California Police Records Access Project

Milan Dujardin [*1] **Abigail Brooks-Ramirez** [*1] **Chetan Goenka** [*1]

Abstract

Often, real-world data is recorded through and stored in various types of documents, including email records, reports, handwritten notes, and images. Analyzing this data presents a challenge, as large amounts of unstructured and dissimilar data, often stored through a shared file method such as PDFs, may be difficult to parse at a large scale. We present an exploratory analysis of a large database of police records, the Californian Police Records Access Project. Given the diversity of the documents and size of the dataset, hand-labeling is not desirable. We therefore present a novel data pipeline that groups these documents by type in an unsupervised manner and extracts information. This will allow us to link and aggregate related information.

1. Introduction

Improving the transparency of police conduct has been the subject of many legal and social justice movements of the 21st century (Vági et al., 2024). Within the United States, data related to policing is constrained by two central barriers: legal access to such documents and the expert knowledge required to interpret these documents. The first barrier is largely constrained by laws and police unions, which may restrict the flow of information, redact data, or determine the boundaries of what constitutes public vs private knowledge.

Our work in this paper is made possible by prior work to lower the barrier to data access, as researchers acquired large amounts of data under the *California "Right to Know" Act* passed in 2018 (SB1, 2018) which postulates that in the case of use of force, discharge of a firearm by an officer, officer dishonesty, or sexual assault by officer all data related to the investigation becomes public record. Researchers in the Police Records Access Project (PRAP) grounded this work in participatory design and human computer interaction principles (Costanza-Chock, 2020), creating an interdisciplinary team of legal experts, data scientists, and journalists who collectively worked with the public to assess the needs and desires of those impacted by policing (Journalism, 2025). This work, conducted over several years, culminated in a 25 terabyte dataset that was released in late 2025 and made available to the public through an interactive user interface (Shah, 2024).

We build off this work to address the second barrier: the interpretation of this data. The audience of this data is vast, spanning from victims of police misconduct to public defenders, and as a result people will have largely different experiences searching and deciphering the data. Police records often use jargon that requires expert knowledge of the legal and criminal domains in order to interpret. Critically, there is a lack of standardization among documents, largely the result of police work being conducted in an increasingly digital world. Though all records are stored as PDFs, records may date as far back as the 1960s – where such documents are entirely handwritten – while more modern records may include images, templatized documents, or interview transcripts. A lack of structure exists between police departments as well; few documents have a standardized template, which means an incident report may have entirely different format at different police departments. The unstructured nature of the documents creates a barrier to retrieving information within documents as well as determining relations between documents.

To address these needs, we introduce a data processing pipeline that generates embeddings based on both text and image data within documents using multimodal models. Utilizing the embeddings, we then group document types together using clustering algorithms which allow us to classify documents based on semantic similarity, though the exact template may vary by document.

2. Related Work

Given the varied nature of records in our dataset, we examine prior work at determining statistical regularities among PDF documents. This work frequently consists of two

[*]Equal contribution [1]Department of Electrical Engineering & Computer Science, University of California, Berkeley, Berkeley, USA. Correspondence to: Milan Liessens Dujardin <milan.ld@berkeley.edu>, Abigail Brooks-Ramirez <a.brooks@berkeley.edu>, Chetan Goenka <cgoenka@berkeley.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

phases: generating embeddings for each document and subsequently applying an algorithm to group documents together based on underlying patterns present in the embeddings.

Document Embeddings Police records often contain a mixture of image and textual data. A common approach to PDF or image-based text extraction is Optical Character Recognition (OCR). OCR converts scanned or photographed text into machine-readable representations which allows further analysis of textual data. Modern OCR methods combine convolutional architectures with sequence models to improve robustness to noise, layout variation, and low-quality scans; a widely used approach is the CRNN-based framework introduced by (Shi et al., 2017).

Utilizing text data extracted from documents, prior work examining legal documents analyzed the efficiency of different text-based embedding methods (Vági et al., 2024). Researchers compared Term Frequency-Inverse Document Frequency (TF-IDF), Doc2Vec, and Latent Semantic Analysis (LSA). TF-IDF weights the occurrences of words based on how often that word appears in the document divided by the number of documents the word occurs in; this minimizes frequently occurring words, such as stop words, while valuing less common terms which hold more weight. Doc2Vec creates document-level embeddings based on the semantic meaning of the words found in the document; it is a generalization of Word2Vec which generates a vector representation for the semantic meaning and relationships of words based on their context within a corpus. LSA compliments TF-IDF by considering the positioning of words, relying on the assumption that words with similar semantic meaning occur in similar positions in text; critically LSA utilizes Singular Value Decomposition, which reduces the dimensionality of the vector, potentially resulting in faster processing time for large datasets (such as ours). The comparative research found that TF-IDF had significantly better performance among subsequent modeling algorithms (k-means and Latent Dirichlet Allocation) in terms of cluster and topic accuracy (Vági et al., 2024).

A more recent embedding model is LayoutLM, which combines text and visual content by concatenating word-token embeddings and image-patch embeddings, providing the result as input to a multimodel transformer (Huang et al., 2022). LayoutLM applies three pre-training tasks, Masked Language Modeling (MLM), Masked Image Modeling (MIM), and Word-Patch Alignment (WPA). Both MLM and MIM mask a percentage of words/images in order to create an objective of predicting the masked object; WPA determines if an unmasked image and text pair correspond to each other. The resulting embedding is one of the highest performing document representations which incorporates layout, textual semantics, and visual structure.

Clustering Given the range of potential tasks and the nature of scanned datasets to be large in size, the majority of prior work employed unsupervised clustering models, which do not require labeled data. Prior research examining legal documents used the k-means clustering algorithm in order to group documents together based on underlying structure and semantic similarity ((Vági et al., 2024)). Their approach emphasizes determining an appropriate number of clusters through the elbow method, where an error function is plotted against increasing numbers of clusters (k); the point at which additional clusters yield diminishing improvement marks the practical "elbow," allowing researchers to determine a stable choice for k.

The same research built upon the clustering by employing Latent Dirichlet Allocation (LDA) which is a topic modeling method that represents topics as a statistical distribution over words. LDA allows more more in-depth semantic understanding of clusters. In grouping k-means and LDA, (Vági et al., 2024) shows how combining unsupervised learning methods can enhance understanding and organization of large datasets of text-based documents.

3. Methods

3.1. Dataset

Our (current) dataset consists of 184 PDF documents downloaded from the Police Records Access Project online portal. The documents come from various California police departments and represent a diverse collection of police documentation spanning multiple years. The documents are significantly different in both format and content. They contain arrest reports, interviews, emails, meeting notes, images, logs, and medical and legal papers. Some are document are digital while others are scanned. Functionally alike documents (e.g., arrest reports) lack standardization and have completely different layouts and formats across police departments. Text length ranges from minimal content (files consisting mostly of images) to extensive reports exceeding 900K characters. Similarly, page counts range from single-page memoranda to comprehensive investigation files spanning nearly 200 pages. This variability makes it really difficult to hand-engineer document-level features based on text or layout, motivating our embedding based approach.

To establish some ground truth for evaluation, we manually labeled a subset of the documents into 10 categories: internal affairs reports, arrest reports, coroner reports, use of force reports, interview transcripts, email/letter communications, meeting notes, standalone images, other reports, and a general "other" category for documents that do not fit specified types.

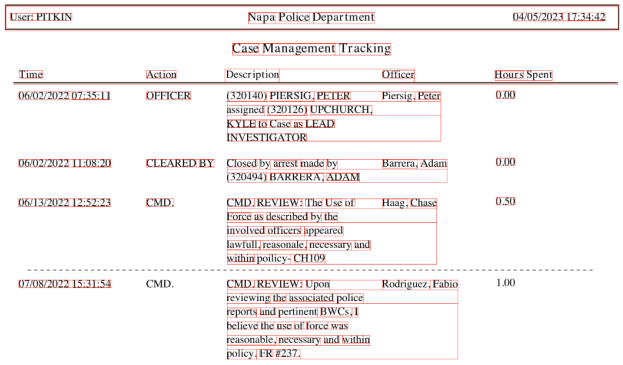

Figure 1. Example of the bounding boxes produced by PyTesseract on the first document of the dataset. These are inputted into LayoutLMv3 together with the extracted text.

3.2. Text Extraction

Given the diverse nature of our dataset, we implemented a hybrid text extraction pipeline to handle both digital and scanned documents. Our approach combines direct text extraction and Optical Character Recognition (OCR) to maximize text extraction across document types. Using PyMuPDF, we first extract machine-identifiable text directly from PDFs. For scanned or image-based documents, where direct extraction fails, we use Pytesseract OCR to extract text from visual character patterns. Each PDF is converted into an image using the pdf2image package and processed with Pytesseract's English language models. We concatenate both direct and OCR extracted text for each document, ensuring text recovery coverage of digital, scanned, and hybrid documents. We store the extracted text in JSON format with metadata such as filename and character count. Of the 184 documents in our dataset, the pipeline was able to extract sufficient text (>50 characters) from 113 of them, which was then retained for further analysis. Failed extractions were usually caused by low-quality scans and files containing only photographs.

3.3. Document Embedding

In order to create an accurate vector representation for each document, we first extract page-level tokens and their spatial coordinates using OCR, which provides both the recognized text and the bounding box for each word. These OCR outputs are converted into token-level features through a pre-processing step that produces (1) a sequence of textual tokens from the text and (2) normalized bounding boxes formatted for compatibility. We employ the LayoutLMv3 tokenizer ((Huang et al., 2022)), which encodes both textual content and layout information, enabling downstream models to consume a unified multimodal embedding.

To convert the 2-dimensional embeddings for each document by a single 1-dimensional vector, we follow the example of (Sampaio & Maxcici, 2025) and average over each of the 768 dimensions in the embedding space.

In addition, we explored using the several embedding models available in the SBERT Python library to generate text embeddings for clustering. Specifically, we evaluated three models: all-MiniLM-L6-v2, all-mpnet-base-v2, and e5-large-v2.

3.4. Clustering

We apply three approaches of document clustering by type.

3.4.1. Using TF-IDF Vectors

Often, the type and content of a document is established on the first (or second) page, whether in the form of a title, cover page, or description. This enables humans to quickly understand what kind of document we are dealing with. Using this approach, we run a TF-IDF vectorizer on the first two pages of the document to infer the most important words. Since the dataset comprises a finite set of document types, we can describe these types through a set of characteristic words. For example, words like 'For:', 'To:', and 'Subject:' suggest a document represents an email. By feeding a vocabulary that contains such predictive words to the vectorizer, we reduce parts of the text that tell us little about the document type.

Using DBSCAN, we cluster the documents. We find that the model is able to cluster most memoranda, images, investigative lead summaries, arrest/detention information documents, and interviews. However, further finetuning is to be performed in the remaining time of the project to ensure appropriate clustering over all documents. We are still awaiting API access to broaden the scope of our analysis, since the 184 documents we worked with already represent a wide range of document types and therefore contain little representation for each type.

3.4.2. Using Aggregated Embeddings

We also fed the per-document embeddings produced by LayoutLM to various clustering algorithms directly. We incorporated Gaussian Mixture, DBSCAN, HBDSCAN, and K-means clustering into the pipeline. Given the large number of file types, we opted for the more flexible HBDSCAN model, so that we don't have to define the number of clusters beforehand. The clusters are shown in the figure below.

3.4.3. Using a Q&A Model

The so-called "impira/layoutlm-document-qa" model allows us to enter a question in natural language and get back part of the text as an answer. We prompted the model with 'What

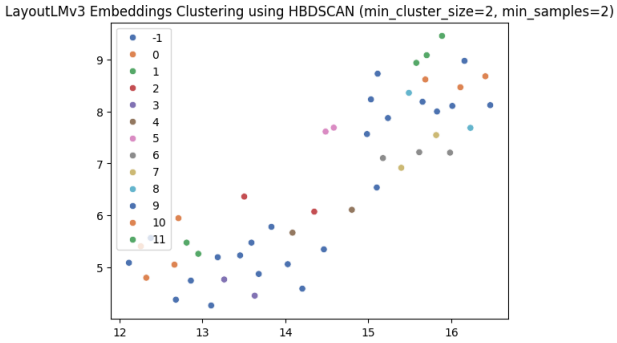

Figure 2. Clustering of the first 50 documents by HBDSCAN.

Figure 3. Word Clouds generated by Top2Vec.

kind of document is this?' and received mostly logical type descriptions (Report, Supplement, Investigative Report, Vehicle Detection Report, Use of force Report, Case Management Tracking, Memorandum...) but also odd type estimates (Merced County District Attorney's Office, ShotSpotter°, Office of the District Attorney, Page 1). For each document, we selected only the answers with confidence above a threshold (10

We will further finetune this approach, potentially in combination with a vocabulary.

3.5. Clustering by Content

3.5.1. Using a Topic Extractor Model

An approach to clustering documents by content (whether it is by case, by topic, or other modes) consists of utilizing a topic extractor and extracting informations or topics. Top2Vec is a promising model that enables topic extraction. However, our attempt to infer topics from the documents using Top2Vec yielded rather generic categories, including:

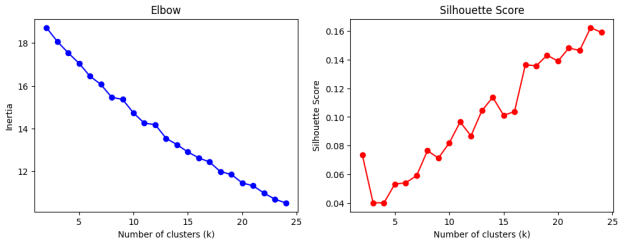

Figure 4. Inertia and silhouette score plot for clusters made on the E5-large-v2 embeddings

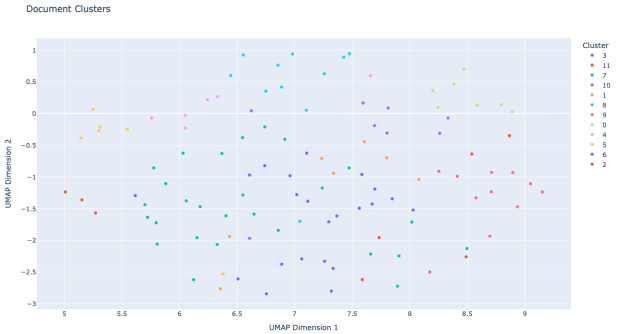

Figure 5. Document clusters using E5-large-v2 embeddings

4. Evaluation & Analysis

4.1. Clustering Quality

Given the unsupervised nature of our pipeline and unlabeled data, the clustering quality is difficult to assess. However, we used metrics that capture the level of coherence within and between clusters without needing a ground truth: the Silhouette Score and the Calinski-Harabasz Score.

We evaluated K-Means clustering on the SBERT-derived embeddings using values of k ranging from 2 to 25, assessing performance with both inertia and silhouette scores. The all-MiniLM-L6-v2 model (384 dimensions) achieved the highest silhouette score of 0.218 at k = 25, while the E5-large-v2 model (1024 dimensions) peaked at 0.162 at k = 23. The absence of a clear elbow in the inertia curves suggests that the documents do not form strongly separated clusters.

4.2. Cluster Analysis

Preliminary results for the clustering by type (0.20 for the former and 21.47 for the latter score) are not yet satisfactory, but should be taken with a grain of salt, as documents of similar type might not relate at all content-wise.

4.3. Model Comparison

The best performance will likely stem from either a very powerful model or an aggregation of different approaches. For the final part of this project, we intend to investigate ways to combine the models we used for more robust clustering.

5. Conclusion

In this paper we present a data processing pipeline for large datasets of PDFs, containing both visual and textual data. Utilizing OCR and a transformer-based embedding model (LayoutLMv3) we create a document-level embedding that combines text and image content from each document. These embedding are then clustered together using the k-means clustering algorithm, which addresses a key challenge of working with large, non-standardized documents. The pipeline provides a scalable approach for processing large amounts of non-standardized data.

Future work will explore a more comprehensive embedding process, such as processing larger documents and further fine tuning. Future work will also expand unsupervised learning methods, potentially using topic modeling to improve semantic understanding of each document. Detailed embeddings may lead to improvement within the clustering algorithm while further topic modeling provides additional semantic information for document metadata and cluster labels.

References

Senate bill no. 1421, peace officers: Release of records (2017–2018), 2018. URL https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1421.

Costanza-Chock, S. *Design Justice: Community-Led Practices to Build the Worlds We Need*. The MIT Press, 03 2020. ISBN 9780262356862. doi: 10.7551/mitpress/12255.001.0001.

Huang, Y., Lv, T., Cui, L., Lu, Y., and Wei, F. Layoutlmv3: Pre-training for document ai with unified text and image masking, 2022. URL https://arxiv.org/abs/2204.08387.

Journalism, U. B. Uc berkeley and stanford join forces on groundbreaking database on police misconduct and use of force, 2025. URL https://journalism.berkeley.edu/berkeley-and-stanford-police-database/.

Sampaio, P. R. and Maxcici, H. Unsupervised document and template clustering using multimodal embeddings. *arXiv preprint arXiv:2506.12116*, 2025. doi: 10.48550/arXiv.2506.12116. URL https://doi.org/10.48550/arXiv.2506.12116.

Shah, T. The use of unstructured data to study police use of force. *International Journal of Police Science Management*, 26(4), 2024. doi: 10.1080/09332480.2024.2434437. URL https://www.tandfonline.com/doi/full/10.1080/09332480.2024.2434437.

Shi, B., Bai, X., and Yao, C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11): 2298–2304, 2017.

Vági, R. et al. Increasing access to legal information with unsupervised solutions. *Hungarian Journal of Legal Studies*, 64(3):456–471, 2024. doi: 10.1556/2052.2023.00473.

Label: 0 - Document 3 - Document 5 - Document 48 Label: 1 - Document 6 - Document 28 - Document 31 Label: 2 - Document 7 - Document 30 Label: 3 - Document 8 - Document 9 - Document 17 - Document 22 - Document 23 - Document 44 Label: 4 - Document 10 - Document 11 - Document 19 - Document 20 - Document 21 - Document 25 - Document 26 - Document 27 - Document 34 - Document 35 - Document 37 - Document 47 - Document 49 Label: 5 - Document 12 - Document 13 Label: 6 - Document 14 - Document 16 Label: 7 - Document 18 - Document 24 Label: 8 - Document 29 - Document 33 Label: -1 - Document 0 - Document 1 - Document 2 - Document 4 - Document 15 - Document 32 - Document 36 - Document 38 - Document 39 - Document 40 - Document 41 - Document 42 - Document 43

Cluster	Number of Documents	Semantic Label
0	3	Incident Report
1	3	Arrest Report
2	2	Shot Fired Report
3	6	Overview of Incident
4	13	Other
5	2	Interview
6	2	Memorandum
7	2	Discovery Package
8	2	Police Commission
-1	13	Email

Table 1. Cluster summary.