Cluster-Aware QUBO Instance Selection with Representativeness Scoring and Noise Pre-Filtering for the QuantumCLEF 2026 Sentiment Classification Task

Notebook for the QuantumCLEF Lab at CLEF 2026

Prateek Awate[1,*]

[1]DS@GT QuantumCLEF, Georgia Institute of Technology, North Ave NW, Atlanta, GA 30332

Abstract

We describe the DS@GT submission to the QuantumCLEF 2026 Task 2 (Instance Selection for Sentiment Analysis) on the noise-injected VADER NYT dataset. Instance selection is formulated as a Quadratic Unconstrained Binary Optimization (QUBO) problem with three innovations beyond prior formulations: a cluster-aware off-diagonal penalty that suppresses within-cluster redundancy, a representativeness-aware diagonal that balances SVM margin scoring with class centroid proximity, and a Cleanlab confident-learning pre-filter that removes likely mislabelled instances before QUBO construction. The hyperparameters were tuned via multi-objective Bayesian optimization (BoTorch). We submitted both Simulated Annealing (SA) and Quantum Annealing (QA) variants of the identical QUBO to D-Wave hardware. On the official QCLEF 2026 evaluation, our QA submission achieved Macro-F1 $= 64.8 \pm 8.9$ at 67.62% reduction, ranking 1st among pure quantum annealing entries in Task 2, and is the only submission across all teams where QA outperforms SA, a result we discuss in the context of energy-landscape structure at aggressive reduction ratios.

Keywords

Quantum annealing, QUBO, Instance selection, Sentiment analysis, Cluster-aware penalty, Cleanlab, QuantumCLEF

1. Introduction

Instance selection (IS) addresses a fundamental tension in supervised learning: more training data generally improve model quality but increase training costs. For Large Language Model (LLM) fine-tuning, the relationship is roughly linear (a 70% reduction in training data translates into a comparable reduction in fine-tuning compute), so the practical question is how aggressively one can prune training data without sacrificing downstream classifier quality.

QuantumCLEF 2026 Task 2 [1, 2, 3] frames this as a challenge. Given a noise-injected variant of the VADER NYT sentiment dataset, participants must select a subset of training instances such that a downstream classifier trained on the subset performs as well as possible. The organizers explicitly stated that approximately 20% of labels were corrupted by design; therefore the subset must be both *representative* of the underlying class structure and *robust* to mislabelled examples. Submissions were evaluated on Macro-F1 and size reduction.

Looking at the 2025 leaderboard [4], every submitted system that performed instance selection ended up below the no-selection baseline on the F1 score (the 2025 edition evaluated submissions with a LLaMA 3.1 8B classifier). DS@GT achieved the best F1 at 65.9% (versus 88.9% baseline) at only 25% reduction; Malto achieved 63.1% at 75.1% reduction; GPLSI achieved the most aggressive reduction (96.2%) but with F1 of just 47.4%. The pattern is clear: prior work traded F1 for reduction. No 2025 system simultaneously achieved both above-baseline-like F1 and substantial reductions.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

awateprateek@gmail.com (P. Awate)

https://github.com/Patrickoo7 (P. Awate)

0009-0002-1878-9700 (P. Awate)

We set out to determine whether QUBO-based instance selection can break this trade-off and arrive at three concrete contributions:

1. A cluster-aware QUBO formulation in which the off-diagonal pairwise penalty is modulated by KMeans cluster membership and per-cluster spread. This is a refinement of the standard cosine similarity penalty in [5] and discourages the selection of near-duplicate instances within the same dense cluster.
2. A representativeness-aware diagonal scoring that mixes SVM-margin hardness with the inverse distance to the per-class centroid. At aggressive reduction ratios (>60%), this term prevents the optimizers from selecting only hard-near-boundary points and discarding representative class interiors.
3. A Cleanlab [6] confident-learning pre-filter run before the QUBO construction. Even at its modest filtering rate (~2.3% of instances), it measurably lowers cross-fold variance and improves the worst fold's F1.

We submit both a Simulated Annealing (SA-local) and a Quantum Annealing (D-Wave Advantage) version of the identical QUBO formulation.

2. Task Description

Task 2 uses the VADER NYT sentiment dataset [7] with injected-label noise. The dataset ships as five pre-defined fold files; for each fold $f \in \{0, 1, 2, 3, 4\}$, the training set is the documents in fold file f (approximately 4,000 instances), and instance selection is performed within that fold. The classification target was binary sentiment (negative vs. positive).

The organizers provided two pre-computed embedding representations: 768-dimensional Sentence-BERT embeddings and 4096-dimensional LLaMA 3.1 8B embeddings, and participants were free to use whichever representation they preferred. The raw text is also provided in the Supplementary Material. Submissions consist of five plain-text files (one per fold), each listing the selected document indices within that fold's file (0-based, sorted).

According to the organizers' Task 2 description, evaluation was based on Macro-F1, and the per-fold reduction was also measured.

3. Related Work

QUBO-based instance selection in QuantumCLEF.

QuantumCLEF lab [8, 4] established the QUBO formulation as the standard approach for quantum instance selection in information retrieval and sentiment classification tasks. The cosine similarity off-diagonal penalty with a class-balance diagonal was introduced in the ICTIR 2024 quantum-annealing instance-selection work [5], providing the baseline QUBO structure that subsequent submissions refined. Our 2025 submission [9] extended this with SVM-margin diagonal scoring and an iterative deletion variant, achieving the best F1 on the 2025 leaderboard at 25% reduction, but without addressing representativeness at aggressive reduction ratios. The present study directly addresses this limitation.

Classical instance selection.

Instance selection has a long history in classical machine learning, with prototype selection and condensed nearest-neighbor methods forming the foundation. The core challenge of selecting a compact, informative, and diverse subset maps naturally to combinatorial optimization, which motivates the QUBO formulation. Noise-robust instance selection adds a further constraint: the selected subset must also resist mislabelled examples, which the classical literature addresses through techniques such as confident learning [6] and label noise filtering.

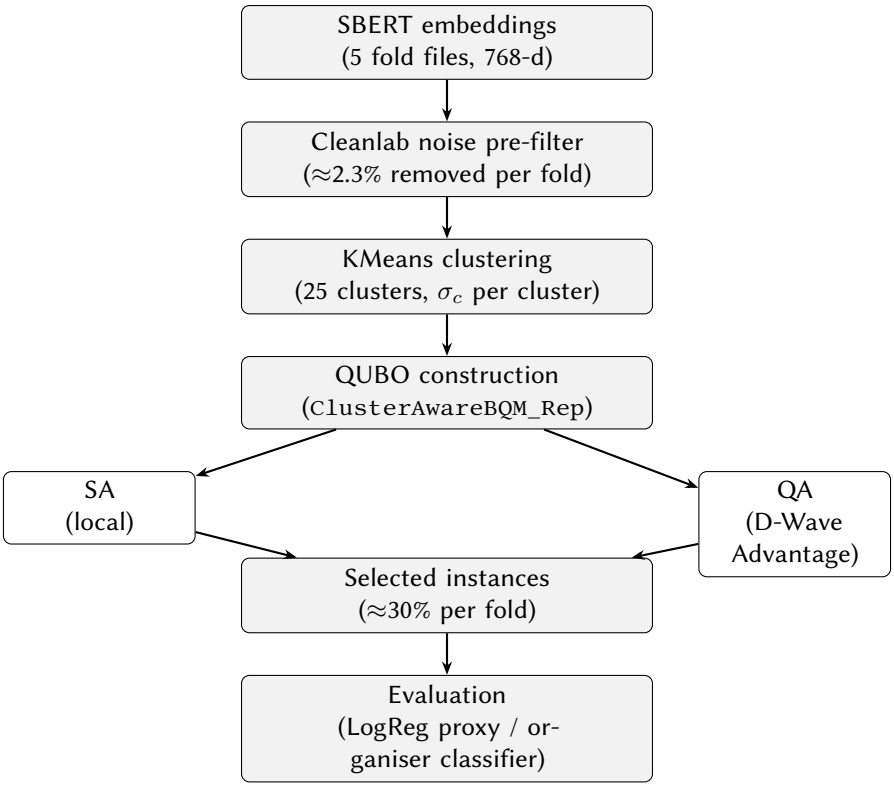

Figure 1: Overview of the DS@GT instance-selection pipeline for the QCLEF 2026 Task 2. Each fold's training documents pass through noise pre-filtering, clustering, and QUBO construction before being sampled by either SA or the D-Wave QPU.

Noise-robust learning.

Cleanlab's confident learning [6] identifies likely mislabelled instances by comparing cross-validated predicted probabilities with given labels. We applied it as a pre-filter before QUBO construction rather than as a training-time correction, which allowed the downstream annealer to operate on a cleaner instance pool without modifying the QUBO formulation itself. Alternative noise-robust approaches such as DivideMix and MixUp augmentation were explored by the other 2026 participants but were not pursued here.

4. System Description

Figure 1 provides an overview of the pipeline, and the following subsections detail each component.

4.1. Embeddings

We encoded the training instances using the `all-mpnet-base-v2` Sentence-BERT model [11], producing 768-dimensional embeddings. These serve as the geometric basis for the QUBO similarity matrix and as inputs to the Logistic Regression proxy evaluator used during the hyperparameter search. We chose Sentence-BERT over the alternative LLaMA 3.1 8B embeddings based on the well-established empirical observation that encoder-based sentence embeddings produce tighter and more discriminative geometry for classification tasks than generative-model hidden states of comparable parameter counts.

4.2. QUBO Formulation

Instance selection over a training set of size N is formulated as a binary optimization problem. Each training instance i is assigned a binary variable $x_i \in \{0, 1\}$, where $x_i = 1$ denotes "select i". The QUBO

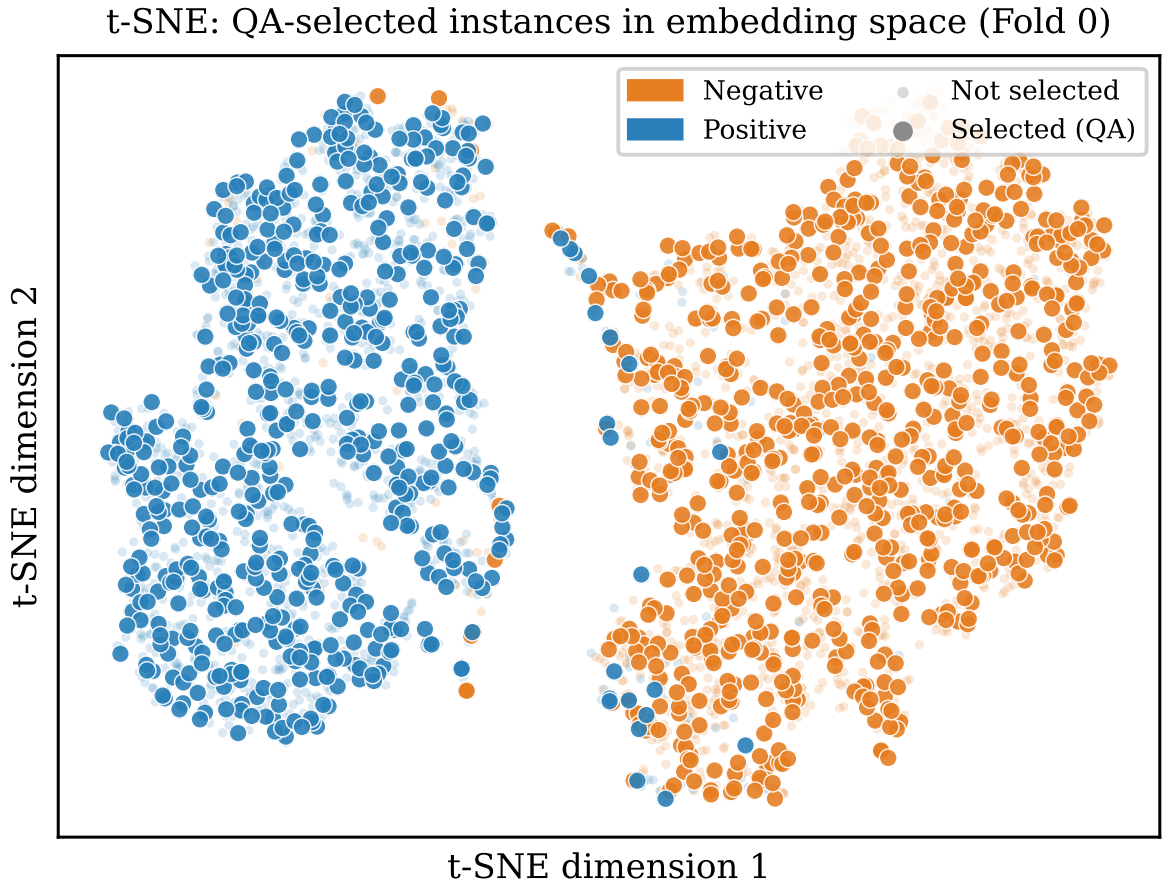

Figure 2: t-SNE [10] projection of Fold 0 training instances (768-d SBERT embeddings reduced via PCA to 50-d before projection). Faded dots: not selected; bold dots: selected by the QA submission. Selected instances cover the full geometry of both classes, confirming that the cluster-aware representativeness diagonal preserves class-interior coverage rather than concentrating selection near the decision boundary.

objective is

$$\min_{\mathbf{x}\in\{0,1\}^N} \mathbf{x}^\top Q\mathbf{x} \tag{1}$$

subject to a cardinality constraint $\sum_i x_i = k$, where $k = \lfloor \texttt{percentage\_keep} \cdot N \rfloor$ and the constraint is implemented as an auto-scaled quadratic penalty added to the diagonal (see Section 4.2.3).

Matrix Q encodes two competing objectives. The off-diagonal terms Q_{ij} $(i \neq j)$ penalize the selection of redundant pairs, encouraging diversity. Diagonal terms Q_{ii} reward informative instances, encouraging the selection of points near the decision boundary or representative of their class.

Because N can reach several thousand per fold, we processed the training set in batches of 80 instances. On the D-Wave Advantage system (Pegasus topology, 15-way qubit connectivity), a fully connected QUBO of up to $\sim$150 variables can be minor-embedded (mapping logical problem variables onto physical qubits), so 80 is not a hard hardware limit but a deliberate tradeoff: smaller batches admit high-quality minor embeddings found quickly, keeping chain lengths short and per-batch QPU time low, at the cost of some cross-batch redundancy. In this study, we used the same batch size for SA to ensure a fair SA-vs-QA comparison. Batches are assigned by stratified splitting on KMeans cluster labels; thus each batch contains instances from all major clusters. This limits but does not eliminate cross-batch redundancy; two near-duplicate instances in different batches can be selected because the off-diagonal penalty only operates within a batch.

4.2.1. Cluster-Aware Off-Diagonal Penalty

The standard off-diagonal penalty in [5] uses signed cosine similarity between embedding pairs; pairs from the same class are rewarded for being distant (encouraging same-class diversity), and pairs from different classes are rewarded for being close (encouraging informative borderline pairs). This works well when the embedding space is roughly uniform; however sentiment data often form tight clusters of paraphrases, and a flat penalty treats every within-cluster duplicate the same as a cross-cluster pair at the same cosine distance.

We modified the penalty to weight within-cluster pairs more heavily. KMeans is first fitted on the full fold's embeddings to produce cluster labels used for stratified batch splitting (Section 4.2); it is then re-fitted independently on each batch's embeddings for QUBO construction. After this per-batch KMeans fit (with 25 clusters in our final configuration), the off-diagonal entry becomes

$$Q_{ij}^{\text{off}} = w_{\text{sim}} \cdot \text{sign}(c_{ij}) \cdot |\cos(e_i, e_j)| + w_{\text{cluster}} \cdot \cos(e_i, e_j) \cdot \exp\left(-\frac{\|e_i - e_j\|}{\sigma_{c_i}}\right) \cdot \mathbf{1}[c_i = c_j], \quad (2)$$

where e_i is the embedding of instance i, c_i is its KMeans cluster, σ_{c_i} is the standard deviation of the intra-cluster pairwise distances (computed via `compute_sigma_dict`), and $\text{sign}(c_{ij})$ flips depending on whether i and j share a class label. Note that the cluster term uses the signed cosine $\cos(e_i, e_j)$ rather than its absolute value: for same-cluster, same-class pairs, the signed cosine is positive (penalizing redundant selection), while same-cluster, cross-class pairs receive a negative value (reducing the penalty, since cross-class pairs near the boundary are informative). The intuition is that within-cluster, same-class near-paraphrases are the primary source of redundancy; the exponential decay envelope concentrates the penalty exactly there, scaling naturally with cluster spread.

4.2.2. Representativeness-Aware Diagonal

The base diagonal in our cluster-aware QUBO is the inverse SVM-margin score: a one-versus-rest SVM is fitted on each batch, the absolute decision-function value gives a margin per instance, and the score is defined as $\text{margin}_i = 1/(\epsilon + |f(x_i)|)$, where $f(x_i)$ is the raw decision-function value. This ensures that instances near the decision boundary (small $|f(x_i)|$) receive high scores and are preferentially selected.

At aggressive reduction ratios ($\geq$60%), the margin-only diagonal systematically over-selects hard examples near the decision boundary and under-selects representative-class interiors. To correct this, we add a representativeness term: the inverse distance from each instance to its class centroid, mixed with the inverse margin score:

$$\text{score}_i = \alpha \cdot \text{margin}_i + (1 - \alpha) \cdot \frac{1}{\epsilon + \|e_i - \mu_{y_i}\|}, \quad (3)$$

where μ_{y_i} is the centroid of the embeddings labelled with class y_i, ϵ is a small numerical stabilizer, and α controls the hardness-versus-representativeness mix. We use $\alpha = 0.5$. Both margin_i and $1/(\epsilon + \|e_i - \mu_{y_i}\|)$ are independently min-max normalized to $[0, 1]$ before mixing, so that α controls the true balance rather than being confounded by differences in the raw magnitude. The representativeness term ensures that the selected subset covers the interior of each class, not only its margin. We refer to the QUBO formulation that uses this augmented diagonal as the `svc_cluster_rep`, and it is the variant used in our final submission.

4.2.3. Cardinality Enforcement

The hard cardinality constraint $\sum_i x_i = k$ is encoded as a quadratic penalty added to the QUBO as follows:

$$P_{\text{card}}(\mathbf{x}) = \lambda \left(\sum_i x_i - k\right)^2, \quad (4)$$

with $\lambda = \max_i \Big(|Q_{ii}| + \sum_{j \neq i} |Q_{ij}| \Big)$, the maximum single-variable energy delta, that is, the largest change in the objective value from flipping any one variable. This auto-scaling ensures that the cardinality penalty dominates the objective whenever the constraint is violated, regardless of the off-diagonal term magnitudes.

4.2.4. QUBO Normalisation

The D-Wave hardware auto-scales the submitted QUBO before annealing, which can crush small-magnitude terms to zero. To prevent the off-diagonal terms (which are bounded by the cosine values in $[-1, 1]$) from being dominated by the larger-magnitude diagonal scores, we normalize the off-diagonal block by the infinity norm and scale it so that $\max |Q^{\text{off}}| = r \cdot \max |Q^{\text{diag}}|$ for a configurable ratio r (we use $r = 0.5$ in the final configuration).

4.3. Noise Pre-Filtering

The organizers state that approximately 20% of the labels are corrupted. We apply Cleanlab's confident-learning pre-filter [6] to the training set of each fold before QUBO construction. Confident learning estimates label noise without access to ground-truth labels: it first obtains an out-of-sample predicted class distribution for every instance (here via a 5-fold cross-validated Logistic Regression), and derives a per-class *self-confidence* threshold, the average predicted probability the model assigns to instances of each given class. An instance is flagged as likely mislabelled when the model confidently assigns it to a different class than its given label, that is, when that class's predicted probability exceeds the threshold while the given label's does not. The flagged instances are ranked by self-confidence, and we remove them before QUBO construction, capping the removal rate at 35% as a safety guard against over-filtering.

In our setting, Cleanlab flagged only ~2.3% of instances per fold. This is substantially below the 20% nominal noise rate, suggesting that the noise injection deliberately targets borderline examples that appear plausible under the embedding geometry. Nevertheless, even this modest filter improves the worst-fold F1 by $+0.0028$ and tightens the cross-fold standard deviation from 0.0122 to 0.0110 in our final configuration, a small but consistent gain that we include in the submission.

4.4. Solver

We used `SimulatedAnnealingSampler` from the D-Wave Ocean SDK [12], running locally on the CPU for SA submissions, and the D-Wave Advantage system via the `LazyFixedEmbeddingComposite` for QA submissions. For both samplers, we drew 400 reads per batch (repeated anneals from different random initializations) and took the lowest-energy sample. The QA composite caches the minor embedding after the first batch; therefore all subsequent batches in a fold reuse the same physical qubit mapping, which keeps the per-batch QPU time near-constant at ~360 ms.

We chose to develop primarily with SA because it is deterministic given a fixed seed, runs locally without quota consumption, and is reliable for our problem size of 80 logical qubits per batch. The QA submission uses the identical QUBO formulation as a methodological demonstration that the cluster-aware design transfers cleanly from classical to quantum solvers.

The full implementation is available at https://github.com/dsgt-arc/QCLEF-2026_Instance_Selection.

5. Hyperparameter Optimization

5.1. Multi-Objective Bayesian Search

The QUBO formulation has four free weights (`similarity_weight` (w_{sim}), `cluster_similarity_weight` (w_{cluster}), `offdiag_to_diag_ratio` (r), and `n_clusters`), in addition to the selection ratio `percentage_keep`. These interact nontrivially; the optimal cluster

Table 1
DistilBERT fine-tuning configuration for our internal validation experiment (Section 7.4). This is a sanity check we ran ourselves and is separate from the organiser's official evaluation.

Hyperparameter	Value
Epochs	10 (early-stop patience 3)
Batch size	32
Learning rate	2×10^{-5}
Warmup ratio	10%
Weight decay	0.01
Optimizer	AdamW

count depends on the embedding geometry, and the optimal weight balance depends on the cluster count and the chosen reduction ratio.

We framed the search as a multi-objective Bayesian optimization problem: jointly maximise Macro-F1 and the size reduction. We use BoTorch's q-Expected Hypervolume Improvement in log-space [13, 14] with a Gaussian Process surrogate (10 Sobol initialization trials + 25 BO iterations) and a reference point of $(\text{F1} = 0.85,\ \text{Reduction} = 0.20)$. The BoTorch sweep identified the following Pareto-frontier configuration:

- `percentage_keep` = 0.30, `n_clusters` = 25, $w_{\text{sim}} = 0.3$, $w_{\text{cluster}} = 0.3$, $r = 0.5$, `num_reads` = 400.

After 5-fold verification, this base configuration achieved F1 = 0.9497 ± 0.0119 at 70.37% reduction with the original cluster-aware diagonal. We then applied two further enhancements that were not in the BoTorch search space: the representativeness-aware diagonal (Section 4.2.2) and the Cleanlab pre-filter (Section 4.3). The combined configuration was submitted.

6. Experimental Setup

Dataset. VADER NYT sentiment with ∼20% injected label noise, 5 pre-defined fold files; each fold uses the documents within that fold file as training data (∼4,000 instances per fold), and the selected indices are written relative to that fold file.

Primary evaluator. Logistic Regression on Sentence-BERT embeddings, `liblinear` solver, fixed seed. Fast (∼1 s per fold) enough to run in the inner loop of multi-objective optimization. Reported below as the headline metric.

Internal validation evaluator (ours). As an additional check that we ran ourselves (separate from the organiser's official evaluation), we fine-tuned DistilBERT [15] (`distilbert-base-uncased`) on the raw text of the selected instances using the configuration in Table 1. This verifies that the QUBO selection generalizes beyond the SBERT embedding space on which the QUBO is built.

Hardware. QUBO selection, KMeans clustering, and LogReg evaluation were run on a virtualized 32-vCPU PACE node. DistilBERT validation runs were attempted on the same CPU node and confirmed at $\texttt{keep} = 0.75$ (the only setting feasible within our compute budget, requiring ∼30 hours per 5-fold sweep without a GPU). QA submissions run on the D-Wave Advantage system through the CINECA-hosted QuantumCLEF API.

7. Results

7.1. Method Ablation at 70% Reduction

To understand whether each QUBO variant adds value over the baseline cluster-aware formulation, we ran all implemented methods at `percentage_keep` = 0.30 on all 5 folds, evaluated with the LogReg

Table 2
Method ablation at 70% reduction, 5-fold LogReg proxy evaluation. Fold 2 is the hardest fold in this dataset; the no-IS LogReg baseline on Fold 2 is only 0.8933 compared to ≥ 0.95 on the other four folds, so Fold 2 is reported separately as a stress-test.

Method	Mean F1	Std F1	Reduction	Fold 2 F1
`bcos` [9] (prior submission)	0.8879	0.1122	70.00%	0.6637
`svc_cluster_noise`	0.9381	0.0331	70.37%	0.8722
`svc_cluster_full`	0.9472	0.0146	70.37%	0.9182
`svc_cluster`	0.9497	0.0119	70.37%	0.9262
`svc_cluster_rep` + **denoise (ours)**	**0.9501**	**0.0110**	**70.88%**	**0.9286**

Table 3
Final `svc_cluster_rep` + denoise submission, 5-fold cross-validation. Both samplers were given the identical QUBO. Internal evaluation uses a LogReg-on-SBERT proxy; official evaluation uses the organiser's fine-tuned text classifier on raw text.

Metric	SA	QA
Mean Macro-F1 (LogReg proxy, internal)	0.9501 ± 0.0110	0.9501 ± 0.0110
Mean Macro-F1 (official evaluation)	$\mathbf{60.4 \pm 7.4}$	$\mathbf{64.8 \pm 8.9}$
Mean Reduction (official)	71.44%	67.62%
Mean annealing time per fold	∼110 s	∼17.6 s
Total annealing time across 5 folds	$551\ s$	$88\ s$ QPU access
Per-batch QPU access time	-	∼360 ms

proxy (Table 2). The base `bcos` formulation from our prior submission [9] was included as a sanity-check baseline.

The representativeness-aware variant wins on every cross-fold metric: best mean F1, lowest variance, highest reduction (Cleanlab removes an extra ∼0.5% on top of the base 70.37%), and best Fold 2 performance. Notably, the prior-year `bcos` submission collapses on Fold 2 (F1 = 0.66), and the noise-aware diagonal variant (`svc_cluster_noise`) shows the second-highest variance; the per-instance label-confidence signal is unstable across CV splits. The full-stack combination (`svc_cluster_full`) underperforms `svc_cluster_rep` alone, confirming that the representativeness signal is the dominant contributor and that stacking correlated signals dilutes rather than helps it. These proxy values are substantially higher than the official evaluation scores (different evaluator model); however, since all variants were evaluated under identical conditions with the same proxy, the relative ordering in Table 2 is expected to be preserved.

7.2. Final Submission Results

We submit the identical `svc_cluster_rep` + denoise QUBO formulation under both SA and QA samplers (Table 3). The QA submission completed all 245 anneals (5 folds $\times$ ∼49 batches per fold) on the D-Wave Advantage system through the CINECA-hosted API. The SA submission has a slightly higher reduction (71.44%) than QA (67.62%), which may seem counter-intuitive since both use the same `percentage_keep` = 0.30. This reflects run-to-run variation in Cleanlab's cross-validated noise detection: the QA run was executed on a separate pass after correcting the submission index format, and Cleanlab flagged a slightly different set of borderline instances (∼0.4% of the fold) in that pass. The resulting QUBO inputs are near-identical; we treat this as a negligible confound given the 4.4-point F1 gap between the two samplers.

Table 4
F1 versus reduction trade-off for `svc_cluster_rep`, 5-fold LogReg proxy.

`percentage_keep`	Mean F1	Std F1	Reduction
0.75	0.9492	0.0120	25.13%
0.50	0.9439	0.0225	50.00%
0.30 (submitted)	**0.9501**	**0.0110**	**70.88%**

Table 5
Official QCLEF 2026 Task 2 results. Best QA and best SA submission per team shown. Type H (R2AI) denotes a Hybrid classical-quantum sampler, not counted as pure QA. GPLSI's best SA and best QA come from different submissions (reduction range 12-82%). DS@GT is our submission.

Group	Best QA F1	Best SA F1	Reduction
R2AI	75.7 (H)	67.8	26%
GPLSI	55.5	67.9	12-82%
Entropy of Void	63.9	65.9	60%
DS@GT (ours)	**64.8**	**60.4**	**71%**
FAST-NUCES	58.5	56.3	70%

7.3. Selection-Ratio Sensitivity

To characterise the F1-versus-reduction trade-off, we ran `svc_cluster_rep` at three selection ratios with the same proxy evaluator (Table 4).

Notably, the lowest `percentage_keep` (most aggressive reduction) yields the *highest* mean F1 with the *lowest* variance. This is consistent with our hypothesis: once Cleanlab removes the most obvious noise and the representativeness diagonal is engaged, the selected subset is denser in clean, informative examples than the full noisy training set itself. The aggressive cardinality constraint also forces the QUBO to concentrate the selection budget on instances that simultaneously satisfy the representativeness, hardness and diversity criteria.

7.4. Independent DistilBERT Validation

To check that the selection generalises beyond the SBERT-embedding evaluator the QUBO was built on, we ran our own full 5-fold DistilBERT fine-tuning pipeline (Table 1) on the selected indices at `percentage_keep` = 0.75. The DistilBERT Macro-F1 was 0.7429 ± 0.0017. This is below the LogReg-proxy F1 (0.9492 at the same setting) because LogReg is evaluating on the same embedding space the QUBO was optimized on, while DistilBERT fine-tunes from scratch on raw text. This internal-evaluator number is the more conservative one for cross-system comparison.

We report this internal DistilBERT result (0.7429 at 25% reduction) as a standalone sanity check on the selected instances; we do not compare it head-to-head against the 2025 leaderboard, which was produced under a different classifier (LLaMA 3.1 8B) rather than DistilBERT. We could not run 5-fold DistilBERT at our submitted `keep` = 0.30 setting because the only available compute was a 32-vCPU virtual machine without a GPU, on which a single fold of DistilBERT fine-tuning takes ~8 hours. We expect the DistilBERT F1 at `keep` = 0.30 to be somewhat lower than at `keep` = 0.75 given less training data, but the LogReg trend (rising from 0.75 to 0.30) suggests it will not collapse.

7.5. Official Competition Results

Table 5 shows the official QCLEF 2026 Task 2 leaderboard. Our QA submission (`svc_cluster_rep`) achieves Macro-F1 $= 64.8$ at 67.62% reduction, ranking 4th overall and 1st among teams using pure quantum annealing (R2AI used a Hybrid classical-quantum sampler, denoted H, not counted as pure QA). Our SA submission achieves 60.4 at 71.44% reduction.

The most notable finding is that our QA submission outperforms our SA submission by 4.4 F1 points (64.8 vs 60.4), while every other participating team shows SA $\geq$ QA. DS@GT is the only team with a pure QA score above its SA score, and with 64.8 it ranks 1st among pure QA submissions. This is the clearest signal in the competition that quantum annealing adds value beyond simulated annealing, and it is consistent with our hypothesis that the cluster-aware QUBO at aggressive reduction ratios (30% keep) creates a more rugged energy landscape in which quantum tunnelling navigates between local minima more effectively than classical simulated annealing. We caution, however, that the gap should not be over-interpreted: the standard deviations are large (± 8.9 for QA, ± 7.4 for SA across five folds); the dataset is small; and a controlled experiment with more folds or a larger dataset would be needed to make a stronger claim of quantum advantage.

A secondary finding is the large gap between our internal LogReg-on-SBERT proxy (0.9501) and the official evaluation (64.8 for QA). The two metrics measure fundamentally different things: the proxy measures linear separability of SBERT embeddings, while the official evaluation measures how well a fine-tuned text classifier generalises from the selected raw texts. A better internal proxy (such as DistilBERT fine-tuned on the selected instances) would have produced more calibrated pre-submission estimates.

8. Discussion

Why the cluster-aware penalty helps.

Sentiment datasets contain many near-paraphrase clusters (reviews expressing the same opinion in slightly different wording). The standard pairwise cosine penalty in [5] treats all similar pairs equally, but within-cluster pairs are more interchangeable than cross-cluster pairs: selecting either member of a within-cluster pair adds roughly the same information. The distance-modulated within-cluster term in Section 4.2.1 explicitly penalises redundancy where redundancy is highest, pushing the optimizer toward selecting one representative per tight cluster. The effect is largest at aggressive reduction ratios where the optimizer has less budget to spend on duplicates.

Why representativeness matters at low keep.

At `keep` = 0.75 the margin-only diagonal (`svc_cluster`) already retains enough class interior simply by virtue of selecting 75% of the data, so the optimizer has room to be choosy about hard examples without losing coverage. At `keep` = 0.30 this changes: with only 30% of the data selected, prioritising hard near-boundary examples leaves the class interiors under-represented and degrades the resulting classifier. The representativeness term in Section 4.2.2 explicitly preserves class-interior coverage and is what allows F1 to *rise* as reduction increases (Table 4).

SA versus QA.

The official evaluation reveals a meaningful difference between samplers that our internal proxy did not detect: the QA submission scores 64.8 ± 8.9 versus 60.4 ± 7.4 for SA, a 4.4-point gap in favour of QA. Among all participating teams, DS@GT is the only one where QA outperforms SA; every other team shows SA $\geq$ QA (see Table 5). One plausible explanation is that the cluster-aware QUBO at aggressive reduction ratios (30% keep) creates an energy landscape with many shallow local minima, corresponding to different "good" subsets of instances that are each internally diverse and representative but differ from each other. Classical simulated annealing at a fixed temperature schedule may settle into whichever basin it first encounters, whereas quantum tunnelling allows the QPU to move between basins and sample from a wider region of low-energy configurations. The selection budget at 30% keep is tight enough that the difference between basins is detectable by the organiser's downstream classifier.

The standard deviations (± 8.9 for QA, ± 7.4 for SA) overlap substantially, and the competition comprised only five folds. We are not claiming quantum supremacy: at this scale, sampling variance

alone could account for a 4-point gap. What we can say is that no other team in this competition showed the same QA-over-SA pattern, which makes the result worth investigating at larger problem sizes with more controlled experimental conditions.

What did not work as well.

We explored two variants that we ultimately did not submit:

- *Noise-aware diagonal* (`svc_cluster_noise`): weighting the QUBO diagonal by cross-validated label confidence instead of (or in addition to) SVM margin. At `keep` = 0.30 this achieved F1 = 0.9381 ± 0.0331, competitive on average but unstable, with Fold 2 collapsing to F1 = 0.87. The per-instance label-confidence signal is too sensitive to the random CV split.
- *Full-stack variant* (`svc_cluster_full`): combining representativeness, label confidence and multi-scale clustering simultaneously. This produced F1 = 0.9472 ± 0.0146, worse than the representativeness alone. The signals are correlated and stacking them dilutes the strongest one.

Additionally, we did not use the LLaMA 3.1 8B embeddings provided by the organizers; the encoder-based Sentence-BERT embeddings (Section 4.1) gave consistently better LogReg-proxy F1 in pilot experiments, consistent with the observation that generative-model hidden states are less well-calibrated for discriminative similarity than encoder embeddings of comparable scale.

9. Conclusion

We presented a cluster-aware QUBO formulation for instance selection that jointly handles diversity (via a within-cluster exponential penalty), informativeness (via SVM margin scoring), representativeness (via inverse distance to per-class centroid), and noise (via Cleanlab pre-filtering). The key internal finding is that, on the noise-injected VADER NYT dataset, the LogReg-proxy Macro-F1 *rises* from 0.9492 at 25% reduction to 0.9501 at 71% reduction, contrary to the conventional intuition that aggressive reduction always degrades F1. This is enabled by the combination of representativeness-aware scoring (which preserves class-interior coverage at low keep ratios) and noise pre-filtering (which removes the most obviously mislabelled examples before they reach the QUBO).

On the official competition evaluation, the QA submission achieves Macro-F1 $= 64.8 \pm 8.9$ at 67.62% reduction, ranking 4th overall and 1st among pure quantum annealing submissions (R2AI used a Hybrid sampler). The SA submission achieves 60.4 ± 7.4 at 71.44% reduction. DS@GT is the only team in the competition where QA outperforms SA, a 4.4-point gap that is consistent with quantum tunnelling navigating the cluster-aware energy landscape more effectively than simulated annealing at aggressive reduction ratios. Given the large per-fold standard deviations, this result warrants further investigation at scale, but DS@GT produced the strongest positive QA-vs-SA signal in the QCLEF 2026 competition. Our own internal DistilBERT validation at `keep` = 0.75 achieves F1 = 0.7429, confirming that the selection generalises beyond the SBERT embedding space on which the QUBO was built.

The result has a practical implication: the usual assumption that aggressive instance selection degrades classifier quality may not hold when the QUBO jointly encodes diversity, representativeness, and noise resistance. The penalty structure, not just the keep ratio, determines what the annealer selects.

Acknowledgments

We thank the QuantumCLEF organizers for providing CINECA-hosted D-Wave access and the noise-injected VADER NYT dataset [2, 3]. This work was supported by the Data Science at Georgia Tech (DS@GT) student organisation. Compute for SA experiments and submission generation ran on the Partnership for an Advanced Computing Environment (PACE) at Georgia Institute of Technology [16].

Declaration on Generative AI

During the preparation of this work, the author used Google Gemini for grammar and spelling checks, converting text from word-processed format to LaTeX, and editorial assistance. After using these tool(s)/service(s), the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] A. Pasin, M. Ferrari Dacrema, P. Cremonesi, W. Cunha, M. Gonçalves, N. Ferro, QuantumCLEF 2026: The third edition of the quantum computing lab at CLEF, in: Advances in Information Retrieval. Proceedings of the 48th European Conference on Information Retrieval (ECIR 2026), Lecture Notes in Computer Science, Springer, Cham, 2026. URL: https://doi.org/10.1007/978-3-032-21321-1_41. doi:10.1007/978-3-032-21321-1_41.

[2] A. Pasin, M. F. Dacrema, W. Cunha, M. A. Gonçalves, P. Cremonesi, N. Ferro, Overview of QuantumCLEF 2026: The third quantum computing challenge for information retrieval and recommender systems at CLEF, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction — 17th International Conference of the CLEF Association, CLEF 2026, Jena, Germany, September 21–24, 2026, Proceedings, Lecture Notes in Computer Science, Springer, 2026.

[3] A. Pasin, M. F. Dacrema, W. Cunha, M. A. Gonçalves, P. Cremonesi, N. Ferro, QuantumCLEF 2026: Overview of the third quantum computing challenge for information retrieval and recommender systems at CLEF, in: Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2026, Jena, Germany, 21–24 September 2026, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[4] A. Pasin, M. Ferrari Dacrema, P. Cremonesi, W. Cunha, M. Gonçalves, N. Ferro, Overview of QuantumCLEF 2025: The second quantum computing challenge for information retrieval and recommender systems at CLEF, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), volume 16133 of *Lecture Notes in Computer Science*, Springer, Cham, 2025. URL: https://doi.org/10.1007/978-3-032-04354-2_22. doi:10.1007/978-3-032-04354-2_22.

[5] A. Pasin, M. Ferrari Dacrema, P. Cremonesi, N. Ferro, A quantum annealing instance selection approach for efficient and effective transformer fine-tuning, in: Proceedings of the 2024 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '24), Association for Computing Machinery, 2024. doi:10.1145/3664190.3672515.

[6] C. G. Northcutt, L. Jiang, I. L. Chuang, Confident learning: Estimating uncertainty in dataset labels, Journal of Artificial Intelligence Research 70 (2021) 1373–1411. doi:10.1613/jair.1.12125.

[7] C. Hutto, E. Gilbert, VADER: A parsimonious rule-based model for sentiment analysis of social media text, in: Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM), AAAI Press, 2014. doi:10.1609/icwsm.v8i1.14550.

[8] A. Pasin, M. Ferrari Dacrema, P. Cremonesi, N. Ferro, Overview of QuantumCLEF 2024: The quantum computing challenge for information retrieval and recommender systems at CLEF, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), volume 14959 of *Lecture Notes in Computer Science*, Springer, Cham, 2024. URL: https://doi.org/10.1007/978-3-031-71908-0_12. doi:10.1007/978-3-031-71908-0_12.

[9] C. Pomeroy, A. Pramov, K. Thakrar, L. Yendapalli, Cluster-aware QUBO instance selection for noise-robust sentiment classification at QuantumCLEF 2025, 2025. URL: https://arxiv.org/abs/2507.15063. doi:10.48550/arXiv.2507.15063. arXiv:2507.15063.

[10] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, Journal of Machine Learning Research 9 (2008) 2579–2605.

[11] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-Networks,

in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. doi:10.18653/v1/d19-1410.

[12] D-Wave Systems, D-Wave Ocean SDK documentation, 2024. URL: https://docs.ocean.dwavesys.com/, accessed: 2026-06-01.

[13] S. Daulton, M. Balandat, E. Bakshy, Differentiable expected hypervolume improvement for parallel multi-objective Bayesian optimization, in: Advances in Neural Information Processing Systems 33, 2020. doi:10.48550/arxiv.1910.06403.

[14] M. Balandat, B. Karrer, D. R. Jiang, S. Daulton, B. Letham, A. G. Wilson, E. Bakshy, BoTorch: A framework for efficient Monte-Carlo Bayesian optimization, in: Advances in Neural Information Processing Systems 33, 2020. doi:10.48550/arxiv.2006.05078.

[15] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019). doi:10.48550/arxiv.1910.01108.

[16] PACE, Partnership for an Advanced Computing Environment (PACE), 2017. URL: http://www.pace.gatech.edu, accessed: 2026-06-01.