CS 289A Graduate Project Peer Review

Paper title: California Police Records Access Project
Authors: Milan Dujardin, Abigail Brooks-Ramirez, Chetan Goenka

1. Summary

Durjadin et al. explore the California Police Records dataset, with the goal to cluster documents by type. By enabling easier data navigation, this should address the lack of standardization and information exchange between police departments. The project combines modern NLP based machine learning with classical unsupervised learning methods, the latter primarily focusing on clustering techniques. The core contribution is a system-level clustering pipeline that uses `OCR` and `LayoutLMv3` (transformer based) to create multimodal document embeddings while attempting to demonstrate effective handling of large, non-standardized documents. The work falls within the applied ML and Systems domains, employing methods such as `HDBSCAN`, `k-means`, and `Top2Vec` extraction. Durjadin et al. evaluate their approach using inertia, silhouette scores, and Calinski-Harabasz metrics but can not identify significant cluster separation criteria like the elbow criterion. They hypothesize that noisy labels and generic `Top2Vec` topics contribute to clusters with room for future optimization. Overall, the authors intend to extend the study with a more precisely defined objective in a future revision based on a larger dataset.

2. Claims and Evidence

- The authors present a data-processing pipeline for large, heterogeneous police-record PDF files and claim that `OCR` together with multimodal transformer architectures such as `LayoutLMv3` provide adequate clustering strength.
- The author claims that these embeddings combined with unsupervised clustering (mainly `k-means`) can group documents into semantically meaningful semantic / content-based clusters without manual labels.
- The paper implies that multimodal `LayoutLMv3` embeddings are more promising for clustering the given dataset than simpler text-only embedding baselines (`SBERT` / `E5`).
- Automatic labeling methods (a `QA` model for document-type guesses and `Top2Vec` for topic extraction) can provide interpretable semantic labels for clusters, helping users navigate and understand the `CPRAP` collection.
- The introduced pipeline is said to be robust and given the context of studying a 25 terabyte dataset, the authors indicate readiness of the architecture to handle multi terabyte datasets.

Evaluation of evidence:

- The authors describe a high-level pipeline description without discussing design challenges and architectural considerations. **Recommendation**: Visualize the pipeline architecture and information flow in a diagram, as this is the paper's central contribution.
- The experiments could benefit from systematic conduction to support claims about meaningful content-based clustering. The "Meaningful clusters" remain undefined and no systematic evaluation methods were used. **Recommendation**: Research state-of-the-art text clustering evaluation frameworks and conduct systematic, controlled explorations.
- As currently written, the study doesn't provide many strong claims and only contains relatively vague assertions about pipeline robustness and performance without clear optimization or benchmarking targets. Some clustering metrics (like inertia, silhouette, Calinski-Harabasz scores) are shown out of context with minimal explanation of the results. No baselines (e.g., regex extraction, bag-of-words) or state-of-the-art comparisons are included. **Recommendation**: Define clear objectives, use established benchmarking datasets, compare against baselines and state of the art methods, and present findings systematically (tables/plots) with quantitative evidence.
- The authors mentioned restricted access to the full multi-terabyte dataset and therefore only use 184 samples for their studies. It's not clear whether this allows generalization of findings. Generally, missing quantitative evidence lets the writeup appear overly reliant on manual feature identification through human inspection. **Recommendation**: Consider obtaining a larger, representative sample of the dataset and implementing automated feature extraction methods to strengthen the ML-driven analysis and improve generalization.
- The authors only evaluate clustering performance using frozen embeddings from pretrained open-source models. **Recommendation**: An additional benchmark baseline worth exploring would be fine-tuning strategies, which could involve self-supervision, reinforcement learning, LLM judges and/or human labelers. Given the small sample size, labeling is feasible. Long documents could be segmented into smaller chunks and treated as independent training samples.

- The authors do not provide detailed or quantitative evidence to clarify whether the pipeline can be used with large scale datasets. **Recommendation**: Describe scalability, robustness or time/space complexity. This is a more a systems contribution, latency and throughput are two core metrics for that.

3. Relation to Prior Works

- Related work is clearly described with explicit connections to the current study. However, some explanations could be better balanced, common methods like the "elbow method" receive lengthy treatment, while more complex techniques such as `CRNN`, `Doc2Vec`, or `LSA` would benefit from deeper review. Related papers inform design choices but this project would benefit from comparative benchmarking against alternative implementations.

4. Other Aspects

Novelty and potential impact (e.g., publishability at ICML/other venues)

- The project addresses a challenging, but highly impactful real-world dataset. Facing data quality problems and lack of labels the authors come up with a pipeline that performs well in this challenging setting.
- The work applies existing methods to a new dataset without introducing methodological novelty. With respect to competitive ICML standards, this work would benefit from a systematic evaluation with proper baselines and benchmarks and provides insufficient technical depth with respect to pipeline architecture, scalability, or optimization objectives.

Clarity and organization of writeup

- The writeup is clear and easy to follow but could be improve by analytical depth and visual presentation of core contributions.
- Minor: Incomplete sentence: "However, our attempt to infer topics from the documents using Top2Vec yielded rather generic categories, including: [missing]"

Clarity and organization of figures

- **Figure 1**: The bounding box figure is too small to read due to the scale and appears to be a low resolution rasterized screenshot. Consider regenerating the plot with larger fonts, potentially showing fewer columns or removing the last table row to create more space. The first three examples already demonstrate the concept effectively.
- **Figures 2-5** also have resolution issues. **Figures 2 and 5** particularly need improvement, the scatter plots are difficult to interpret, and the excessive color palette creates accessibility problems for readers with color vision deficiencies. **Recommendations**: increase marker size, encode clusters using marker shapes in addition to color, add text labels directly in plots to reduce legend dependence, and include captions summarizing the key findings to emphasize.
- **Table 1**: The table itself is fine, the label to document mapping above it is not very clear and visually very noisy. The same information can be visualized in a table-based form with labels on the left (index column) and a list of documents (IDs only) in the document column. Even if presented in a clear way, we are not entirely convinced that the information provides strong evidence of your claims and supports your thesis.

Additional comments / Minor Remarks

- **Paper title**: The current title *"California Police Records Access Project"* undersells the technical contribution and could be made more informative and enticing; for example, you might consider something like *"Unsupervised Multimodal Clustering of California Police Records for Document Type Discovery."*

5. Questions for Authors

- Is there no known benchmark for this setting that you could use for making quatitative assessments of your method?
- How do you ensure that samples in the same cluster receive exactly the same label (e.g., all as "police report" rather than "police report" / "crime report")? Or is it sufficient in this setting that their labels are only close in embedding space (i.e., synonyms) rather than exactly identical?

Overall
This is a very interesting project with high practical relevance and an adequate choice of methods and tools. To reach an ICML level standard the work requires major revision and extension of the work as indicated by the authors themselves addressing analytical depth, novelty, quantitative evidence, and clear definition of optimization metrics.