

Summary

The paper reports an exploratory study on the California Police Records Access Project (PRAP), using a 184-PDF working subset from a 25 TB corpus of heterogeneous police records. The goal is to enable navigation and organization without large-scale manual labeling. The pipeline (i) performs hybrid text extraction (PyMuPDF for digital text; PyTesseract OCR for scans), (ii) builds document embeddings with multimodal LayoutLMv3 (text + layout) and several text-only encoders (all-MiniLM-L6-v2, all-mpnet-base-v2, e5-large-v2), and (iii) applies clustering (DBSCAN, HDBSCAN, Gaussian Mixture Models, k-means). The authors also try TF-IDF and DBSCAN on “first pages” for type cues, a LayoutLM QA prompt for auto type descriptions, and Top2Vec for topics. Reported clustering metrics include inertia, silhouette, and Calinski–Harabasz, supplemented by qualitative inspection of clusters and word clouds.

Claims and Evidence

Claim 1: A hybrid OCR + multimodal/text embedding pipeline yields usable document-level representations for heterogeneous PDFs.

Evidence: The paper details a pragmatic extraction setup and the construction of per-document vectors. Qualitative examples and handling of edge cases support feasibility; quantitative evidence is indirect via downstream clustering metrics.

Claim 2: These embeddings, with standard clustering methods, can group documents by type/content to aid corpus organization.

Evidence: Multiple clustering experiments are presented. TF-IDF + DBSCAN on first pages groups some memoranda/interviews/forms. For embedding-based clustering, k-means/HDBSCAN are evaluated using silhouette and Calinski–Harabasz; best silhouette scores are modest, and no clear inertia elbow is observed, indicating weak global separation. A final qualitative cluster summary suggests partial coherence, but alignment to ground-truth types is not quantified.

Claim 3: Simple topic modeling and QA labeling are limited in this setting.

Evidence: Top2Vec yields generic topics on the small/heterogeneous subset; the LayoutLM QA prompt produces plausible but noisy type strings, indicating that off-the-shelf prompts are insufficient without adaptation.

Methods/Design soundness: Reasonable for an exploratory study: fixed subset, diverse encoders, and standard clustering metrics. The biggest evaluation gap is the absence of external clustering metrics against the paper’s own 10 manual document types. No theoretical claims apply. No supplementary material is cited.

Relation to Prior Works

The paper situates itself within document AI and legal/document embedding work. This is appropriate for the multimodal PDF setting. Two additional strands would strengthen context: (i) supervised document type classification and weak-labeling learning for legal/government records; (ii) embedding-based retrieval and case linking for large legal corpora, as a bridge from clustering to user-facing search/navigation.

Reviewer familiarity: Moderate with multimodal document models and legal text embeddings; not exhaustive on police-records analytics.

Other Aspects

Originality: Integrative and application-driven. Components are standard, but tailoring a multimodal pipeline to PRAP's heterogeneous PDFs (and reporting negative results) is a useful contribution. Significance: Potentially high for practitioners (journalists, defenders, community groups) if scaled and validated; current technical significance is limited by small sample size and modest cluster separation. Generally clear and well structured.

Questions for Authors

How were the 184 documents sampled from the full PRAP corpus? Are they representative of the broader distribution of document types? And if the sample is biased, how might that affect the conclusions about clustering performance?

Some documents are essentially images or have very poor OCR. Do you have any separate pipeline for these, and how do they affect cluster quality?