CLEF 2025 JOKER Track: Enhancing Humor-Aware Information Retrieval with Relevance-Aware Classification

Notebook for the Joker Lab at CLEF 2025

Bo Chen, ChangLe Zhong and LeiLei Kong*

[1]*Foshan University, Foshan, Guangdong, China*

Abstract

This study investigates the humor-aware information retrieval, within the JOKER Lab at CLEF 2025. This task involves retrieving short humorous texts from a document collection and subsequently identifying texts containing puns. A Enhancing Humor-Aware Information Retrieval with Relevance-Aware Classification approach is introduced, utilizing the retrieval model and the classification model to fulfill task objectives with each model adopting its unique methodology and cooperating to enhance overall performance. This collaborative strategy not only achieves the primary task objectives but also diminishes the task's complexity, rendering it more feasible to implement. Experimental results demonstrate that our method achieved second place ranked by MAP (0.16) on the English dataset, with P@5 at 0.36 and NDCG@5 at 0.41. On the Portuguese dataset, it attained first place in both P@5 (0.44) and NDCG@5 (0.83), while achieving second place in MAP (0.40).. These findings indicate the potential effectiveness of the proposed method. Nevertheless, significant scope for improvement remains, warranting further exploration and research to enhance the overall methodology.

Keywords

pun detection, humor-aware information retrieval, Relevance-Aware Classification,

1. Introduction

Humor serves as a fundamental mechanism for relationship building and has demonstrably positive effects on performance and motivation [3]. Within Natural Language Processing (NLP), the automatic analysis of humor, with a specific focus on detecting puns and wordplay, represents a significant research challenge. Puns, defined as a form of humor relying on linguistic ambiguity, introduce substantial complexity into natural language understanding due to their inherent reliance on dual meanings. Specifically, puns exploit polysemy (multiple meanings of a single word) or homophony/homonymy (similar sounds between different words) to generate humorous or witty effects, often employing double entendre. Illustrative examples include word pairs such as "profit" / "prophet" or "check" / "Czech".

Our work addresses the first task of the JOKER Lab shared Task 1:Humour-aware Information Retrieval, which requires retrieving short humorous texts exhibiting wordplay from a document collection, ensuring relevance to a given query. Inspired by the objectives of relevance and humor detection, we propose a Enhancing Humor-Aware Information Retrieval with Relevance-Aware Classification method, aiming to be effective for both English and Portuguese. Our methodology leverages two distinct pre-trained models: one fine-tuned RoBERTa-base model for text classification (pun detection) and another (the paraphrase-multilingual-mpnet-base- v2 model) for semantic similarity calculation (query relevance). The first model, which calculates similarity, processes the data to produce a training set that is subsequently passed to the second model, a classification model, for fine-tuning. Meanwhile, a separate dataset is reserved for testing and will be passed to the second model once its fine-tuning is completed. Ultimately, these two models work in tandem to achieve the task. According to the final experimental results, the method can achieve certain effects on both languages. Overall, the method is more effective for Portuguese.

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

Corresponding author.

rasion086@gmail.com (B. Chen); clzhong@fosu.edu.cn (C. Zhong); kongleilei@fosu.edu.cn (L. Kong)

0009-0009-6504-6169 (B. Chen); 0009-0008-3044-2383 (C. Zhong); 0000-0002-4636-3507 (L. Kong)

The paper is structured as follows: Section 2 reviews recent advancements in pun detection methodologies. Section 3 details our proposed approach. Section 4 presents and discusses the experimental results. Concluding remarks and future directions are provided in Section 5.

2. Related Work

Pun detection has been a topic of computational linguistics research for many years[10]. Recent advancements have primarily leveraged deep learning techniques. A dominant approach frames pun detection and localization as a sequence labeling task. Notably, BiLSTM-CRF models have been effectively applied in this context. These models typically employ specialized annotation schemes, such as NP (Normal Phrase) and BPA (Bi-directional Pun Annotation), where the BPA scheme is particularly designed to capture structural constraints by ensuring at most one word per context is labeled as a pun[6, 7].

Within this BiLSTM-CRF framework, a bidirectional long short-term memory network (BiLSTM) learns contextual information from the input sequence, while a conditional random field (CRF) layer handles sequence label prediction and models dependencies between labels. To enhance performance, these models often incorporate diverse input features. These include pre-trained word embeddings, character-level features processed through character-level LSTMs and highway networks, and positional features indicating word locations[6, 7]. By treating both detection (identifying the presence of a pun) and localization (identifying the pun word) as a unified sequence labeling problem, this approach learns to perform both tasks concurrently during training, outputting label probabilities for each word and decoding the final sequence via the CRF layer[6, 7].

More recently, the advent of large language models (LLMs) has opened new avenues for pun detection, shifting focus towards leveraging their inherent semantic understanding and generation capabilities. Initial explorations of LLMs for this task, such as using ChatGPT and the SimpleT5 model, were conducted during CLEF 2023, demonstrating promising results in understanding and identifying puns in texts[7, 8]. Building on this potential, subsequent research has adapted powerful sequence-to-sequence transformers like T5 specifically for pun detection[4].

The T5-based methodology involves several key steps[4]. First, queries are augmented by sourcing synonyms for query terms using resources like WordNet to improve document matching potential. The augmented queries and document texts are then transformed into embedding vectors using efficient tokenizers (e.g., all-MiniLM-L6-v2). Similarity scores between query and document embeddings are computed, and a threshold (e.g., 0.35) is applied to filter relevant documents. Finally, a T5 model (e.g., flan-T5-base) is fine-tuned on task-specific training data to detect puns within the filtered documents[4]. This approach highlights the trend of utilizing pre-trained LLMs and fine-tuning them for the specific nuances of pun identification.

3. Our Method

Our task aims to extract all relevant and humorous texts from the given text dataset D based on a query q_i. To achieve this goal, we propose a Enhancing Humor-Aware Information Retrieval with Relevance-Aware Classification method, which consists of two main steps. The first step involves constructing a simple, fast, and efficient retrieval model. The purpose of this model is to address the relevance requirement of the task, without considering whether the text is humorous. Through this step, we can obtain the set of humor source texts $T_{i\text{-}rel}$ related to each query q_i. The second step involves building a binary classification model to further screen the humor source texts $T_{i\text{-}rel}$ related to each query q_i obtained from the first step and merge all texts, ultimately yielding the set of relevant humorous texts R for all queries q_i.Overall process is shown in Figure 1:

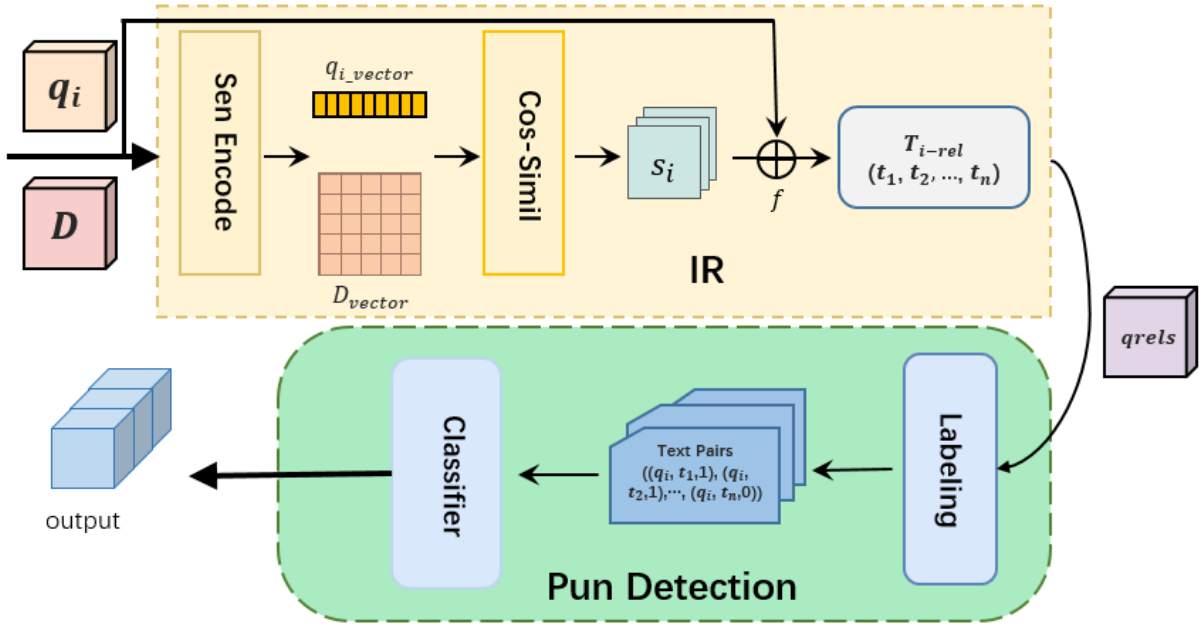

Figure 1: The entire processing flow of the method.

3.1. Information Retrieval Step

Text retrieval is the process of finding documents relevant to a user's query from large collections. We employs a vector space model for relevant document retrieval. Within this model, documents and queries are treated as points or vectors in a vector space. By calculating the similarity between a document vector and a query vector, the relevance of the document to the query can be determined. Similarity measures, such as cosine similarity, are used for this calculation. This approach effectively transforms the text retrieval problem into a geometric problem within a vector space, enabling efficient document retrieval.

In the construction of this vector model, we utilize a pre-trained model called "paraphrase-multilingual-mpnet-base-v2" as the encoding model. This model encodes each query q_i and the entire text dataset D, generating vectors for each query q_i and text t_j. We then compute the similarity score s between q_i and t_j using the cosine similarity method (Equation 1). If the similarity score s exceeds the threshold f, we deem q_i and t_j to be similar, and consequently, add t_j to the humor source text set $T_{i\text{-}rel}$.

$$s = \cos(\theta) = \frac{q_i \cdot t_j}{\|q_i\| \|t_j\|} \tag{1}$$

Parameters:
q_i: Query vector corresponding to the query q_i.
t_j: Text vector corresponding to the text t_j.
$q_i \cdot t_j$: Dot product of vectors q_i and t_j.
$\|q_i\|$: Euclidean norm (magnitude) of vector q_i.
$\|t_j\|$: Euclidean norm (magnitude) of vector t_j.
s: Cosine similarity score, which ranges from -1 to 1. A score of 1 indicates identical direction, -1 indicates opposite direction, and 0 indicates orthogonality.

3.2. Pun Detection Step

We formalize the identification of desired relevant humorous texts as a binary classification task: given a pair of texts (q_i, t_j), a classification model determines whether they represent a relevant humorous text pair.

In machine learning, classification is a supervised learning technique aimed at assigning input data to predefined categories or classes. It involves training a model on a labeled dataset, where each instance is associated with a specific class label. Given that the official data lacks directly suitable fine-tuning training data, we first construct a training set D_{train} using data from the retrieval phase.

In the initial step, we acquire the humorous source text $T_{i\text{-}rel}$ for each query q_i. Subsequently, leveraging the "query," "docid," "qrel," and other details from dataset Q, we match the corresponding q_i and t_j, and label the text pairs according to "qrel". The first-stage retrieval returned numerous irrelevant documents. Consequently, annotating this data to build the training set yielded a large number of negative samples (label 0), creating dataset imbalance. To mitigate this, we employed a method where we randomly select negative samples from a large pool, with a quantity 50% higher than the number of positive samples, to be used together with the positive samples as training data.

In the construction of the binary classification model, akin to the first step, we employ the pre-trained "xlm-roberta-base" model as the foundation. We fine-tune this model using the cross-entropy function (Equation 2) as the objective loss function. Throughout the training process, we retain the model iteration that achieves the highest F1 score. This top-performing model is then applied to the final humorous text recognition task, resulting in the desired relevant humorous texts R.

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i)] \tag{2}$$

Parameters:
$\mathcal{L}$: Cross-entropy loss.
N: Number of training samples.
y_i: True label of the i-th sample.
$\hat{y}_i$: Predicted probability of the i-th sample.

4. Experiments And Evaluation

4.1. Data Description

This year's dataset not only includes English data but also adds Portuguese data, and the English dataset has more content than last year's.tent. For the English dataset, there are a total of 77,658 corpus texts, with 5,198 being humorous texts and the rest being non-humorous texts but related to certain queries. There are 12 queries with labeled information that can be used for model training, and 219 queries that we ultimately need to use for testing the output results.In the Portuguese dataset, the quantity is significantly smaller than that of the English dataset. There are only 45,126 text entries, with 1,199 being humorous texts. Additionally, there are 98 queries used for generating output results and 29 queries used to assist in model training. The distribution of data across both languages is shown in Table 1.

Table 1
The distribution of the two types of data

Language	Num corpus	Num qrels	Num train-query	Num test-query
English	77658	660	12	219
Portuguese	45126	1894	29	98

Furthermore, apart from the differences in quantity, the composition structure and information of the data are consistent. All data files are individual JSON files, and the key ID information is explained as follows[1]:

- **docid**:a unique document identifier
- **text**: the text of the instance, which may or may not contain wordplay
- **qid**:a unique query identifier from the input file

- **query**:the search query
- **qrel**:indication the document docid is relevant to the query qid and is a wordplay instance

4.2. Experimental Setting

In the first step, we selected the multilingual pre-trained model paraphrase-multilingual-mpnet-base-v2 as our encoding model for information retrieval. This model is capable of handling both English and Portuguese simultaneously, thereby reducing the time costs associated with switching between different models. To maximize the retrieval of relevant texts, we employed a simple static threshold method, setting the threshold f to 0. In the second step, we continued with this approach, choosing an appropriate multilingual classification pre-trained model, specifically xlm-roberta-base. All experiments were conducted on a single NVIDIA A800 GPU (80GB memory). We utilized the HuggingFace transformers library for model training, with AdamW as the optimizer. The learning rate was set to 2e-5, and the weight decay was 0.1. Training was carried out for 8 epochs with a batch size of 16, and a linear warmup strategy was applied with 10% warmup ratio. The maximum input sequence length was set to 128.

- **paraphrase-multilingual-mpnet-base-v2**:The "Paraphrase-Multilingual-MPNet-Base-V2" model is a powerful tool for generating paraphrases across multiple languages. It's built on the MPNet architecture, which combines masked and permuted language modeling to capture context effectively. The model is trained on multilingual data, allowing it to handle text in various languages with high accuracy. It is an improved version (V2) of an earlier model, offering better performance and more reliable paraphrasing.[9]
- **xlm-roberta-base**:The "XLM-RoBERTa-Base" model is a multilingual version of the RoBERTa architecture, designed to handle text in multiple languages. It is trained on a massive multilingual corpus, enabling it to understand and generate text across different languages effectively. This model uses byte-level BPE tokenization, which helps in better handling of out-of-vocabulary words and diverse scripts. As a "base" model, it offers a good balance between computational efficiency and performance, making it suitable for a wide range of multilingual NLP tasks.

4.3. Evaluation Metrics

To directly evaluate the method proposed in the paper, we adopt the official evaluation metrics for assessment, focusing on extracting a subset of these metrics to showcase its performance. [1]

- **map**:mean average precision – i.e., the mean of the average precision scores for each query
- **ndcg**:normalised discounted cumulative gain, the gain of each document based on its relevance,discounted logarithmically by its position in the ranking normalised over the ideal ranking
- **bpref**:binary preference, a sum-based metric showing how many relevant documents are ranked before irrelevant documents
- **MRR**:mean reciprocal rank, the average of the multiplicative inverse of the ranks of the first correct answer of results for a sample of queries
- **P5, P10, P15**:precision – i.e. the ability of a system to present only relevant items, at different numbers of top ranked results

4.4. Experimental Results

This section primarily presents the evaluation metrics of our method across two languages and compares it with methods proposed by other teams. Table 2 shows the evaluation results for English, while Table 3 displays the evaluation results for Portuguese.

In the Table 2, our approach demonstrated both strengths and challenges. Compared to the "Cryptpix_SBERT" team, it showed a slight advantage in Rprec (Recall Precision), with scores of 0.19 versus 0.15. This indicates a relative strength in retrieving relevant documents within the English context. However, compared to "UAds_team_3", our method was at a disadvantage across key metrics including

Table 2
Performance reports of some teams for EN.

Run ID	map	Rprec	P@5	P@10	P@100	NDCG@5
pjmathematician_Q14Q8Q32	0.35	0.39	0.55	0.41	0.09	0.61
UAms_RM3RoBERTa_drop60	0.17	0.23	0.31	0.23	0.06	0.02
Rasion_SenTransf+Roberta	0.16	0.21	0.36	0.24	0.04	0.41
Cryptix_SBERT	0.15	0.19	0.29	0.21	0.05	0.33
UAms_RM3	0.15	0.20	0.24	0.20	0.06	0.26
UAms_Anserini_rm3	0.08	0.27	0.50	0.09	0.11	0.06
CCC_Entity_RoBERTa_RM3	0.14	0.16	0.17	0.18	0.07	0.18
UAms_en_bm25	0.12	0.12	0.13	0.13	0.06	0.14
CCC_TFIDF_Rerank	0.11	0.15	0.21	0.17	0.06	0.22
CCC_Ensemble_COIBERT_RM3	0.05	0.07	0.08	0.07	0.04	0.09
Skommarkhos_BM25_E5_MiniLM	0.05	0.03	0.01	0.03	0.04	0.01

map (mean average precision), P@5, P@10, P@100, and NDCG@5. For instance, map was 0.16 vs. 0.13 and P@5 was 0.21 vs. 0.24. This highlights significant room for improvement in overall retrieval precision, the accuracy of the top 5, 10, and 100 results, and ranking quality. Furthermore, compared to "UAds_RM3", while our approach held a marginal edge in P@5 (0.21 vs. 0.20), performance was generally comparable across other metrics: map (0.16 vs. 0.15), P@10 (0.26 vs. 0.24), P@100 (0.36 vs. 0.40), and NDCG@5 (0.41 vs. 0.43). Nevertheless, "UAds_RM3" performed slightly better in the more critical areas of accuracy for the top 100 results and ranking quality for the top 5 results.

Table 3
Performance reports of some teams for PT

Run ID	map	Rprec	P@5	P@10	P@100	NDCG@5	bpref
pjmathematician_Q14-Q4-R	0.42	0.41	0.44	0.34	0.09	0.51	0.58
Rasion_SenTransF+Roberta	0.40	0.41	0.44	0.38	0.08	0.51	0.83
UAds_pt_bm25	0.08	0.06	0.05	0.06	0.03	0.06	0.11
Skommarkhos_BM25_E5_MiniLM	0.07	0.05	0.06	0.06	0.03	0.07	0.08
pjmathematician_Q06-gist	0.07	0.04	0.05	0.06	0.03	0.08	0.07
Skommarkhos_BM25_E5_MiniLM	0.07	0.05	0.05	0.06	0.03	0.06	0.07
results_pt_pt_finetuned	0.07	0.06	0.06	0.07	0.03	0.07	0.32
UAds_pt_rm3	0.07	0.05	0.09	0.05	0.03	0.04	0.10
myteam_BERT	0.06	0.06	0.08	0.06	0.03	0.08	0.07
duth_xanthi_pt	0.06	0.05	0.07	0.08	0.03	0.07	0.15
pjmathematician_Q06-gist32	0.04	0.02	0.07	0.03	0.02	0.07	0.04
UAds_pt_rm3_CE1K	0.04	0.02	0.05	0.03	0.02	0.07	0.04
UAds_pt_bm25_CE1K	0.04	0.02	0.04	0.01	0.02	0.04	0.04

Transitioning to the Portuguese-language environment, our approach performed notably well. It achieved an Rprec score of 0.41, second only to "pjmathematician_Q14-Q-R"'s 0.42. This underscores its strong capability for retrieving relevant documents in Portuguese. However, the map score was 0.40, slightly lower than "pjmathematician_Q14-Q-R"'s 0.42, indicating potential for improvement in overall retrieval precision. Performance on P@5, P@10, and P@100 was largely on par with "pjmathematician_Q14-Q-R". Yet, it fell short in NDCG@5, suggesting room for enhancing ranking quality.

Comparing performance across the two languages reveals that our approach performs better overall in the Portuguese environment. In Portuguese, it achieved relatively high scores on key metrics like map, Rprec, and P@5. Its Rprec score was particularly close to the leading team, reflecting better model adaptation to Portuguese datasets and more effective handling of Portuguese text features and semantic information. In contrast, performance metrics in the English environment were generally lower. The comparative disadvantage against teams like "UAds_team_3" and "UAds_RM3" suggests potential

shortcomings in handling English vocabulary, the depth of semantic understanding, or optimization on English training data. Further refinement is needed to enhance retrieval performance in English.

5. Conclusion

This study introduces a Enhancing Humor-Aware Information Retrieval with Relevance-Aware Classification approach, conceptually inspired by the inherent duality of the task requirements. Crucially, these distinct requirements are addressed separately and processed independently through dedicated screening modules, prior to their synergistic integration for generating the final output.

However, the observed suboptimal performance metrics, specifically low precision and recall, highlight significant areas necessitating improvement. While leveraging state-of-the-art pre-trained models followed by appropriate fine-tuning undoubtedly yields substantial performance gains, our investigation reveals that the method encounters persistent and significant challenges primarily within the data processing pipeline. Given the paramount importance of data quality and quantity for model efficacy, future research efforts should prioritize exploring effective strategies for data enhancement under resource-constrained conditions. This includes investigating techniques such as targeted data augmentation, intelligent sampling, synthetic data generation, or advanced data cleaning methodologies to mitigate data-related bottlenecks and unlock further performance potential.

Acknowledgments

This work is supported by the Quality Engineering Projects for Teaching Quality and Teaching Reform in Undergraduate Colleges and Universities of Guangdong Province (No.xxx).

Declaration on Generative AI

During the preparation of this work, the author(s) used DeepSeek, Kimi in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] Liana Ermakova, Ricardo Campos, Anne-Gwenn Bosser, Tristan Miller (2025). Overview of CLEF 2025 Joker Task 1: Humour-aware Information Retrieval. In: Guglielmo Faggioli, Nicola Ferro, Paolo Rosso, Damiano Spina (Eds). 2025. Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025). CEUR Workshop Proceedings, CEUR-WS.org.

[2] Liana Ermakova, Ricardo Campos, Anne-Gwenn Bosser, and Tristan Miller. Overview of the CLEF 2025 JOKER Lab: Humour in Machine. In Jorge Carrillo-de-Albornoz, Julio Gonzalo, Laura Plaza, Alba García Seco de Herrera, Josiane Mothe, Florina Piroi, Paolo Rosso, Damiano Spina, Guglielmo Faggioli, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025)*, 2025.

[3] B. M. Savage, H. L. Lujan, R. R. Thipparthi, S. E. DiCarlo. Humour, laughter, learning, and health! a brief review. Advances in Physiology Education (2017).

[4] Arina Gepalova, Adrian-Gabriel Chifu, and Sébastien Fournier. CLEF 2024 Joker Task 1: Exploring Pun Detection Using The T5 Transformer Model. Laboratoire d'Informatique et des Systèmes (LIS) UMR 7020, Aix-Marseille Université, Université de Toulon, CNRS, LIS, France.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Wennie Yang, Jieyu Zhao, Guangxuan Wang, Peizhen Lou, Nianzu Zhang, Sinno Jialin Pan. Measuring and Reducing Gendered Correlations in Pre-trained Models. In: Proceedings of the 2023 Conference on Fairness, Accountability, and Transparency (FAT* '23). New York, NY, USA: Association for Computing Machinery, 2023, pp. 123–134.

[10] Liana Ermakova et al. 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.

[11] 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).