This highlighted the difference in retrieval performance but does not fully represent the scale, noise, and complex internal structure of a real-world, large-scale police report database. The models were evaluated using only a single, high-stakes metric, Hit@3. While this is a practical metric for assessing a system's ability to retrieve the report, the evaluation may be improved by considering more metrics, such as normalized discounted cumulative gain or mean average precision.

Future Research

The embedding models used were trained on general web and language datasets. Domain-specific fine-tuning, adapted to the unique terminology and semantic relationships found in law enforcement documents, might perform even better (Chalkidis et al., 2020). Additionally, future research may investigate the magnitude of improvement achievable on a larger dataset (millions of crime reports). Future research may also explore hybrid approaches that combine the

strengths of BM25 for exact matching with the semantic capabilities of embedding models.

Conclusion

Our comparative analysis suggests that embedding models may offer a measurable advantage over traditional keyword-based models in terms of retrieval accuracy. The MPNet model achieved significantly higher retrieval accuracy (Hit@3 of 100%) than BM25 (Hit@3 of 44%), a traditional keyword-based model, a difference of 56 percentage points. The other traditional keyword-based model, TF-IDF showed the lowest performance among the evaluated models, with a Hit@3 of 28%. This work points to the relatively limited retrieval accuracy of traditional keyword-based models due to vocabulary mismatch in complex law enforcement reports. Moreover, it emphasizes the need for smarter and more advanced search models in law enforcement. Interdisciplinary collaboration among computational linguists, criminologists, and legal experts may help guide the development of more effective retrieval systems that serve the pursuit of justice.

References

Al-Haddad, R., Al-Zubaidy, S., & Al-Taee, M. (2025). A hybrid information retrieval system for regulatory documents using BM25 and fine-tuned sentence transformers. *arXiv*. https://doi.org/10.48550/arXiv.2502.16767

Böhlin, R. (2024). LLM-based PII detection and remediation: Challenges and opportunities. *arXiv preprint*. https://arxiv.org/abs/2501.12465

Boubekeur, F., & Azzoug, W. (2013). Concept-based indexing in text information retrieval. *arXiv*. https://arxiv.org/abs/1303.1703

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). Legal-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. https://doi.org/10.18653/v1/2020.findings-emnlp.261

Dixon, A., & Birks, D. (2021). Improving policing with natural language processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact* (pp. 115-124). https://aclanthology.org/2021.nlp4posimpact-1.13.pdf

Güss, C. D., Tuason, M. T., & Devine, A. (2020). Problems with police reports as data sources: A researchers' perspective. Frontiers in Psychology, 11, Article 582428. https://doi.org/10.3389/fpsyg.2020.582428

Lovell, R. E., Klingenstein, J., Du, J., Overman, L., Sabo, D., Flannery, D., & Ye, X. (2022). *Using sentiment analysis and topic modeling in assessing the impact of police signaling on investigative and prosecutorial outcomes in sexual assault reports* (NCJ 306955). U.S. Department of Justice, National Institute of Justice. https://www.ojp.gov/pdffiles1/nij/grants/306955.pdf

Opitz, J., Möller, L., Michail, A., & Clematide, S. (2025). Interpretable text embeddings and text similarity explanation: A primer. *arXiv.* https://arxiv.org/abs/2502.14862

Parker, S. T. (2025). Assessing Supervised Natural Language Processing (NLP) Classification of Violent Death Narratives: Development and Assessment of a Compact Large Language Model (LLM) Approach. *medRxiv*, 2025-01. https://doi.org/10.1101/2025.01.16.25320680

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 3980–3990). https://arxiv.org/abs/1908.10084

Reimers, N., Gurevych, I., & Muennighoff, N. (2025). Sparse encoder—SentenceTransformers documentation. https://sbert.net/docs/package_reference/sparse_encoder/index.html

Robertson, S. E., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, *3*(4), 333–389. https://doi.org/10.1561/1500000019

Sarzaeim, P., Mahmoud, Q. H., Azim, A., Bauer, G., & Bowles, I. (2023). A systematic review of using machine learning and natural language processing in smart policing. *Computers, 12*(12), 255. https://doi.org/10.3390/computers12120255

Schutz, P., & Hupfeld, T. (2021). BERT fine-tuning strategies for dense retrieval efficiency. *arXiv*. https://arxiv.org/abs/2109.10739

Shrestha, Y. R., & Yang, Y. (2019). Fairness in algorithmic decision-making: Applications in multi-winner voting, machine learning, and recommender systems. *Algorithms*, *12*(9), 199. https://doi.org/10.3390/a12090199

Steck, H., Ekanadham, C., & Kallus, N. (2024). Is cosine similarity of embeddings really about similarity? *arXiv*. https://arxiv.org/abs/2403.05440

Xing, X., & Chen, P. (2024). Entity Extraction of Key Elements in 110 Police Reports Based on Large Language Models. *Applied Sciences, 14*(17), 7819. https://doi.org/10.3390/app14177819

Zhelezniak, V., Parker, J., Smith, S., Saphra, N., Tsvetkov, Y., & Cotterell, R. (2019). Correlation coefficients and semantic textual similarity. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1000–1010). https://aclanthology.org/N19-1100.pdf