Feature Selection Using Quantum Annealing: A Mutual Information Based QUBO Approach

Muhammad Talha Shaikh[1,*,†], Muhammad Hamza[1,†], Syed Bilal Ali[1,†], Muhammad Rafi[1,†] and Sumaiyah Zahid[1,†]

[1]National University of Computer and Emerging Sciences -FAST, St-4, Sector 17-D, N-5, Karachi, Pakistan

Abstract

This paper presents a quantum-inspired feature selection approach to Learning to Rank (LTR) on the MQ2007 collection of Quantum CLEF 2025 Task 1A. The goal is to find a minimal yet informative subset of features that maximizes ranking performance with respect to NDCG@10. We cast feature selection as a Binary Quadratic Model (BQM), where coefficients are derived from Mutual Information (MI) and Conditional Mutual Information (CMI). The resulting Quadratic Unconstrained Binary Optimization (QUBO) problems are optimized using Quantum Annealing (QA) on quantum hardware and Simulated Annealing (SA) on traditional processors. Experimental results show that QA obtains an optimal NDCG@10 value of 0.44, outperforming SA and with considerable computation time reduction.

Keywords

Quantum Annealing, Feature Selection, QUBO, Simulated Annealing, Quantum Machine Learning,

1. Introduction

In high-dimensional information retrieval tasks, such as Learning to Rank (LTR), the selection of a subset of features plays a crucial role in determining model effectiveness, generalizability, and efficiency. Feature selection strives to discover a subset of input features that convey the most information about relevant downstream tasks. If solved in general, this is a combinatorially intractable problem that belongs to NP-Hard problems [1]. Thus, heuristic or approximate solutions are generally used. Mutual Information (MI) and Conditional Mutual Information (CMI) have been extensively used in feature selection due to their ability to measure feature relevance and feature redundancy [2, 3]. MI measures the level of mutual dependence between two variables, while CMI measures the level of information shared between two variables conditioned on a third variable. These measurements allow for features with individual informativeness as well as collective diversity to be chosen.

Simulated Annealing (SA), a well-known classical metaheuristic, has been found effective in feature selection for Learning to Rank (LTR) with low data. Advanced forms of SA have been proposed in recent research involving a range of neighborhood selection methods, temperature cooling policies, and hyperparameters like the progress parameter, all for effective search space exploration and ranking performance [4]. Experimental comparisons with local beam search have noted SA-based models to be similarly performing on most benchmark LTR datasets, hence justifying its effectiveness in real-world scenarios. Although traditional techniques such as Simulated Annealing (SA) are somewhat flexible and are generally simple to apply, they tend to struggle with avoiding local minima, particularly if the optimization landscape is complex. Quantum Annealing (QA) presents a very different approach by leveraging quantum tunneling phenomena to tunnel through neighboring humps. QA has proven to be extremely effective on problems that can be formulated as Binary Quadratic Models (BQMs) or Quadratic Unconstrained Binary Optimization (QUBO) problems [5]. The method enables exponentially large solution spaces to be explored by mapping the energy landscape onto a quantum framework.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

†These authors contributed equally.

k214564@nu.edu.pk (M. T. Shaikh); k213293@nu.edu.pk (M. Hamza); k213153@nu.edu.pk (S. B. Ali); muhammad.rafi@nu.edu.pk (M. Rafi); sumaiyah@nu.edu.pk (S. Zahid)

Current research suggests that QA can be superior to traditional methods for specific combinatorial optimization problems, particularly if the problem is amenable to the targeted quantum hardware [6]. Experimental efficiency in QA relies significantly on problem definition and on mapping quality to quantum hardware, with encouraging initial results seen in industrial applications [7].

Herein, we investigate the application of both QA and SA for feature subset selection on the MQ2007 LTR dataset. Our approach maps MI- and CMI-based relevance and redundancy estimation into QUBO formulation and uses quantum and simulated annealing to identify highly informative and diverse subsets of features. We then evaluate the effect of selected subset using the NDCG@10 metric together with visualization and comparative analysis to study feature importance and interrelationships. This dual investigation is intended to better understand the trade-offs between quantum and classical annealing in practical feature selection pipelines.

2. Background and Motivation

2.1. Feature Selection and Its Significance

Feature selection involves reducing dimensionality by identifying the most impactful features from a larger set. Beyond boosting model accuracy, its advantages include:

- **Avoiding the Curse of Dimensionality:** As the dimensionality grows, the feature space becomes more and more sparse exponentially, and models find it hard to generalize. Feature selection assists in focusing learning on the most relevant variables.
- **Less Computational Cost:** The reduction of feature numbers leads to a reduction of input vector sizes, which in turn enables accelerated model training, decreases inference time, and decreases memory consumption—this is especially beneficial for large ranking tasks.
- **Improved Generalization:** Models that are trained on data of high dimensionality will overfit, particularly if there are irrelevant or redundant features. Feature selection improves generalization by concentrating on the essential predictive features.
- **Improved Model Explainability:** A streamlined, carefully curated set of features is simpler to understand the model's decision process and behavior, which is critical in ranking applications where explainability is important.
- **Noise Reduction:** Elimination of features that are irrelevant or weakly related eliminates noise in learning and yields stronger and more stable models.

2.2. MI-Based Formulation of Feature Selection

The objective is to choose features that hold maximal information about the target label, quantified using **Mutual Information (MI)** and **Conditional Mutual Information (CMI)**. MI measures how much knowing a feature reduces uncertainty about the label. CMI extends this to measure additional information gained from one feature conditional on another.

Given $X = \{X_1, X_2, \ldots, X_n\}$ and label Y, the goal is to maximize:

$$\sum_{i \in S} I(X_i; Y) + \sum_{\substack{i,j \in S \\ i \neq j}} I(X_j; Y | X_i)$$

Where $S \subseteq X$ is the selected feature subset.

2.3. QUBO Encoding via MI and CMI

The above MI-based objective can be converted into a QUBO as follows:

- Linear terms $q_i \leftarrow -I(X_i; Y)$
- Quadratic terms $q_{ij} \leftarrow -I(X_j; Y | X_i)$

This yields the QUBO:

$$\min_{\mathbf{x}\in\{0,1\}^n} \quad \mathbf{x}^T\mathbf{Q}\mathbf{x}$$

Where $\mathbf{x}$ is the binary vector indicating selected features.

2.4. Intuition Behind Quantum Annealing

Traditional feature selection algorithms have to search 2^n subsets of n features, which is intractable for moderately sized n. Quantum Annealing, through the application of superposition, allows us to search an exponentially large search space in polynomial time in certain instances. It's important to know, however, that although a quantum system can exist in all possible states at once during the evolution, it collapses to one result when measured. Therefore, the construction of good QUBOs that steer the system to high quality optima is essential.

3. Feature Selection Methodology

3.1. Problem Formulation

The MQ2007 dataset consists of query-document feature vectors $X = \{x_1, x_2, \ldots, x_n\}$ with corresponding relevance labels Y. The objective is to identify a subset $S \subseteq X$ of fixed size k that collectively maximizes feature relevance to the target while minimizing redundancy among selected features, thereby improving overall ranking performance. This problem is cast as a Binary Quadratic Model (BQM), which is then minimized by the quantum annealer.

Formally, given $X = \{X_1, X_2, ..., X_n\}$ and label Y, the goal is to minimize the following objective function, which directly maps to the BQM formulation:

$$\min_S \left[\sum_{x_i \in S} -I(x_i; Y) + \sum_{\substack{x_i, x_j \in S \\ i<j}} -I(x_i; x_j \mid Y) \right]$$

Here, $-I(x_i; Y)$ encourages the inclusion of features with high relevance to Y (as minimizing a large negative value makes it less negative, closer to zero, or positive), and $-I(x_i; x_j \mid Y)$ penalizes redundancy among selected features (as minimizing a large negative value for redundant features contributes less to the overall minimum). The fixed cardinality $|S| = k$ is imposed using a soft constraint, typically enforced via a penalty term in the BQM framework.

3.2. Mutual and Conditional Mutual Information Estimation

The information-theoretic quantities $I(x_i; Y)$ and $I(x_i; x_j \mid Y)$ are computed using histogram-based entropy estimation. First, joint probability distributions are computed by discretizing feature vectors into bins. Entropy is calculated using the Shannon formulation:

$$H(X) = -\sum_x p(x) \log_2 p(x)$$

For a feature x_i and target Y, the mutual information is computed as:

$$I(x_i; Y) = H(x_i) + H(Y) - H(x_i, Y)$$

Conditional mutual information for feature pairs (x_i, x_j) is defined as:

$$I(x_i; x_j \mid Y) = H(x_i, Y) + H(x_j, Y) - H(x_i, x_j, Y) - H(Y)$$

In practice, joint histograms are estimated using `np.histogramdd`, and entropies are computed directly from normalized joint probabilities. A bin size of 10 is used for 2D MI estimation, while a smaller bin size of 6 is used for 3D CMI to mitigate sparsity.

3.3. Binary Quadratic Model Construction

The optimization is mapped onto a Binary Quadratic Model (BQM), where each binary variable $z_i \in \{0, 1\}$ indicates whether feature x_i is selected. The energy function is defined as:

$$E(z) = \sum_i a_i z_i + \sum_{i<j} b_{ij} z_i z_j + \alpha \left(\sum_i z_i - k \right)^2$$

where:

- $a_i = -I(x_i; Y)$ encourages inclusion of features with high relevance to Y.
- $b_{ij} = -I(x_i; x_j \mid Y)$ penalizes redundancy among selected features (as minimizing a large negative value for redundant features contributes less to the overall minimum).
- The last term is a soft cardinality constraint with strength $\alpha = 10.0$.

The BQM is constructed by iterating over all features and feature pairs to compute MI and CMI values, which are stored for downstream analytics. The constraint enforcing exactly k features is generated and merged into the BQM.

3.4. Simulated Annealing

Simulated Annealing (SA) is accessed by using the `neal` sampler. Simulated annealing simulates temperature variations, traversing through the Binary Quadratic Model's (BQM) energy landscape by stochastically modifying binary variables according to an imposed temperature schedule. A single run is 2,000 reads, and the lowest energy sample is used as the best subset of features.

3.5. Quantum Annealing

Quantum Annealing (QA) is run on the D-Wave quantum processing unit (QPU) by creating the `EmbeddingComposite` structure. The quantum sampler attempts to find the global optimum of the Binary Quadratic Model (BQM) by evolving an initial quantum superposition according to an energy Hamiltonian derived from the problem parameters.

3.6. Feature Extraction and Submission Formatting

Once the annealing process concludes, the resulting samples are processed to extract the selected feature indices. Features corresponding to binary values of 1 are extracted, and their original 1-based indices are saved. A submission file is generated for each annealing method, containing:

- A sorted list of selected feature indices.
- A trailing line containing the problem identifier (submission ID) for traceability.

3.7. Summary

The provided methodology enables effective feature selection through hybrid classical-quantum optimisation. The integration of MI-based relevance, CMI-based redundancy management, and cardinality-enforcing BQM formulation enables fine-grained feature subset control. The joint use of SA and QA enables submission or hardware resilience, and generates competitive and analytically interpretable feature selections.

3.8. Experimental Setup

Data Preparation: The MQ2007 dataset consists of 46 numerical features per query-document pair, along with relevance labels. The data preprocessing steps include:

- **Min-max normalization:** Each feature is scaled to the $[0, 1]$ range independently to ensure uniform contribution to mutual information calculations and compatibility with the binary nature of the BQM formulation.
- **Discretization:** To compute entropy-based metrics, all continuous features are discretized into a fixed number of bins (default is 10). This binning enables the estimation of empirical probability distributions required for MI and CMI computations.
- **MI and CMI estimation:** Mutual Information (MI) is computed between each feature and the target label (relevance). Conditional Mutual Information (CMI) is also estimated for each feature pair conditioned on the label, which helps capture joint relevance. These scores are later used to form the coefficients of the BQM.

BQM Construction: A Binary Quadratic Model (BQM) is constructed where:

- **Linear terms:** The linear coefficients represent the negative Mutual Information scores, i.e., higher MI is encouraged by minimizing the BQM objective.
- **Quadratic terms:** The quadratic coefficients are based on the negative Conditional Mutual Information between feature pairs. This penalizes the selection of redundant features.
- **Soft constraint on feature count:** To enforce approximately $k = 10$ features to be selected, an auxiliary quadratic penalty term is added to the BQM. This term penalizes any solution where the sum of selected binary variables is not equal to k, using a high penalty coefficient (e.g., 5.0).

3.9. Evaluation Pipeline

Once the BQM is constructed, the feature selection and evaluation proceed as follows:

- **Sampling:** The BQM is solved by two approaches: Quantum Annealing on D-Wave and Simulated Annealing based on the Neal sampler. For both samplers, the lowest-energy solution is employed, a binary vector of the chosen features.
- **Feature subset selection:** The feature indices of the corresponding active bits (1s) of the lowest energy sample are calculated. These are used to build a compact feature matrix by selecting the corresponding columns of the original data.
- **Model training:** The RankSVM model is trained on the smaller training set. Default hyperparameters are used in the RankSVM configuration, which is trained using query-specific splits in the training data.
- **Evaluation:** It is assessed with NDCG@10 (Normalized Discounted Cumulative Gain at position 10) on the provided test set, in accordance with the provided Quantum CLEF Task 1A [8, 9] evaluation protocol. This specific measure assesses the quality of ranking obtained from the top 10 documents retrieved per query.

3.10. Submission and Visualization

Final submission files include the 1-based indices of selected features and the associated problem ID (from D-Wave or fallback SA runs). The analytics module produces the following visualizations:

- **Top-20 Mutual Information Bar Chart:** Displays the MI scores of the most relevant features.
- **Correlation Heatmap:** Plots Pearson correlation between features selected by SA and QA.
- **Feature Selection Comparison:** Venn-style bar chart showing overlap and divergence in features selected by QA vs. SA.

These diagnostics enable qualitative assessment of feature quality and algorithmic behavior under quantum and classical regimes.

4. Results

Table 1 summarizes the final submission results. Each submission used 15 features selected via SA or QA, with reported annealing times and NDCG@10 scores.

Table 1
Summary of Submission Results

Team	Submission ID	NDCG@10	Time (μs)	Type	#Features
FAST-NU	MQ2007_SA_FAST-NU_SA-2918	0.4212	4,072,677	S	15
FAST-NU	1A_MQ2007_SA_FAST-NU_SA-2915	0.3358	4,163,986	S	15
FAST-NU	1A_MQ2007_QA_FAST-NU_ae194be3	0.4311	338,763	Q	15
FAST-NU	1A_MQ2007_QA_FAST-NU_26065450	0.4409	286,966	Q	15
FAST-NU	1A_MQ2007_QA_FAST-NU_1bba5207	0.4375	274,721	Q	15

4.1. Performance Comparison

Quantum Annealing (QA) consistently outperformed Simulated Annealing (SA) across all metrics. The best QA submission (NDCG@10 = **0.4409**) surpassed the best SA score (NDCG@10 = 0.4212) while also achieving substantially lower annealing times (286,966 μs vs. 4,072,677 μs). All QA-based submissions clustered around high performance, showcasing the robustness of the quantum annealing approach. The results highlight QA's ability to explore the feature space more efficiently and effectively than classical SA methods.

5. Visualization and Analysis

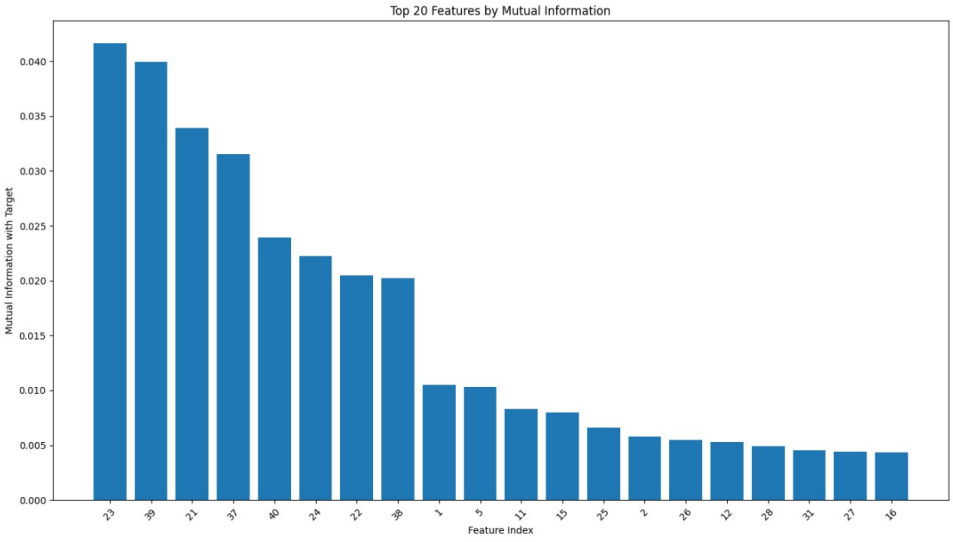

Figure 1: Top 20 features by average Mutual Information with relevance labels.

This bar chart (Figure 1) illustrates the varying relevance of individual features, showing that some features (e.g., Feature 23, 39, 21) exhibit significantly higher mutual information with the target label, indicating their strong individual predictive power. This helps in understanding the baseline information content of the most relevant features.

Figure 2 visualizes the overlap and unique feature selections between Quantum Annealing (QA) and Simulated Annealing (SA). The "Common Features" bar indicates the set of features identified by both methods, while "QA Only" and "SA Only" show features uniquely selected by each. This comparison

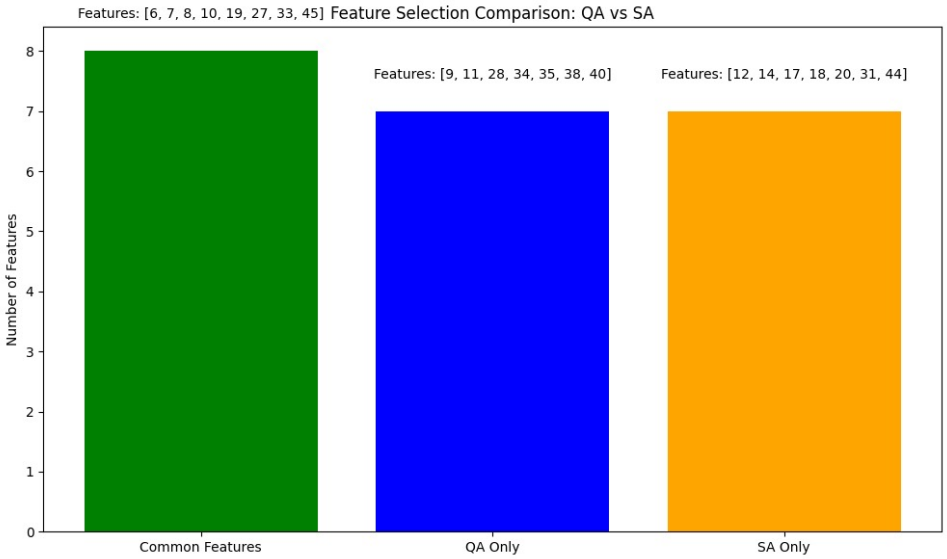

Figure 2: Feature selection overlap between QA and SA.

highlights that while there's a significant common subset, both annealing approaches also identify distinct features, suggesting different exploration patterns or optima found within the feature space.

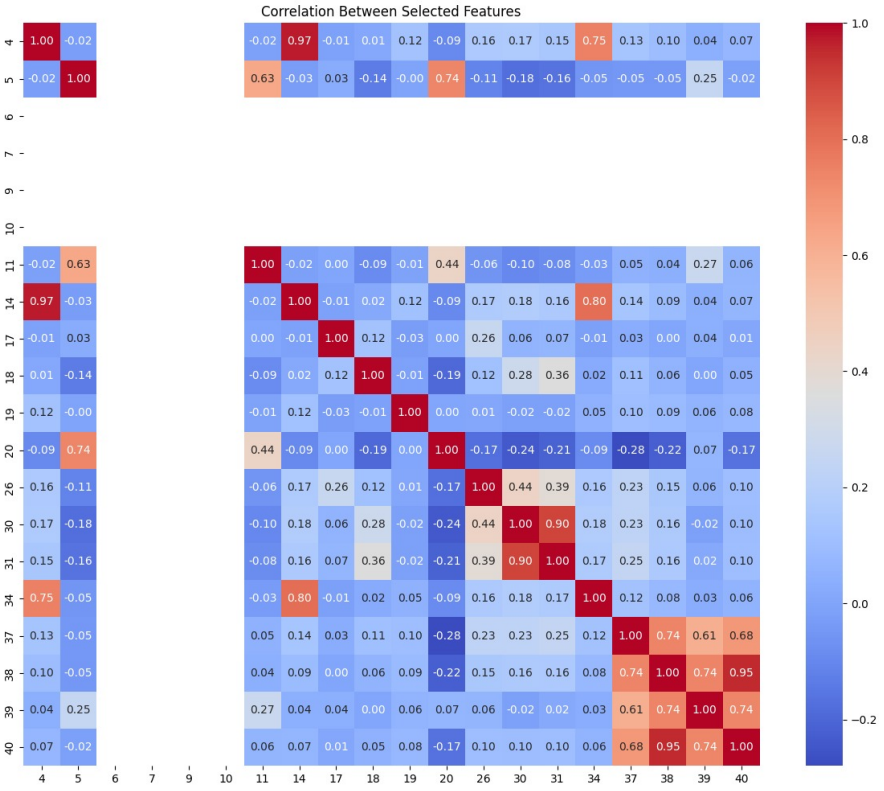

Figure 3: Correlation matrix of selected features showing redundancy.

The QA and SA methods selected partially overlapping sets of features, but QA demonstrated better diversity and lower intra-set redundancy, as evident in the correlation matrix (Figure 3). This heatmap displays the Pearson correlation coefficients between the selected features. Lighter areas (approaching white) within the heatmap indicate correlation values close to zero, meaning there is

minimal linear relationship and thus low redundancy between those specific feature pairs. This is a desirable characteristic for a feature subset, as it suggests the selected features contribute unique and non-overlapping information, leading to improved model robustness.

The QA and SA methods selected partially overlapping sets of features, but QA demonstrated better diversity and lower intra-set redundancy, as evident in the correlation matrix (Figure 3).

6. Discussion

Our findings demonstrate the promise of quantum annealing for effective and efficient feature selection in ranking tasks. Our speeds-up in annealing time and enhanced ranking performance indicate that quantum approaches may augment or even supersede conventional metaheuristics as quantum hardware grows.

However, we also notice some limitations including variation in QA outcomes between runs, limited control of embedding on the QPU, and sensitivity to the accuracy of MI estimation. All these aspects are deserving of further study.

6.1. Conclusion and Future Work

We demonstrated a hybrid quantum-classical pipeline for feature selection in learning-to-rank tasks using the MQ2007 dataset. Our Quantum Annealing-based approach showed promising results both in performance and computation time. Future work will explore scaling the method to larger feature spaces, integrating alternative quantum algorithms (e.g., QAOA), and improving mutual information estimators tailored for sparse ranking labels.

Acknowledgements

We thank the Quantum CLEF 2025 organizers and the D-Wave team for providing access to quantum annealing hardware.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4 to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, Journal of Machine Learning Research 3 (2003) 1157–1182.

[2] H. Peng, F. Long, C. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (2005) 1226–1238.

[3] F. Fleuret, Fast binary feature selection with conditional mutual information, Journal of Machine Learning Research 5 (2004) 1531–1555.

[4] M. S. Haque, M. Fahim, M. Ibrahim, An exploratory study on simulated annealing for feature selection in learning-to-rank, https://arxiv.org/abs/2310.13269, 2023. ArXiv preprint arXiv:2310.13269.

[5] A. Lucas, Ising formulations of many NP problems, Frontiers in Physics 2 (2014) 5.

[6] S. Mücke, R. Heese, S. Müller, M. Wolter, N. Piatkowski, Feature selection on quantum computers, Quantum Machine Intelligence 4 (2022) 1–11.

[7] S. Yarkoni, E. Raponi, T. Bäck, S. Schmitt, Quantum annealing for industry applications: introduction and review, Reports on Progress in Physics 85 (2022) 104001.

[8] A. Pasin, M. F. Dacrema, W. Cuhna, M. A. Gonçalves, P. Cremonesi, N. Ferro, QuantumCLEF 2025: Overview of the second quantum computing challenge for information retrieval and recommender systems at CLEF, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[9] A. Pasin, M. F. Dacrema, W. Cuhna, M. A. Gonçalves, P. Cremonesi, 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), Lecture Notes in Computer Science, Springer, 2025.