Mining Unstructured Medical Texts With Conformal Active Learning

Juliano Genari[1], Guilherme Tegoni Goedert[1]

[1]Escola de Matemática Aplicada, Fundação Getúlio Vargas, Rio de Janeiro, RJ, Brazil

FGV

EMAp

Challenges in Mining Medical Texts

- Medical data often exists in **free-text form**, which is usually underused.
- Symptoms and patterns in medical texts are described in variable and **non-uniform terms**.
- Extracting insights from them **requires substantial time, expertise, and resources**.
- Many institutions **lack the personnel** to routinely analyze vast amounts of textual data.

We propose a **Conformal Active Learning framework** combining active learning with label-conditional conformal prediction to automate epidemiological surveillance. Key contributions include: (1) a novel Conformal Active Learning framework that combines **active learning with label-conditional conformal prediction**, offering reliable predictions while minimizing manual labelling; (2) a **model-agnostic design** that works with any classification model capable of generating embeddings; (3) a clustering-based selection process that improves performance by **ensuring diversity on the texts selected for manual labelling**; and (4) the release of **open-source and user-friendly web interface**, OLIM to facilitate deployment and accessibility.

Conformal Active Learning

Goal: Infer accurate labels Y (e.g., whether a patient has a specific symptom) for unstructured text data X, such as texts from Electronic Health Records (EHRs), while minimizing the amount of manual labelling required.

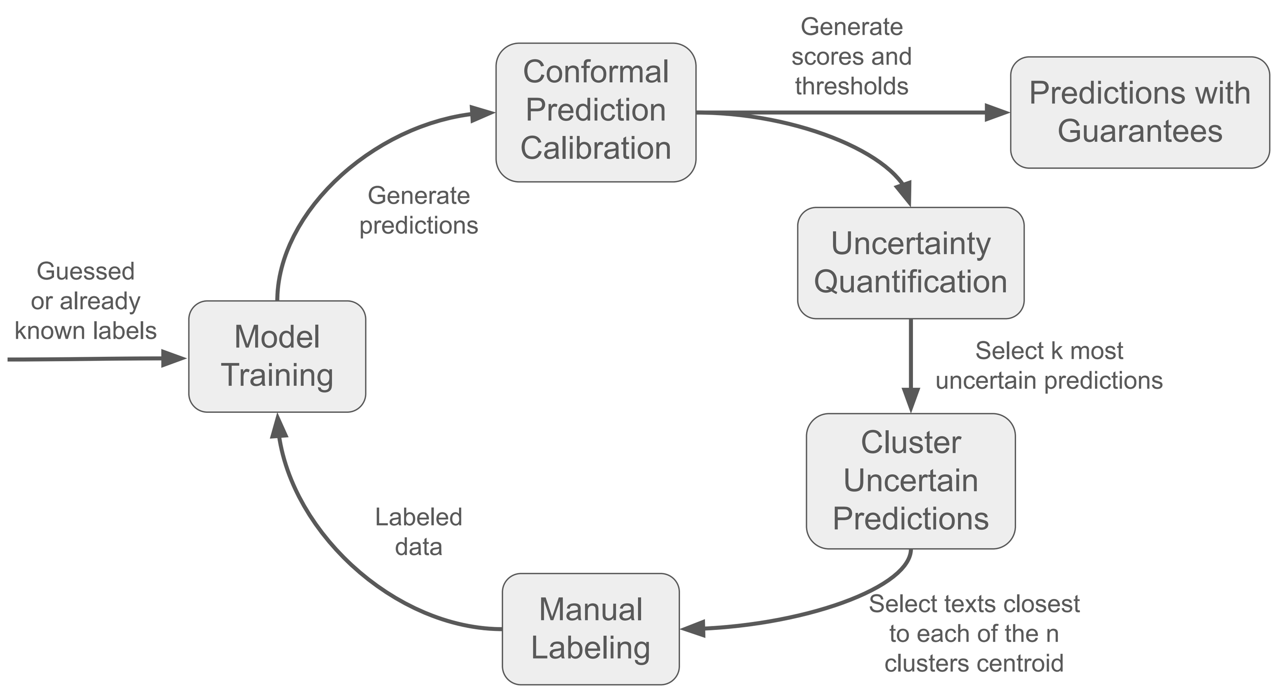

Figure 1. Active learning workflow.

Conformity Scores: For each data point x, the classification model estimates the probability $\hat{p}(y|x)$ that x belongs to category y. Than we calibrate a label-conditional conformal model on the validation dataset. This allows each point x to be associated with a conformity score:

$$s(x, y) = 1 - \hat{p}(y|x).$$

Ranking Samples by Uncertainty: After predictions, for each sample X, we calculate the **mean conformity score** across its predicted label set $C_\alpha(X)$:

$$S_X = \frac{1}{|C_\alpha(X)|} \sum_{y \in C_\alpha(X)} s(X, y)$$

Data points are **ranked based on their scores**, with higher scores indicating greater uncertainty.

Clustering Selection for Manual labelling: To ensure diversity, we select the k_{top} samples with the highest uncertainty scores. Using the classification model's embeddings we apply k-means clustering to group these samples into k_{cluster} clusters ($k_{\text{cluster}} < k_{\text{top}}$). From each cluster, select the sample closest to the centroid as the most representative data point for manual labelling.

Mixing high- and low-uncertainty: Optionally, we can include a fraction of low-uncertainty points in k_{top} before clustering to validate model performance on straightforward cases. By combining uncertainty-based ranking with clustering, the framework maximizes the value of manually labeled data and accelerates model improvement.

Deployability

Our framework is classification model-agnostic, requiring only text embeddings for operation. On-premise deployment preserves privacy by processing sensitive EHR data locally, even on low-resource hardware. Open-source code and Docker containers enable seamless installation. Compatible with lightweight models or more advanced architectures (such as transformers), our framework generates de-identified, structured insights for epidemiological analysis and monitoring while keeping raw patient data secure.

OLIM interface

We also developed OLIM (Open Labeller for Iteractive Machine Learning, Figures 2 and 3), it provides a web-based interface for collaborative text labelling, featuring role-based access, Elasticsearch-powered text filtering, and bulk export and import operations. Tightly integrated with the active learning framework, it prioritizes uncertain samples for annotation. Dockerized deployment supports both cloud or on-premise, and even mixed setups.

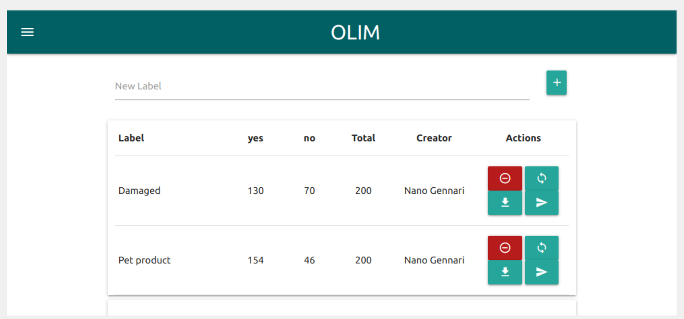

Figure 2. Label management dashboard with active learning controls.

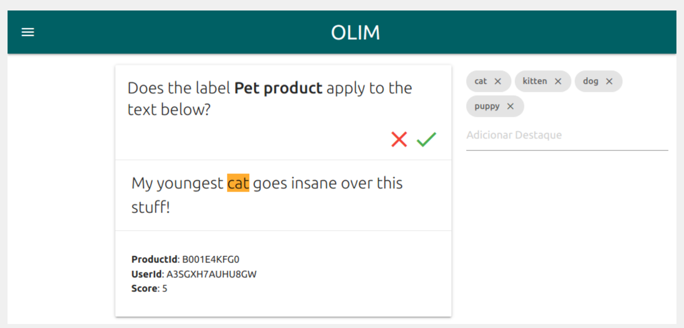

Figure 3. Interaction page for domain specialists (in development).

Experiments

Experimental Setup We evaluated our framework on Amazon product reviews—*as a proxy for unavailable public medical text databases, sharing many of the same challenges*—using four labels: *Pet/Drinkable Product* (common), *Low Quality* (subjective), and *Damaged* (rare). Experiments used 100–200 manual labels, $k_{\text{top}} = 500$, $k_{\text{cluster}} = 6$, and 90% confidence. Classification models included lightweight (**XGBoost+TF-IDF**) and transformers (**DeBERTaV3**) architectures.

Key Results With 200 labels, XGBoost achieved 92% and 85% accuracy on common labels (Table 1). Mixing high/low uncertainty samples boosted AUC-ROC and stabilized convergence (Figure 4). Rare labels (*Damaged*) required 40 pre-labels to reach AUC-ROC of 75%. Surprisingly, DeBERTaV3 underperformed (44% accuracy and 66% AUC-ROC for *Pet product*), suggesting simpler models suffice for resource-constrained settings.

Label	Accuracy	AUC-ROC	Yes/No
Pet product	0.92 ± 0.01	0.94 ± 0.06	62/138
Drinkable product	0.85 ± 0.01	0.82 ± 0.04	70/130
Low quality	0.77 ± 0.01	0.79 ± 0.04	46/154
Damaged[1]	0.91 ± 0.01	0.75 ± 0.08	39/161

Table 1. Final performance with **XGBoost+TF-IDF** for the proposed labels after 200 manual labels using our framework, with k_{top} split 30/70 on high and low uncertainty, started with 20 pre-labelled texts. ([1]Started with 40 pre-labelled texts.)

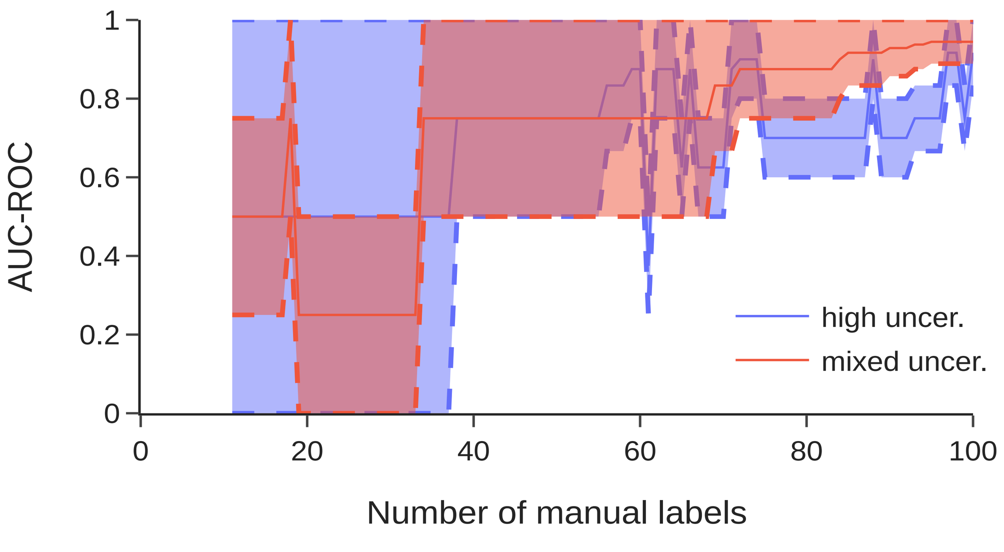

Figure 4. Convergence of AUC-ROC for the *Pet product* label, using **XGBoost+TF-IDF**.