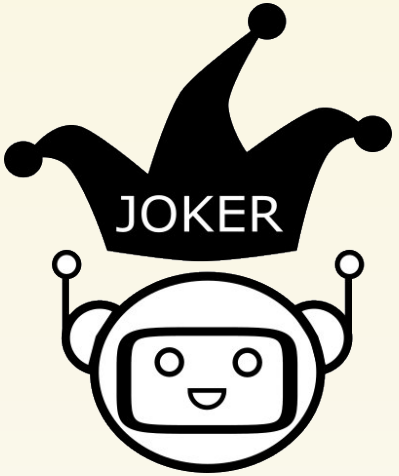
ORPAILLEUR & SyNaLP at JOKER 2024 Task 2

Good Old Cross Validation for Large Language Models Yields the Best Humorous Detection.



Pierre Epron[1], Gaël Guibon[1,3], Miguel Couceiro[1,2]

[1] LORIA, Université de Lorraine
[2] INESC-ID, IST, U.Lisboa
[3] LIPN, Université Sorbonne Paris Nord

Institut des sciences du Digital Management & Cognition

UNIVERSITÉ DE LORRAINE

Loria
Laboratoire lorrain de recherche en informatique et ses applications

Context

- **CLEF JOKER 2024 Task 2 [4, 7]**:
  Humour classification w.r.t. genre and technique
- I made an internship at LORIA (Nancy) on **Irony classification with LLMs**
- Opportunity to **assess our method** on a relatively close task

Context

- **CLEF JOKER 2024 Task 2 [4, 7]**:
  Humour classification w.r.t. genre and technique
- I made an internship at LORIA (Nancy) on **Irony classification with LLMs**
- Opportunity to **assess our method** on a relatively close task

Main questions:

- Can we leverage LLMs' **implicit knowledge** to better classify humor genres?
- Are classification **results consistent** across different LLMs?

Methods: Classification

- Add a **Feed Forward Layer** on top of the **last hidden layer** of the **last token**.
- Trained with a **Cross Entropy Loss**

Methods: LLMs

- Llama2-7b [8], Mistral-7b [6], Llama3-8b [1].
- **Instruction-tuned** version for the three models.
- 4-bits quantization [2]
- QLoRa [3]

Methods: Cross-Validation

- 5 stratified split.
- **Leave-one-out strategy**: 4 splits for training and 1 split for evaluation.
- Submitted runs:
  - **Ensemble** the 5 splits.
  - **Best** split and **Worst** split.

Method: Important Parameters

- Adapter Learning rate: $1.5e-4$
- FeedForward Learning rate: $1e-3$
- LoRa [5] paper: **low rank with high alpha** performs better.
- QLoRa [3] paper: **high rank with low alpha** performs better.
- We tried both:
  - 64 rank for 16 alpha.
  - 16 rank for 64 alpha.

Results: Before Submission

- Balanced cross-entropy ↓↓
- Linear scheduling strategy ↑↑
- QLoRa:
  - Rank=16, Alpha=64. ↑↑
  - Rank=64, Alpha=16. ↓↓
- Llama3 and Llama2 ↑↑. Mistral ↓↓.
- IR (Irony) and SC (Sarcasm) were challenging to differentiate.

Results: Label Results

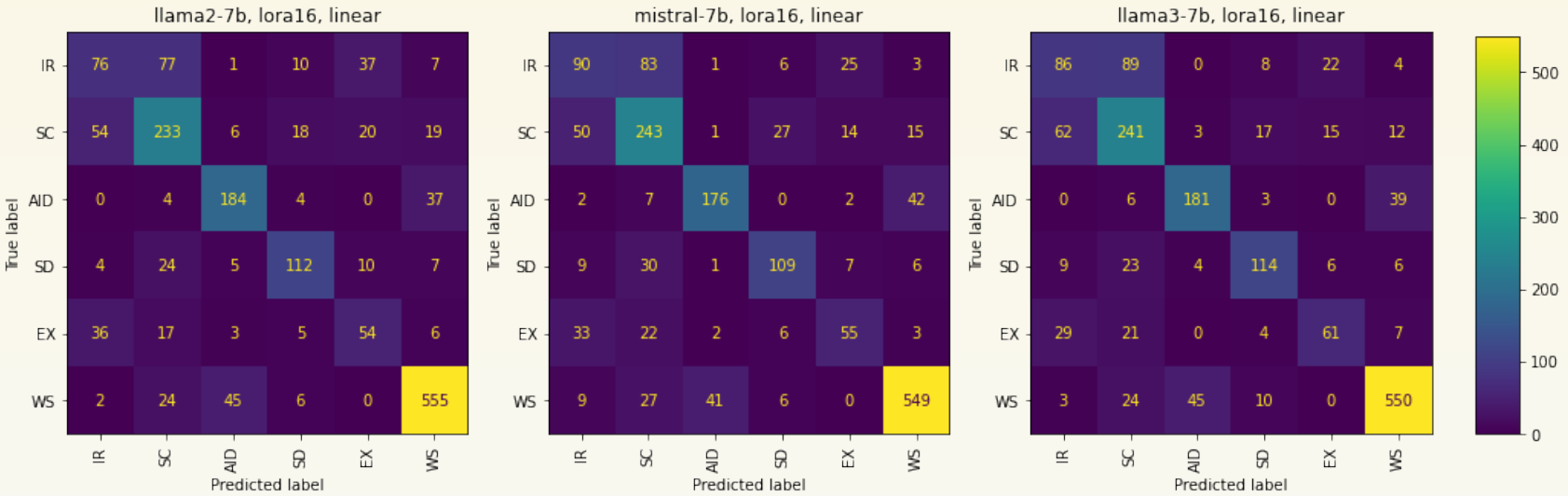


Figure 1: Confusion matrix aggregated over the five splits of the best run for each LLM.

Results: Submission Results

Model	Strategy	Macro↑			Weighted↑			Accuracy↑
		P	R	F	P	R	F	
llama2	ens	0.684	0.672	0.659	0.737	0.738	0.721	0.738
	best	0.632	0.646	0.635	0.711	0.711	0.708	0.711
	worst	0.649	0.635	0.617	0.708	0.701	0.683	0.701
mistral	ens	**0.714**	**0.697**	**0.700**	**0.753**	**0.756**	**0.749**	**0.756**
	best	0.669	0.657	0.660	0.719	0.723	0.719	0.723
	worst	0.650	0.606	0.604	0.694	0.673	0.661	0.673
llama3	ens	0.675	0.652	0.659	0.724	0.727	0.723	0.727
	best	0.630	0.611	0.614	0.689	0.701	0.691	0.701
	worst	0.638	0.626	0.629	0.696	0.699	0.695	0.699

Table 1: Our submission results for each LLM. The best overall results are in **bold**. The best results for each LLM are underlined. ↑ indicates that higher is better.

Results: Submission Results

Run ID	Macro	Weighted	SD	WS	EX	IR	SC	AID
* ORPAILLEUR_mistral-7b-ens	**0.700**	**0.749**	0.772	0.556	0.517	**0.742**	0.739	0.875
* ORPAILLEUR_mistral-7b-high	0.660	0.719	0.727	0.460	0.521	0.704	0.695	0.853
* ORPAILLEUR_llama2-7b-ens	0.659	0.721	0.798	0.494	0.365	0.676	0.722	**0.901**
* ORPAILLEUR_llama3-8b-ens	0.659	0.723	**0.802**	0.452	0.441	0.650	0.709	0.898
CYUT_llama3-fine-tuning	0.639	0.695	0.696	0.521	0.457	0.618	0.672	0.868
* ORPAILLEUR_llama2-7b-high	0.635	0.708	0.783	0.404	0.500	0.628	0.610	0.885
* ORPAILLEUR_llama3-8b-low	0.629	0.695	0.783	0.409	0.434	0.591	0.678	0.881
* ORPAILLEUR_llama2-7b-low	0.617	0.683	0.765	0.442	0.312	0.634	0.676	0.874
PunDerstand_DeBERTaSampled	0.616	0.677	0.777	0.500	0.424	0.515	0.600	0.880
* ORPAILLEUR_llama3-8b-high	0.614	0.691	0.794	0.308	0.404	0.611	0.685	0.885
PunDerstand_GuidedAnnotation	0.606	0.668	0.667	0.615	0.222	0.400	**0.900**	0.833
* ORPAILLEUR_mistral-7b-low	0.604	0.661	0.741	0.529	0.309	0.614	0.594	0.838
PunDerstand_DeBERTa	0.603	0.673	0.774	0.476	0.261	0.597	0.624	0.889
DadJokers_bert_base_uncased	0.601	0.669	0.750	0.395	0.402	0.588	0.618	0.851
NLPalma_BERTd	0.594	0.665	0.714	0.395	0.488	0.589	0.543	0.835
CodingRangers_bert_uncased	0.594	0.658	0.756	0.446	0.335	0.594	0.590	0.841
Code Rangers_roberta	0.590	0.659	0.472	**0.697**	**0.727**	0.027	0.737	0.881
Demonteam_BERTM	0.580	0.650	0.719	0.463	0.345	0.569	0.537	0.850
UAms_BERT_ft	0.522	0.602	0.755	0.418	0.054	0.589	0.484	0.832
NLPalma_PREDCNN	0.515	0.587	0.648	0.301	0.396	0.527	0.464	0.753

Table 2: Top 15 results submitted for the shared task Joker number 2. All values are the F1-score. Our submission starts with a star: "* ORPAILLEUR_..."

Qualitative Results: SD (Self-Deprecation)

- Wrongly classified examples as SD (Self-Deprecation)

a. **(IR)** My poo is green, how festive.
b. **(WS)** I'm mad at myself for not taking karate sooner.
c. **(WS)** My name is Bet. I am a cutter.
d. **(AID)** I always pronounce one word wrong. Wrong.

Qualitative Results: SD (Self-Deprecation)

- Wrongly labeled examples of SD (Self-Deprecation)

a. **(SD)** Self-deprecating humor is my cardio.

b. **(SD)** I hate this pandemic. If I wanted to waste my early 20s, I would have gotten married.

Qualitative Results: SD (Self-Deprecation)

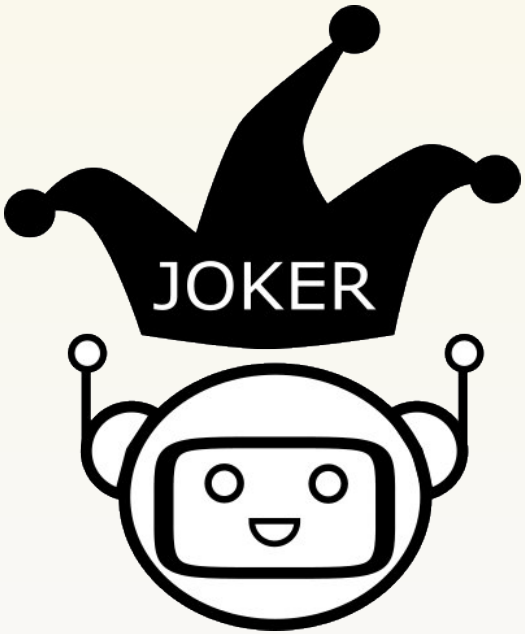
- Lexical Features vs Semantic Features
- With a "perfectly annotated dataset", is the model **always biased by the first person**?
- Study the **error ratio** to see how it evolves with the **amount of data**.

Conclusion

- We obtain first place!
- **LLMs hidden representation** works well for humor classification.
- **Ensemble strategy** is an important key for subjective task.
- Some categories of the **dataset have potential** but may benefit from **refined annotation**.

[plain,noframenumbering]

References



References ii

[1] AI@Meta. **Llama 3 Model Card.** 2024. URL: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[2] Tim Dettmers et al. **"8-bit Optimizers via Block-wise Quantization".** In: *ArXiv* abs/2110.02861 (2021).

[3] Tim Dettmers et al. **"QLoRA: Efficient Finetuning of Quantized LLMs".** In: *ArXiv* abs/2305.14314 (2023). URL: https://api.semanticscholar.org/CorpusID:258841328.

References iii

[4] Liana Ermakova et al. **"CLEF 2024 JOKER lab: Automatic humour analysis".** In: *Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, Proceedings, Part VI*. Ed. by Nazli Goharian et al. Vol. 14613. Lecture Notes in Computer Science. Cham: Springer, 2024, pp. 36–43. ISBN: 978-3-031-56072-9. DOI: 10.1007/978-3-031-56072-9_5.

[5] J. Edward Hu et al. **"LoRA: Low-Rank Adaptation of Large Language Models".** In: *ArXiv* abs/2106.09685 (2021). URL: https://api.semanticscholar.org/CorpusID:235458009.

References iv

[6] Albert Qiaochu Jiang et al. **"Mistral 7B".** In: *ArXiv* abs/2310.06825 (2023).

[7] Victor Manuel Palma Preciado et al. **"Overview of the CLEF 2024 JOKER Task 2: Humour classification according to genre and technique".** In: *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)*. Ed. by Guglielmo Faggioli et al. CEUR Workshop Proceedings. CEUR-WS.org, 2024.

[8] Hugo Touvron et al. **"Llama 2: Open Foundation and Fine-Tuned Chat Models".** In: *ArXiv* abs/2307.09288 (2023). URL: https://api.semanticscholar.org/CorpusID:259950998.