# Overview of QuantumCLEF 2025: The Second Quantum Computing Challenge for Information Retrieval and Recommender Systems at CLEF

Andrea Pasin[1(✉)], Maurizio Ferrari Dacrema[2], Washington Cunha[3], Marcos André Gonçalves[3], Paolo Cremonesi[2], and Nicola Ferro[1]

[1] University of Padua, Padua, Italy
andrea.pasin.1@phd.unipd.it, nicola.ferro@unipd.it
[2] Politecnico di Milano, Milan, Italy
{maurizio.ferrari,paolo.cremonesi}@polimi.it
[3] Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
{washingtoncunha,mgoncalv}@dcc.ufmg.br

**Abstract.** Quantum Computing (QC) is an emerging research field that is attracting significant interest from the scientific community due to its potential to solve complex problems more efficiently than traditional computers by leveraging the principles of quantum physics. Even though real quantum computers exist, at the moment we are still in the early stages of development of these innovative technologies, and many of their capabilities and limitations are yet to be discovered.

In this work, we present an overview of the second edition of QuantumCLEF, a lab that focuses on the application of Quantum Annealing (QA), a specific QC paradigm, for different tasks related to IR and RS. The main objective of the QuantumCLEF lab is to investigate QC, raise awareness, and develop and evaluate new QC algorithms for different applications. This lab represents a great chance for researchers and industry practitioners to understand more about this new field by having access to real quantum computers, which are still not easily accessible nowadays.

This edition consisted of three different tasks: Feature Selection for IR and RS systems, Instance Selection for IR systems, and Clustering for IR systems. There have been a total of 44 teams that registered for this lab, and eventually, 5 teams managed to successfully submit their runs following the lab guidelines. Participants have been provided with examples, tutorials, and comprehensive materials due to the novelty of the QC field, allowing them to understand how QA works and how to program quantum annealers.

## 1 Introduction

Quantum Computing (QC) is a new computational paradigm that focuses on leveraging the quantum mechanical principles such as superposition, entanglement, and tunnelling to perform calculations. Quantum computers have the


J. Carrillo-de-Albornoz et al. (Eds.): CLEF 2025, LNCS 16089, pp. 412–435, 2026.
https://doi.org/10.1007/978-3-032-04354-2_22

potential to revolutionize the way we solve tasks in several fields, especially when dealing with complex combinatorial problems, due to their capabilities in exploring large search spaces very efficiently [20].

Nowadays, there already exist real quantum computers. However, we are still in their early stages of development, and researchers are studying the capabilities of these machines while, at the same time, trying to overcome some of their current limitations. In fact, quantum computers are really delicate to work with, since noise (e.g., thermal fluctuations, electromagnetic interferences) can easily break computation. Furthermore, existing quantum computers have a limited number of qubits, which limits the size of problems that can be targeted. Nevertheless, it is already possible to have access to these technologies to learn how they can be applied and what benefits they offer.

In 2024, the QuantumCLEF lab [31–34] was started with the goal of studying how QC technologies could be used in the fields of IR and RS. It represented the first evaluation challenge, giving participants access to real quantum computers to develop and evaluate algorithms for different practical tasks. This year, a second edition of the lab was conducted to further shed light on the potential of QC for IR and RS, with the following goals:

- develop and evaluate new QC algorithms for IR and RS, comparing the results (efficiency and effectiveness) with traditional approaches;
- gather all resources and data for future researchers to compare their results with the ones achieved during the lab;
- allow participants to learn more about QC through comprehensive materials and to use real quantum computers, which are still not easily accessible to the public;
- raise the awareness of the potential of QC and form a new research community around this new field.

In this paper, we present the overview of the second edition of QuantumCLEF held in 2025 [30]. Similarly to the previous edition, this edition has focused on the usage of Quantum Annealing (QA), a specific QC paradigm that can be used to tackle optimization problems. We have granted participants access to the state-of-the-art QA devices (quantum annealers) produced by D-Wave, one of the leading companies in this sector. The QA paradigm is easier to understand with respect to the Universal Gate-Based paradigm. Furthermore, D-Wave provides several tools and libraries to program quantum annealers without requiring a particularly deep knowledge of the quantum physics governing these devices.

QuantumCLEF 2025 was composed of three main tasks:

- **Task 1**: Feature Selection for IR and RS;
- **Task 2**: Instance Selection [11,12,29] for IR;
- **Task 3**: Clustering for IR.

Participants were invited to design and implement their own algorithms to address the proposed tasks using both QA and Simulated Annealing (SA). SA is a well-established optimization technique that shares some similarities with QA,

but it operates entirely on classical hardware and does not exploit or simulate any quantum phenomena. Given the novelty and technical complexity, participants were provided with extensive support materials, including instructional videos, slides, tutorials, and code examples, to facilitate their understanding of QA and to help them solve the tasks using quantum annealers.

To ensure easy accessibility, the KIMERA infrastructure [35] has been used. This platform not only allowed participants to run their algorithms on actual quantum annealers but also helped improve reproducibility, comparability, and simplification of the workflow management.

A total of 44 teams registered for the lab, out of which 6 teams actively participated. Finally, 5 teams submitted at least one run. Due to an unforeseen circumstance that prevented access to quantum computers in the last 5 days of the challenge, many of the teams' submissions involve the usage of SA only. Despite this issue, the submitted solutions featured innovative and well-structured techniques that remain highly relevant to the proposed problems. Many of these approaches are readily adaptable to QA, underlining their potential for future quantum implementations. Overall, the experimental results show similar trends to the previous 2024 edition: QA or Hybrid (H) approaches usually perform on par with SA and other classical techniques, with the added advantage of generally higher efficiency in terms of annealing time.

The paper is organized as follows: Sect. 2 discusses related works; Sect. 3 presents the tasks of the QuantumCLEF 2024 lab while Sect. 4 introduces the lab's setup; Sect. 5 shows and discusses the results achieved by the participants; finally, Sect. 6 draws some conclusions and outlooks some future work.

## 2 Related Works

In this section, we provide an introduction to QA and SA. Furthermore, we summarize the tasks and the main outcomes of the previous QuantumCLEF edition.

### 2.1 Background on Quantum and Simulated Annealing

**Quantum Annealing.** QA is a QC paradigm based on special-purpose devices, known as quantum annealers, designed to solve optimization problems that are properly formulated using specific formats. The fundamental idea behind a quantum annealer is to encode the problem into the energy configuration of a physical system. Quantum-mechanical phenomena such as superposition, entanglement, and tunneling are then exploited to guide the system toward its lowest energy state, which corresponds to the optimal solution of the original problem.

To use quantum annealers for a given optimization problem, first, it must be formulated as a minimization problem using the Quadratic Unconstrained Binary Optimization (QUBO) formulation [19], a well-established mathematical framework. The QUBO problem is defined as:

$$\min \quad y = x^T Q x \tag{1}$$

where $x$ is a vector of binary decision variables, and $Q$ is a matrix of constant values encoding the problem to be solved.

An additional step, known as *minor embedding*, is required to map this mathematical formulation onto the physical hardware of the quantum annealer, taking into account the limited number of available qubits and their connectivity. In fact, each Quantum Processing Unit (QPU) has a specific hardware topology that can be represented as a graph: vertices correspond to qubits, and edges correspond to couplers (interactions) between qubits. Minor embedding involves selecting physical qubits to represent logical decision variables. If a variable requires more connections than physically available on the QPU, a chain of qubits is used to represent it, and its connections are distributed across these qubits.

As a consequence, the number of physical qubits required to represent a problem can be significantly larger than the number of logical variables. Minor embedding is, in itself, an *NP*-hard problem, typically solved using heuristic algorithms [9].

If the problem is too large to fit directly onto the QPU, D-Wave offers a H approach, which decomposes the problem into smaller sub-problems, and solves them using a combination of classical and quantum methods.

Constraints can be incorporated into the problem formulation using penalty terms $P(x)$ [45], which assign a cost to infeasible solutions. These represent *soft constraints*, meaning they push quantum annealers to favor feasible solutions but do not strictly enforce compliance. In other words, solutions violating these constraints are penalized within the optimization process and thus unlikely to be returned by quantum annealers. These penalties are added to the original objective function to obtain the final objective function:

$$\min \quad C(x) = y + P(x) \tag{2}$$

The relative influence of these penalties can be tuned via hyperparameters, depending on how *strict* the constraints need to be.

In summary, the process of solving a problem using a quantum annealer involves the following stages [45]:

1. **Formulation**: express the problem as a QUBO.
2. **Embedding**: generate a minor embedding of the QUBO onto the specific QPU architecture.
3. **Data Transfer**: submit the problem and its embedding to the quantum annealer via the global network.
4. **Annealing**: execute the quantum annealing process, which involves programming the QPU, sampling multiple solutions, and reading the results. Due to the stochastic nature of the process, it is typically repeated hundreds of times to obtain a distribution of candidate solutions. These are then filtered and evaluated to select the most optimal and feasible one.

Once the QUBO has been successfully embedded and transferred to the quantum annealer, solving the problem typically takes only a few milliseconds.

**Simulated Annealing.** SA is a consolidated meta-heuristic that can be run on traditional hardware [6,7,44]. It is a probabilistic algorithm that can be used to find the global minimum of a given cost function, even in the presence of many local minima. It is based on an iterative process that starts from an initial solution and tries to improve it by randomly perturbing it. The cost function is represented by the QUBO problem formulation, similar to what would be used for QA. In SA, there is no minor embedding phase since the problem is directly solved on a traditional machine.

We underline that SA is an optimization algorithm different from QA; it is not a simulation of QA on a traditional machine, and, therefore these two algorithms are not equivalent. However, SA can be used for benchmarking purposes to compare the performance of QA with respect to a traditional hardware counterpart.

Moreover, access to quantum annealers in QuantumCLEF is limited to ensure a fair distribution of resources. Therefore, SA can also be used to perform initial experiments to assess a QUBO formulation feasibility without affecting the available quota in the quantum environment.

### 2.2 QuantumCLEF 2024

The QuantumCLEF 2024 lab [32,34], presented at CLEF 2024, explored the application of QA for IR and RS. The lab consisted in two tasks:

- **Feature Selection:** this task focused on selecting optimal feature subsets for training IR and RS systems using QA.
- **Clustering:** Leveraging document embeddings, this task involved grouping similar documents via QA to speed up dense retrieval.

Participants were given access to D-Wave hardware through CINECA and used the KIMERA [35] infrastructure for enhanced resource access, comparability, and reproducibility.

Out of 26 registered teams, 7 submitted official runs [1,2,17,26,28,36,41]. The results demonstrated the feasibility of using real quantum annealers in IR and RS, encouraging interdisciplinary research and offering benchmarks for further exploration of QC in real-world applications.

## 3 Tasks

QuantumCLEF 2025, addresses three different tasks involving computationally intensive problems: Feature Selection, Instance Selection, and Clustering. The main goals for each task are:

- finding some possible QUBO formulations of the considered problem;
- map the QUBO formulation to the physical QPU;
- evaluating the QA approach compared to corresponding traditional approaches to assess both efficiency and effectiveness.

For each task, we provided Jupyter Notebooks as entry points for participants, helping them to learn how to program quantum annealers and to successfully complete the tasks following the submission guidelines. Additionally, we supplied the tutorial slides presented at ECIR and SIGIR [15,16], which introduce the core concepts of QC and QA. A video tutorial[1] was also made available, demonstrating the usage of our KIMERA infrastructure and the provided notebooks.

For each tasks, participants are asked to submit their runs using both QA and SA. It was recommended that participants use SA for testing purposes during the development phase of their algorithms, due to the limited availability of QA resources.

### 3.1 Task 1 Quantum Feature Selection

This task focuses on formulating the well-known *NP-Hard* feature selection problem to make it solvable through a quantum annealer, similarly to what has already been done in previous works [14,27].

**Objectives.** Feature Selection is a widespread problem for both IR and RS. It consists of identifying a subset of the available features (e.g., the most informative, less noisy, less redundant, etc.) to train a learning model for improved efficiency and, in some cases, effectiveness. This problem is very impacting since many IR and RS systems involve the optimization of complex learning models, and reducing the dimensionality of the input data can improve their performance. Therefore, in this task, we aim to understand if QA can be applied to solve this problem more efficiently and effectively, exploiting its capability of exploring large problem spaces in a short amount of time.

Feature selection fits very well the QUBO formulation, in which there is one variable $x$ per feature and its value indicates whether it should be selected or not. The challenge lies in designing the objective function, i.e., the matrix $Q$ (see Eq. 1).

**Sub-Tasks.** Task 1 is divided into two sub-tasks:

- **Task 1A**: Feature Selection for IR. This task involves selecting the optimal subset of features using QA and SA that will be used to train a LambdaMART [8] model according to a Learning-To-Rank framework;
- **Task 1B**: Feature Selection for RS. This task involves selecting the optimal subset of features using QA and SA that will be used to train a kNN recommendation system model. The item-item similarity is computed with cosine on the feature vectors, a shrinkage of 5 is added to the denominator and the number of selected neighbors for each item is 100.

[1] https://www.youtube.com/watch?v=fKrnaJn40Kk/ (accessed June 17, 2025).

**Datasets.** For Task 1A, we decided to employ the well-known MQ2007 [38] and the Istella S-LETOR [22] datasets. MQ2007 represents an easier challenge since it has 46 features, allowing direct embedding of the problem formulations inside the QPU of quantum annealers. Istella instead has 220 features, making it impossible to embed problem formulations directly, thus requiring some further processing steps for the participants to fit the problem into the current physical QPU hardware.

For Task 1B instead, we decided to employ a custom dataset of music recommendations containing 1.9 thousand users and 18 thousand items. The dataset contains both collaborative data, with 92 thousand implicit user-item interactions, as well as two different sets of item features that are derived from item descriptions and user-provided tags, called Item Content Matrix (ICM). The small set, ICM_100, includes 100 features and can be embedded directly on the QPU with small adjustments, the large set, ICM_400, has 400 features and requires significant pruning to fit in the QPU or the use of Hybrid methods. Both sets of features contain noisy and redundant features.

**Evaluation Measures.** The official evaluation measure for both Task 1A and Task 1B is nDCG@10.

**Baseline.** For sub-task 1A the baseline is a Feature Selection model that uses a Recursive Feature Elimination approach paired with a Linear Regression model to select the most relevant subset of features.

For sub-task 1B the baseline is a kNN recommendation system model that uses all the available features. The hyperparameters are the same used for the model computed on the selected features, i.e., the item-item similarity is computed with cosine, adding a shrink term of 5 to the denominator, and the number of neighbors is 100.

**Runs Format.** Participants in both tasks 1A and 1B can submit a maximum of 5 runs per dataset using QA or Hybrid methods and a maximum of 5 runs using SA. Each run that uses QA or Hybrid methods should correspond to a run that employs SA to make a fair comparison between them.

The results of the run must be a text file which lists the features that were selected, one per line. The discarded features are not reported in the run file. Furthermore, the last line must report the list of IDs associated with the problems solved using QA, SA, or Hybrid to obtain the final subset of features by the considered approach.

Each run file must be left in each team's workspace in a specific directory called */config/workspace/submissions*, which is already available.

The submission file name should comply with the format *[Task]_[Dataset]_[Method]_[Groupname]_[SubmissionID].txt*, where:

- **[Task]**: it should be either *1A* or *1B* based on the task the submission refers to;

- **[Dataset]**: it should be either *MQ2007*, *Istella*, *100_ICM* or *400_ICM* based on the dataset used;
- **[Method]**: it should be either *QA* or *SA* based on the method used;
- **[Groupname]**: the team name;
- **[SubmissionID]**: a custom submission ID that must be the same for the submissions using the same algorithm but performed with different methods (e.g., QA or SA).

### 3.2 Task 2 Quantum Instance Selection

This task focuses on formulating the Instance Selection problem and solving it with quantum annealers. Instance Selection focuses on identifying and selecting the most representative samples in a dataset to train a learning model. In this way, the training phase will be more efficient and the trained model will achieve a comparable level of effectiveness.

**Objectives.** Instance Selection is an important task in the fields of IR and RS, particularly when dealing with large-scale datasets. By selecting a representative subset of instances from a larger dataset, Instance Selection aims to reduce the time required to train a learning model and maintain (or even improve) its effectiveness [11–13]. It has already been shown that using QA for this task is possible and it allowed to reduce the training dataset size by up to 28% without compromising model performance [29].

In this task, the participants should identify subsets of the considered datasets that are then used to train a Llama3.1 7B model [42] for Text Classification and Sentiment Analysis. A good Instance Selection approach should improve the training efficiency while retaining the model's effectiveness.

**Datasets.** For this task, we considered two different datasets:

- Vader NYT: A sentiment-labeled dataset from New York Times articles;
- Yelp Reviews: A collection of customer reviews labeled with sentiment scores.

Both datasets have been split into training and test sets. The training sets have also been split into 5 folds following a 5-fold cross-validation approach.

**Evaluation Measures.** The performance will be evaluated on the test sets from the 5-fold cross-validation splits using the Macro-F1 score, ensuring a fair and comprehensive assessment of effectiveness. Furthermore, the evaluation will also consider the time required to fine-tune the model and the dataset reduction rate.

**Baseline.** For this task, the baseline is the Llama3.1 7B model trained on the whole initial datasets.

**Runs Format.** Participants can submit a maximum of 5 runs per dataset using QA or Hybrid methods and a maximum of 5 runs using SA. Each run that uses QA or Hybrid methods should correspond to a run that employs SA. In this way, it is possible to make a fair comparison between them.

The results of the run must be a text file which lists the documents that were selected, one per line. The discarded documents are not reported in the run file. Furthermore, the last line must report the list of IDs associated with the problems solved using QA, SA, or Hybrid to obtain the final subset of features by the considered approach.

Each run file must be left in each team's workspace in a specific directory called */config/workspace/submissions*, which is already available.

The submission file name should comply with the format *[Dataset]_[FoldNumber]_[Method]_[Groupname]_[SubmissionID].txt*, where:

- **[Dataset]**: it should be either *Vader* or *Yelp* based on the dataset used;
- **[FoldNumber]**: it should be a number in $[0, 4]$, representing the considered fold;
- **[Method]**: it should be either *QA* or *SA* based on the method used;
- **[Groupname]**: the team name;
- **[SubmissionID]**: a custom submission ID that must be the same for the submissions using the same algorithm but performed with different methods (e.g., QA or SA).

### 3.3 Task 3 Quantum Clustering

This task focuses on the formulation of the Clustering problem in such a way that it can be solved with a quantum annealer. It involves grouping the items according to their characteristics. Thus, "similar" items fall into the same group while different items belong to distinct groups.

**Objectives.** Clustering is important for IR and RS as it helps organize large collections, assists users in exploring content, and provides similar search results for a query.

This task focuses on IR in a document retrieval setting where documents are represented as embeddings from a Transformer model. Each document is a vector, and clusters are based on distances between vectors, representing dissimilarity. Participants should use QA and SA methods to cluster documents into 10, 25, and 50 clusters, reporting the centroids and their associated documents.

Clustering enables faster search by matching the query to the nearest centroid and retrieving documents only from that cluster instead of the entire collection.

Clustering fits the QUBO formulation, and various methods have already been proposed [3,4,43]. Most of these methods involve the usage of one variable per document, thus making it very hard to consider large datasets due to the limited number of physical qubits and interconnections between them. There are ways to overcome this issue, such as by applying a coarsening or a hierarchical approach.

**Datasets.** For this task, we considered a custom split of the ANTIQUE [21] dataset containing 6486 documents, 200 queries, and manual relevance judgments. Each document and each query have been transformed into a corresponding embedding with the pre-trained **all-mpnet-base-v2** model[2]. The queries are divided into 50 for the Training Dataset and 150 for the Test Dataset.

**Evaluation Measures.** The official evaluation measures for Task 2 are:

- the Davies-Bouldin Index to measure the overall cluster quality without considering the document retrieval phase;
- nDCG@10 to measure the retrieval effectiveness based on the clusters found.

**Baseline.** For this task, the baseline is a traditional k-Medoids approach using the cosine distance as a distance function.

**Runs Format.** Participants can submit a maximum of 5 runs for each number of clusters (i.e., 10, 25, 50) using QA or Hybrid methods and a maximum of 5 runs using SA. Each run that uses QA or Hybrid methods should correspond to a run that employs SA. In this way, it is possible to make a fair comparison between them.

The run file must be a text file (JSON formatted) with a list of 10, 25, and 50 vectors that represent the final centroids achieved through their clustering algorithm. Each centroid should also be followed by the list of documents that belong to the given cluster. Furthermore, the last line must report the list of IDs associated with the problems solved using QA, SA, or Hybrid to obtain the final clusters by the considered approach.

Each run file must be left in each team's workspace in a specific directory called */config/workspace/submissions*, which is already available.

The submission file name should comply with the format *[Centroids]_[Method]_[Groupname]_[SubmissionID].txt*, where:

- **[Centroids]**: it should be either 10, 25, or 50 based on the number of centroids;
- **[Method]**: it should be either *QA* or *SA* based on the method used;
- **[Groupname]**: the team name;
- **[SubmissionID]**: a custom submission ID that must be the same for the submissions using the same algorithm but performed with different methods (e.g., QA or SA).

## 4 Lab Setup

In this section, we detail the infrastructure used during this lab, and we present the guidelines the participants had to comply with to submit their runs.

[2] https://huggingface.co/sentence-transformers/all-mpnet-base-v2.

**Table 1.** The hardware resources corresponding to the machine where KIMERA was deployed and to the participants' workspaces.

| Hardware resources. | | | |
|---|---|---|---|
| **-** | **CPU** | **RAM** | **Hard Drive** |
| Infrastructure | 32 cores | 128 GB RAM | 1 TB HDD |
| Workspace | 1200 millicores | 12 GB RAM | 20 GB HDD |

**Table 2.** The monthly quotas to use quantum resources according to the tasks.

| Monthly quotas for the tasks. | | | | |
|---|---|---|---|---|
| Task | February | March | April | May |
| Task 1: Feature Selection | 30 s | 30 s | 30 s | 30 s* |
| Task 2: Instance Selection | 120 s | 120 s | 120 s | 120 s* |
| Task 3: Clustering | 120 s | 120 s | 120 s | 120 s* |

* From 05/05/2025 to 10/05/2025 (submissions deadline), quantum annealers were unavailable

### 4.1 Infrastructure

Access to quantum annealers is limited. D-Wave enforces it through monthly time quotas and the use of API keys to monitor and control usage. Since sharing our API key with participants is not possible, we decided to use KIMERA, which enables participants to access quantum annealers without requiring their own API keys or agreements with D-Wave. It also ensures fairness and reproducibility, since all participants are provided with identical computational workspaces, each equipped with the same CPU and RAM resources. This uniform environment facilitates consistent efficiency measurements. Moreover, it also allows participants to monitor quotas and develop, test, and run code directly from their browsers, eliminating the need for local experiments and for owning computational resources. Finally, all participants' submissions are stored in a database to track quota usage and collect data for analysis and reporting.

The final infrastructure was deployed on a machine hosted at the Department of Information Engineering at the University of Padua. Table 1 reports the specifications of the hardware resources corresponding to the machine and to each team's workspace. All participants were given the same monthly quota to use quantum resources. Table 2 reports the monthly quotas according to the three tasks.

### 4.2 General Guidelines

Each team has access to its personal area inside our infrastructure with the credentials that have been provided to them. All runs must be executed by using the workspaces that have been created for each one of the participating teams, thus ensuring a fair comparison and easy reproducibility.

**Table 3.** The teams that participated and submitted to QuantumCLEF 2025.

| Team | Affiliation | Country |
|---|---|---|
| DS@GT qClef [37] | Georgia Institute of Technology | United States |
| FAST-NU [40] | National University of Computer and Emerging Sciences | Pakistan |
| GPLSI [10] | Language Processing and Information Systems (GPLSI), University of Alicante. | Spain |
| Malto [18] | Politecnico di Torino | Italy |
| SINAI-UJA [24] | Universidad de Jaén | Spain |

All participants cannot exceed their given quotas (see Table 2) to execute problems on quantum devices. The quotas can be monitored by each participating team through a dashboard that is constantly being automatically updated, reporting usages of the different methods (i.e., QA, H, and SA) and some general statistics.

All participants' runs must follow the file formats that are detailed in Section 3.1,3.2, and 3.3 for an easy automated evaluation with the prepared scripts.

## 5 Results

In this section, we present the results achieved by the participants and we discuss their approaches. Out of the 44 registered teams, 5 teams managed to upload some final runs. In total, the number of runs is 69, considering both SA, QA, and H(H was introduced in Sect. 2.1). Table 3 reports the 5 teams that correctly participated and submitted some final runs.

In total, throughout the entire lab, participants have submitted 7183 problems (vs 976 in QuantumCLEF 2024). Specifically, 6333 of them were solved with SA, while 848 were solved using QA and 2 with the H method. The total execution time of SA has been more than 4 h, while the total QA and H execution time has been roughly 1 min.

The QA execution time in this whole Section refers to the *Annealing* phase as described in Sect. 2.1, therefore it includes the time required to program the QPU, sampling, and reading the result. The embedding time and network latencies are not taken into account and are left to be considered for possible future editions of the QuantumCLEF lab.

### 5.1 Task 1A: Quantum Feature Selection for IR

Here we present the results achieved by the teams participating in task 1A, considering each dataset separately.

**MQ2007 Dataset.** As it is possible to see in Table 4, teams considered different numbers of features in their submissions. In general, we can observe that most of the submissions achieve similar nDCG@10 values, especially when considering a number of features $n_f \geq 20$.

**Table 4.** The results for Task 1A on the MQ2007 dataset. Rows marked in grey (●) represent the results achieved with QA/H, rows marked in yellow (●) refer to the baselines' results, and the remaining refer to results achieved with SA.

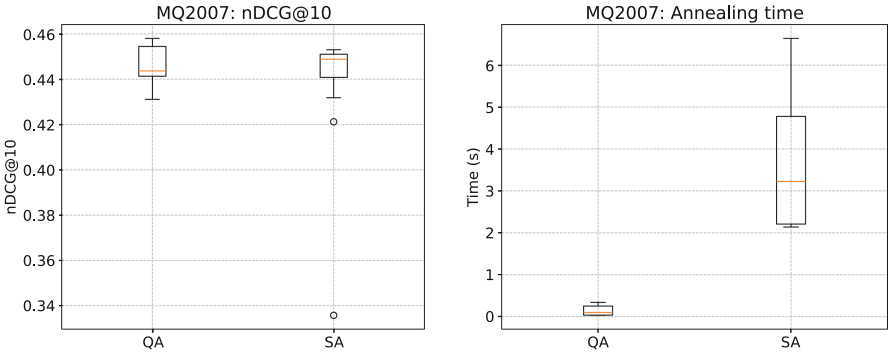
| Group | Submission id | nDCG@10 | Annealing time (ms) | Type | N° features |
|---|---|---|---|---|---|
| DS@GT qClef | 1A_MQ2007_SA_DS@GT qClef_pfi-k-25-cmi | 0.4500 | 2219 | SA | 25 |
| DS@GT qClef | 1A_MQ2007_SA_DS@GT qClef_pfi-k-20-cpfi | 0.4318 | 2185 | SA | 20 |
| DS@GT qClef | 1A_MQ2007_SA_DS@GT qClef_mi-k-25 | 0.4510 | 2214 | SA | 25 |
| DS@GT qClef | 1A_MQ2007_SA_DS@GT qClef_mi-k-15 | 0.4485 | 2136 | SA | 15 |
| DS@GT qClef | 1A_MQ2007_SA_DS@GT qClef_pfi-k-30-cmi | 0.4523 | 2157 | SA | 30 |
| DS@GT qClef | 1A_MQ2007_QA_DS@GT qClef_mi-1 | 0.4436 | 183 | QA | 15 |
| DS@GT qClef | 1A_MQ2007_QA_DS@GT qClef_mi-2 | 0.4552 | 160 | QA | 13 |
| FAST-NU | MQ2007_SA_FAST-NU_SA-2918 | 0.4212 | 4073 | SA | 15 |
| FAST-NU | 1A_MQ2007_SA_FAST-NU_SA-2915 | 0.3358 | 4164 | SA | 15 |
| FAST-NU | 1A_MQ2007_QA_FAST-NU_ae194be3-5267-45dd-aa0e-36a58579d719 | 0.4311 | 339 | QA | 15 |
| FAST-NU | 1A_MQ2007_QA_FAST-NU_26065450-e42a-4d92-bfb9-ff367d132142 | 0.4409 | 287 | QA | 15 |
| FAST-NU | 1A_MQ2007_QA_FAST-NU_1bba5207-9919-4048-b4a0-80f89b03f603 | 0.4375 | 275 | QA | 15 |
| SINAI-UJA | response_k21_nr3000 | 0.4530 | 3448 | SA | 21 |
| SINAI-UJA | response_k23_nr3000 | 0.4478 | 6632 | SA | 23 |
| SINAI-UJA | response_k25_nr3000 | 0.4510 | 2998 | SA | 25 |
| SINAI-UJA | response_k27_nr3000 | 0.4438 | 6637 | SA | 27 |
| SINAI-UJA | response_k29_nr3000 | 0.4491 | 6614 | SA | 29 |
| SINAI-UJA | response_k21_nr100 | 0.4580 | 34 | QA | 21 |
| SINAI-UJA | response_k23_nr100 | 0.4437 | 37 | QA | 23 |
| SINAI-UJA | response_k25_nr100 | 0.4550 | 31 | QA | 25 |
| SINAI-UJA | response_k27_nr100 | 0.4425 | 34 | QA | 27 |
| SINAI-UJA | response_k29_nr100 | 0.4528 | 34 | QA | 29 |
| BASELINE | ALL_FEATURES | 0.4473 | - | - | 46 |
| BASELINE | RFE HALF_FEATURES | 0.4450 | - | - | 23 |

Figure 1 shows the nDCG@10 values and Annealing timings of the runs that used QA and SA. From this figure it is possible to see that, in terms of efficiency (i.e., Annealing time), runs using QA required a substantially shorter amount of time with respect to SA. On average, QA required $\approx 26.83$ times less compared to SA, thus representing a more efficient alternative.

Considering effectiveness, QA seems to be performing more consistently. In fact, on average it performs $\approx 1.02$ times better compared to SA. Moreover, SA led to 2 outliers that underperformed with respect to the overall results.
Teams adopted different approaches to address this task:

- Team **DS@GT qClef**, inspired by a previous work [25], employed QUBO formulations that focused on different combinations of importance measures and redundancy measures for feature selection [37];
- Team **FAST-NU** adopted a Mutual Information-based approach, selecting features that maximize the information associated with relevance [40];
- Team **SINAI-UJA** also leveraged a Mutual Information-based approach, but also focused on post-processing the returned samples through normalization and projection techniques, aiming to find better solutions [24].

**Istella Dataset.** As it is possible to see in Table 5, also in this case, different numbers of features were considered among the submissions. It is interesting to




(a) nDCG@10 of SA and QA runs (b) Annealing time of SA and QA runs

**Fig. 1.** The box plots of the nDCG@10 values and Annealing timings associated with the runs using QA and SA on the MQ2007 dataset.

**Table 5.** The results for Task 1A on the Istella dataset. Rows marked in grey (●) represent the results achieved with QA/H, rows marked in yellow (●) refer to the baselines' results, and the remaining refer to results achieved with SA.

| Group | Submission id | nDCG@10 | Annealing time (ms) | Type | N° features |
|---|---|---|---|---|---|
| DS@GT qClef | 1A_Istella_SA_DS@GT qClef_mi_25 | 0.6025 | 126631 | SA | 25 |
| DS@GT qClef | 1A_Istella_SA_DS@GT qClef_mi_30 | 0.5104 | 133814 | SA | 30 |
| DS@GT qClef | 1A_Istella_SA_DS@GT qClef_mi_50 | 0.6524 | 159964 | SA | 50 |
| DS@GT qClef | 1A_Istella_SA_DS@GT qClef_mi_60 | 0.6682 | 173222 | SA | 60 |
| DS@GT qClef | 1A_Istella_SA_DS@GT qClef_mi_70 | 0.6523 | 184047 | SA | 70 |
| DS@GT qClef | 1A_Istella_QA_DS@GT qClef_mi_50 | 0.5586 | 9987 | H | 50 |
| BASELINE | ALL_FEATURES | 0.7146 | - | - | 220 |
| BASELINE | RFE HALF_FEATURES | 0.5560 | - | - | 110 |

see that the baseline method employing Recursive Feature Elimination for the extraction of 110 features performed worse than most of the participants' runs that considered a much lower number of features. Furthermore, running Recursive Feature Elimination to keep the top 110 features required a considerable amount of time (almost 2 h of computation) and a considerable amount of RAM (24 GB), which is much higher than the teams' workspace specifications.

Overall, it is possible to see how a different choice of features can lead to very different outcomes. In particular, the baseline *RFE_HALF_FEATURES*, performed poorly compared to other submissions that considered a lower number of kept features. This can probably be due to a poor feature choice from the baseline. Team **DS@GT qClef** leveraged QUBO formulations including importance measures and redundancy measures for feature selection [37].

**Table 6.** The results for Task 1B. Rows marked in grey (●) represent the results achieved with QA/H, rows marked in yellow(●) refer to the baselines' results, and the remaining refer to results achieved with SA.

| Dataset | Group | Submission id | nDCG@10 | Annealing time (ms) | Type | N° features |
|---|---|---|---|---|---|---|
| ICM_100 | Malto | 1B_100_ICM_SA_MALTO_1B - 100_ICM submission | 0.0207 | 6149 | SA | 51 |
| | BASELINE | ALL_FEATURES | 0.0226 | - | - | 100 |
| ICM_400 | Malto | 1B_400_ICM_SA_MALTO_1B - 400_ICM submission - 200 | 0.0294 | 80781 | SA | 200 |
| | Malto | 1B_400_ICM_SA_MALTO_1B - 400_ICM submission | 0.0182 | 70269 | SA | 53 |
| | BASELINE | ALL_FEATURES | 0.0328 | - | - | 400 |

In general, we can also see that the H approach required a much lower Annealing time than the SA counterparts, making use of a combination of QA and traditional hardware computations.

### 5.2 Task 1B: Quantum Feature Selection for RS

Here we present the results achieved in task 1B. Results are divided according to the two feature sets.

Table 6 presents the performance of different submissions for Task 1B, focusing on the impact of feature selection on the recommendation quality, as measured by nDCG@10, as well as on the computational cost, measured via annealing time.

For the ICM_100 dataset, the SA-based submission from the Malto group using 51 features achieved an nDCG@10 of 0.0207, which is slightly below the baseline score of 0.0226 obtained using all 100 features. This suggests a marginal performance degradation due to feature reduction, but with a 49% reduction in the number of features used could represent a good trade-off for improving efficiency at run-time.

In contrast, for the ICM_400 dataset, results are more varied. The best SA-based configuration (with 200 features) achieved an nDCG@10 of 0.0294, closer to the baseline performance of 0.0328, though still lower. Another SA configuration using only 53 features underperformed (nDCG@10 of 0.0182), indicating that excessive feature reduction can harm recommendation quality. The annealing times were considerably higher for this dataset (around 70–80 s), due to the increased search space associated with the larger feature set.

Team **Malto** addressed Task 1B by computing feature importance using a Random Forest classifier trained on the full feature set. They constructed a QUBO objective function that incorporated these importance scores along with pairwise Pearson correlations to penalize redundant features [18].

Overall, these results highlight that SA can effectively reduce the number of features while maintaining competitive recommendation quality, especially when a moderate number of features are retained. However, aggressive feature selection may lead to substantial performance degradation, and the method incurs non-trivial computational costs, particularly with larger datasets. These observations underscore the importance of balancing performance, computational efficiency,

**Table 7.** The results for Task 2 on the Yelp dataset averaged over 5 folds. Rows marked in grey (●) represent the results achieved with QA/H, rows marked in yellow(●) refer to the baselines' results, and the remaining refer to results achieved with SA.

| Group | Submission id | Avg Macro F1 | Avg Reduction | Avg Fine-Tuning time (s) | Avg Annealing time (ms) | Type |
|---|---|---|---|---|---|---|
| DS@GT qClef | Yelp_SA_qclef_bcos_075 | 99.5(0.2) | 0.25 | 1548.5(2.8) | 25997 | SA |
| DS@GT qClef | Yelp_SA_qclef_it_del_075 | 99.3(0.3) | 0.25 | 1549.2(1.5) | 25784 | SA |
| DS@GT qClef | Yelp_SA_qclef_svc_075 | 99.3(0.4) | 0.25 | 1550.5(2.6) | 25917 | SA |
| DS@GT qClef | Yelp_QA_qclef_bcos | 99.4(0.2) | 0.274 | 1500(54.7) | 1767 | QA |
| GPLSI | Yelp_SA_gplsi_2-SentimentPairs(docs****Remove***)=just-final... | 90.8(5.7) | 0.963 | 170.8(3.8) | 35810 | SA |
| GPLSI | Yelp_SA_gplsi_2-SentimentPairs(docs****Remove***)=pair-related... | 99.2(0.3) | 0.627 | 822.2(395) | 35810 | SA |
| GPLSI | Yelp_SA_gplsi_2-LocalSets | 99.4(0.2) | 0.512 | 1045.5(5.3) | 28789 | SA |
| GPLSI | Yelp_SA_gplsi_2-SentimentKmeansCard | 98.5(1.1) | 0.875 | 338.8(21) | 17823 | SA |
| GPLSI | Yelp_SA_gplsi_2-emoconflictCard | 98.6(0.5) | 0.728 | 628.2(65.9) | 34024 | SA |
| GPLSI | Yelp_QA_gplsi_2-SentimentKmeansCard | 98.7(0.2) | 0.869 | 351(25.1) | 553 | QA |
| GPLSI | Yelp_QA_gplsi_2-emoconflictCard | 98.8(0.6) | 0.702 | 678.8(80.9) | 549 | QA |
| Malto | Yelp_SA_MALTO_2 - vader_nyt_2L_0 | 99.2(0.2) | 0.751 | 582(2) | 142949 | SA |
| BASELINE | BASELINE_ALL | 99.4(0.1) | - | 2027.1(1.1) | - | - |

and the level of feature reduction when applying feature selection techniques in real-world recommender systems.

### 5.3 Task 2: Quantum Instance Selection for IR

Here we present the results achieved by the teams participating in Task 2, divided by dataset.

**Yelp Dataset.** Table 7 reports the results achieved by the different teams on the Yelp dataset for Task 2. As it is possible to see, the teams approached the task by trying different reduction rates (from $\approx 25\%$ to $\approx 96\%$). In particular, the submission *Yelp_SA_qclef_bcos_075* managed to improve the effectiveness of the Llama3.1 7b model with respect to the baseline. This could be due to the removal of noisy documents in that dataset, which lowered the performance of the model if used during the fine-tuning.

Notably, the submission *Yelp_QA_gplsi_2-SentimentKmeansCard* shows how QA was able to produce a reduction rate of $\approx 87\%$ and a high level of effectiveness (i.e., 98.7 vs 99.4 of the full dataset). Generally, it is possible to observe that QA requires less Annealing time than SA, while its performance is overall on par with SA. We list here some of the approaches considered by the teams:

- Team **GPLSI** [10] considered different approaches, such as prioritizing diversity by selecting pairs with very high or low similarity and minimizing semantic overlap (Sentiment Pairs) or selecting training instances using local set geometry through the combination of noise filtering and clustering with Euclidean distance (Local Sets);
- Team **DS@GT qClef** extended a previous approach [29] for selecting document embeddings. The team used cosine similarity for off-diagonal $Q$-matrix entries, but introduced two new strategies for diagonal terms: weighting instances by their distance to an Support Vector Machine (SVM )decision

**Table 8.** The results for Task 2 on the Vader dataset averaged over 5 folds. Rows marked in grey (●) represent the results achieved with QA/H, rows marked in yellow(●) refer to the baselines' results, and the remaining refer to results achieved with SA.

| Group | Submission id | Avg Macro F1 | Avg Reduction | Avg Fine-Tuning time (s) | Avg Annealing time (ms) | Type |
|---|---|---|---|---|---|---|
| DS@GT qClef | Vader_SA_qclef_svc_075 | 65.4(7.1) | 0.25 | 1529(2.4) | 25530 | SA |
| DS@GT qClef | Vader_SA_qclef_combined_075 | 65.9(4.7) | 0.25 | 1529.4(3) | 25300 | SA |
| DS@GT qClef | Vader_SA_qclef_it_del_075 | 65.6(3) | 0.25 | 1529.5(2.3) | 25348 | SA |
| DS@GT qClef | Vader_SA_qclef_bcos_075 | 62.5(10.4) | 0.25 | 1528.6(2.2) | 25735 | SA |
| DS@GT qClef | Vader_QA_qclef_bcos | 62.6(7.5) | 0.283 | 1493.3(83) | 1874 | QA |
| GPLSI | Vader_SA_gplsi_2-LocalSets | 63.3(4.9) | 0.505 | 1048.3(6.7) | 29110 | SA |
| GPLSI | Vader_SA_gplsi_2-SentimentPairs-docs=just-final... | 47.4(5.4) | 0.962 | 172.8(5.7) | 42408 | SA |
| GPLSI | Vader_SA_gplsi_2-SentimentPairs-docs=pair-related... | 62.2(4.1) | 0.7 | 671.8(352.8) | 42408 | SA |
| GPLSI | Vader_QA_gplsi_2-SentimentPairs-docs=just-final... | 50(64)* | 0.835* | 172.9(26.9)* | 545* | QA |
| GPLSI | Vader_QA_gplsi_2-SentimentPairs(docs***Remove**)=pair-related... | 62.1(1.8)* | 0.658* | 750.7(2653.2)* | 545* | QA |
| Malto | Vader_SA_MALTO_2 - vader_nyt_2L | 63.1(2.5) | 0.751 | 574.5(1.7) | 126087 | SA |
| BASELINE | BASELINE_ALL | 88.9(0.8) | - | 1997.3(5.7) | - | - |

* The submission did not include all 5 folds

boundary and measuring their influence using a logistic regression leave-one-out approach. The team tested each method individually and in combination, using batching to handle the datasets [37].

Overall, the results clearly highlight the critical importance of Instance Selection, particularly in the current context where computational efficiency and sustainability have a constantly increasing impact. By carefully selecting representative subsets of data, it is possible to significantly reduce the computational cost associated with fine-tuning large language models. Specifically, the fine-tuning time for the Llama3.1 7B model was reduced by approximately up to a factor of 9, with a negligible performance trade-off (i.e., losing less than 1 absolute point in macro-F1 score).

This demonstrates that substantial gains in efficiency can be achieved without severely compromising model effectiveness, making Instance Selection a key technique for training or fine-tuning models, especially when they are computationally expensive.

**Vader Dataset.** Table 8 reports the results achieved by the different teams on the Vader dataset for Task 2. Also in this case, the teams considered different reduction rates (from $\approx 25\%$ to $\approx 96\%$). However, differently from the previous results, all the subsets produced by the different approaches lead to a higher loss in terms of effectiveness of the fine-tuned Llama3.1 7b model.

It is possible to observe that submission *Vader_SA_MALTO_2 - vader nyt_2L* produced a subset of $\approx 25\%$ of the original dataset size while allowing to achieve a Average Macro F1 score which is higher than the subset produced by *Vader_SA_qclef_bcos_075* subset that instead was $\approx 75\%$ of the original dataset size, showing how different datasets can yield to potentially very different results.

For this task, the participating teams adopted similar approaches to the ones used for the other Yelp dataset. Also in this case, the QA approaches required

**Table 9.** The results for Task 3. Rows marked in grey (●) represent the results achieved with QA/H, rows marked in yellow(●) refer to the baselines' results, and the remaining refer to results achieved with SA.

| N° centroids | Team | Submission id | nDCG@10 | DBI | Annealing Time (ms) | Type |
|---|---|---|---|---|---|---|
| 10 | GPLSI | 10_SA_gplsi_3-FPS-Medoids | 0.5783 | 7.5147 | 15375 | SA |
| | GPLSI | 10_SA_gplsi_3-SubMedoidsQUBO | 0.5579 | 6.8779 | 15305 | SA |
| | GPLSI | 10_SA_gplsi_CLARA-CLARANS | 0.5444 | 6.6710 | 15395 | SA |
| | GPLSI | 10_SA_gplsi_MBK-Medoids | 0.5600 | 6.4258 | 15510 | SA |
| | DS@GT qClef | 10_SA_DS@GT qClef_1 | 0.5800 | 7.4776 | 83 | SA |
| | DS@GT qClef | 10_SA_DS@GT qClef_2 * | 0.0172 | 4.4706 | 83 | SA |
| | BASELINE | BASELINE_10 | 0.5509 | 7.9892 | - | - |
| 25 | GPLSI | 25_SA_gplsi_3-FPS-Medoids | 0.5475 | 5.5577 | 20875 | SA |
| | GPLSI | 25_SA_gplsi_3-SubMedoidsQUBO | 0.5298 | 5.6255 | 40687 | SA |
| | GPLSI | 25_SA_gplsi_CLARA-CLARANS | 0.5310 | 5.6507 | 20532 | SA |
| | GPLSI | 25_SA_gplsi_MBK-Medoids | 0.5193 | 5.3755 | 20758 | SA |
| | BASELINE | BASELINE_25 | 0.5284 | 6.1201 | - | - |
| 50 | GPLSI | 50_SA_gplsi_3-FPS-Medoids | 0.5592 | 4.4531 | 9869 | SA |
| | GPLSI | 50_SA_gplsi_3-SubMedoidsQUBO | 0.5148 | 4.9325 | 23719 | SA |
| | GPLSI | 50_SA_gplsi_CLARA-CLARANS | 0.5017 | 5.1703 | 9976 | SA |
| | GPLSI | 50_SA_gplsi_MBK-Medoids | 0.5383 | 4.5025 | 24004 | SA |
| | DS@GT qClef | 50_SA_DS@GT qClef_3 * | 0.0064 | 3.4217 | 228 | SA |
| | BASELINE | BASELINE_50 | 0.4656 | 5.3679 | - | - |

*Dimensionality reduction was applied

considerably less time with respect to the SA approaches while achieving similar trends in terms of effectiveness.

### 5.4 Task 3: Quantum Clustering for IR

Here we present the results achieved by the teams participating in Task 3. Table 9 reports the results achieved in this task.

It is possible to see that in this task, teams focused only on the usage of SA to solve the clustering problem. From the achieved results, we can notice how both GPLSI and DS@GT qCLEF teams managed to provide some submissions that performed better with respect to a traditional $k$-medoids baseline in terms of nDCG@10 and Davies-Bouldin Index. This suggests that their proposed approaches managed to successfully identify representative clusters of vectors that could help efficiently and effectively retrieve documents corresponding to queries.

We briefly detail some of the approaches considered by the teams:

– Team **GPLSI** [10] developed a method that filters the dataset embeddings to 150 pivots, optimizes centroid selection through annealing, and assigns documents to improve retrieval efficiency. The pivot selection was carried out in different ways, using heuristic approaches (FPS and CLARA–CLARANS), $k$-Means, and a technique inspired by the qIIMAS team from the previous QuantumCLEF edition [36] (SubMedoids). These techniques try to choose pivots that provide good coverage of the whole dataset;

– Team **DS@GT qClef** used a two-step approach. First, they applied classical clustering methods ($k$-Medoids, HDBSCAN [23], GMM, and GMM-HDBSCAN), optionally with dimensionality reduction (e.g., UMAP [5], PaCMAP), to select a manageable subset of instances. In the second step, they applied the QUBO-based $k$-medoids formulation on this reduced subset [37].

It is really interesting to see how the submission with id *50_SA_gplsi_3-FPS-Medoids* managed to achieve a higher level of nDCG@10 with respect to *BASELINE_10*. This improvement is especially impressive given that the method utilized a substantially larger number of clusters, 50 in total, versus the 10 used by the baseline. Despite the increase in cluster count, which could potentially lead to over-segmentation and reduced retrieval performance, the clusters generated by the GPLSI's approach proved to be more effective.

Submissions *10_SA_DS@GT_qClef_2* and *50_SA_DS@GT_qClef_3* marked with an asterisk (*), involved the usage of UMAP [5] dimensionality reduction technique. The dimensionality of the document vector embeddings was reduced to only 2 dimensions, potentially causing a big loss of information. Due to this aggressive reduction, the effectiveness of these runs was negatively impacted, leading to low nDCG@10 values.

## 6 Conclusions and Future Work

In this paper, we have presented the overview of the second edition of the QuantumCLEF lab that was held in 2025. QuantumCLEF represents the first lab at CLEF focusing on the study, development, and evaluation of QC algorithms using real quantum computers. This lab was composed of three tasks concerning the problems of Feature Selection, Instance Selection, and Clustering, focusing on computationally complex problems faced by IR and RS systems. Participants used the KIMERA [35] infrastructure for a smooth workflow. The infrastructure has granted participants access to both computational resources and cutting-edge quantum annealers provided by D-Wave, thus giving the possibility of experimenting with real quantum computers.

A total of 44 teams registered for the lab, and 5 of them successfully managed to submit their runs. The results have shown that QA and H managed to achieve comparable results in terms of effectiveness with respect to SA while achieving a higher level of efficiency in terms of Annealing time. This shows that QC is starting to become a powerful technology that could help in the resolution of complex problems, especially in the future once it has matured enough. Furthermore, the QA results are competitive with respect to traditional baselines, showing that QA solutions are able to achieve good levels of effectiveness.

This second edition of the QuantumCLEF lab represented a great opportunity not only to develop and evaluate QC algorithms on **real quantum computers** (quantum technologies are still not easily accessible to the general public) but also to raise awareness of the potential of QC, which is likely to become a

powerful technology in the future. In fact, participants were provided with comprehensive materials such as videos, slides, and examples that allowed them to learn how QC and QA work. Moreover, we opted for maximum transparency, allowing participants to work with the actual D-Wave libraries. In this way, participants familiarized themselves with them and, thus, are now able to program quantum annealers even outside our infrastructure to solve other problems in their research field.

In the future, we plan to organize a third edition of QuantumCLEF with different tasks and more challenges. We would like to invest in a more powerful infrastructure that will grant access to more participants and that will provide more resources (in terms of CPU and RAM) to each workspace. If possible, we would also like to extend the infrastructure to include gate-based quantum computers [39], in addition to the already available quantum annealers.

**Acknowledgments.** We acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources and support.

## References

1. Almeida, T.M., Matos, S.: Towards a hyperparameter-free QUBO formulation for feature selection in IR. In: Faggioli, G., Ferro, N., Galuscáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, CEUR Workshop Proceedings, vol. 3740, pp. 3054–3063, CEUR-WS.org (2024). https://ceur-ws.org/Vol-3740/paper-298.pdf
2. Alvarez-Giron, W., Tellezz-Torres, J., Tovar-Cortes, J., Gómez-Adorno, H.: Team qiimas on task 2 - clustering. In: Faggioli, G., Ferro, N., Galuscáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, CEUR Workshop Proceedings, vol. 3740, pp. 3064–3074, CEUR-WS.org (2024). https://ceur-ws.org/Vol-3740/paper-299.pdf
3. Arthur, D., Date, P.: Balanced $k$-means clustering on an adiabatic quantum computer. Quantum Inf. Process. **20**(9), 1–30 (2021). https://doi.org/10.1007/s11128-021-03240-8
4. Bauckhage, C., Piatkowski, N., Sifa, R., Hecker, D., Wrobel, S.: A QUBO formulation of the k-medoids problem. In: Jäschke, R., Weidlich, M. (eds.) Proceedings of the Conference on "Lernen, Wissen, Daten, Analysen", Berlin, Germany, September 30 - October 2, 2019, CEUR Workshop Proceedings, vol. 2454, pp. 54–63, CEUR-WS.org (2019). https://ceur-ws.org/Vol-2454/paper_39.pdf
5. Becht, E., McInnes, L., Healy, J., Dutertre, C.A., Kwok, I.W., Ng, L.G., Ginhoux, F., Newell, E.W.: Dimensionality reduction for visualizing single-cell data using umap. Nat. Biotechnol. **37**(1), 38–44 (2019)
6. Bertsimas, D., Nohadani, O.: Robust optimization with simulated annealing. J. Glob. Optim. **48**(2), 323–334 (2010). https://doi.org/10.1007/S10898-009-9496-X
7. Bertsimas, D., Tsitsiklis, J.: Simulated annealing. Stat. Sci. **8**(1), 10–15 (1993)
8. Burges, C.J.C.: From RankNet to LambdaRank to LambdaMART: An Overview. Technical Report, Microsoft Research, MSR-TR-2010-82 (2010)

9. Cai, J., Macready, W.G., Roy, A.: A practical heuristic for finding graph minors. CoRR **abs/1406.2741** (2014). http://arxiv.org/abs/1406.2741
10. Consuegra-Ayala, J.P., Morote-Martínez, A., Valero-Abellón, F., Lloret, E., Moreda, P., Palomar, M.: Team gplsi at qclef 2025: Quantum-inspired instance selection and clustering. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)
11. Cunha, W., Fernández, A.M., Esuli, A., Sebastiani, F., Rocha, L., Gonçalves, M.A.: A noise-oriented and redundancy-aware instance selection framework. ACM Trans. Inf. Syst. **43**(2), 45:1–45:33 (2025). https://doi.org/10.1145/3705000
12. Cunha, W., França, C., Fonseca, G., Rocha, L., Gonçalves, M.A.: An effective, efficient, and scalable confidence-based instance selection framework for transformer-based text classification. In: Chen, H., Duh, W.E., Huang, H., Kato, M.P., Mothe, J., Poblete, B. (eds.) Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023, pp. 665–674, ACM (2023). https://doi.org/10.1145/3539618.3591638
13. Cunha, W., Viegas, F., França, C., Rosa, T., Rocha, L., Gonçalves, M.A.: A comparative survey of instance selection methods applied to non-neural and transformer-based text classification. ACM Comput. Surv. **55**(13s), 265:1–265:52 (2023). https://doi.org/10.1145/3582000
14. Ferrari Dacrema, M., Moroni, F., Nembrini, R., Ferro, N., Faggioli, G., Cremonesi, P.: Towards feature selection for ranking and classification exploiting quantum annealers. In: Amigó, E., Castells, P., Gonzalo, J., Carterette, B., Culpepper, J.S., Kazai, G. (eds.) SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022, pp. 2814–2824, ACM (2022). https://doi.org/10.1145/3477495.3531755
15. Ferrari Dacrema, M., Pasin, A., Cremonesi, P., Ferro, N.: Quantum computing for information retrieval and recommender systems. In: Goharian, N., et al. (eds.) Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part V, LNCS, vol. 14612, pp. 358–362, Springer (2024). https://doi.org/10.1007/978-3-031-56069-9_47
16. Ferrari Dacrema, M., Pasin, A., Cremonesi, P., Ferro, N.: Using and evaluating quantum computing for information retrieval and recommender systems. In: Yang, G.H., Wang, H., Han, S., Hauff, C., Zuccon, G., Zhang, Y. (eds.) Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024, pp. 3017–3020, ACM (2024). https://doi.org/10.1145/3626772.3661378
17. Fröbe, M., Alexander, D., Hendriksen, G., Schlatt, F., Hagen, M., Potthast, M.: Team openwebsearch at CLEF 2024: quantumclef. In: Faggioli, G., Ferro, N., Galuscáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, CEUR Workshop Proceedings, vol. 3740, pp. 3075–3081, CEUR-WS.org (2024). https://ceur-ws.org/Vol-3740/paper-300.pdf
18. Giobergia, F., Savelli, C., Koudounas, A., Baralis, E.: Quantum feature selection from interpretable models using QUBO formulation. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)

19. Glover, F., Kochenberger, G., Hennig, R., Du, Y.: Quantum bridge analytics I: a tutorial on formulating and using QUBO models. Ann. Oper. Res. **314**, 141–183 (2022)
20. Gyongyosi, L., Imre, S.: A survey on quantum computing technology. Comput. Sci. Rev. **31**, 51–71 (2019)
21. Hashemi, H., Aliannejadi, M., Zamani, H., Croft, W.B.: ANTIQUE: a non-factoid question answering benchmark. In: Jose, J.M., et al. (eds.) Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part II, LNCS, vol. 12036, pp. 166–173, Springer (2020). https://doi.org/10.1007/978-3-030-45442-5_21
22. Lucchese, C., Nardini, F.M., Orlando, S., Perego, R., Silvestri, F., Trani, S.: Post-learning optimization of tree ensembles for efficient ranking. In: Perego, R., Sebastiani, F., Aslam, J.A., Ruthven, I., Zobel, J. (eds.) Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016, pp. 949–952, ACM (2016). https://doi.org/10.1145/2911451.2914763
23. McInnes, L., Healy, J., Astels, S., et al.: hdbscan: Hierarchical density based clustering. J. Open Source Softw. **2**(11), 205 (2017)
24. Molino-Piñar, L., Collado-Montañez, J., Montejo-Ráez, A.: Sinai team at quantumclef 2025: Quantum feature selection based on energy with d-wave. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)
25. Mücke, S., Heese, R., Müller, S., Wolter, M., Piatkowski, N.: Feature selection on quantum computers. Quantum Mach. Intell. **5**(1), 11 (2023)
26. Naebzadeh, A., Eetemadi, S.: NICA at quantum computing CLEF tasks 2024. In: Faggioli, G., Ferro, N., Galuscáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, CEUR Workshop Proceedings, vol. 3740, pp. 3087–3095, CEUR-WS.org (2024). https://ceur-ws.org/Vol-3740/paper-302.pdf
27. Nembrini, R., Ferrari Dacrema, M., Cremonesi, P.: Feature selection for recommender systems with quantum computing. Entropy **23**(8), 970 (2021). https://doi.org/10.3390/E23080970
28. Niu, J., Li, J., Deng, K., Ren, Y.: CRUISE on quantum computing for feature selection in recommender systems. In: Faggioli, G., Ferro, N., Galuscáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, CEUR Workshop Proceedings, vol. 3740, pp. 3096–3104, CEUR-WS.org (2024). https://ceur-ws.org/Vol-3740/paper-303.pdf
29. Pasin, A., Cunha, W., Gonçalves, M.A., Ferro, N.: A quantum annealing instance selection approach for efficient and effective transformer fine-tuning. In: Oosterhuis, H., Bast, H., Xiong, C. (eds.) Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2024, Washington, DC, USA, 13 July 2024, pp. 205–214, ACM (2024). https://doi.org/10.1145/3664190.3672515
30. Pasin, A., Ferrari Dacrema, M., Cremonesi, P., Cunha, W., Gonçalves, M.A., Ferro, N.: Quantumclef 2025 - the second edition of the quantum computing lab at CLEF. In: Hauff, C., Macdonald, C., Jannach, D., Kazai, G., Nardini, F.M., Pinelli, F., Silvestri, F., Tonellotto, N. (eds.) Advances in Information Retrieval - 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6-10, 2025, Proceedings, Part V, LNCS, vol. 15576, pp. 450–458, Springer (2025). https://doi.org/10.1007/978-3-031-88720-8_66

31. Pasin, A., Ferrari Dacrema, M., Cremonesi, P., Ferro, N.: qclef: A proposal to evaluate quantum annealing for information retrieval and recommender systems. In: Arampatzis, A., et al. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction - 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18-21, 2023, Proceedings, LNCS, vol. 14163, pp. 97–108, Springer (2023). https://doi.org/10.1007/978-3-031-42448-9_9
32. Pasin, A., Ferrari Dacrema, M., Cremonesi, P., Ferro, N.: Overview of quantumclef 2024: the quantum computing challenge for information retrieval and recommender systems at CLEF. In: Goeuriot, L., et al. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction - 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9-12, 2024, Proceedings, Part II, Lecture Notes in Computer Science, vol. 14959, pp. 260–282, Springer (2024). https://doi.org/10.1007/978-3-031-71908-0_12
33. Pasin, A., Ferrari Dacrema, M., Cremonesi, P., Ferro, N.: Quantumclef - quantum computing at CLEF. In: Goharian, N., et al. (eds.) Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part V, Lecture Notes in Computer Science, vol. 14612, pp. 482–489, Springer (2024). https://doi.org/10.1007/978-3-031-56069-9_66
34. Pasin, A., Ferrari Dacrema, M., Cremonesi, P., Ferro, N.: Quantumclef 2024: overview of the quantum computing challenge for information retrieval and recommender systems at CLEF. In: Faggioli, G., Ferro, N., Galuscáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, CEUR Workshop Proceedings, vol. 3740, pp. 3032–3053, CEUR-WS.org (2024). https://ceur-ws.org/Vol-3740/paper-297.pdf
35. Pasin, A., Ferro, N.: Kimera: From evaluation-as-a-service to evaluation-in-the-cloud. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padova, Italy, July 13-18, 2025, ACM (2025). https://doi.org/10.1145/3726302.3730298
36. Payares, E., Puertas, E., Santos, J.C.M.: Team QTB on feature selection via quantum annealing and hybrid models. In: Faggioli, G., Ferro, N., Galuscáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, CEUR Workshop Proceedings, vol. 3740, pp. 3105–3114, CEUR-WS.org (2024). https://ceur-ws.org/Vol-3740/paper-304.pdf
37. Pomeroy, C., Pramov, A., Thakrar, K., Yendapalli, L.: Quantum annealing for machine learning: Applications in feature selection, instance selection, and clustering. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)
38. Qin, T., Liu, T.: Introducing LETOR 4.0 datasets. CoRR **abs/1306.2597** (2013). http://arxiv.org/abs/1306.2597
39. Rieffel, E., Polak, W.: An introduction to quantum computing for non-physicists. ACM Comput. Surv. (CSUR) **32**(3), 300–335 (2000)
40. Shaikh, M.T., Hamza, M., Ali, S.B., Rafi, M., Zahid, S.: Feature selection using quantum annealing: a mutual information based QUBO approach. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)

41. Shimi, G., C, J.M., Thenmozhi, D.: Quantum feature selection. In: Faggioli, G., Ferro, N., Galuscáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, CEUR Workshop Proceedings, vol. 3740, pp. 3082–3086, CEUR-WS.org (2024). https://ceur-ws.org/Vol-3740/paper-301.pdf
42. Touvron, H., et al.: Llama: Open and efficient foundation language models. CoRR **abs/2302.13971** (2023). https://doi.org/10.48550/ARXIV.2302.13971
43. Ushijima-Mwesigwa, H., Negre, C.F.A., Mniszewski, S.M.: Graph partitioning using quantum annealing on the d-wave system. CoRR **abs/1705.03082** (2017). http://arxiv.org/abs/1705.03082
44. Van Laarhoven, P.J., Aarts, E.H., van Laarhoven, P.J., Aarts, E.H.: Simulated annealing. Springer (1987)
45. Yarkoni, S., Raponi, E., Bäck, T., Schmitt, S.: Quantum annealing for industry applications: introduction and review. Reports on Progress in Physics **85**(10), 104001:1–104001:27 (2022)