Overview of the CLEF 2025 JOKER Task 1: Humour-aware Information Retrieval

Liana Ermakova[1], Ricardo Campos[2,3], Anne-Gwenn Bosser[4] and Tristan Miller[5,6,*]

[1]Université de Bretagne Occidentale, HCTI, France
[2]INESC TEC, Porto, Portugal
[3]University of Beira Interior, Covilhã, Portugal
[4]Bretagne INP – ENIB, Lab-STICC CNRS UMR 6285, France
[5]Department of Computer Science, University of Manitoba, Winnipeg, Canada
[6]Austrian Research Institute for Artificial Intelligence (OFAI), Vienna, Austria

Abstract

This paper presents the details of Task 1 of the JOKER-2025 Track, an information retrieval task where the goal is to find relevant humorous in a collection of text documents. The intended use case is retrieving jokes on a specific topic, something that may benefit humanities research, second-language learning, and the writing or translation of comedic texts. We provide two document collections: one in English and another in European Portuguese. The English collection consists of 77,658 documents, of which 5,198 are annotated as humorous, and 219 queries with relevance judgments. The Portuguese collection contains 45,126 texts, including 1,199 humorous documents along with 98 queries. Together, these collections support cross-linguistic studies in humour detection and contribute to the development of more inclusive and language-aware retrieval systems. Nine teams submitted 62 runs in total for this task.

Keywords

Wordplay, Puns, Humour-aware Information Retrieval, Computational Humour, wordplay detection, test collection,

1. Introduction

This paper provides an overview of Task 1 of the JOKER-2025 Track[1], which was held as part of the 16th Conference and Labs of the Evaluation Forum (CLEF 2025)[2]. The overall objective of the JOKER track series [1, 2, 3, 4], which began in 2022, is to facilitate collaboration among linguists, translators, and computer scientists to advance the development of automatic humour analysis. In each edition of the JOKER track, we construct and publish reusable, quality-controlled datasets to serve as training and test data for various humour processing tasks. In Task 1, run since 2024 [5], participants build systems to retrieve short humorous texts from a document collection based on a given query. For details on JOKER-2025's other two tasks, we refer the reader to their respective overview papers [6, 7]. Further information and insights are also presented in the Track's overview paper [1].

Search engines generally do not account for humour, ambiguity, or subversion of linguistic rules as features for selecting relevant documents to be returned. However, humour-aware retrieval, such as retrieval of wordplay-containing passages, can be useful for certain use cases or for user groups who appreciate or are interested in humorous qualities of text [8, 9]. For example, humour retrieval could benefit humour scholars in the humanities (as a research tool), second-language learners (as a learning aid), advertisement copywriters and professional comedians (as a writing aid), or even for translators who might need help adapting certain jokes to other cultures. Formally, for Task 1, the objective is to retrieve short humorous texts from a document collection based on a given query. The retrieved texts

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
0000-0002-7598-7474 (L. Ermakova); 0000-0002-8767-8126 (R. Campos); 0000-0002-0442-2660 (A. Bosser); 0000-0002-0749-1100 (T. Miller)

[1]https://www.joker-project.com
[2]https://clef2025.clef-initiative.eu/

Figure 1: CLEF 2025 JOKER Task 1 on Codabench

should fulfill two criteria: they must be relevant to the query, which encodes a topic, and they must be humorous, which for our purposes means being instances of wordplay. For example, a search query of "math" would mean that the goal is to find math jokes (e.g., "Why don't mathematicians argue? Because they always try to find common denominators!"), while the query "Tom" would mean that the goal is to find jokes about some person or entity named Tom (e.g., "Why did Tom bring a ladder to the bar? Because he heard the drinks were on the house!").

The data for our task builds upon the English corpora constructed in previous editions of JOKER [2, 3], and has been expanded with a substantial set of humorous texts in Portuguese.

This year, nine teams, out of the total thirteen active JOKER participants, submitted 62 runs for Task 1 out of the 136 runs submitted to the track. (See run statistics in Table 1.) The English subtask attracted nearly twice as many participants, with nine teams submitting 41 runs, compared to five teams submitting 21 runs for Portuguese.

This year saw a significant change to the task infrastructure, now hosted at Codabench [10], a Free Software web-based platform for organising AI benchmarks. We provided separate Codabench benchmarks for English[3] (see Fig. 1 and Portuguese[4]. Codabench facilitated the organisation of the 2025 track and attracted many new participants, who registered on the platform and gained full access to the competition, including submission and leaderboard pages. We continue to receive new registrations and post-competition submissions; however, this paper presents only those runs submitted prior to the official release of results to participants.

The remainder of this paper is structured as follows: Section 2 describes the test and train data in English and Portuguese as well as it format, Section 3 presents the evaluation metrics, Section 4 describes the participants' runs, and Section 5 presents an analysis of their results on the training and test data. Finally, Section 6 provides some concluding remarks.

2. Dataset construction and characterisation

The data for this task consist of documents in both English and Portuguese, allowing for cross-lingual research and evaluation. The following two sections describe the procedures carried out to construct and prepare the dataset.

[3]https://www.codabench.org/competitions/8686/
[4]https://www.codabench.org/competitions/8736/

Table 1
Number of runs submitted to CLEF JOKER 2025 Task 1 by language

team	EN	PT
arampaegos	3	9
cryptix [11]	3	0
fhelms [12]	4	0
igoranchik [13]	3	2
kamps [12]	4	4
pjmathematician [14]	7	4
rasion [15]	2	2
sarath_kumar [11]	1	0
tanishc228 [16]	14	0
Total	41	21

2.1. English data

In the 2025 edition, the English data is an extension of that used in Task 1: Humour-aware Information Retrieval from JOKER 2024 [2, 5], which was constructed based on an English wordplay detection corpus [17, 18] and valid translations [2, 19]. We grouped the humorous texts into clusters of related topics and created queries based on these clusters. We added a significant number of topically relevant but non-humorous texts by extracting relevant passages from Wikipedia and by generating passages using Meta's Llama-2 (7B) models. Due to the number of queries, the corpus contains a large fraction of non-relevant content. Both positive and negative examples included a mix of generated and human-written texts to prevent the task from being reduced to simply detecting generated content.

In 2024, the total number of documents in the corpus was 61,268, with 4,492 humorous texts and 56,776 non-humorous ones. For 57 queries, 11,831 documents were considered topically relevant.

For the 2025 edition, we expanded this data with new manually created jokes and texts generated by the LLMs Bard, Claude, ChatGPT, and Phi-3 Mini. The resulting corpus contains 77,658 texts in total, of which 5,198 are humorous. Detailed statistics on the English-language data sources for Task 1: Humour-aware Information Retrieval is given in Table 2.

For creating the set of queries, we harnessed data from CLEF 2023 JOKER Task 2: Pun Location and Interpretation [20, 3, 21], and in particular, the locations of wordplay in texts – i.e. words or phrases carrying multiple meanings. In CLEF 2023 JOKER Task 2, puns were either homographic (identical spelling as in "I used to be a banker but I lost interest") or heterographic (i.e., exploiting paronymy as in *propane/profane* in "When the church bought gas for their annual barbecue, proceeds went from the sacred to the propane.") To expand the queries, we used the semantic annotations of pun locations (pun interpretation) – i.e., pairs of lemmatised word sets, containing the synonyms (or, if absent, hypernyms) of the two words involved in the pun, excluding any that share the same spelling as the pun. The lists of query expansions were manually checked. The document was deemed humorous and relevant to the query if it came from the positive examples of the JOKER corpus and included the query term or its expansions.

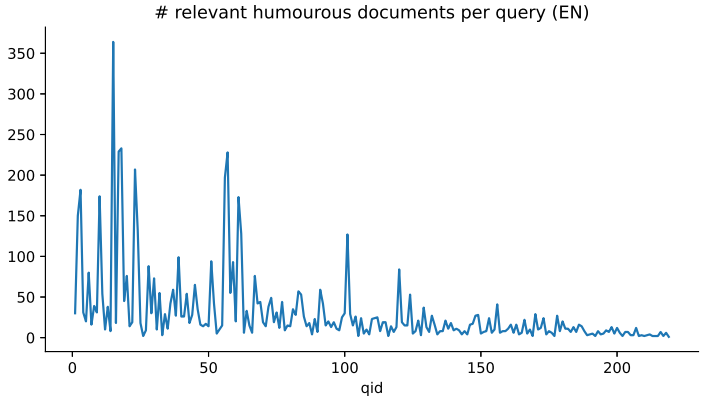
In this edition for 219 queries, 6,655 documents were judged humourous and topically relevant. As in 2024, we used 11 queries for the train and the rest for the test. The detailed statistics on the number of relevant humourous documents per query for the English dataset is given in Figure 2.

2.2. Portuguese data

To extend the multilingual scope of the task, we introduced a substantial new dataset in European Portuguese (PT-PT). This collection consists of 45,126 documents, of which 1,199 are humorous and 43,927 are non-humorous. The humorous texts were compiled through a three-stage process. First, 660 humorous instances from the English dataset were automatically translated into Portuguese using DeepL. Second, 421 texts were manually curated from various Portuguese-language websites. Finally,

Table 2
Number of humorous and non-humorous documents in the English Task 1 data by source

source	non-humorous	humorous	total
Bard	36	4	40
Claude	0	74	74
ChatGPT	149	381	530
JOKER	4,954	3,507	8,461
Llama-2	12,523	0	12,523
Phi-3 Mini	8,204	0	8,204
manual	2	247	249
translations	985	0	985
Wikipedia	46,592	0	46,592
total	72,460	5,198	77,658



count	219
mean	30
std	48
min	1
25%	7
50%	15
75%	29
max	364

Figure 2: Statistics on the number of relevant humourous documents per query in the English Task 1 data

118 humorous texts were generated using ChatGPT (4o-mini model). All texts underwent manual validation to ensure quality and conformity to the PT-PT variant. Queries for this collection were derived through a systematic topic-grouping procedure. Using GPT-3.5-turbo, the puns were clustered by theme, e.g., "grapes" and "oranges" were grouped under the broader category "fruit". Puns without a clear thematic link were marked as irrelevant. A manual curation process refined these groupings into 98 distinct queries associated with the 1,199 humorous texts.

To compile the 43,927 non-humorous documents, we employed a two-step process. First, 41,028 sentences were retrieved from Wikipedia using the same API-based approach as in the English dataset. Then, 2,899 additional non-humorous texts were generated using GPT-3.5-turbo. To ensure consistency with the European Portuguese variant, all texts were passed through the PtVId model [22] to detect Brazilian Portuguese (PT-BR) entries. Any PT-BR texts were automatically translated into PT-PT using ChatGPT-4o-mini, followed by manual validation.

Twenty-nine queries with their judgments (qrels) were created for training or validating participants' systems. Then, another 69 queries were created as a test set.[5]. For all 98 queries (combined test and training), 21,636 documents were considered topically relevant (i.e., they matched the query or its expansions). Of these, 1,334 were humorous.[6]

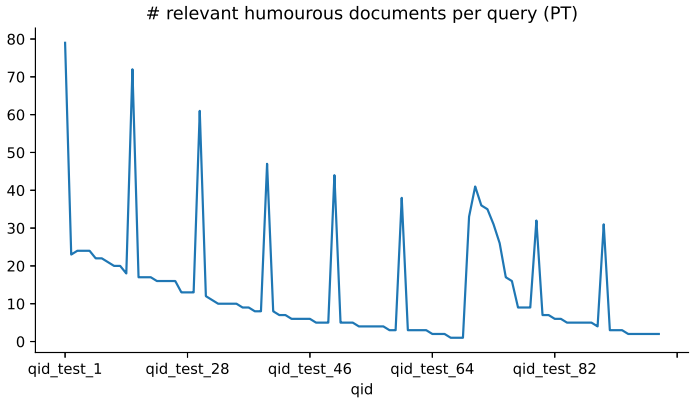
The descriptive statistics of the Portuguese data sources are provided in Table 3, while Figure 3 shows the distribution of relevant humorous texts per query. For Portuguese, the average is 14, with a median of 8, reflecting a more compact distribution aligned with the smaller dataset size.

[5]Note that we also included all the training-set queries in the test input file; however, they are excluded from the resulting scores.

[6]Note that this number is higher than the 1,199 humorous documents collected, as a document may be associated with more than one query.

Table 3
Number of documents in the Portuguese Task 1 data by source

source	#
chatgpt-3.5turbo	1615
joker	972
joker-2	227
wikipedia	15720
wikipedia-non-relevant	26592



count	98
mean	14
std	15
min	1
25%	4
50%	8
75%	17
max	79

Figure 3: Statistics on the number of relevant humourous documents per query in the Portuguese Task 1 data

2.3. Input formats

As described in the following subsections, the input formats for the document collection, queries, and training/validation data (qrels) generally follows those used for the 2024 edition of the task.

2.3.1. Document collection

We provide the training and test data in a JSON format with the following fields:

- **docid** a unique document identifier
- **text** the text of the instance, which may or may not contain wordplay

Input example:

```
[
  {
    "docid": "1",
    "text":  "Good laws have sprung from bad customs."
  },
  {
    "docid": "2",
    "text":  "The musical score to Topsyturveydom does not survive, but amateur
        productions in recent decades have used newly composed scores or performed the
        work as a non-musical play."
  },
  {
    "docid": "3",
    "text": "The organic compound primarily responsible for the characteristic odor of
        musk is muscone."
  },
  {
```

```
    "docid": "51135",
    "text":  "I've inherited a fortune, said Tom, willfully"
  },
  {
    "docid": "591",
    "text":  "My name is Will, I'm a lawyer."
  }
]
```

2.3.2. Queries

The train and test queries are also JSON files, this time with the following fields:

- **qid** a unique query identifier from the input file
- **query** the search query, e.g. "math" means that the goal is to find math jokes, while the query "Tom" means that the goal is to find jokes about Tom

Input example:

```
[
  {"qid":"qid_train_1","query":"steps"},
  {"qid":"qid_train_3","query":"math"},
  {"qid":"qid_train_4","query":"Tom"}
]
```

2.3.3. Qrels

Finally, we provide training/validation data in the format of JSON qrels files with the following fields:

- **qid** a unique query identifier from the query input file
- **docid** a unique document identifier from the corpus
- **qrel** indication the document **docid** is relevant to the query **qid** and is a wordplay instance

Example of a qrel file:

```
[
  {
    "qid":   "qid_train_0",
    "docid": "27260",
    "qrel":  0
  },
  {
    "qid":   "qid_train_0",
    "docid": "591",
    "qrel":  1
  },
  {
    "qid":  "qid_train_0",
    "docid": "51135",
    "qrel":1
  }
]
```

2.4. Output format

As with the input formats, the output format is identical to that used in CLEF 2024 JOKER Task 1. That is, we required results to be provided in a JSON format with the following fields:

- **run_id** run ID starting with <team_id>_<task_id>_<method_used>, e.g. UBO_task_1_TFIDF
- **manual** flag indicating if the run is manual {0,1}
- **qid** a unique identifier from the input file
- **docid** an identifier of the document retrieved from the corpus to the **qid** query
- **rank** retrieved document rank
- **score** normalised document relevance score (in the [0–1] scale)

For each query, the maximum allowed number of distinct documents (**docid** field) is 1000. A sample output file is as follows:

```
[
  {
    "run_id":"team1_task_1_TFIDF",
    "manual":0,
    "qid":"qid_train_0",
    "docid":"27260",
    "rank":1,
    "score":0.97
  },
  {
    "run_id":"team1_task_1_TFIDF",
    "manual":0,
    "qid":"qid_train_0",
    "docid":"591",
    "rank":2,
    "score":0.8
  },
  {
    "run_id":"team1_task_1_TFIDF",
    "manual":0,
    "qid":"qid_train_1",
    "docid":"27261",
    "rank":1,
    "score":0.7
  }
]
```

3. Evaluation measures

Performance was measured with standard information retrieval metrics as implemented in TrecTools, a Free Software Python library for information retrieval [23]. For each run we report the number of documents retrieved (#ret), the number of relevant documents retrieved (#rel), mean average precision (MAP; the mean of average precision scores across queries), geometric mean average precision (GMAP), precision at the number of relevant documents (P@R), mean reciprocal rank (MRR; the average of the reciprocal rank of the first relevant item across queries), precision (P@n; the proportion of relevant items retrieved at the top $n = 5, 10, 100, 1000$ positions), normalised discounted cumulative gain (NDCG; accounting for the relevance and position of documents in the ranking, normalised against the ideal ranking), and (for Portuguese only) the binary preference score (bpref).

4. Participants' approaches

In total, nine teams submitted 62 runs (see Table 1), with five of these teams contributing 21 runs to the Portuguese subtask. Every team participating in the Portuguese subtask also submitted runs for English. The approaches used by the participating teams are as follows:

rasion [15] This team proposed a dual-screening architecture that separates humour-aware information retrieval into two distinct stages. The first employs a semantic similarity model that uses the paraphrase-multilingual-mpnet-base-v2 model to encode queries and documents into dense vector representations, and distance-based metrics and cosine similarity to quantify semantic alignment and filter query-relevant documents. This step is followed by a transformer-based classifier (xlm-roberta-base) that identifies humorous texts containing puns. The method, applied to both English and Portuguese datasets, aims to reduce task complexity through modularisation. Their system achieved strong performance in Portuguese, highlighting the effectiveness of separating relevance and humour detection subtasks.

cryptix and sarath_kumar [11] These participants employed a fine-tuned Sentence-BERT (SBERT) model to generate semantic embeddings of queries and documents. They trained the model using a cosine similarity loss on humour-labelled query–document pairs, aiming to capture implicit humour such as irony or exaggeration. The resulting vectors were indexed using the Facebook AI Similarity Search (FAISS) for efficient retrieval, and results were re-ranked using human-annotated humour intensity scores.

igoranchik [13] This team implemented a hybrid retrieval pipeline combining dense and lexical retrieval, followed by cross-encoder reranking. They fine-tuned the intfloat/multilingual-e5-small model using contrastive objectives – Multiple Negative Ranking Loss (MNRL) and an Adaptive Margin Loss – on humour-annotated data, including synthetic queries generated with GPT-4o-mini. BM25 was used for lexical retrieval via Anserini, while dense vectors were stored in Qdrant. The top 1000 documents from both retrieval methods were merged using reciprocal rank fusion and re-ranked using the cross-encoder/ms-marco-MiniLM-L12-v2.

pjmathematician [14] This team implemented a two-stage pipeline using the Qwen family of large language models (LLMs). First, they applied large Qwen models (Qwen3-14B and Qwen3-32B) to analyse the entire document corpus, generating humour-related metadata such as a binary 'isJoke' flag and textual explanations for each document. These enriched representations were then used in a dense retrieval step, where smaller Qwen embedding models (Qwen3-4B and Qwen3-8B) indexed either the original text or the explanation-augmented versions. Retrieval was performed using both generic and humour-specific query prompts.

tanishc228 [16] This participant proposed a multi-stage ensemble retrieval system combining traditional IR methods with neural rerankers (ColBERT and a BERT-based cross-encoder), complemented by handcrafted wordplay features. Their pipeline retrieves documents using both lexical and semantic methods, followed by contextual reranking and score fusion. The system aims to capture humorous content by incorporating features such as punctuation, repetition, and alliteration.

kamps and fhelms [12] These teams submitted baseline runs using Anserini BM25 or BM25+RM3 and zero-shot MSMARCO-trained neural cross-encoder rerankings of the top 100 results.

All participants who submitted runs also submitted system description papers to the Working Notes volume [24]. Two teams from the same university (alecs and kamps) submitted a single joint report, as did teams cryptix and sarath_kumar, resulting in a total of seven Working Notes from the participants

of Task 2. Despite the requirement to include the team ID in the run name, participants' submissions often differed in their run names, registration details, and Codabench IDs. We manually matched the Working Notes with the submitted runs and report the results using the team names provided in those submissions.

5. Results

5.1. Test data

Tables 4 and 5 report the official Task 1 results for English and Portuguese, respectively.

For the English subtask, we received 39 distinct valid non-zero scored runs. The top-performing results in terms of MAP and NDCG@5 for both English and Portuguese were delivered by the team pjmathematician [14], which used the Qwen model for retrieval and filtering, and by the team Rasion [15], which employed dense retrieval combined with transformer-based humour detection.

Across all metrics, team pjmathematician [14], who applied two-stage Qwen LLM filter–explainer and dense retriever, obtained the best scores. In terms of MAP, their best approach (MAP = 0.3501) outperformed the next-best team by a factor of two. This latter team, from the University of Amsterdam [12], applied RM3 RoBERTa with drop 60 (MAP = 0.1672). The difference in NDCG@5 is even more significant, with 0.608 for pjmathematician compared to just 0.0152 for the University of Amsterdam. Close results were achieved by team rasion [15], who applied pre-trained models for a semantic matching network for relevance and a humour classification network for wordplay detection by RoBERTa. This approach had comparable NDCG@5 scores to the run pjmathematician_Q14-Q8-R. The application of cross-encoders by the University of Amsterdam [12] achieved significantly lower results than the RM3 and BM25 baselines in terms of MAP (0.0027 vs. 0.1237) and NDCG@5 (0.0038 vs. 0.14). This year's best run nearly tripled last year's top MAP =0.12 from the University of Amsterdam, who used RM3 with a T5 filter [2, 5]. The run UAms_RM3RoBERTa_drop60 has MAP = 0.17, being in the third batch of the results according to this metric. However, NDCG@5 = 0.015 is surprisingly low. This may be explained by the uneven performance drop across queries when applying a simple threshold. A more detailed per-query analysis could help determine whether this effect is due to unstable qrels, particularly for queries with very few relevant documents.

This year, the best results are comparable with those of topical relevance only in 2024 [5]. However, the baseline BM25 and RM3 runs by the University of Amsterdam this year and in 2024 [25] show comparable performance, with only slight improvements. This suggests that the core properties of the dataset have remained largely stable and we can assume that the improvement might be explained by the participants' approaches rather than changes in the data.

For the Portuguese subtask, we received 19 distinct valid non-zero scored runs. The best-scoring teams pjmathematician [14] and rasion [15] have very close results in terms of MAP (around 0.4) and NDCG@5 (around 0.5); they are followed by the University of Amsterdam [12] submitting the BM25 baseline with MAP = 0.08.

For the best runs, we observe weakly opposite trends for the English and Portuguese subtasks – namely, better results in terms of MAP for Portuguese but lower results in terms of P@5 and NDCG@5. This might be related to the higher average number of relevant humorous documents per query in English. This hypothesis is supported by the fact that for English, the best P@10 was 0.4 with a median of 15 relevant humorous documents per query, while for Portuguese, P@10 reaches 0.34 with a median of 8.

5.2. Training data

As in previous years, runs were submitted for both the training and test datasets in order to analyse potential overfitting and related effects. Tables 6 and 7 report the Task 1 results on the training data for English and Portuguese, respectively.

For the English subtask, we observe that the top four runs according to MAP (0.33 to 0.35) and NDCG@5 (0.56 to 0.61) submitted by pjmathematician [14] remain at the top of the table with also closed values of MAP (0.44 to 0.48) and NDCG@5 (0.53 to 0.68). Both runs Rasion_SenTransF+Roberta [15] have better scores on the train data than the University of Amsterdam with the best-scored runs achieving MAP = 0.59 and NDCG@5 = 0.61. The run UAms_RM3RoBERTa_drop60 [12] shows similar performance in terms of MAP on the train and the test sets, but with an improvement of NDCG@5 for the train data. However, many runs achieved higher scores on the training set, which lowered the ranking of the University of Amsterdam's run. Two runs Rasion_SenTransF+Roberta [15] and the run Cryptix_SBERT achieved more than double the MAP and at least 50% improvement in terms of NDCG@5 on the training data compared to the test data which might be a result of overfitting. The RM3 and BM25 runs from the University of Amsterdam [12] remain mid-ranked, showing stable scores without signs of overfitting, suggesting similar properties between the training and test data and confirming their strength as baselines. This also suggests that the improvement of other approaches may be attributed to their quality rather than differences between this year's data and the 2024 test collection. Cross-encoders performed poorly on both the training and test sets, likely because they are not designed to detect humour.

Teams rasion [15] and pjmathematician [14] submitted the highest-scoring runs on the English collections, also achieving the best results on the Portuguese training and test collections. Note that they achieved better results on the training data than on the official test data. They are followed by the BM25 run from the University of Amsterdam [12], which ranked fifth on both the Portuguese training and test collections, showing a 2–3 times drop in MAP and a 4–10 times drop in NDCG@5. However, the high ranking of BM25 may be partly due to the fact that the Portuguese subtask had roughly half as many runs as the English subtask. Note that on the test sets, the drop in terms of MAP and NDCG@5 is even higher. Cross-encoders remain low and stable among test and training collections, as for English.

6. Conclusions

In this paper, we have presented an overview and discussion of the results of Task 1 of the JOKER-2025 challenge on the retrieval of humorous texts relevant to a search query. Based on the data for wordplay detection and interpretation previously constructed within the CLEF JOKER track [1, 26, 20, 3, 21], we constructed a unique reusable test collection for wordplay retrieval in English. We manually created new jokes to avoid potential LLM data contamination. To prevent the task from being reduced to generated text detection, both positive and negative examples comprised a combination of human-written and machine-generated texts. The English collection consists of 77,658 documents, of which 5,198 are annotated as humorous, and 219 queries with relevance judgments.

In addition to this, this year we also expanded the dataset with Portuguese data collected from Portuguese-language websites, translated from the English corpus and generated by Chat-GPT (4o-mini model). The Portuguese collection contains 98 queries and 45,126 texts, including 1,199 humorous documents.

This year, the track setup was updated, with submissions managed through Codabench. Nine teams submitted 41 runs for the English subtask, of which five also submitted 21 runs for the Portuguese subtask, resulting in 62 valid distinct runs in total.

The teams applied diverse methods, ranging from traditional approaches rankers such as TF–IDF, BM25, and RM3 to cross-encoders with and without filtering, to more modern ones, including fine-tuned transformers and LLMs. The best results both for English and Portuguese were achieved by the team pjmathematician [14], which applied the Qwen model for retrieval and filtering, and the team rasion [15], which applied dense retrieval and transformer-based detection of humorous texts. These results might testify AI progress in pun detection. Further analysis is needed to assess the impact of potential LLM data contamination on this performance.

This year's English task showed remarkable progress, with the best run by team pjmathematician achieving a MAP of 0.3501 – nearly triple last year's top score – and outperforming all competitors by

a wide margin across the various metrics. In contrast, the University of Amsterdam's cross-encoder approaches performed substantially worse than their RM3 and BM25 baselines, confirming the effectiveness of simpler retrieval strategies for this dataset. For the Portuguese subtask, results were more balanced, with pjmathematician and raison achieving similar MAP and NDCG@5 scores around 0.4 to 0.5, far ahead of the BM25 baseline. Interestingly, while the Portuguese runs achieved higher MAP scores, they trailed the English runs in precision and NDCG@5, likely due to the smaller pool of relevant humorous documents per query. Overall, these findings suggest that while the dataset's core properties have remained stable, combining retrieval and filtering remains key to advancing performance.

The University of Amsterdam's RM3 and BM25 runs remained stable and reliable baselines, showing no overfitting and similar performance across test/training and English/Portuguese datasets. Improvements by other methods likely reflect their quality rather than dataset differences. Cross-encoders performed poorly, likely due to their unsuitability for humour detection.

In general, our results confirm that retrieval models are humour-agnostic and humour detection is still a challenge for machine learning models and LLMs despite improvement over the last year edition.

For more information about the JOKER lab this year, please refer to the overview paper [1], and the Working Notes papers for Task 2: Pun Translation [6] and Task 3: Onomastic Wordplay Translation [7]. Visit the JOKER website at https://joker-project.com for any other information related to the track.

Acknowledgments

This work has received a government grant managed by the National Research Agency under the program Investissements d'avenir integrated into France 2030, with the Reference ANR-19-GURE-0001. It was also financed by National Funds through the Portuguese funding agency FCT through the project LA/P/0063/2020 (DOI 10.54499/LA/P/0063/2020). Ricardo Campos would also like to acknowledge project StorySense, with reference 2022.09312.PTDC (DOI 10.54499/2022.09312.PTDC). We thank all other colleagues and students who participated in data construction, the translation contests, and the CLEF JOKER track.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Grammarly in order to: Grammar and spelling check and Paraphrase and reword. Further, the authors used Gemini in order to: Generate images. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] L. Ermakova, A.-G. Bosser, T. Miller, R. Campos, Overview of JOKER: Humour in the machine, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. G. S. de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), 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.

[2] L. Ermakova, A.-G. Bosser, T. Miller, V. M. Palma Preciado, G. Sidorov, A. Jatowt, Overview of the CLEF 2024 JOKER track: Automatic humour analysis, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. D. Nunzio, P. Galuščáková, A. G. S. de Herrera, G. Faggioli, N. Ferro (Eds.), 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, pp. 165–182. doi:10.1007/978-3-031-71908-0_8.

[3] L. Ermakova, T. Miller, A.-G. Bosser, V. M. Palma Preciado, G. Sidorov, A. Jatowt, Overview of JOKER – CLEF-2023 track on automatic wordplay analysis, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction: Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023), volume 14163 of *Lecture Notes in Computer Science*, Springer, Cham, 2023, pp. 397–415. doi:10.1007/978-3-031-42448-9_26.

[4] L. Ermakova, T. Miller, F. Regattin, A.-G. Bosser, C. Borg, Élise Mathurin, G. L. Corre, S. Araújo, R. Hannachi, J. Boccou, A. Digue, A. Damoy, B. Jeanjean, Overview of JOKER@CLEF 2022: Automatic wordplay and humour translation workshop, in: A. Barrón-Cedeño, G. D. S. Martino, M. D. Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction: Proceedings of the Thirteenth International Conference of the CLEF Association (CLEF 2022), volume 13390 of *Lecture Notes in Computer Science*, Springer, Cham, 2022, pp. 447–469. doi:10.1007/978-3-031-13643-6_27.

[5] L. Ermakova, A.-G. Bosser, T. Miller, A. Jatowt, Overview of the CLEF 2024 JOKER task 1: Humour-aware information retrieval, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. Seco de Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), volume 3740 of *CEUR Workshop Proceedings*, 2024, pp. 1775–1785.

[6] L. Ermakova, A.-G. Bosser, T. Miller, R. Campos, Overview of the CLEF 2025 JOKER Task 2: Wordplay Translation from English into French, in: [24], 2025.

[7] L. Ermakova, T. Miller, Y. Naud, A.-G. Bosser, R. Campos, Overview of the CLEF 2025 JOKER Task 3: Onomastic Wordplay Translation, in: [24], 2025.

[8] D. Gupta, M. Digiovanni, H. Narita, K. Goldberg, Jester 2.0 (demonstration abstract): Collaborative filtering to retrieve jokes, in: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '99, Association for Computing Machinery, New York, NY, USA, 1999, p. 333. doi:10.1145/312624.312770.

[9] L. Friedland, J. Allan, Joke retrieval: Recognizing the same joke told differently, in: Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM '08, Association for Computing Machinery, New York, NY, USA, 2008, p. 883–892. doi:10.1145/1458082.1458199.

[10] Z. Xu, S. Escalera, A. Pavão, M. Richard, W.-W. Tu, Q. Yao, H. Zhao, I. Guyon, Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform, Patterns 3 (2022). doi:10.1016/j.patter.2022.100543.

[11] S. K. P, B. A, S. M, T. S, REC_Cryptix at JOKER CLEF 2025: Teaching Machines to Laugh: Multilingual Humor Detection and Translation, in: [24], 2025.

[12] A. Kreefft-Libiu, F. Helms, C. Selçuk, J. Bakker, J. Kamps, University of Amsterdam at the CLEF 2025 JOKER Track, in: [24], 2025.

[13] I. Kuzmin, CLEF 2025 JOKER track: No pun left behind, in: [24], 2025.

[14] P. Vachharajani, pjmathematician at the CLEF 2025 JOKER Lab Tasks 1, 2 & 3: A Unified Approach to Humour Retrieval and Translation using the Qwen LLM Family, in: [24], 2025.

[15] B. Chen, C. Zhong, L. Kong, CLEF 2025 JOKER track enhancing humor-aware information retrieval with relevance-aware classification, in: [24], 2025.

[16] T. Chaudhari, A. Vora, S. Hotha, S. Sonawane, PICT at CLEF 2025 JOKER Task 1: BERT-Enhanced Ensemble Methods, in: [24], 2025.

[17] L. Ermakova, T. Miller, A.-G. Bosser, V. M. Palma Preciado, G. Sidorov, A. Jatowt, Overview of JOKER – CLEF-2023 track on automatic wordplay analysis, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, volume 14163, Springer Nature Switzerland, Cham, 2023, pp. 397–415. doi:10.1007/978-3-031-42448-9_26.

[18] L. Ermakova, A.-G. Bosser, A. Jatowt, T. Miller, The JOKER Corpus: English–French parallel data for multilingual wordplay recognition, in: SIGIR '23: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, New York, NY, 2023, pp. 2796–2806. doi:10.1145/3539618.3591885.

[19] L. Ermakova, A.-G. Bosser, T. Miller, A. Jatowt, Overview of the CLEF 2024 JOKER Task 3: Translate puns from English to French, in: G. Faggioli, N. Ferro, P. Galuscakova, A. G. Seco de Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 1800–1810.

[20] L. Ermakova, T. Miller, A.-G. Bosser, V. M. Palma Preciado, G. Sidorov, A. Jatowt, Overview of JOKER 2023 Automatic Wordplay Analysis Task 2 – pun location and interpretation, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum, volume 3497 of *CEUR Workshop Proceedings*, 2023, pp. 1804–1817.

[21] L. Ermakova, A.-G. Bosser, A. Jatowt, T. Miller, The JOKER Corpus: English–French parallel data for multilingual wordplay recognition, in: SIGIR '23: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, New York, NY, 2023, pp. 2796–2806. doi:10.1145/3539618.3591885.

[22] H. Sousa, R. Almeida, P. Silvano, I. Cantante, R. Campos, A. Jorge, Enhancing Portuguese variety identification with cross-domain approaches, in: Proceedings of the 39th Annual AAAI Conference on Artificial Intelligence (AAAI'25), volume 39, 2025, pp. 25192–25200.

[23] J. a. Palotti, H. Scells, G. Zuccon, TrecTools: an open-source Python library for information retrieval practitioners involved in TREC-like campaigns, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, New York, 2019, pp. 1325–1328. doi:10.1145/3331184.3331399.

[24] G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2025.

[25] E. Schuurman, M. Cazemier, L. Buijs, J. Kamps, University of Amsterdam at the CLEF 2024 JOKER track, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, 2024, pp. 1909–1922.

[26] L. Ermakova, T. Miller, A.-G. Bosser, V. M. Palma Preciado, G. Sidorov, A. Jatowt, Overview of JOKER 2023 Automatic Wordplay Analysis Task 1 – pun detection, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum, volume 3497 of *CEUR Workshop Proceedings*, 2023, pp. 1785–1803.

Table 4: Results for Task 1 English (test data)

Run ID	#ret	#rel	MAP	GMAP	P@R	MRR	P@5	P@10	P@100	P@1000	NDCG@5
pjmathematician_Q14-Q8-Q32	207000	3007	35.01	24.65	38.68	79.04	54.88	40.63	9.05	1.45	60.80
pjmathematician_Q14-Q8-Q14	207000	2954	34.86	24.31	39.07	78.74	54.49	40.29	9.00	1.43	60.59
pjmathematician_Q32-Q8-Q14	207000	2932	34.38	23.98	38.98	80.78	54.49	41.01	8.82	1.42	60.94
pjmathematician_Q32-Q8-Q32	207000	3011	32.91	23.03	36.99	76.20	50.34	39.95	9.00	1.45	55.98
pjmathematician_Q14-Q8-R	207000	2835	23.88	16.03	27.10	64.27	36.62	28.70	7.84	1.37	41.23
UAms_RM3RoBERTa_drop60	82818	1448	16.72	7.04	23.05	54.46	30.82	23.09	5.98	0.70	1.52
Rasion_SenTransF+Roberta	4588	811	16.21	NaN	20.59	64.93	35.92	24.47	3.92	0.39	41.34
Rasion_SenTransF+Roberta	53552	1475	15.79	5.53	20.21	55.94	30.82	22.75	5.72	0.71	34.02
Cryptix_SBERT	207000	1914	15.07	5.52	19.44	56.94	28.70	20.97	4.75	0.92	33.46
UAms_RM3	207000	1864	15.02	7.22	19.53	40.87	24.35	20.00	6.22	0.90	25.66
UAms_RM3RoBERTa	186164	1798	14.94	7.16	19.56	42.47	25.51	19.61	6.07	0.87	26.76
CCC_Ensemble_ColBERT_RM3	103500	1967	14.15	7.44	16.29	33.69	16.71	17.73	6.70	0.95	18.17
CCC_Ensemble	206091	2050	14.03	7.12	17.01	40.33	20.48	18.65	5.87	0.99	22.55
UAms_RM3	207000	1872	12.16	5.77	15.36	33.30	18.55	18.07	5.72	0.90	19.57
UAms_en_bm25	41270	1884	11.91	5.64	12.23	26.28	12.95	12.71	5.94	0.91	14.00
CCC_TFIDF_Rerank	100185	1764	11.26	NaN	15.25	40.59	20.59	17.07	5.75	0.86	22.04
UAms_Anserini	207000	2134	10.76	5.35	10.56	25.03	11.88	12.22	6.14	1.03	12.37
UAms_en_rm3	207000	2134	10.76	5.35	10.56	25.03	11.88	12.22	6.14	1.03	12.37
CCC_ColBERT_Enhanced	207000	1879	9.93	5.17	12.47	33.35	15.75	14.15	5.56	0.91	16.53
CCC_XLM_R_Rerank	207000	2227	9.66	5.03	9.83	36.41	13.82	11.98	5.52	1.08	16.19
CCC_ColBERT_Enhanced	207000	1418	6.69	2.06	9.34	31.24	13.43	11.40	3.82	0.69	14.55
CCC_XLM_R_Rerank	10350	918	6.30	2.56	9.65	23.88	9.18	9.08	4.43	0.44	19.10
CCC_ColBERT_Enhanced	207000	1367	6.21	1.82	8.67	28.52	11.30	10.34	3.64	0.66	12.42
CCC_Advanced_Ensemble_LTR	165600	2122	6.20	3.65	6.38	11.54	2.90	5.89	5.67	1.03	2.67
CCC_TFIDF	207000	1321	5.79	1.56	8.41	25.29	9.47	9.03	3.79	0.64	10.50
CCC_Ensemble_ColBERT_RM3	103500	1904	5.44	2.24	6.52	21.03	8.12	6.96	3.79	0.92	8.77
CCC_TF-IDF_Ensemble_ColBERT_RM3	103500	1922	5.31	2.22	5.98	20.16	7.15	6.47	3.79	0.93	7.90
Skommarkhos_BM25_E5_MiniLM	207000	2182	5.02	2.98	3.03	6.47	0.87	3.24	4.44	1.05	0.65
UAms_en_bm25_CE1K	41270	1884	4.88	2.60	2.47	5.68	0.48	2.75	4.30	0.91	0.38
UAms_en_rm3_CE1K	207000	2134	4.78	2.67	2.32	5.48	0.39	2.51	4.32	1.03	0.27
UAms_Anserini	10350	849	4.76	1.41	3.92	7.67	0.87	3.67	4.10	0.41	0.75
cryptix_crossencoder	20700	999	3.78	1.55	2.43	5.64	0.48	2.51	4.83	0.48	0.34
CCC_Ensemble_RoBERTa_RM3	41400	1718	3.33	1.69	4.02	11.01	3.77	3.82	4.23	0.83	23.65
Skommarkhos_BM25_E5_MiniLM	20700	517	2.49	0.49	2.62	6.15	0.77	3.19	2.50	0.25	0.57
CCC_pipeline	5175	211	2.43	0.05	3.93	11.81	4.54	4.15	1.02	0.10	0.80
team_reranker_EN	20700	824	1.38	0.23	2.40	6.36	2.13	2.03	3.98	0.40	2.05
yourteam_xlm_roberta_large	20700	271	0.42	0.02	1.29	4.71	1.35	1.50	1.31	0.13	1.30
duth_xanthi_en	20700	62	0.04	0.00	0.21	0.75	0.10	0.10	0.30	0.03	0.08
cryptix_crossencoder	414000	336	0.02	0.00	0.01	0.08	0.00	0.00	0.01	0.05	0.00

Table 5: Results for Task 1 Portuguese (test data)

Run ID	#ret	#rel	MAP	GMAP	P@R	MRR	P@5	P@10	P@100	P@1000	NDCG@5	bpref
pjmathematician_Q32-Q4-R	69000	932	42.21	30.78	42.01	69.07	43.77	34.35	8.80	1.35	42.14	58.40
pjmathematician_Q14-Q4-R	69000	938	42.17	30.81	41.65	68.98	43.77	34.49	8.83	1.36	51.69	58.65
Rasion_SenTransF+Roberta	69000	905	40.51	28.90	40.17	66.57	44.93	38.41	8.61	1.31	50.15	83.68
Rasion_SenTransF+Roberta	62576	904	40.51	28.90	40.17	66.57	44.93	38.41	8.61	1.31	50.12	83.62
UAms_pt_bm25	12856	229	7.89	0.19	5.96	9.83	5.22	6.09	3.03	0.33	5.13	11.52
Skommarkhos_BM25_E5_MiniLM	69000	503	7.42	1.65	5.74	11.91	6.38	6.23	2.87	0.73	6.44	7.60
pjmathematician_Q06-gist	69000	562	6.95	1.75	4.99	11.20	5.51	6.38	2.64	0.81	5.46	7.16
Skommarkhos_BM25_E5_MiniLM	6900	228	6.90	0.28	5.58	12.65	5.22	5.94	3.30	0.33	5.35	6.90
results_pt_pt_finetuned	6900	199	6.71	0.41	6.74	20.21	7.54	7.10	2.88	0.29	9.29	32.38
UAms_pt_rm3	67994	262	6.54	0.25	5.91	9.51	4.64	5.65	2.78	0.38	4.47	10.28
myteam_BERT	69000	496	6.13	1.26	6.38	19.54	8.12	6.38	2.54	0.72	8.78	7.35
duth_xanthi_pt	6900	225	5.95	0.37	6.76	15.65	7.54	8.41	3.26	0.33	7.03	15.13
pjmathematician_Q06-gist-exp32	69000	512	4.91	1.35	2.92	7.15	2.61	3.48	2.42	0.74	2.73	4.03
UAms_pt_rm3_CE1K	67994	262	4.16	0.19	2.47	5.20	1.45	3.19	2.41	0.38	1.34	4.80
UAms_pt_bm25_CE1K	12856	229	3.84	0.12	1.99	4.47	1.16	3.04	2.35	0.33	0.91	4.31
team_xlmr_PT	6900	133	2.96	0.11	5.33	12.03	4.64	5.94	1.93	0.19	4.57	17.66
results_pt_large_pt_finetuned	6900	65	0.31	0.01	0.02	0.73	0.00	0.00	0.94	0.09	1.72	4.00
yourteam_pt_zeroshot	6900	46	0.27	0.01	0.28	1.13	0.00	0.29	0.67	0.07	0.00	5.25
xlm-roberta-triplet-pt	6900	28	0.22	0.00	0.33	2.48	0.58	0.43	0.41	0.04	0.70	2.19

Table 6: Results for Task 1 English (training data)

Run ID	#ret	#rel	MAP	GMAP	P@R	MRR	P@5	P@10	P@100	P@1000	NDCG@5
Rasion_SenTransF+Roberta	1376	642	59.28	56.10	61.41	69.72	60.00	52.50	23.75	5.35	60.87
pjmathematician_Q32-Q8-Q14	12000	395	48.10	35.02	47.15	78.47	66.67	53.33	18.50	3.29	67.55
pjmathematician_Q32-Q8-Q32	12000	439	45.91	32.72	45.17	73.06	60.00	49.17	17.00	3.66	59.80
pjmathematician_Q14-Q8-Q14	12000	411	44.90	30.77	47.98	70.74	56.67	50.83	18.75	3.43	56.99
pjmathematician_Q14-Q8-Q32	12000	440	44.28	30.38	45.55	69.63	51.67	50.83	18.42	3.67	53.27
Rasion_SenTransF+Roberta	1811	596	42.91	39.31	45.86	72.50	45.00	46.67	20.33	4.97	48.37
Cryptix_SBERT	12000	571	39.21	27.13	40.63	79.32	56.67	42.50	15.25	4.76	59.59
pjmathematician_Q14-Q8-R	12000	532	32.35	20.79	36.25	53.19	40.00	36.67	13.83	4.43	38.45
CCC_Ensemble	11844	477	18.54	4.46	18.89	47.78	21.67	25.83	11.25	3.98	24.70
CCC_XLM_R_Rerank	12000	538	18.52	3.67	18.45	36.79	18.33	15.00	9.42	4.48	20.91
CCC_Ensemble_ColBERT_RM3	6000	366	17.57	4.77	19.48	36.52	23.33	20.00	10.83	3.05	22.86
CCC_TFIDF_Rerank	5617	359	17.14	4.10	25.53	52.51	25.00	24.17	9.25	2.99	27.53
UAms_RM3	12000	465	17.09	2.63	16.62	33.53	20.00	16.67	9.08	3.88	21.79
UAms_RM3RoBERTa_drop60	4802	252	16.33	1.48	19.89	49.65	31.67	27.50	10.00	2.10	6.75
CCC_ColBERT_Enhanced	12000	473	15.99	3.17	17.19	31.34	21.67	16.67	10.75	3.94	20.88
UAms_RM3RoBERTa	10932	459	15.69	2.17	19.62	39.93	23.33	22.50	10.75	3.83	21.20
UAms_RM3	12000	459	15.53	2.08	19.44	39.07	21.67	21.67	10.50	3.83	20.11
UAms_Anserini	12000	476	15.51	3.39	13.61	29.91	16.67	17.50	9.75	3.97	16.86
UAms_en_rm3	12000	476	15.51	3.39	13.61	29.91	16.67	17.50	9.75	3.97	16.86
UAms_en_bm25	4204	471	14.22	3.31	13.71	21.40	6.67	13.33	10.83	3.93	7.24
CCC_ColBERT_Enhanced	12000	391	10.97	0.53	14.48	32.95	15.00	16.67	8.17	3.26	16.62
CCC_TFIDF	12000	383	10.16	0.49	12.31	34.58	15.00	14.17	8.42	3.19	17.11
CCC_ColBERT_Enhanced	12000	382	9.97	0.47	11.84	37.55	16.67	14.17	8.17	3.18	18.93
CCC_Ensemble_ColBERT_RM3	6000	371	9.79	1.94	10.41	21.90	13.33	12.50	8.67	3.09	12.62
CCC_TF-IDF_Ensemble_ColBERT_RM3	6000	370	9.59	1.84	10.77	19.15	13.33	12.50	8.17	3.08	12.06
CCC_Advanced_Ensemble_LTR	9600	487	9.46	2.17	7.65	8.18	3.33	6.67	8.33	4.06	2.31
Skommarkhos_BM25_E5_MiniLM	12000	442	9.29	5.13	6.65	9.03	6.67	6.67	7.67	3.68	5.35
UAms_en_rm3_CE1K	12000	476	8.63	1.76	7.17	8.24	5.00	6.67	8.17	3.97	4.04
UAms_en_bm25_CE1K	4204	471	8.59	1.77	7.63	8.24	5.00	6.67	8.00	3.93	4.04
CCC_XLM_R_Rerank	600	88	6.04	1.36	6.55	18.71	10.00	9.17	7.33	0.73	17.14
UAms_Anserini	600	74	5.31	0.60	4.83	12.67	8.33	8.33	6.17	0.62	6.97
CCC_Ensemble_RoBERTa_RM3	2400	207	4.94	1.46	5.19	20.98	15.00	12.50	9.00	1.73	29.57
cryptix_crossencoder	1200	122	4.45	1.08	3.38	7.79	5.00	4.17	10.17	1.02	4.09
Skommarkhos_BM25_E5_MiniLM	1200	62	2.59	0.71	2.68	8.64	5.00	4.17	5.17	0.52	4.29
team_reranker_EN	1200	46	1.37	0.28	2.42	7.90	5.00	3.33	3.83	0.38	4.09
yourteam_xlm_roberta_large	1200	16	0.95	0.04	2.66	10.07	5.00	3.33	1.33	0.13	4.04
CCC_pipeline	300	17	0.31	0.01	2.05	8.20	1.67	4.17	1.42	0.14	2.31
duth_xanthi_en	1200	9	0.04	0.01	0.54	4.78	1.67	1.67	0.75	0.08	1.78
cryptix_crossencoder	24000	18	0.01	0.00	0.02	0.09	0.00	0.00	0.00	0.03	0.00

Table 7: Results for Task 1 Portuguese (training data)

Run ID	#ret	#rel	MAP	GMAP	P@R	MRR	P@5	P@10	P@100	P@1000	NDCG@5	bpref
Skommarkhos_BM25_E5_MiniLM	29000	219	8.26	2.91	4.70	8.38	4.14	5.52	2.79	0.76	4.04	21.31
UAms_pt_bm25_CE1K	4474	115	4.69	0.24	3.00	4.11	2.07	3.10	2.03	0.40	1.54	15.00
results_pt_pt_finetuned	2900	100	9.84	0.55	10.56	19.19	8.28	6.90	3.45	0.34	10.33	44.87
duth_xanthi_pt	2900	107	5.86	0.46	7.07	15.95	5.52	5.52	3.69	0.37	7.83	13.26
yourteam_pt_zeroshot	2900	25	0.39	0.02	0.27	1.52	0.00	0.00	0.86	0.09	0.00	12.90
pjmathematician_Q32-Q4-R	29000	286	39.20	24.77	35.86	60.49	36.55	26.90	6.45	0.99	38.88	64.49
Rasion_SenTransF+Roberta	26477	296	49.00	29.84	49.37	65.47	42.07	32.07	7.59	1.02	52.24	98.09
yourteam_xlm-roberta-en	2900	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
pjmathematician_Q06-gist-exp32	29000	268	4.75	2.56	2.38	6.32	2.07	2.76	2.17	0.92	1.59	11.26
myteam_BERT	29000	191	8.13	2.11	7.21	18.23	7.59	4.83	2.45	0.66	9.63	22.35
xlm-roberta-triplet-pt	2900	6	0.08	0.00	0.11	0.68	0.00	0.00	0.21	0.02	0.00	2.16
results_pt_large_pt_finetuned	2900	47	0.61	0.01	0.27	4.32	0.69	0.34	1.62	0.16	5.28	7.48
pjmathematician_Q14-Q4-R	29000	278	39.46	24.51	36.17	60.41	36.55	26.55	6.34	0.96	45.40	64.26
Skommarkhos_BM25_E5_MiniLM	2900	116	7.62	0.51	3.92	9.81	3.45	6.21	4.00	0.40	3.47	19.66
pjmathematician_Q06-gist	29000	263	8.41	4.33	7.53	16.05	6.90	5.86	3.28	0.91	8.02	19.32
UAms_pt_bm25	4474	115	16.24	0.61	14.65	15.79	9.66	9.66	3.83	0.40	12.19	33.38
UAms_pt_rm3_CE1K	28905	124	4.68	0.39	3.00	4.12	2.07	3.10	2.03	0.43	1.54	15.03
team_xlmr_PT	2900	71	5.08	0.13	4.67	10.48	4.83	6.21	2.45	0.24	4.68	25.37
Rasion_SenTransF+Roberta	29000	297	49.00	29.84	49.37	65.47	42.07	32.07	7.59	1.02	52.24	98.66
yourteam_xlm-roberta-pt	2900	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
UAms_pt_rm3	28905	124	8.86	0.64	7.00	9.25	4.83	6.55	3.59	0.43	4.85	21.25