Overview of JOKER 2023 Automatic Wordplay Analysis Task 1 – Pun Detection

Liana Ermakova[1,*], Tristan Miller[2], Anne-Gwenn Bosser[3], Victor Manuel Palma Preciado[1,4], Grigori Sidorov[4] and Adam Jatowt[5]

[1]*Université de Bretagne Occidentale, HCTI, France*
[2]*Austrian Research Institute for Artificial Intelligence (OFAI), Vienna, Austria*
[3]*École Nationale d'Ingénieurs de Brest, Lab-STICC CNRS UMR 6285, France*
[4]*Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Mexico City, Mexico*
[5]*University of Innsbruck, Austria*

Abstract

This paper presents details of Task 1 of the JOKER-2023 Track, which aims to detect sentences in English, French, and Spanish that contain wordplay. With applications in humour generation, sentiment analysis, conversational agents, content filtering, and linguistic creativity, this task is still challenging despite significant recent progress in information retrieval and natural language processing. Building on the lessons learned from last year's edition of the JOKER track, our overall goal is to foster progress in the automatic interpretation, generation, and translation of wordplay in English, Spanish, and French. In this paper, we define our task and describe our approaches to corpus creation and evaluation in the three languages. We then present an overview of the participating systems, including summaries of their approaches and a comparison of their performance.

Keywords

wordplay, puns, computational humour, wordplay detection, text classification, corpus, data collection,

1. Introduction

This paper presents details of Task 1 of the JOKER-2023 Track[1], which was held as part of the 14th Conference and Labs of the Evaluation Forum (CLEF 2023)[2]. The JOKER-2023 track is an evaluation lab aiming to foster progress in the automatic interpretation, generation, and translation of wordplay in English, Spanish, and French. It follows the JOKER-2022 workshop [1]. JOKER-2023 Task 1 aims to detect sentences in English, French and Spanish that contain a pun, a specific form of humorous wordplay. For details on the track's other two tasks, please refer to their respective overview [2, 3]; further information and insights are also presented in the Track overview paper [4].

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
*Corresponding author.
0000-0002-7598-7474 (L. Ermakova); 0000-0002-0749-1100 (T. Miller); 0000-0002-0442-2660 (A. Bosser); 0000-0001-8711-1106 (V. M. Palma Preciado); 0000-0003-3901-3522 (G. Sidorov); 0000-0001-7235-0665 (A. Jatowt)

[1]https://www.joker-project.com
[2]https://clef2023.clef-initiative.eu/

A *pun* is a form of wordplay in which a word or phrase evokes the meaning of another word or phrase with a similar or identical pronunciation [5]. Pun detection contributes to the field of natural language processing by improving language understanding capabilities. Recognizing and identifying puns requires understanding the multiple meanings or wordplay involved, which helps in building more sophisticated language models and algorithms. By identifying puns in text, researchers can study humour patterns, comedic structures, and linguistic devices. This can lead to advancements in humour generation, joke understanding, and humour-based applications. Pun detection can be valuable for applications such as sentiment analysis, opinion mining, conversational agents, and content filtering, where recognizing humour is essential for accurate interpretation and analysis. Detecting puns helps linguists and cultural researchers study wordplay, linguistic creativity, and the use of humour across different languages and cultural contexts. It provides insights into how language is manipulated for humorous effect and can contribute to cross-cultural understanding and communication. However, pun detection is still a challenging task.

More formally, *pun detection* is a binary classification task where the goal is to distinguish between texts containing a pun (the *positive examples*) and texts not containing any pun (the *negative examples*) [6]. Performance on this task is evaluated using the standard precision, recall, accuracy, and F-score metrics from text classification and information retrieval [7, Ch. 8.3].

Humour is a component of our social coexistence and, by extension, interpersonal interactions, and is characterized by its ambiguous nature and dependence on a number of subjective factors. Coping with humour, even in its written form, becomes a relatively complicated task. Still, there are a variety of studies on the topic within computer science, including humour translation and determining whether the humorous intent or interpretability is maintained, humour detection, and humour classification.

When it comes to written humour, various computational tasks have been proposed, which help to gain a broader understanding of its structure, aspects, and nuances. Recently transformers, including BERT-like models, have been reported to achieve good results on humour detection [8, 9]; this also holds for Bi-LSTMs, CNNs, and RNNs [10, 11]. However, as we showed previously, these results might be explained by the corpus heterogeneity rather than the capacity of the models to recognize wordplay [12].

In the remainder of this paper, we present an overview of our data preparation process (Section 2), describe our participants' runs (Section 3), and present the analysis of their results (Section 4). Section 5 concludes the paper.

2. Dataset

The English- and French-language data used for our tasks is described in detail in our resource paper, published at SIGIR 2023 [12]. Here, we provide a concise overview of the overall data collection procedure. For Task 1, the data consists of positive and negative texts. The positive examples are all short jokes (one-liners), each containing a single pun. In contrast to previously published punning datasets, our negative examples have been generated by data augmentation techniques that involve manually or semi-automatically modifying positive examples to remove the pun while retaining the majority of the meaning. In each positive text, we generally replaced

Table 1
Task 1 data statistics

	Train			Test		
Language	Positive	Negative	Total	Positive	Negative	Total
English	3,085	2,207	5,292	809	2,374	3,183
French	1,998	2,001	3,999	5,308	7,565	12,873
Spanish	855	1,139	1,994	952	1,289	2,241

a single word, which may or may not have been the pun. We employed this strategy to minimize the differences in length, vocabulary, style, etc. that existed between the positive and negative subsets of previous pun detection datasets and on which classifiers could, inadvertently or otherwise, distinguish those subsets. For the French subcorpus, additional negative examples were derived from the English positive examples via machine translation, a procedure in which ambiguity is almost always lost.

The Spanish data was collected mainly using two methodologies. The initial step consisted of scraping a manually seeded set of web pages known to collect jokes, followed by manually removing non-puns and other inappropriate texts. Our data source was Twitter, from which we extracted approximately 195K tweets containing the hashtags `#humor`, `#juegodepalabras`, and `#chiste` (meaning "humour", "pun", and "joke", respectively). Similarly, we manually filtered out examples that lacked puns or contained irrelevant information (images, URLs, emoticons, additional hashtags, etc.). We were able to compile approximately one thousand examples, approximately a quarter of which came from web pages and the rest from Twitter. Then, negative examples were generated with essentially the same data augmentation technique as the English and French data.

Each language's data was divided into test and training sets and provided to Task 1 participants in simple JSON and delimited text formats with fields containing a unique ID, the text to be classified, and (for training data) a boolean value signifying whether or not the text contains a pun. Participants could select the language(s) for which they would submit classification runs. The expected output format was a similarly straightforward, delimited text file containing columns for the run ID, the text ID, the boolean classification result, and a boolean indicator indicating whether the classification was performed manually or automatically.

The statistics for the Task 1 data are presented in Table 1. These statistics report the numbers actually used for the evaluation (the effective numbers for test and training data). However, in the files shared with the participants we included the training data in the input of the test file.

The inclusion of the training data within the dataset for which participants had to make predictions allows for a comprehensive evaluation of system performance on both the training and testing data. By incorporating the training data, it becomes possible to assess how well the systems generalize to unseen test data based on their performance on familiar training examples.

Input format. The train and the test data are provided in JSON and CSV formats with the following fields:

id a unique identifier

text the text of the instance, which may or may not contain wordplay

Input example:

```
[{"id":"en_135",
"text":"Cleopatra was the Pharaohs one of all."},

{"id":"en_7942",
"text":"Strike while the iron is hot."}]
```

Qrels. We provide training data in the format of JSON or TSV qrels files with the following fields:

id a unique identifier from the input file

wordplay yes/no

Example of a qrel file:

```
[{"id":"en_135",
"wordplay":"yes"},
{"id":"en_7942",
"wordplay":"no"}]
```

Output format. Results should be provided in a TREC style's JSON or TSV 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

id a unique identifier from the input file

wordplay yes/no

Example of output format:

```
[{"run_id":"team1_task_1_TFIDF",
"manual":0,
"id":"en_135",
"wordplay":"yes"},
{"run_id":"team1_task_1_TFIDF",
"manual":0,
"id":"en_7942",
"wordplay":"no"}]
```

Table 2
Statistics on the runs submitted for Task 1: Pun detection

Team	EN	FR	ES	Total
Croland	1	1	1	3
LJGG	3	3	3	9
Les_miserables	3	3	3	9
MiCroGerk	6	—	—	6
Smroltra	7	7	7	21
TeamCAU	6	—	—	6
TheLangVerse	1	—	—	1
ThePunDetectives	6	—	—	6
UBO	1	1	1	3
AKRaNLU	2	2	2	6
Innsbruck	3	—	—	3
NPalma	1	—	1	2
Total	40	17	18	75

3. Participants' Approaches

Twelve teams submitted 75 runs in total (see Table 2). More than half of the submissions (40) were done for English. The approaches used by the twelve teams are as follows:

1. The AKRaNLU team [13] described two methods for pun detection based on sentence embeddings with a binary classifier and sequence classification with XLM-Roberta. They utilized a six-layer neural network with a classifier head for each of the three languages.
2. The NLPalma [14] team experimented with models based on the BERT architecture designed for multiple languages. The authors concluded that this approach is promising, but suggested that additional model fine-tuning should lead to improved outcomes.
3. The MiCroGerk participants [15] utilized six distinct runs to determine sentences' classes, with runs based on FastText, T5, BLOOM alone, MLP (multilayer perceptron), Naïve Bayes, and Ridge in addition to experimenting also with the TF–IDF vectorizer and Count vectorizer. T5 achieved the best result among the other methods, as we discuss later.
4. As a classifier layer, TheLangverse teamutilized a combination of FastText and MLP, obtaining moderately good results.
5. ThePunDetectives team [16] used multiple models for the classification task, including Random, FastText, Ridge, Naïve Bayes, T5, and RoBERTa. The RoBERTa base model produced the best results.
6. The members of the UBO team [17] used the T5 model, with varied results across languages, French having the highest success rate.
7. TeamCAU [18] employed various models, such as multiple Large Language Models (LLMs), FastText, and a TF–IDF based model. In the case of LLMs, BLOOM, Jurassic-2 (via AI21's inference API), and T5 were used. The authors reported that using LLMs yielded the best results among their runs.

8. The Smroltra team [19] experimented extensively with various classification methods, including Random Forest, FastText, Naïve Bayes, Logistic Regresion, TF–IDF, MLP, and a T5 transformer. The results for the Spanish and French datasets were very similar. Despite the fact that the majority of methods are primarily used for English, their approaches were not as effective at detecting English puns.
9. Using OpenAI's GPT-3, the Croland team [20] tackled Task 1 with the assumption that LLMs should have a decent sense of humour.
10. In order to enhance humour recognition in English, the Innsbruck team [21] experimented with various data augmentation techniques to extend the original training dataset, including synonym replacement, back-translation, and shortening. However, the performance improvements were only moderate.
11. The LJGG team experimented with various training techniques for the T5 model.
12. Finally, the Les_miserables team (no system description paper submitted) provided predictions based on T5 and FastText embeddings, as well as random baseline results.

4. Results

4.1. Comparison of Results on Test Data

Tables 3, 5, and 7 show participants' results for the wordplay detection task in English, French, and Spanish, respectively, when using test data. Since some participants submitted only partial runs, we provide precision, recall, F_1, and accuracy for the total number of instances in the test set (P, R, F_1, A) and also for just the number of instances (#) where a classification was attempted (P^*, R^*, F_1^*, A^*). Overall, the results suggest that wordplay detection is quite challenging for most of the models and for all three languages. When compared to random runs, the improvement of the best runs are less than 15 points based on F_1 score for all three languages.

The best results according to the F_1 metric for English are obtained with the T5 model (used by two teams, LJGG and UBO). Still, the performance is lower than 60%. We also note that the results depend heavily on the implementation choices, fine-tuning, and/or the prompts used.

For French, the best results were also achieved by teams using T5, with F_1 being 66.45. The higher score compared to English might be due to higher similarity between training and test data for the French dataset, as this data is derived through the translation of the overlapping sets of English puns. On the other hand, the test set of English contains different puns which have no or little semantic or vocabulary similarity. Interestingly, Logistic Regression and TF–IDF classifiers achieved comparable results for French data, which also indicates that careful training of "lighter" models can lead to results comparable to those of large pre-trained models.

The T5 model used by the AKRaNLU team, which applied sentence embeddings with a binary classifier and a sequence classification using XLM-Roberta, was the winner for Spanish data ($F_1 = 59.64$). However, this should be interpreted in light of Smroltra's random prediction runs, which achieve $F_1 = 51.92$ on the very same data.

The Les_miserables team obtained the $P = R = F_1 = 0$ for both Spanish datasets; however, the accuracy of the predictions is $A > 50$. The confusion matrices in Tables 12 and 13 give more insights into this phenomenon. It can be seen that the system assigned non-pun labels to all instances. Recall that the precision and recall metrics have only true positive labels on

their numerators while accuracy's numerator is based on both true positive and true negative predictions.

The confusion matrices for the test and training sets for English and French can be found in Tables 10, 11, 14, and 15.

4.2. Comparison of Results on Training Data

As we have incorporated the training data into the input file, it becomes possible to compare runs both on the test and training sets. This comparison provides valuable insights into the effectiveness and robustness of the systems across different datasets and helps gauge their ability to apply learned knowledge to new instances.

For completeness and to provide more insights into model performance, we also investigate the results using training data (see Table 4). Looking at the English language runs, the training data results are significantly higher than the ones on testing data, which is of course not surprising. Some runs achieve close to 100% F_1 score. Three different teams (UBO, MiCroGerk, and TeamCAU) achieved an F_1 score higher than 90%.

For French, the situation is similar, as shown in Table 6. The UBO team recorded 93.81% F_1 and in total, two teams out of the six participating in the French version of Task 1 had F_1 scores over 90%.

Finally, for Spanish (Table 8), we observe that the maximum F_1 increased from 61.22% to 89.02% when training data was used. (No team managed to surpass 90%.)

The results reveal potential overfitting issues, where the system performs well on the training data but struggles with generalization to unseen test data.

4.3. Additional Analysis

Table 9 reports the number of wordplay instances correctly predicted by a given number of runs ("True count"). The table suggests that the most challenging texts, indicated by the lowest number of correct predictions, tend to have a higher occurrence of false positives in both the test set and the training set for English and French. This trend is less apparent for Spanish.

One potential explanation for this observation is the presence of augmented data in English and French, where non-pun instances may closely resemble pun instances from the training data. This similarity between non-puns and puns in the training data could result in a higher likelihood of false positives during prediction, particularly for the more challenging texts. It emphasizes the need for language-specific training data and the careful selection and augmentation of non-pun instances to ensure that they sufficiently capture the distinguishing features of puns. These findings underscore the importance of data quality, diversity, and representativeness in training models for pun detection, as the distinction between puns and non-puns may be subtle. This finding also proves that the state-of-the-art models still struggle to distinguish wordplay.

5. Conclusions

In this paper, we have provided the overview and detailed results of Task 1 of the JOKER-2023 challenge, which consists of detecting sentences with wordplay in English, French, and Spanish.

Table 3
Results for Task 1 (pun detection) in English on the test data

run ID	#	P	R	F_1	A	P^*	R^*	F_1^*	A^*
Croland_EN_GPT3	3,183	**100.00**	0.86	1.71	**74.80**	**100.00**	0.86	1.71	**74.80**
LJGG_t5_large_easy_en	3,183	42.73	71.94	**53.61**	68.36	42.73	71.94	**53.61**	68.36
LJGG_t5_large_label_en	3,183	25.41	**100.00**	40.53	25.41	25.41	**100.00**	40.53	25.41
LJGG_t5_large_no_label_en	3,183	25.41	**100.00**	40.53	25.41	25.41	**100.00**	40.53	25.41
Les_miserables_fasttext	3,183	25.78	80.96	39.11	35.94	25.78	80.96	39.11	35.94
Les_miserables_random	3,183	26.43	51.29	34.88	51.33	26.43	51.29	34.88	51.33
Les_miserables_simplet5	3,183	28.13	88.75	42.72	39.52	28.13	88.75	42.72	39.52
MiCroGerk_EN_BLOOM	13.00	8.33	0.12	0.24	74.26	8.33	**100.00**	15.38	15.38
MiCroGerk_EN_FastText	3,183	25.87	82.94	39.44	35.28	25.87	82.94	39.44	35.28
MiCroGerk_EN_MLP	3,183	29.04	72.92	41.54	47.84	29.04	72.92	41.54	47.84
MiCroGerk_EN_NB	3,183	25.98	95.42	40.84	29.75	25.98	95.42	40.84	29.75
MiCroGerk_EN_Ridge	3,183	26.74	85.16	40.70	36.94	26.74	85.16	40.70	36.94
MiCroGerk_EN_SimpleT5	3,183	30.75	83.06	44.88	48.16	30.75	83.06	44.88	48.16
Smroltra_EN_FastText	3,183	25.62	80.34	38.85	35.72	25.62	80.34	38.85	35.72
Smroltra_EN_Logistic-Regression	3,183	26.14	86.15	40.11	34.62	26.14	86.15	40.11	34.62
Smroltra_EN_MLP	3,183	27.78	72.43	40.16	45.14	27.78	72.43	40.16	45.14
Smroltra_EN_NBC	3,183	26.12	95.55	41.02	30.19	26.12	95.55	41.02	30.19
Smroltra_EN_Random	3,183	25.54	66.99	36.98	41.97	25.54	66.99	36.98	41.97
Smroltra_EN_SimpleT5	3,183	31.97	83.68	46.27	50.61	31.97	83.68	46.27	50.61
Smroltra_EN_TFIDF	3,183	26.90	84.05	40.76	37.92	26.90	84.05	40.76	37.92
TeamCAU_EN_AI21	40	27.58	0.98	1.90	74.17	27.58	80.00	41.02	0.43
TeamCAU_EN_BLOOM	40	30.00	0.37	0.73	74.45	30.00	30.00	30.00	65.00
TeamCAU_EN_FastText	3,183	25.71	80.84	39.02	35.78	25.71	80.84	39.02	35.78
TeamCAU_EN_Random-ForestWithTFidfEncoding	3,183	25.69	83.43	39.28	34.46	25.69	83.43	39.28	34.46
TeamCAU_EN_ST5	3,183	26.99	93.32	41.87	34.15	26.99	93.32	41.87	34.15
TeamCAU_EN_TFidfRidge	3,183	26.74	85.16	40.70	36.94	26.74	85.16	40.70	36.94
TheLangVerse_fasttext-MLP	3,183	26.31	75.40	39.01	40.08	26.31	75.40	39.01	40.08
ThePunDetectives_Fasttext	3,183	26.07	80.22	39.35	37.16	26.07	80.22	39.35	37.16
ThePunDetectives_NaiveBayes	3,183	25.43	99.62	40.52	25.66	25.43	99.62	40.52	25.66
ThePunDetectives_Random	3,183	25.96	50.55	34.31	50.80	25.96	50.55	34.31	50.80
ThePunDetectives_Ridge	3,183	27.44	88.75	41.92	37.51	27.44	88.75	41.92	37.51
ThePunDetectives_Roberta	3,183	26.11	91.96	40.67	31.82	26.11	91.96	40.67	31.82
ThePunDetectives_SimpleT5	3,183	29.21	93.20	44.48	40.87	29.21	93.20	44.48	40.87
UBO_SimpleT5	3,183	36.51	85.53	51.18	58.52	36.51	85.53	51.18	58.52
AKRaNLU_sentemb	3,183	26.29	86.40	40.32	34.99	26.29	86.40	40.32	34.99
AKRaNLU_seqclassification	3,183	25.41	**100.00**	40.53	25.41	25.41	**100.00**	40.53	25.41
Innsbruck_DS_backtranslation	3,183	27.35	84.91	41.38	38.86	27.35	84.91	41.38	38.86
Innsbruck_DS_r1	3,183	27.32	86.89	41.57	37.92	27.32	86.89	41.57	37.92
Innsbruck_DS_synonym	3,183	27.15	86.89	41.37	37.41	27.15	86.89	41.37	37.41

We have also described the Spanish-language addition to the JOKER corpus previously described in our SIGIR 2023 paper [12]. We hope our dataset will foster research on wordplay analysis in various domains, including AI, information retrieval, natural language processing, humour studies, and second-language learning.

Twelve participating teams submitted 75 runs in total for Task 1. These teams applied various methods, from the traditional classifiers used in natural language processing to more modern, large-scale pre-trained models. The participants' results confirm that pun detection is still a challenging task. There is still much room of improvement, with the maximum results of 60–65% F_1, especially considering that random prediction achieves 34–51% F_1. The T5 model appeared to be most effective for all the languages and especially for English. The fact that

even state-of-the-art models exhibit limitations in distinguishing wordplay suggests that further research and development are required to improve performance in this domain.

We observed that the performance of the same or similar methods varied depending on factors such as implementation, fine-tuning, and the specific prompts used. This indicates the sensitivity of the results to the specific choices made during the implementation process. Fine-tuning models and carefully selecting appropriate prompts can have a substantial impact on the effectiveness of the methods employed.

Some participants were constrained by limitations on tokens or time when using language models. As a result, they had to submit incomplete runs. This highlights the importance of considering the efficiency in addition to the effectiveness of the approaches. Even still, a relatively simplistic Logistic Regression classifier exhibits good results on French and Spanish data.

The results also indicate the presence of overfitting issues in the systems.

Additional information on the track is available on the JOKER website: http://www.joker-project.com/

Acknowledgments

This project 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. JOKER is supported by *La Maison des sciences de l'homme en Bretagne*. For their help and support in the first Spanish pun translation contest, we thank Carolina Palma Preciado, Leopoldo Jesús Gutierrez Galeano, Khatima El Krirh, Nathalie Narváez Bruneau, and Rachel Kinlay. We also thank all other colleagues and students who participated in data construction, the translation contests, and the CLEF JOKER track.

References

[1] 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.

[2] L. Ermakova, T. Miller, A.-G. Bosser, V. M. P. Preciado, G. Sidorov, A. Jatowt, Overview of JOKER 2023 Automatic Wordplay Analysis Task 2 - Pun location and interpretation, in: [22], 2023.

[3] L. Ermakova, T. Miller, A.-G. Bosser, V. M. P. Preciado, G. Sidorov, A. Jatowt, Overview of JOKER 2023 Automatic Wordplay Analysis Task 3 - Pun translation, in: [22], 2023.

[4] L. Ermakova, T. Miller, A.-G. Bosser, V. M. P. Preciado, G. Sidorov, A. Jatowt, Overview of JOKER – CLEF-2023 Track on Automatic Wordplay Analysis, in: A. Arampatzis, E. Kan-

oulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), CLEF'23: Proceedings of the Fourteenth International Conference of the CLEF Association, Lecture Notes in Computer Science, Springer, 2023.

[5] W. Kolb, T. Miller, Human–computer interaction in pun translation, in: J. L. Hadley, K. Taivalkoski-Shilov, C. S. C. Teixeira, A. Toral (Eds.), Using Technologies for Creative-Text Translation, Routledge, 2022, pp. 66–88. doi:10.4324/9781003094159-4.

[6] T. Miller, C. F. Hempelmann, I. Gurevych, SemEval-2017 Task 7: Detection and interpretation of English puns, in: Proceedings of the 11th International Workshop on Semantic Evaluation, 2017, pp. 58–68. doi:10.18653/v1/S17-2005.

[7] C. D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, Cambridge, 2008.

[8] F. Dhanani, M. Rafi, M. A. Tahir, FAST_MT participation for the JOKER CLEF-2022 automatic pun and human translation tasks, in: Proceedings of the Working Notes of CLEF 2022 – Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th to 8th, 2022, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022, p. 14.

[9] V. M. P. Preciado, G. Sidorov, C. P. Preciado, Assessing Wordplay-Pun classification from JOKER dataset with pretrained BERT humorous models, in: Proceedings of the Working Notes of CLEF 2022 – Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th to 8th, 2022, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022, p. 6.

[10] L. S. M. Altin, À. Bravo, H. Saggion, Lastus/taln at haha: Humor analysis based on human annotation, in: M. Á. G. Cumbreras, J. Gonzalo, E. M. Cámara, R. Martínez-Unanue, P. Rosso, J. C. de Albornoz, S. Montalvo, L. Chiruzzo, S. Collovini, Y. Gutiérrez, S. M. J. Zafra, M. Krallinger, M. M. y Gómez, R. Ortega-Bueno, A. Rosá (Eds.), Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, 2019.

[11] A. Epimakhova, Using machine learning to classify and interpret wordplay, in: Proceedings of the Working Notes of CLEF 2022 – Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th to 8th, 2022, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022, p. 6.

[12] 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. doi:10.1145/3539618.3591885, to appear.

[13] R. R. Dsilva, AKRaNLU @ CLEF JOKER 2023: Using sentence embeddings and multilingual models to detect and interpret wordplay, in: [22], 2023.

[14] V. M. P. Preciado, C. P. Preciado, G. Sidorov, NLPalma @ CLEF 2023 JOKER: A BLOOMZ and BERT approach for wordplay detection and translation, in: [22], 2023.

[15] A. Prnjak, D. R. Davari, K. Schmitt, CLEF 2023 JOKER Task 1, 2, 3: pun detection, pun interpretation, and pun translation, in: [22], 2023.

[16] F. Ohnesorge, M. Á. Gutiérrez, J. Plichta, CLEF 2023 JOKER Tasks 2 and 3: using NLP models for pun location, interpretation and translation, in: [22], 2023.

[17] Q. Dubreuil, UBO Team @ CLEF JOKER 2023 track for Task 1, 2 and 3 - applying AI models

in regards to pun translation, in: [22], 2023.
[18] A. Anjum, N. Lieberum, Exploring Humor in Natural Language Processing: A Comprehensive Review of JOKER Tasks at CLEF Symposium 2023, in: [22], 2023.
[19] O. Popova, P. Dadić, Does AI have a sense of humor? CLEF 2023 JOKER tasks 1, 2 and 3: Using BLOOM, GPT, SimpleT5, and more for pun detection, location, interpretation and translation, in: [22], 2023.
[20] J. Komorowska, I. Čatipović, D. Vujica, CLEF2023' JOKER Working Notes, in: [22], 2023.
[21] S. Reicho, A. Jatowt, Innsbruck @ CLEF JOKER 2023 Track Task 1: Data Augmentation Techniques for Humor Recognition in Text, in: [22], 2023.
[22] Proceedings of the Working Notes of CLEF 2023: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2023.

Table 4
Results for Task 1 (pun detection) in English on the training data

run ID	#	P	R	F_1	A	P*	R*	F_1^*	A*
Croland_task_1.1/2_EN_GPT3	5,292	**100.00**	2.69	5.24	43.27	**100.00**	2.69	5.24	43.27
LJGG_task1_t5_large_easy_en_-test_auto	5,292	94.12	80.39	86.71	85.64	94.12	80.39	86.71	85.64
LJGG_task1_t5_large_label_en_-test_auto	5,292	58.30	**100.00**	73.65	58.30	58.30	**100.00**	73.65	58.30
LJGG_task1_t5_large_no_label_-en_test_auto	5,292	58.30	**100.00**	73.65	58.30	58.30	**100.00**	73.65	58.30
Les_miserables_fasttext	5,292	74.40	86.68	80.07	74.85	74.40	86.68	80.07	74.85
Les_miserables_random_model	5,292	58.27	51.02	54.41	50.15	58.27	51.02	54.41	50.15
Les_miserables_simplet5	5,292	78.65	87.16	82.69	78.72	78.65	87.16	82.69	78.72
MiCroGerk_task_1.1_EN_BLOOM	17	53.33	0.26	0.52	41.72	53.33	**100.00**	69.57	58.82
MiCroGerk_task_1.1_EN_FastText	5,292	75.51	89.17	81.78	76.83	75.51	89.17	81.78	76.83
MiCroGerk_task_1.1_EN_MLP	5,292	**100.00**	99.97	**99.98**	**99.98**	**100.00**	99.97	**99.98**	**99.98**
MiCroGerk_task_1.1_EN_NB	5,292	76.45	97.44	85.68	81.01	76.45	97.44	85.68	81.01
MiCroGerk_task_1.1_EN_Ridge-Classification	5,292	87.53	97.18	92.10	90.29	87.53	97.18	92.10	90.29
MiCroGerk_task_1.1_EN_SimpleT5	5,292	82.36	85.67	83.98	80.95	82.36	85.67	83.98	80.95
Smroltra_task_1.1_EN_FastText	5,292	73.92	87.20	80.01	74.60	73.92	87.20	80.01	74.60
Smroltra_task_1.1_EN_LogisticRe-gression	5,292	75.97	92.12	83.27	78.42	75.97	92.12	83.27	78.42
Smroltra_task_1.1_EN_MLP	5,292	92.68	90.76	91.71	90.44	92.68	90.76	91.71	90.44
Smroltra_task_1.1_EN_NBC	5,292	74.01	95.72	83.48	77.91	74.01	95.72	83.48	77.91
Smroltra_task_1.1_EN_Random	5,292	58.97	67.75	63.06	53.72	58.97	67.75	63.06	53.72
Smroltra_task_1.1_EN_SimpleT5	5,292	85.68	85.74	85.71	83.33	85.68	85.74	85.71	83.33
Smroltra_task_1.1_EN_TFIDF	5,292	84.85	92.22	88.38	85.87	84.85	92.22	88.38	85.87
TeamCAU_task_1.1_EN_AI21	60	51.11	0.75	1.47	41.72	51.11	67.65	58.23	45.00
TeamCAU_task_1.1_EN_BLOOM	60	58.06	0.58	1.16	41.80	58.06	52.94	55.38	51.67
TeamCAU_task_1.1_EN_FastText	5,292	72.39	84.99	78.19	72.35	72.39	84.99	78.19	72.35
TeamCAU_task_1.1_EN_Random-ForrestWithTFidfEncoding	5,292	99.77	99.84	99.81	99.77	99.77	99.84	99.81	99.77
TeamCAU_task_1.1_EN_ST5	5,291	74.73	92.51	82.68	77.40	74.73	92.51	82.68	77.40
TeamCAU_task_1.1_EN_TFid-fRidge	5,292	87.53	97.18	92.10	90.29	87.53	97.18	92.10	90.29
TheLangVerse_task_1_fasttext-MLP	5,292	73.11	75.98	74.52	69.71	73.11	75.98	74.52	69.71
ThePunDetectives_task_1.1_Fast-text	5,292	74.64	84.73	79.37	74.32	74.64	84.73	79.37	74.32
ThePunDetectives_task_1.1_Naive-Bayes	5,292	63.78	99.90	77.86	66.87	63.78	99.90	77.86	66.87
ThePunDetectives_task_1.1_Ran-dom	5,292	58.92	51.60	55.02	50.81	58.92	51.60	55.02	50.81
ThePunDetectives_task_1.1_Ridge	5,292	83.49	96.73	89.62	86.94	83.49	96.73	89.62	86.94
ThePunDetectives_task_1.1_-Roberta	5,292	74.34	90.83	81.76	76.38	74.34	90.83	81.76	76.38
ThePunDetectives_task_1.1_Sim-pleT5	5,292	83.74	96.82	89.81	87.19	83.74	96.82	89.81	87.19
UBO_task_1_SimpleT5	5,292	95.55	96.08	95.81	95.11	95.55	96.08	95.81	95.11
AKRaNLU_task_1_sentemb	5,292	71.76	85.41	77.99	71.90	71.76	85.41	77.99	71.90
AKRaNLU_task_1_seqclassification	5,292	58.30	**100.00**	73.65	58.30	58.30	**100.00**	73.65	58.30
Innsbruck_DS_backtranslation	5,292	79.67	92.12	85.45	81.71	79.67	92.12	85.45	81.71
Innsbruck_DS_r1	5,292	78.45	91.09	84.30	80.22	78.45	91.09	84.30	80.22
Innsbruck_DS_synonym	5,292	79.31	91.35	84.91	81.07	79.31	91.35	84.91	81.07

Table 5
Results for Task 1 (pun detection) in French on the test data

run ID	#	P	R	F_1	A	P*	R*	F_1^*	A*
Croland_FR_GPT3	12,873	**100.00**	01.14	02.27	59.24	**100.00**	01.14	02.27	59.24
LJGG_t5_large_easy_fr	12,873	55.13	64.29	59.36	63.70	55.13	64.29	59.36	63.70
LJGG_t5_large_label_fr	12,873	41.23	**100.00**	58.39	41.23	41.23	**100.00**	58.39	41.23
LJGG_t5_large_no_label_fr	12,873	41.23	**100.00**	58.39	41.23	41.23	**100.00**	58.39	41.23
Les_miserables_fasttext	12,873	58.57	19.76	29.55	61.15	58.57	19.76	29.55	61.15
Les_miserables_random	12,873	41.14	49.81	45.06	49.92	41.14	49.81	45.06	49.92
Les_miserables_simplet5	12,873	59.72	74.88	**66.45**	68.82	59.72	74.88	**66.45**	68.82
Smroltra_FR_FastText	12,873	55.24	25.00	34.42	60.72	55.24	25.00	34.42	60.72
Smroltra_FR_Logistic-Regression	12,873	58.43	60.39	59.40	65.95	58.43	60.39	59.40	65.95
Smroltra_FR_MLP	12,873	56.49	62.88	59.52	64.73	56.49	62.88	59.52	64.73
Smroltra_FR_NBC	12,873	56.73	63.18	59.78	64.94	56.73	63.18	59.78	64.94
Smroltra_FR_Random	12,873	42.14	67.70	51.95	48.36	42.14	67.70	51.95	48.36
Smroltra_FR_SimpleT5	12,873	61.21	67.69	64.29	**68.99**	61.21	67.69	64.29	**68.99**
Smroltra_FR_TFIDF	12,873	58.77	62.09	60.38	66.41	58.77	62.09	60.38	66.41
UBO_SimpleT5	12,871	67.80	58.76	62.95	71.49	67.80	58.76	62.95	71.48
AKRaNLU_sentemb	12,873	41.18	73.88	52.88	45.71	41.18	73.88	52.88	45.71
AKRaNLU_seqclassification	12,873	41.23	**100.00**	58.39	41.23	41.23	**100.00**	58.39	41.23

Table 6
Results for Task 1 (pun detection) in French on the training data

run ID	#	P	R	F_1	A	P*	R*	F_1^*	A*
Croland_task_1.1/2_FR_GPT3	3,999	**100.00**	1.45	2.86	50.76	**100.00**	1.45	2.86	50.76
LJGG_task1_t5_large_easy_fr_-test_auto	3,999	70.61	71.67	71.14	70.94	70.61	71.67	71.14	70.94
LJGG_task1_t5_large_label_fr_-test_auto	3,999	49.96	**100.00**	66.63	49.96	49.96	**100.00**	66.63	49.96
LJGG_task1_t5_large_no_label_fr_-test_auto	3,999	49.96	**100.00**	66.63	49.96	49.96	**100.00**	66.63	49.96
Les_miserables_fasttext	3,999	76.68	24.52	37.16	58.56	76.68	24.52	37.16	58.56
Les_miserables_random_model	3,999	50.07	50.25	50.16	50.11	50.07	50.25	50.16	50.11
Les_miserables_simplet5	3,999	86.81	91.89	89.28	88.97	86.81	91.89	89.28	88.97
Smroltra_task_1.1_FR_FastText	3,999	74.58	31.13	43.93	60.29	74.58	31.13	43.93	60.29
Smroltra_task_1.1_FR_LogisticRe-gression	3,999	85.71	85.54	85.62	85.65	85.71	85.54	85.62	85.65
Smroltra_task_1.1_FR_MLP	3,999	92.81	93.74	93.28	93.25	92.81	93.74	93.28	93.25
Smroltra_task_1.1_FR_NBC	3,999	89.55	90.09	89.82	89.80	89.55	90.09	89.82	89.80
Smroltra_task_1.1_FR_Random	3,999	49.07	64.51	55.74	48.81	49.07	64.51	55.74	48.81
Smroltra_task_1.1_FR_SimpleT5	3,999	85.74	86.94	86.33	86.25	85.74	86.94	86.33	86.25
Smroltra_task_1.1_FR_TFIDF	3,999	92.14	92.14	92.14	92.15	92.14	92.14	92.14	92.15
UBO_task_1_SimpleT5	3,999	95.87	91.84	**93.81**	**93.95**	95.87	91.84	**93.81**	**93.95**
AKRaNLU_task_1_sentemb	3,999	49.90	73.97	59.60	49.89	49.90	73.97	59.60	49.89
AKRaNLU_task_1_seqclassification	3,999	49.96	**100.00**	66.63	49.96	49.96	**100.00**	66.63	49.96

Table 7
Results for Task 1 (pun detection) in Spanish on the test data

run ID	#	P	R	F_1	A	P*	R*	F_1^*	A*
Croland_ES_GPT3	2,241	**98.07**	05.35	10.15	**59.75**	**98.07**	05.35	10.15	59.75
LJGG_t5_large_easy_es	2,230	50.34	54.09	52.15	57.83	50.34	54.26	52.23	57.75
LJGG_t5_large_label_es	2,230	42.55	**99.68**	**59.64**	42.70	42.55	**100.00**	**59.70**	42.55
LJGG_t5_large_no_label_es	2,230	42.55	**99.68**	**59.64**	42.70	42.55	**100.00**	**59.70**	42.55
Les_miserables_fasttext	2,230	0.00	0.00	0.00	57.51	0.00	0.00	0.00	57.44
Les_miserables_random	2,230	43.43	51.78	47.24	50.87	43.43	51.94	47.31	50.76
Les_miserables_simplet5	2,230	51.10	17.01	25.53	57.83	51.10	17.07	25.59	57.75
NLPalma_BERT	2,230	55.94	40.54	47.01	61.17	55.94	40.67	47.10	61.12
Smroltra_ES_FastText	2,238	40.75	0.625	49.33	45.47	40.75	0.625	49.33	45.39
Smroltra_ES_Logistic-Regression	2,238	0.50	49.05	49.52	57.51	0.50	49.05	49.52	57.46
Smroltra_ES_MLP	2,238	55.45	44.32	49.27	**61.22**	55.45	44.32	49.27	61.17
Smroltra_ES_NBC	2,238	47.69	56.40	51.68	55.19	47.69	56.40	51.68	55.13
Smroltra_ES_Random	2,241	42.05	67.85	51.92	46.63	42.05	67.85	51.92	46.63
Smroltra_ES_SimpleT5	2,238	44.31	46.21	45.24	52.47	44.31	46.21	45.24	52.41
Smroltra_ES_TFIDF	2,238	53.34	46.11	49.46	59.97	53.34	46.11	49.46	59.91
UBO_SimpleT5	2,230	51.28	62.92	56.50	58.85	51.28	63.11	56.58	58.78
AKRaNLU_sentemb	2,230	41.39	72.26	52.63	44.75	41.39	72.49	52.70	44.61
AKRaNLU_seqclassification	2,230	42.55	**99.68**	**59.64**	42.70	42.55	**100.00**	**59.70**	42.55

Table 8
Results for Task 1 (pun detection) in Spanish on the training data

run ID	#	P	R	F_1	A	P*	R*	F_1^*	A*
Croland_task_1.1/2_ES_GPT3	1,994	**100.00**	4.56	8.72	59.08	**100.00**	4.56	8.72	59.08
LJGG_task1_t5_large_easy_es_-test_auto	1,994	68.82	72.28	70.51	74.07	68.82	72.28	70.51	74.07
LJGG_task1_t5_large_label_es_-test_auto	1,993	42.90	**100.00**	60.04	42.93	42.90	**100.00**	60.04	42.90
LJGG_task1_t5_large_no_label_es_-test_auto	1,993	42.90	**100.00**	60.04	42.93	42.90	**100.00**	60.04	42.90
Les_miserables_fasttext	1,993	0.00	0.00	0.00	57.12	0.00	0.00	0.00	57.10
Les_miserables_random_model	1,993	44.31	51.46	47.62	51.45	44.31	51.46	47.62	51.43
Les_miserables_simplet5	1,993	78.99	32.98	46.53	67.50	78.99	32.98	46.53	67.49
NLPalma_BERT	1,994	90.62	82.46	86.34	88.82	90.62	82.46	86.34	88.82
Smroltra_task_1.1_ES_FastText	1,992	41.75	64.56	50.71	46.19	41.75	64.56	50.71	46.13
Smroltra_task_1.1_ES_LogisticRe-dression	1,992	49.10	47.60	48.34	56.37	49.10	47.60	48.34	56.33
Smroltra_task_1.1_ES_MLP	1,992	54.77	43.63	48.57	60.38	54.77	43.63	48.57	60.34
Smroltra_task_1.1_ES_NBC	1,992	47.66	56.02	51.51	54.76	47.66	56.02	51.51	54.72
Smroltra_task_1.1_ES_Random	1,994	42.77	67.13	52.25	47.39	42.77	67.13	52.25	47.39
Smroltra_task_1.1_ES_SimpleT5	1,992	44.19	45.85	45.01	51.96	44.19	45.85	45.01	51.91
Smroltra_task_1.1_ES_TFIDF	1,992	52.39	44.80	48.30	58.88	52.39	44.80	48.30	58.84
UBO_task_1_SimpleT5	1,993	86.70	91.46	**89.02**	**90.32**	86.70	91.46	**89.02**	**90.32**
AKRaNLU_task_1_sentemb	1,994	43.45	73.33	54.57	47.64	43.45	73.33	54.57	47.64
AKRaNLU_task_1_seqclassification	1,994	42.88	**100.00**	60.02	42.88	42.88	**100.00**	60.02	42.88

Table 9
Number of wordplay instances assigned true predictions per language, dataset, and class. True count refers to the number of runs that made a correct prediction per wordplay

	EN				FR				ES			
	Test		Train		Test		Train		Test		Train	
True count	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
1	39	—	—	—	28	—	—	—	1	—	—	—
2	153	—	—	—	84	—	—	—	5	6	—	—
3	260	—	—	—	158	11	3	—	8	37	1	4
4	246	—	2	—	279	71	12	1	42	56	8	19
5	215	2	22	—	377	248	20	2	55	113	25	66
6	175	—	44	—	422	334	44	15	98	135	81	105
7	130	1	61	—	557	350	54	36	141	90	138	124
8	130	2	69	—	621	410	100	39	136	79	104	83
9	106	1	66	1	736	504	136	63	160	103	106	69
10	107	4	57	1	927	512	214	117	199	114	132	70
11	106	4	65	3	1397	567	524	231	233	91	210	89
12	95	4	62	6	1314	689	607	397	144	74	190	112
13	67	5	56	7	569	705	239	456	49	46	121	84
14	68	5	52	5	96	518	48	351	15	8	22	30
15	61	2	57	8	—	279	—	224	—	—	—	—
16	63	9	54	7	—	109	—	65	—	—	—	—
17	40	8	56	11	—	1	—	1	—	—	—	—
18	42	7	42	19	—	—	—	—	—	—	—	—
19	39	8	45	13	—	—	—	—	—	—	—	—
20	38	8	45	18	—	—	—	—	—	—	—	—
21	41	13	47	33	—	—	—	—	—	—	—	—
22	22	14	50	36	—	—	—	—	—	—	—	—
23	28	12	45	40	—	—	—	—	—	—	—	—
24	21	20	51	59	—	—	—	—	—	—	—	—
25	23	25	66	50	—	—	—	—	—	—	—	—
26	17	23	74	65	—	—	—	—	—	—	—	—
27	11	31	91	94	—	—	—	—	—	—	—	—
28	12	34	112	95	—	—	—	—	—	—	—	—
29	8	40	166	105	—	—	—	—	—	—	—	—
30	6	46	236	136	—	—	—	—	—	—	—	—
31	4	73	251	212	—	—	—	—	—	—	—	—
32	1	107	146	398	—	—	—	—	—	—	—	—
33	—	147	17	703	—	—	—	—	—	—	—	—
34	—	115	—	682	—	—	—	—	—	—	—	—
35	—	37	—	267	—	—	—	—	—	—	—	—
36	—	1	—	7	—	—	—	—	—	—	—	—
37	—	1	—	2	—	—	—	—	—	—	—	—
38	—	—	—	2	—	—	—	—	—	—	—	—

Table 10
Confusion matrix for the English test set

run ID	TP	TN	FN	FP
Croland_task_1.1/2_EN_GPT3	7	2374	802	0
LJGG_task1_t5_large_easy_en_test_auto	582	1594	227	780
LJGG_task1_t5_large_label_en_test_auto	809	0	0	2374
LJGG_task1_t5_large_no_label_en_test_auto	809	0	0	2374
Les_miserables_fasttext	655	489	154	1885
Les_miserables_random_model	415	1219	394	1155
Les_miserables_simplet5	718	540	91	1834
MiCroGerk_task_1.1_EN_BLOOM	1	2363	808	11
MiCroGerk_task_1.1_EN_FastText	671	452	138	1922
MiCroGerk_task_1.1_EN_MLP	590	933	219	1441
MiCroGerk_task_1.1_EN_NB	772	175	37	2199
MiCroGerk_task_1.1_EN_RidgeClassification	689	487	120	1887
MiCroGerk_task_1.1_EN_SimpleT5	672	861	137	1513
Smroltra_task_1.1_EN_FastText	650	487	159	1887
Smroltra_task_1.1_EN_LogisticRegression	697	405	112	1969
Smroltra_task_1.1_EN_MLP	586	851	223	1523
Smroltra_task_1.1_EN_NBC	773	188	36	2186
Smroltra_task_1.1_EN_Random	542	794	267	1580
Smroltra_task_1.1_EN_SimpleT5	677	934	132	1440
Smroltra_task_1.1_EN_TFIDF	680	527	129	1847
TeamCAU_task_1.1_EN_AI21	8	2353	801	21
TeamCAU_task_1.1_EN_BLOOM	3	2367	806	7
TeamCAU_task_1.1_EN_FastText	654	485	155	1889
TeamCAU_task_1.1_EN_RandomForrestWithTFidfEncoding	675	422	134	1952
TeamCAU_task_1.1_EN_ST5	755	332	54	2042
TeamCAU_task_1.1_EN_TFidfRidge	689	487	120	1887
TheLangVerse_task_1_fasttext-MLP	610	666	199	1708
ThePunDetectives_task_1.1_Fasttext	649	534	160	1840
ThePunDetectives_task_1.1_NaiveBayes	806	11	3	2363
ThePunDetectives_task_1.1_Random	409	1208	400	1166
ThePunDetectives_task_1.1_Ridge	718	476	91	1898
ThePunDetectives_task_1.1_Roberta	744	269	65	2105
ThePunDetectives_task_1.1_SimpleT5	754	547	55	1827
UBO_task_1_SimpleT5	692	1171	117	1203
akranlu_task_1_sentemb	699	415	110	1959
akranlu_task_1_seqclassification	809	0	0	2374
Innsbruck_DS_backtranslation	687	550	122	1824
Innsbruck_DS_r1	703	504	106	1870
Innsbruck_DS_synonym	703	488	106	1886

Table 11
Confusion matrix for the English training set

run ID	TP	TN	FN	FP
Croland_task_1.1/2_EN_GPT3	83	2207	3002	0
LJGG_task1_t5_large_easy_en_test_auto	2480	2052	605	155
LJGG_task1_t5_large_label_en_test_auto	3085	0	0	2207
LJGG_task1_t5_large_no_label_en_test_auto	3085	0	0	2207
Les_miserables_fasttext	2674	1287	411	920
Les_miserables_random_model	1574	1080	1511	1127
Les_miserables_simplet5	2689	1477	396	730
MiCroGerk_task_1.1_EN_BLOOM	8	2200	3077	7
MiCroGerk_task_1.1_EN_FastText	2751	1315	334	892
MiCroGerk_task_1.1_EN_MLP	3084	2207	1	0
MiCroGerk_task_1.1_EN_NB	3006	1281	79	926
MiCroGerk_task_1.1_EN_RidgeClassification	2998	1780	87	427
MiCroGerk_task_1.1_EN_SimpleT5	2643	1641	442	566
Smroltra_task_1.1_EN_FastText	2690	1258	395	949
Smroltra_task_1.1_EN_LogisticRegression	2842	1308	243	899
Smroltra_task_1.1_EN_MLP	2800	1986	285	221
Smroltra_task_1.1_EN_NBC	2953	1170	132	1037
Smroltra_task_1.1_EN_Random	2090	753	995	1454
Smroltra_task_1.1_EN_SimpleT5	2645	1765	440	442
Smroltra_task_1.1_EN_TFIDF	2845	1699	240	508
TeamCAU_task_1.1_EN_AI21	23	2185	3062	22
TeamCAU_task_1.1_EN_BLOOM	18	2194	3067	13
TeamCAU_task_1.1_EN_FastText	2622	1207	463	1000
TeamCAU_task_1.1_EN_RandomForrestWithTFidfEncoding	3080	2200	5	7
TeamCAU_task_1.1_EN_ST5	2854	1242	231	965
TeamCAU_task_1.1_EN_TFidfRidge	2998	1780	87	427
TheLangVerse_task_1_fasttext-MLP	2344	1345	741	862
ThePunDetectives_task_1.1_Fasttext	2614	1319	471	888
ThePunDetectives_task_1.1_NaiveBayes	3082	457	3	1750
ThePunDetectives_task_1.1_Random	1592	1097	1493	1110
ThePunDetectives_task_1.1_Ridge	2984	1617	101	590
ThePunDetectives_task_1.1_Roberta	2802	1240	283	967
ThePunDetectives_task_1.1_SimpleT5	2987	1627	98	580
UBO_task_1_SimpleT5	2964	2069	121	138
akranlu_task_1_sentemb	2635	1170	450	1037
akranlu_task_1_seqclassification	3085	0	0	2207
Innsbruck_DS_backtranslation	2842	1482	243	725
Innsbruck_DS_r1	2810	1435	275	772
Innsbruck_DS_synonym	2818	1472	267	735

Table 12
Confusion matrix for the Spanish test set

run ID	TP	TN	FN	FP
Croland_task_1.1/2_ES_GPT3	51	1288	901	1
LJGG_task1_t5_large_easy_es_test_auto	515	781	437	508
LJGG_task1_t5_large_label_es_test_auto	949	8	3	1281
LJGG_task1_t5_large_no_label_es_test_auto	949	8	3	1281
Les_miserables_fasttext	0	1289	952	0
Les_miserables_random_model	493	647	459	642
Les_miserables_simplet5	162	1134	790	155
NLPalma_BERT	386	985	566	304
Smroltra_task_1.1_ES_FastText	595	424	357	865
Smroltra_task_1.1_ES_LogisticRedression	467	822	485	467
Smroltra_task_1.1_ES_MLP	422	950	530	339
Smroltra_task_1.1_ES_NBC	537	700	415	589
Smroltra_task_1.1_ES_Random	646	399	306	890
Smroltra_task_1.1_ES_SimpleT5	440	736	512	553
Smroltra_task_1.1_ES_TFIDF	439	905	513	384
UBO_task_1_SimpleT5	599	720	353	569
akranlu_task_1_sentemb	688	315	264	974
akranlu_task_1_seqclassification	949	8	3	1281

Table 13
Confusion matrix for the Spanish training set

run ID	TP	TN	FN	FP
Croland_task_1.1/2_ES_GPT3	39	1139	816	0
LJGG_task1_t5_large_easy_es_test_auto	618	859	237	280
LJGG_task1_t5_large_label_es_test_auto	855	1	0	1138
LJGG_task1_t5_large_no_label_es_test_auto	855	1	0	1138
Les_miserables_fasttext	0	1139	855	0
Les_miserables_random_model	440	586	415	553
Les_miserables_simplet5	282	1064	573	75
NLPalma_BERT	705	1066	150	73
Smroltra_task_1.1_ES_FastText	552	369	303	770
Smroltra_task_1.1_ES_LogisticRedression	407	717	448	422
Smroltra_task_1.1_ES_MLP	373	831	482	308
Smroltra_task_1.1_ES_NBC	479	613	376	526
Smroltra_task_1.1_ES_Random	574	371	281	768
Smroltra_task_1.1_ES_SimpleT5	392	644	463	495
Smroltra_task_1.1_ES_TFIDF	383	791	472	348
UBO_task_1_SimpleT5	782	1019	73	120
akranlu_task_1_sentemb	627	323	228	816
akranlu_task_1_seqclassification	855	0	0	1139

Table 14
Confusion matrix for the French test set

run ID	TP	TN	FN	FP
Croland_task_1.1/2_FR_GPT3	61	7565	5247	0
LJGG_task1_t5_large_easy_fr_test_auto	3413	4788	1895	2777
LJGG_task1_t5_large_label_fr_test_auto	5308	0	0	7565
LJGG_task1_t5_large_no_label_fr_test_auto	5308	0	0	7565
Les_miserables_fasttext	1049	6823	4259	742
Les_miserables_random_model	2644	3783	2664	3782
Les_miserables_simplet5	3975	4885	1333	2680
Smroltra_task_1.1_FR_FastText	1327	6490	3981	1075
Smroltra_task_1.1_FR_LogisticRegression	3206	5285	2102	2280
Smroltra_task_1.1_FR_MLP	3338	4995	1970	2570
Smroltra_task_1.1_FR_NBC	3354	5007	1954	2558
Smroltra_task_1.1_FR_Random	3594	2632	1714	4933
Smroltra_task_1.1_FR_SimpleT5	3593	5289	1715	2276
Smroltra_task_1.1_FR_TFIDF	3296	5253	2012	2312
UBO_task_1_SimpleT5	3119	6084	2189	1481
akranlu_task_1_sentemb	3922	1963	1386	5602
akranlu_task_1_seqclassification	5308	0	0	7565

Table 15
Confusion matrix for the French training set

run ID	TP	TN	FN	FP
Croland_task_1.1/2_FR_GPT3	29	2001	1969	0
LJGG_task1_t5_large_easy_fr_test_auto	1432	1405	566	596
LJGG_task1_t5_large_label_fr_test_auto	1998	0	0	2001
LJGG_task1_t5_large_no_label_fr_test_auto	1998	0	0	2001
Les_miserables_fasttext	490	1852	1508	149
Les_miserables_random_model	1004	1000	994	1001
Les_miserables_simplet5	1836	1722	162	279
Smroltra_task_1.1_FR_FastText	622	1789	1376	212
Smroltra_task_1.1_FR_LogisticRegression	1709	1716	289	285
Smroltra_task_1.1_FR_MLP	1873	1856	125	145
Smroltra_task_1.1_FR_NBC	1800	1791	198	210
Smroltra_task_1.1_FR_Random	1289	663	709	1338
Smroltra_task_1.1_FR_SimpleT5	1737	1712	261	289
Smroltra_task_1.1_FR_TFIDF	1841	1844	157	157
UBO_task_1_SimpleT5	1835	1922	163	79
akranlu_task_1_sentemb	1478	517	520	1484
akranlu_task_1_seqclassification	1998	0	0	2001