Comparing Pseudo-Labeling Criteria for Active Learning on a Service-Desk Dataset

Rok Klančič
Jožef Stefan Institute
Ljubljana, Slovenia
rok.klancic@gmail.com

Klemen Kenda
Jožef Stefan Institute
Ljubljana, Slovenia
klemen.kenda@ijs.si

Abstract

Recent advances in large language models and human-machine interfaces have made it increasingly feasible to automate aspects of IT support ticket management. Since manual ticket routing can be slow and error-prone, ticket-routing models can help mitigate these issues. Training such models, however, requires organisations to label IT support tickets with the team responsible for handling them. This labelling is expensive because it requires expert knowledge, the tickets are unevenly distributed among the specialised teams and ticket content is sensitive. We study pseudo-labeling criteria that, during a simulated active-learning process, decide which unlabelled tickets the model can label itself. We compare nine criteria, including raw and calibrated confidence, k-nearest-neighbour agreement with and without calibration, SVM decision margins, outlier rejection, class-conditional thresholds and ensemble agreement, at checkpoints of 100, 250 and 500 labelled tickets. The experiments use a real IT support ticket dataset. With about half of the tickets labelled, the best-performing criteria label the entire remaining pool with around 97% accuracy. Agreement-based criteria are the most dependable throughout, whereas post hoc calibration collapses below roughly a quarter of the labels and becomes competitive with the best agreement- and margin-based criteria once the underlying model is sufficiently accurate.

Keywords

active learning, pseudo-labeling, IT support ticket routing, confidence calibration

1 Introduction

Many organisations operate internal service desks where incoming tickets must be routed to the support team best able to resolve them. When a ticket is sent to the wrong queue, it must be reassigned manually and often takes longer to resolve. An accurate automatic router would therefore help mitigate this issue [1]. Building such a router with supervised methods, however, requires a collection of tickets labelled with their handling team. This labelling is expensive for several reasons: experts must understand the responsibilities of many teams, the distribution of team labels is long-tailed with many rare classes, and the ticket content is sensitive. Labelling therefore cannot be outsourced and must be performed by the organisation's own staff.

Active learning (AL) can reduce the number of labels required by repeatedly selecting the tickets that are most informative to the current model [6, 9]. It does not, however, remove the human from the loop: every selected ticket still has to be labelled by an expert. Our initial experiments on service-desk dataset indicate that this effort is concentrated on the most difficult tickets. During an AL run using uncertainty sampling [5], the model predicted the correct team for only about 30% of the queried tickets before their labels were revealed, compared with about 80% using random sampling. Consequently, a labelling process leaves behind a pool of unlabelled tickets that, by construction, the model already finds easier.

This paper asks how much of the remaining labelling work the model can perform itself. We study pseudo-labeling [4] criteria that decide, for each unlabelled ticket, whether the model's prediction is sufficiently trustworthy to be accepted as a label, so that the tickets that pass can be labelled automatically while the remainder continues in the AL process. We compare nine such criteria in an active-learning simulation on an IT support ticket dataset (Section 4) and ask which criterion labels tickets most accurately and how much labelling effort it saves at that accuracy.

Our contribution is a systematic evaluation of these criteria on a real service-desk dataset in the low-label regime that a labelling process actually faces.

2 Related Work

2.1 Active Learning

AL reduces annotation costs by allowing a learning algorithm to select the most informative instances. In the common pool-based setting, an unlabelled pool is available and a query strategy iteratively selects instances for labelling by an oracle, typically a human annotator [6]. Strategies include uncertainty sampling [5], query-by-committee, expected model change, variance reduction, estimated error reduction and density-weighted methods [6]. Their common goal is to achieve competitive predictive performance with substantially fewer labelled examples. Surveys do not identify a single best strategy, however the best practical recommendation is to combine informativeness (e.g. uncertainty) with representativeness (e.g. density) strategies. They also note that AL is the most effective when labels are expensive, initial labeled set is sufficient and the model's uncertainty estimates are reliable [6, 9].

2.2 Pseudo-Labeling and Self-Training

Pseudo-labeling [4] and self-training instead exploit unlabelled data by assigning labels to examples for which the model's prediction is considered sufficiently reliable. FixMatch, for example, retains only high-confidence pseudo-labels and uses them as training targets [8]. FlexMatch extends this approach with class-dependent, dynamically adjusted confidence thresholds to account for differences in class difficulty [11]. More directly related to AL, in [3] confidence-based AL is combined with self-training to reduce the amount of data requiring human annotation. AdaPL integrates adaptive pseudo-label selection into deep AL, setting

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2026.sikdd.117

thresholds according to the uncertainty of labelled and unlabelled data [2]. Together, these approaches show the potential of combining human and automatic labelling, while highlighting the need to identify trustworthy predictions reliably.

2.3 Confidence Calibration and Agreement

The reliability of model confidence is especially important when confidence scores determine which examples are labelled automatically. Raw classifier confidence do not necessarily correspond to the empirical probability of correctness. Post hoc methods such as Platt scaling and isotonic regression can produce more reliable probability estimates [7]. Research on pseudo-labeling likewise shows that better-calibrated confidence can produce more reliable pseudo-labels [10]. Agreement between multiple predictors offers a complementary measure of label reliability: Tri-training, for example, uses agreement between independently trained classifiers to label previously unlabelled instances [12].

2.4 Positioning of This Work

Our work brings these lines of research together by studying how the remaining unlabelled pool can be labelled automatically during active learning. We evaluate raw and calibrated confidence, classifier margins, neighbourhood agreement, combinations of confidence and neighbourhood information, ensemble agreement, outlier rejection, and class-conditional criteria. Unlike approaches that use pseudo-labels primarily as an additional training signal, we focus on the reliability of automatically assigned labels and on how it changes as AL acquires additional human-labelled instances. What remains insufficiently addressed by prior work is precisely this comparison: no systematic study evaluates the reliability of such labelling criteria across the stages of an AL process, which is the gap this paper fills.

3 Methodology

The central object of our study is a pseudo-labeling criterion. Given a classifier fitted on the labelled tickets and the unlabelled pool, the criterion decides which tickets can be labelled automatically and which must be left to human annotators. We compare nine such criteria within the process described in Section 4.

3.1 Pseudo-Labeling Criteria

3.1.1 Confidence and Neighbourhood Criteria. The first three criteria threshold the model's confidence: raw maximum posterior probability, maximum posterior after Platt scaling, and maximum posterior after isotonic recalibration. Both calibrators are fitted on the held-out split. A fourth criterion uses k-nearest-neighbour agreement instead of probabilities: a ticket is accepted when at least $k-1$ of its k nearest labelled neighbours, measured by cosine distance in the full feature space of the MiniLM text embedding concatenated with the service and subcategory indicators, share the model's predicted label. We sweep this criterion over $k \in \{3, 5, 7\}$ as simple defaults, with the accept rule set to $k-1$ or k, and report the best operating point of the sweep. The fifth criterion requires both Platt confidence above the threshold and neighbour agreement with the central setting $k = 5$ (four of five agreeing), so that only its confidence threshold is swept.

3.1.2 Margins and Outlier Rejection. The SVM margin criterion uses the distance to the one-vs-rest decision boundary of the predicted class rather than a probability. Its thresholds are quantiles of the margins on the held-out split. The outlier-rejection criterion combines a confidence threshold with an Isolation Forest (contamination 0.05) fitted on the training features, so that only in-distribution tickets are labelled.

3.1.3 Class and Ensemble Criteria. The class-conditional criterion selects a separate confidence threshold for each class on the held-out split, targeting a class precision of 90%, 95% or 99%. Classes with too little support are never automatically labelled. The ensemble-agreement criterion labels a ticket only when all three models (Section 3.2) predict the same class and their average confidence exceeds the threshold.

3.2 Models

All criteria use the same three classifiers: logistic regression ($C = 10$), a support vector machine with an RBF kernel ($C = 10$, with probability estimates enabled), and a random forest with 200 trees. These values are simple defaults chosen a priori rather than the result of a systematic tuning study; because every criterion shares the same frozen models at each checkpoint, however, their exact values shift all criteria jointly and cannot affect the comparison between criteria, and absolute operating points under other settings may differ. Each model is refit on all labelled tickets at every labelling step. At a given checkpoint, all criteria use the same frozen models, so differences between criteria cannot be attributed to different underlying models. The ensemble-agreement criterion combines all three.

4 Experiment Setting

4.1 Dataset

We use anonymised IT support tickets collected and provided by GFT. Each ticket contains a title and description, the relevant service and subcategory, the conversation log, and the team that resolved it. The task is to predict the handling team from the text and metadata of an incoming ticket. After preprocessing, the dataset contains 1,172 tickets labelled with 21 teams.

4.2 Data Preprocessing

We apply three filters. First, we remove tickets reassigned to another team before their final team, leaving each ticket with one unambiguous label. Second, we discard teams with fewer than ten tickets because they provide too few examples for efficient learning. Third, we drop tickets missing a title and description or a team label. Each feature vector concatenates a 384-dimensional all-MiniLM-L6-v2 embedding of the title and description with one-hot indicators for the service and subcategory. A stratified 80/20 split produces 937 training tickets and 235 held-out tickets. The training split is the AL pool, while the held-out split is used for calibration when required.

4.3 Active Learning Process

The process starts with no labelled tickets, so its first tickets are selected by the query strategy rather than a manually constructed seed set. At this point the classifier cannot be fitted and produces uniform probabilities, so the first least-confidence queries are tied and resolved by random tie-breaking at a fixed seed; i.e. the process starts from a small effectively random seed. Unlabelled tickets from the training split are queried in batches of two, labelled with their true team, and the models are refit on all labelled data after each batch. We compare random sampling, which selects tickets uniformly from the unlabelled pool,

Table 1: Best-achievable (retrospective) operating points: Coverage (Accuracy) at the configuration with the highest coverage subject to accuracy ≥ 0.90, per checkpoint and query strategy, averaged over the three base classifiers. † indicates that the accuracy target is unattainable at that checkpoint and that the best-accuracy result is reported instead. A dash indicates that the criterion produced no labels. SVM margin* operates on the SVM only.

	Random sampling			Uncertainty sampling		
Criterion	100	250	500	100	250	500
Raw confidence	0.40 (0.87)†	0.64 (0.92)	0.73 (0.93)	0.34 (0.94)	0.61 (0.95)	0.95 (0.96)
Platt	0.03 (0.20)†	0.79 (0.92)	0.85 (0.92)	0.07 (0.61)†	0.73 (0.68)†	1.00 (0.96)
Isotonic	—	**0.89 (0.90)**	0.85 (0.91)	0.06 (0.46)†	0.78 (0.70)†	1.00 (0.96)
KNN agreement	**0.64 (0.93)**	0.80 (0.92)	0.84 (0.92)	0.38 (0.94)	**0.89 (0.95)**	0.99 (0.97)
Platt+KNN	0.08 (0.79)†	0.74 (0.94)	0.76 (0.95)	0.10 (0.94)	0.39 (0.95)	0.95 (0.98)
Class-conditional	0.53 (0.91)	0.67 (0.93)	0.74 (0.94)	0.50 (0.94)	0.83 (0.96)	0.93 (0.97)
Outlier rejection	0.39 (0.87)†	0.62 (0.92)	0.70 (0.94)	0.33 (0.95)	0.58 (0.95)	0.93 (0.96)
Ensemble agreement	0.56 (0.94)	0.72 (0.93)	0.76 (0.95)	0.31 (0.94)	0.84 (0.94)	1.00 (0.96)
SVM margin*	0.63 (0.91)	0.80 (0.91)	**0.88 (0.91)**	**0.54 (0.91)**	0.87 (0.92)	**1.00 (0.97)**

with least-confidence uncertainty sampling [5], which selects tickets whose most probable class has the lowest probability. Both strategies label the same number of tickets, so differences in pseudo-labelling quality reflect which tickets were labelled rather than how many. Models are frozen at checkpoints of 100, 250 and 500 labelled tickets, and every criterion is applied to the remaining unlabelled pool. Since the process is simulated, the ground truth of every pool ticket is known and each criterion can thus be scored.

4.4 Metrics

At each checkpoint, let P denote the remaining unlabelled pool and $A \subseteq P$ the subset of tickets that a criterion labels, with $\hat{y}_i$ the label assigned to ticket i and y_i its true team.

Coverage is the fraction of the unlabelled pool that a criterion labels, and therefore the manual effort it saves:

$$\text{Coverage} = \frac{|A|}{|P|}. \tag{1}$$

Accuracy is the proportion of pseudo-labelled tickets with the correct label:

$$\text{Accuracy} = \frac{1}{|A|} \sum_{i \in A} \mathbb{1}[\hat{y}_i = y_i]. \tag{2}$$

For every criterion, we sweep its confidence parameter and report the result with the highest coverage subject to accuracy of at least 90%. Because the simulation knows the ground-truth labels of the remaining pool, this sweep is retrospective: each reported operating point is the best achievable configuration for that criterion at that checkpoint and forms an upper bound rather than a threshold that could be set in a live deployment, where the labels of the remaining tickets are unavailable. In such a deployment, thresholds would have to be fixed in advance, e.g. on held-out labelled data; the class-conditional criterion is the only one whose thresholds are estimated this way, and a validation-based selection for the remaining criteria is left to future work.

5 Results

Table 1 reports coverage and achieved accuracy at each criterion's configuration, that achieved the highest coverage subject to the accuracy of at least 90%. Results are shown for the three checkpoints and averaged over the three base classifiers; all operating points are retrospective upper bounds (Section 4). Daggered cells use the best-accuracy operating point when the 90% target cannot be met. A dash means that the criterion labelled no ticket.

At 100 labelled tickets, only some criteria reach the accuracy floor: k-nearest-neighbour agreement performs best under random sampling with 0.64 coverage at 0.93 accuracy, and the SVM margin leads under uncertainty sampling with 0.54 coverage at 0.91 accuracy. At 250 labels, k-nearest-neighbour agreement leads under uncertainty sampling with 0.89 coverage at 0.95 accuracy. At 500 labels, Platt and isotonic calibration, ensemble agreement and the SVM margin all reach full pool coverage at 0.96 to 0.97 accuracy under uncertainty sampling, while the combined criterion achieves the highest accuracy overall at 0.98 with 0.95 coverage.

The most dependable criterion overall is k-nearest-neighbour agreement. It is the only criterion that meets the 90% target at 100 labels under both strategies and still covers at least 80% of the pool at 250 and 500 labels. The composite criteria add complexity without improving results: outlier rejection consistently covers less of the pool than the raw confidence criterion it extends, and the combined criterion never improves on its k-nearest-neighbour parent.

5.1 Coverage and Accuracy

Figure 1 shows the Coverage-Accuracy trade-off over the whole sweep, complementing the single selected operating points of Table 1. The 100-label panel is sparse because most criteria contain no threshold reaching 90% accuracy, while the remaining criteria cover at most two-thirds of the pool. This is the only checkpoint where random sampling outperforms uncertainty sampling. From 250 labels onwards, the panels fill in and the dashed uncertainty curves consistently lie above the solid random-sampling curves at the same accuracy. Two effects may contribute: the model improves faster per acquired label under uncertainty sampling, and its queries focus on difficult tickets, leaving an increasingly easy unlabeled pool. Random sampling leaves more difficult tickets unlabelled. Pseudo-labeling and AL are therefore complementary: as the process queries more tickets, the model can label a larger share of the remainder itself. While the plot contains averaged results, the per-model frontiers can differ substantially.

5.2 Calibration at Low Label Budgets

While the calibration criteria are one of the best performing at 500 labels, reaching full coverage at 0.96 accuracy, they are not

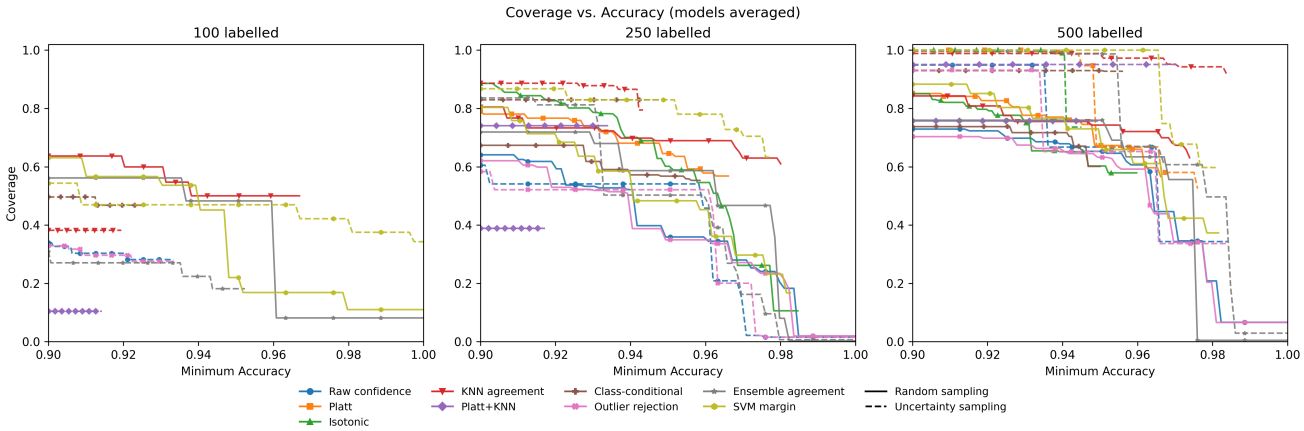


Figure 1: Coverage versus minimum accuracy for the nine criteria, one panel per checkpoint. Each curve averages the three base classifiers; solid lines show random sampling and dashed lines uncertainty sampling. At accuracy level x, the curve height is the highest coverage achievable subject to an accuracy of at least x; criteria whose accuracy never reaches x are absent from the panel.

the most robust. At the 100-label checkpoint, they collapse completely. The pseudo-labeled subset of the SVM is almost always wrong (accuracy close to zero), and logistic regression collapses similarly under random sampling. Isotonic regression fails differently. Under random sampling it auto-labels no ticket at all at 100 labels, and for logistic regression it collapses at 100 and 250 labels under uncertainty sampling. Both failures have the same root: the per-class recalibrations are estimated from a 235-ticket held-out split with 21 classes. Rare teams contribute only two or three positive examples, while a model trained on 100 tickets produces scores that are too weakly separated for a stable fit. Calibration therefore pays off only after the underlying model is accurate; agreement-based signals are safer while at the beginning of the process.

6 Conclusion

We studied nine pseudo-labeling criteria that decide which unlabelled tickets a model may label during an AL process, evaluated at checkpoints of 100, 250 and 500 labelled tickets on an IT support ticket dataset. Automatic labelling pays off once enough human labels have been collected: with 500 labelled tickets, the best-performing criteria label the entire remaining pool with 96 to 98% accuracy. Agreement-based signals remain dependable throughout, while probability calibration is unstable below roughly 250 labels and becomes competitive with the best agreement- and margin-based criteria once the underlying model is accurate.

6.1 Future Work

Two extensions follow naturally. First, the auto-labelled tickets are currently scored but not used for training. A next step is to retrain the model on the union of human and automatic labels and measure whether this accelerates the process or whether incorrect labels compound over iterations. The thresholds would also need to be adapted to prevent this. Second, all nine criteria are heuristics whose accuracy is observed but not bounded. Conformal prediction could replace them with criteria that provide a distribution-free guarantee on the proportion of wrong labels. Whether such guarantees survive the distribution shift introduced by AL remains to be seen.

Acknowledgements

This work was supported by the European Union's funded Project HUMAINE [grant number 101120218]. The dataset was provided by the project partner GFT, as a part of a pilot led by Michele Canteri.

References

[1] Shahan Ahmed, Anisur Rahman, and Md Ashrafuzzaman. 2023. A systematic review of AI and machine learning-driven IT support systems: enhancing efficiency and automation in technical service management. *American Journal of Scholarly Research and Innovation*, 2, 2, 75–101. doi:10.63125/fd34sr03.

[2] Qiang Fang and Xin Xu. 2026. Adapl: adaptive pseudo labeling for deep active learning in image classification. *Pattern Recognition Letters*, 199, 185–190. doi:10.1016/j.patrec.2025.11.024.

[3] Wenjing Han, Eduardo Coutinho, Huabin Ruan, Haifeng Li, Bjorn Schuller, Xiaojie Yu, and Xuan Zhu. 2016. Semi-supervised active learning for sound classification in hybrid learning environments. *PLoS ONE*, 11, 9, e0162075. doi:10.1371/journal.pone.0162075.

[4] Dong-Hyun Lee. 2013. Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop on Challenges in Representation Learning*.

[5] David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM/Springer, 3–12.

[6] Burr Settles. 2009. Active Learning Literature Survey. Tech. rep. 1648. University of Wisconsin–Madison.

[7] Telmo Silva Filho, Hao Song, Miquel Perello-Nieto, Raul Santos-Rodriguez, Meelis Kull, and Peter Flach. 2023. Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Machine Learning*, 112, 3211–3260. doi:10.1007/s10994-023-06336-7.

[8] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Han Zhang, Zizhao Zhang, Colin Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. 2020. Fixmatch: simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*. Vol. 33.

[9] Alaa Tharwat and Wolfram Schenck. 2023. A survey on active learning: state-of-the-art, practical challenges and research directions. *Mathematics*, 11, 4, 820. doi:10.3390/math11040820.

[10] Masahito Toba, Seiichi Uchida, and Hideaki Hayashi. 2024. Pseudo-label learning with calibrated confidence using an energy-based model. In *Proceedings of the International Joint Conference on Neural Networks*. doi:10.1109/IJCNN60899.2024.10650805.

[11] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. 2021. Flexmatch: boosting semi-supervised learning with curriculum pseudo labeling. In *Advances in Neural Information Processing Systems*. Vol. 34.

[12] Zhi-Hua Zhou and Ming Li. 2005. Tri-training: exploiting unlabeled data using three classifiers. *IEEE Transactions on Knowledge and Data Engineering*, 17, 11, 1529–1541. doi:10.1109/TKDE.2005.186.