CLEF 2026 JOKER Track: Relevance-Aware Multilingual Retrieval for Humorous Wordplay

Notebook for the JOKER Lab at CLEF 2026

Xinrui Qiu[1], Huihui Chen[1,*] and Jinghe Zhai[1]

[1]*Foshan University, Foshan, China*

Abstract

Humour-aware information retrieval requires retrieval methods to identify texts that are both query-relevant and humorous. We address this dual requirement in CLEF 2026 JOKER Task 1 with a modular two-stage pipeline that separates topical matching from humour-aware filtering. The first stage uses `paraphrase-multilingual-mpnet-base-v2` to encode queries and candidate texts in a shared multilingual embedding space and retrieves candidates by cosine similarity. The second stage applies a fine-tuned `roberta-base` cross-encoder to estimate whether each query–candidate pair contains puns or wordplay. Retrieval and humour scores are combined with a fixed 0.3/0.7 weighted fusion, after which candidates predicted as humorous are retained for the final ranking. With the same retriever checkpoint for English and Hinglish and language-specific classifier fine-tuning, `RIVER_task_1_MPNet+RoBERTa` reports a `MAP` of 0.3063 on English and 0.3897 on Hinglish, an `NDCG@10` of 0.4129 and 0.5199, and a reciprocal rank of 0.6015 and 0.8382, respectively. On Hinglish the `NDCG@10` margin over the next entry in Table 3 is 0.06 in absolute terms. The results support separating semantic relevance from humour detection as a practical design strategy for multilingual humour-aware retrieval; the `P@10` gap observed for both languages is a concrete target for future reranking and calibration work.

Keywords

humour-aware information retrieval, pun detection, multilingual retrieval, relevance-aware classification,

1. Introduction

Humour is difficult to model computationally because the humorous effect of a short text rarely follows from surface words alone; it often depends on implicit knowledge, ambiguity, and pragmatic interpretation [1, 2]. The CLEF 2026 JOKER Lab formalises this challenge as Task 1, Humour-aware Information Retrieval, in which systems must retrieve short texts that satisfy two coupled requirements: the text must be topically relevant to a given query, and it must contain humorous wordplay [1]. Standard information retrieval methods are not sufficient for this dual criterion because they optimise topical relevance alone, while standalone pun detectors ignore the query-document relation.

To address this dual criterion, we use a modular two-stage design that separates semantic relevance from humour detection. In the first stage, we adopt the multilingual Sentence-Transformer model `paraphrase-multilingual-mpnet-base-v2` to map both queries and candidate texts into a shared dense semantic embedding space. Subsequently, we calculate the cosine similarity to retrieve candidate sentences with similar semantics [3, 4]. In the second stage, we fine-tune a `roberta-base` cross-encoder with `qrels` triples, and employ it to score candidate samples containing humorous wordplay [5, 2]. We fuse the two scores via fixed weighted aggregation and preserve only candidates predicted to contain humorous content. The aforementioned multilingual retriever is shared across English and Hinglish, while we fine-tune the classifier independently for each language variant.

The primary contribution of this work is a lightweight, reproducible pipeline for humour-aware retrieval that supports both English and Hinglish. Our submitted system `RIVER_task_1_MPNet+RoBERTa` achieves a MAP of 0.3063 for English and 0.3897 for Hinglish,

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

*Corresponding author.

qxruber@163.com (X. Qiu); ddchh@163.com (H. Chen); alger2020@163.com (J. Zhai)

0000-0003-4465-2716 (H. Chen)

as shown in Tables 2 and 3. The framework reuses a unified retriever checkpoint while fine-tuning distinct classifiers tailored to each language. The retrieval and classification branches are modularized as independent components, enabling researchers to substitute either module without overhauling the entire pipeline. We fully document our system configuration and language-wise evaluation scores to facilitate easy reproduction.

The remainder of this paper is organised as follows. Section 2 reviews related work on pun detection, pretrained language models, and multilingual dense retrieval. Section 3 presents the proposed two-stage method. Section 4 describes the datasets, experimental settings, evaluation metrics, and results. Section 5 concludes the paper and discusses future directions.

2. Related Work

Computational humour has often been studied as a classification problem in which a model determines whether a short text contains humorous content [6]. The psychology of laughter offers a complementary perspective on why humorous text is recognisable to human readers [8], whereas most computational approaches focus on observable textual signals such as wordplay, incongruity, and surprise. Earlier systems treated the task as a binary or fine-grained classification, often using hand-crafted features and contextual word representations.

Pun detection has frequently been formulated as a sequence-labeling problem. In this line of work, BiLSTM-CRF and related neural architectures are used to identify whether a sentence contains a pun and, in some cases, to locate the pun word that triggers the humorous effect [9, 10, 6]. These approaches are useful because the humorous trigger is often a single word whose interpretation changes in context, but they are typically monolingual, sentence-level, and decoupled from any query-level information need.

Transformer-based language models have further improved humour and wordplay modelling. Large pretrained models such as T5 and RoBERTa encode richer contextual information than earlier neural architectures and can be fine-tuned on task-specific datasets [11, 2]. Multilingual encoders such as XLM-RoBERTa have extended these gains to non-English settings [12], and prior CLEF JOKER submissions have explored T5 fine-tuning, query expansion, and semantic similarity filtering for pun detection and retrieval [11]. Cross-encoder reranking, in which a transformer is fine-tuned to score a query–document pair as a single sequence, has become a standard second-stage component in modern information retrieval pipelines [5]. Other JOKER participants have explored related tasks such as pun translation, providing a complementary multilingual perspective [7].

Dense multilingual retrieval with sentence encoders has recently become a practical alternative to sparse retrieval. The Sentence-BERT family of models, including the multilingual MPNet checkpoint we use in this paper, produces language-portable dense embeddings that can be indexed once and queried many times [3, 4]. Neural information retrieval has been surveyed more broadly by [13], who situate dense retrieval with respect to sparse baselines. Hybrid pipelines that combine a sparse first stage with a cross-encoder reranker have been shown to perform strongly, especially with query expansion methods such as RM3 [14]. However, cross-encoder rerankers are often fine-tuned under specific language assumptions and may require additional adaptation when the target data includes code-switching or a different script distribution.

Humour-aware information retrieval differs from standalone pun detection because it also requires query relevance. A retrieval method must avoid returning humorous texts that are unrelated to the query, while also avoiding relevant texts that do not contain wordplay. This dual requirement motivates a modular pipeline in which retrieval and classification are handled in separate stages. Our method follows this direction by combining a multilingual sentence encoder for retrieval with a relevance-aware cross-encoder for humour classification, and by using the same retriever checkpoint for English and Hinglish.

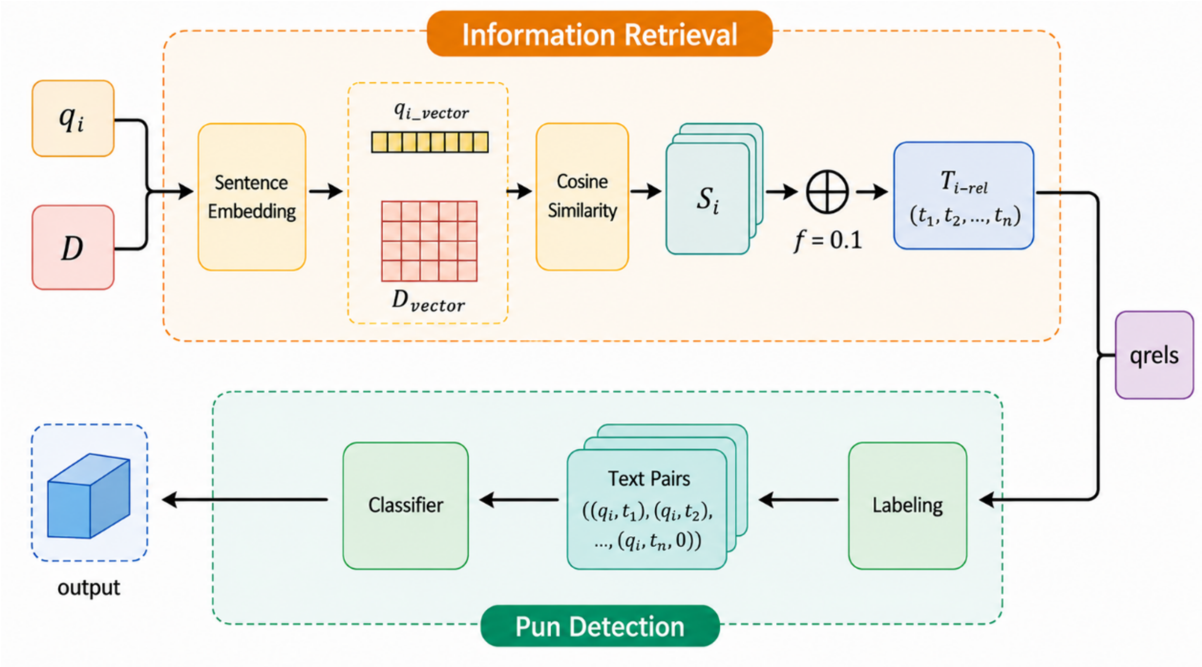

Figure 1: Overall workflow of the proposed two-stage method. The first stage encodes queries and documents with a multilingual sentence encoder and retrieves candidates whose cosine similarity with the query exceeds 0.1. The second stage scores each candidate with a fine-tuned cross-encoder and combines the two signals with a $0.3/0.7$ weighted fusion; only candidates classified as humorous are retained in the final run.

3. Our Method

As shown in Figure 1, the set of queries is denoted by $Q = \{q_i\}$ and the document collection is denoted by $D = \{d_j\}$. The Humour-aware Information Retrieval's goal is to retrieve texts from D that are relevant to each query in Q and contain humorous wordplay. Our method's framework simply divides the task into two stages. In the first stage, i.e., information retrieval stage, candidate texts are found according to semantic relevance, and in the second stage, i.e., pun detection stage, these texts are filtered and re-ranked according to the probability of containing puns or wordplay.

3.1. Information Retrieval Step

The information retrieval stage is designed to obtain a broad candidate set for each query. We use `paraphrase-multilingual-mpnet-base-v2` as the sentence embedding model because it supports multilingual semantic representations and can be applied to both English and Hinglish without changing the architecture [3, 4]. Each query q_i and candidate text t_j is embedded as a dense vector of the same dimensionality.

After encoding, topical relevance is estimated with cosine similarity. For a query vector $\boldsymbol{q}$ and a document vector $\boldsymbol{t}$, cosine similarity is computed as:

$$\cos(\boldsymbol{q}, \boldsymbol{t}) = \frac{\boldsymbol{q} \cdot \boldsymbol{t}}{\|\boldsymbol{q}\| \, \|\boldsymbol{t}\|}. \tag{1}$$

A candidate is retained when its similarity score exceeds a permissive threshold of 0.1. This choice favours recall at the first stage: the second stage performs the humour-aware filter, so removing candidates too early would propagate that loss to the final ranking. The retrieved set for query q_i is denoted as $T_{i\text{-}rel}$ and is passed to the second stage together with the original similarity scores.

3.2. Pun Detection Step

The second stage distinguishes humorous wordplay from non-humorous or weakly humorous candidates. For this purpose, we fine-tune `roberta-base` as a binary classifier in cross-encoder fashion [5, 2]. The classifier takes the query and candidate as a single concatenated input sequence, allowing the attention layers to model the interaction between the information need and the candidate text directly.

During training, positive examples are query–document pairs labelled as relevant wordplay instances in `qrels`, while negative examples are non-wordplay candidates drawn from the same collection. We balance the training set by downsampling the negatives to a 1.5:1 ratio against the positives. The classifier is optimised with cross-entropy over two logits, using AdamW with a learning rate of 4×10^{-5}, a batch size of 16, and 10 epochs. The maximum input length is 256 tokens, and a linear warm-up of 10% is applied to the learning rate. The training script monitors the validation F_1 at the end of each epoch and saves the corresponding checkpoint for inference. We note that, in the released code, the validation split coincides with the training split; the reported F_1 should therefore be read as an in-sample indicator rather than a held-out validation estimate.

At inference time, the classifier is applied to the top-3,000 candidates for each query, sorted by cosine similarity. The humour probability produced by the classifier is combined with the retrieval score by a fixed weighted fusion

$$\mathrm{score}(q,t) \;=\; 0.3 \cdot \cos(\boldsymbol{q},\boldsymbol{t}) \;+\; 0.7 \cdot p(\mathrm{humour} \mid q,t), \tag{2}$$

and only candidates classified as humorous are retained. The retained candidates are sorted by the fused score, truncated to the top 1,000 per query, and written to the run file under the identifier `RIVER_task_1_MPNet+RoBERTa`. In this way, the method separates query matching from humour detection while allowing both components to influence the final ranking.

4. Experiments and Evaluation

4.1. Data Description

The CLEF 2026 JOKER Task 1 dataset includes English and Hinglish data. The English dataset was updated for the 2026 edition, and Hinglish is used as the non-English setting in this edition of the lab [1]. The English evaluation set contains 967 queries, 2,008 retrieved relevant documents, and 4,572 relevant documents in total; the ratio of retrieved relevant documents to all relevant documents is therefore $2{,}008/4{,}572 \approx 0.44$. The Hinglish evaluation set contains 464 queries, 1,490 retrieved relevant documents, and 2,910 relevant documents in total, with a retrieved-relevant to relevant ratio of $1{,}490/2{,}910 \approx 0.51$. Table 1 summarises the available evaluation statistics.

Table 1
Evaluation statistics for the two languages.

Language	Num q	Num rel ret	Num rel
English	967	2008	4572
Hinglish	464	1490	2910

The data files follow the official task format and include document identifiers, query identifiers, text fields, and relevance labels. Document identifiers refer to candidate texts, while query identifiers link retrieved documents to specific search queries. The relevance label indicates whether a document is query-relevant and contains wordplay. These fields allow the model to learn both semantic relevance and humour-related classification signals.

4.2. Experimental Setting

In the retrieval stage, we use `paraphrase-multilingual-mpnet-base-v2` to encode both queries and candidate texts [3, 4]. We use the same model for English and Hinglish so that the retrieval component remains language-independent. Documents are encoded with a maximum length of 256 tokens and a batch size of 128; queries are encoded with a maximum length of 128 tokens. Cosine similarity is used as the relevance score, and the retrieval threshold is set to 0.1 to preserve a wide candidate set for the classifier.

In the classification stage, we fine-tune `roberta-base` as a binary cross-encoder classifier [2, 5]. All experiments were conducted on a single NVIDIA A800 GPU with 80 GB of memory. We used the Hugging Face Transformers library for model training and AdamW as the optimiser. The learning rate was set to 4×10^{-5}, no explicit weight decay was applied, and the model was trained for 10 epochs with a batch size of 16. We applied a linear warm-up with a 10% warm-up ratio and set the maximum input sequence length to 256 tokens. At inference, the top 3,000 candidates per query are rescored, combined with the retrieval score by the 0.3/0.7 fusion in Equation 2, filtered to the candidates classified as humorous, and truncated to the top 1,000 per query. The run identifier written to the output file is `RIVER_task_1_MPNet+RoBERTa`.

4.3. Evaluation Metrics

We evaluate the submitted runs using the official retrieval metrics provided for the task [1, 15]. The main metric is mean average precision (MAP), which measures retrieval precision averaged across queries. We also report normalised discounted cumulative gain at different cutoffs, including NDCG@5, NDCG@10, and NDCG@100, to evaluate ranking quality at both shallow and deeper positions. Reciprocal rank measures how early the first relevant item appears in the ranking, and P@10 measures the proportion of relevant items among the top ten retrieved results.

4.4. Experimental Results

This section reports the evaluation results for the English and Hinglish datasets. Table 2 reports the English results, and Table 3 reports the Hinglish results. In both tables, our submitted run is shown alongside selected team runs to contextualise its performance.

Table 2
Performance of selected team runs for English.

Run ID	MAP	NDCG@5	NDCG@10	NDCG@100	recip_rank	P@10	num_q	num_rel_ret	num_rel
RIVER_task_1_MPNet+RoBERTa	**0.3063**	**0.4143**	**0.4129**	0.4426	**0.6015**	0.1656	967	2008	4572
UAds_pt_rm3_CE1K	0.3042	0.3330	0.3868	**0.5041**	0.5211	**0.1943**	971	3477	4572
UAds_pt_bm25_CE1K	0.2996	0.3807	0.3981	0.4446	0.5621	0.1731	971	2846	4572
myteam_BERT	0.2658	0.2556	0.3411	0.4593	0.3879	0.1982	971	3435	4572
LTR_ensemble_rankavg	0.2485	0.2806	0.3140	0.4273	0.4430	0.1440	971	3586	4572
ENSEMBLE_rankavg	0.2323	0.2667	0.2842	0.4036	0.4226	0.1214	971	3586	4572
HGB_D	0.2177	0.2477	0.2673	0.3847	0.3978	0.1184	971	3586	4572
TeamX_Task1_BM25__test	0.1202	0.1065	0.1149	0.2933	0.2242	0.0595	971	3348	4572
Skommarkhos_BM25_E5_MiniLM	0.1087	0.1035	0.1063	0.2715	0.2080	0.0472	967	3554	4572
UAds_pt_rm3	0.1035	0.0913	0.1008	0.2668	0.1869	0.0481	967	3668	4572
UAds_pt_bm25	0.0968	0.0813	0.0849	0.2752	0.1749	0.0407	971	2940	4572
duth_xanthi_pt	0.0813	0.0762	0.0867	0.2212	0.1479	0.0378	971	2172	4572
pjmathematician_Q06-gist32	0.0530	0.0151	0.0324	0.2082	0.0633	0.0231	971	2637	4572

For the English dataset, Table 2 lists our run with a MAP of 0.3063, NDCG@5 of 0.4143, NDCG@10 of 0.4129, and reciprocal rank of 0.6015. The NDCG@10 margin over the next entry in the table is 0.01 in absolute terms; the NDCG@100 score of 0.4426 is exceeded by three other pipelines that retrieve a larger candidate pool (more than 2,800 relevant documents, against our 2,008), indicating that deeper-rank coverage is a current limitation of the first stage. The P@10 score of 0.1656 is exceeded by three other runs in the table; the next P@10 entry is reported by the `UAds_pt_rm3_CE1K` pipeline, which combines BM25 with RM3 query expansion and a 1K cross-encoder reranker, at 0.1943. Together these observations suggest that the current 0.3/0.7 fusion yields good ranking quality across the list but does not yet concentrate relevant humorous items densely enough in the top ten.

Table 3
Performance of selected team runs for Hinglish.

Run ID	MAP	NDCG@5	NDCG@10	NDCG@100	recip_rank	P@10	num_q	num_rel_ret	num_rel
RIVER_task_1_MPNet+RoBERTa	**0.3897**	**0.5395**	**0.5199**	0.5446	**0.8382**	0.2735	464	1490	2910
myteam_BERT	0.3830	0.3959	0.4523	**0.6113**	0.6596	**0.2804**	485	2493	2910
UAds_pt_rm3	0.3707	0.3811	0.4380	0.6010	0.6383	0.2744	485	2493	2910
UAds_pt_bm25	0.3516	0.3599	0.4180	0.5854	0.6101	0.2678	485	2493	2910
results_pt_finetuned	0.3506	0.3582	0.4173	0.5849	0.6102	0.2676	485	2493	2910
pjmathematician_Q06-gist	0.3372	0.4487	0.4616	0.5091	0.7607	0.2565	485	1542	2910
pjmathematician_Q06-gist32	0.3148	0.4187	0.4342	0.4840	0.7135	0.2449	485	1501	2910
duth_xanthi_pt	0.3047	0.2989	0.3635	0.5416	0.4934	0.2503	485	2493	2910
Skommarkhos_BM25_E5_MiniLM	0.2183	0.2850	0.2881	0.4501	0.5541	0.1575	485	2405	2910
UAds_pt_rm3_CE1K	0.0876	0.1783	0.1591	0.1591	0.5258	0.0526	485	255	2910

For the Hinglish dataset, Table 3 lists our run with a MAP of 0.3897, NDCG@5 of 0.5395, NDCG@10 of 0.5199, and reciprocal rank of 0.8382. The NDCG@10 margin over the next entry in the table is 0.06 in absolute terms, and the reciprocal-rank margin is 0.08. The P@10 score of 0.2735 is exceeded by only the myteam_BERT run, which reports 0.2804. The cross-encoder reranking pipeline that performed well on English (UAds_pt_rm3_CE1K) reports a MAP of 0.0876 on Hinglish, indicating that its reranker is not language-portable. Our pipeline does not suffer from this issue because the multilingual retriever is shared between the two languages and the classifier is fine-tuned per language.

Looking across the two languages, our pipeline shows a wider performance margin on Hinglish than on English. We attribute this to the multilingual retriever transferring well across the language boundary and to the per-language fine-tuning of the classifier, although a controlled ablation is needed to separate these two factors. The NDCG@100 gap on both languages is small, while the P@10 gap is a concrete weakness of the current pipeline. Future improvements should therefore focus on reranking and candidate-filtering strategies that increase the density of relevant humorous texts in the top ten positions, rather than on changes to the multilingual retriever.

5. Conclusion

This paper has presented a two-stage approach for the CLEF 2026 JOKER Task 1 on Humour-aware Information Retrieval. The method combines multilingual semantic retrieval with relevance-aware pun classification. The retrieval stage uses `paraphrase-multilingual-mpnet-base-v2` to identify query-related candidates, while the classification stage uses a fine-tuned `roberta-base` cross-encoder to filter candidates according to humour and wordplay evidence. The two scores are combined by a fixed 0.3/0.7 fusion, and only candidates classified as humorous are retained in the run.

Experimental results on English and Hinglish show that the proposed approach is practical and adaptable for multilingual humour-aware retrieval. The method reports a MAP of 0.3063 on English and 0.3897 on Hinglish among the runs listed in Tables 2 and 3, using the same retriever checkpoint and language-specific classifier fine-tuning. The larger margin observed on Hinglish, where our run exceeds the next entry in Table 3 by 0.06 in absolute NDCG@10, provides empirical support for separating semantic relevance from humour detection as a design principle for this task.

Future work will focus on improving top-ranked precision, motivated by the observed gap in P@10. Promising directions include: calibrating the 0.3/0.7 fusion weight on a held-out subset of the labelled data; replacing the multilingual retriever with BM25+RM3 or a hybrid first stage to improve candidate density in the top ten positions; and augmenting the positive training set with paraphrases of labelled wordplay instances to increase its diversity.

Acknowledgments

We thank the CLEF 2026 JOKER organisers, the `sentence-transformers` team, the `RoBERTa` team, and the `Hugging Face Transformers` developers for releasing the resources used in this work.

Declaration on Generative AI

During the preparation of this work, the author used GitHub Copilot Chat (Anthropic Claude family)in order to: Grammar and spelling check, paraphrase and reword, and code refactoring. After usingthese tool(s), the author reviewed and edited the content as needed and takes full responsibility for thepublication's content.

References

[1] Liana Ermakova, Igor Kuzmin, Poojan Vachharajani, Tristan Miller, Anne-Gwenn Bosser, Jaap Kamps. CLEF 2026 JOKER Track: Humour Detection, Search, and Translation. In: R. Campos *et al.* (Eds.), *Advances in Information Retrieval: 48th European Conference on Information Retrieval (ECIR 2026), Delft, The Netherlands, March 29–April 2, Proceedings, Part IV*, Lecture Notes in Computer Science, vol. 16486, Springer, Cham, 2026, pp. 242–250. doi:10.1007/978-3-032-21321-1_34.

[2] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692 (2019).

[3] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3982–3992.

[4] Nils Reimers and Iryna Gurevych. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. arXiv preprint arXiv:2004.09813 (2020).

[5] Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. Multi-Stage Document Ranking with BERT. CoRR abs/1910.14424 (2019). arXiv:1910.14424.

[6] 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. Springer Nature Switzerland, Cham, 2023, pp. 397–415.

[7] Q. Dubreuil. UBO Team @ CLEF JOKER 2023 Track For Task 1, 2 and 3 – Applying AI Models In Regards To Pun Translation. 2023. URL: https://ceur-ws.org/Vol-3497/paper-155.pdf.

[8] Brandon M. Savage, Heidi L. Lujan, Raghavendar R. Thipparthi, and Stephen E. DiCarlo. Humor, laughter, learning, and health! A brief review. Advances in Physiology Education 41(3) (2017) 341–347. doi: https://doi.org/10.1152/advan.00030.2017.

[9] L. Ren, B. Xu, H. Lin, L. Yang. ABML: attention-based multi-task learning for jointly humor recognition and pun detection. Soft Computing 25 (2021) 14109–14118. doi: https://doi.org/10.1007/s00500-021-06136-y.

[10] Y. Zou, W. Lu. Joint detection and location of English puns. CoRR abs/1909.00175 (2019). URL: http://arxiv.org/abs/1909.00175. arXiv:1909.00175.

[11] Arina Gepalova, Adrian-Gabriel Chifu, and Sébastien Fournier. CLEF 2024 Joker Task 1: Exploring Pun Detection Using The T5 Transformer Model. In: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org, 2024.

[12] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, Ves Stoyanov. Unsupervised

Cross-lingual Representation Learning at Scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 8440–8451.
[13] Bhaskar Mitra and Nick Craswell. An Introduction to Neural Information Retrieval. Foundations and Trends® in Information Retrieval 13(1) (2018) 1–126.
[14] V. Lavrenko and W. B. Croft. Relevance-Based Language Models. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001), pp. 120–127, 2001.
[15] Chris Buckley and Ellen M. Voorhees. Retrieval System Evaluation. In: TREC: Experiment and Evaluation in Information Retrieval. MIT Press, 2005.