Humour Classification by Fine-tuning LLMs: CYUT at CLEF 2024 JOKER Lab Subtask Humour Classification According to Genre and Technique

Notebook for the CYUT Lab at CLEF 2024

Shih-Hung Wu[1,*,†], Yu-Feng Huang[2,†] and Tsz-Yeung Lau[3,†]

Chaoyang University of Technology, Taichung, Taiwan (R.O.C)

Abstract

This paper reports how we attend the CLEF 2024 JOKER lab, Humour classification according to genre and technique subtask. The system will classifying short texts of humor among the six classes such as irony, sarcasm, exaggeration, incongruity-absurdity, self-deprecating and wit-surprise. This year, CYUT team sent three runs based on 3 deep learning models. Run 1 is based on a fine-tuned Llama 3 model, run 2 is based on a fine-tuned RoBERTa model and run 3 is using the GPT4.0 api provided by OpenAI with a zero-shot and CoT prompt. During the system developing phrase, our Llama 3 model can achieve an 89.68% accuracy, however, the offical result is 69.78%.

Keywords

Deep Learning, Humour Classification, Large Language Models (LLMs), Llama 3, GPT-4

1. Introduction

The subtask Humour Classification According to Genre and Technique of JOKER Track @ CLEF 2024 is a multiclass classification task. [1] The system automatically classify each given sentence into the following classes: irony, sarcasm, exaggeration, incongruity-absurdity, self-deprecating and wit-surprise.

The organizers provide manually annotated training and test data from existing corpora, including the positive examples of the JOKER-2023 pun detection corpus as well as new data.

Humor is a complex and ambiguous emotional concept unique to natural language[2]. Humorous language cannot exist independently, as language gains meaning only when accompanied by context, situation, and cultural background[3]. Discourse analysis has the capability to interpret humor. Language itself becomes the subject of humor[4]. Humor recognition is a challenging issue in natural language processing (NLP) for several reasons. Firstly, humor often stems from the use of figurative language, such as irony and sarcasm. Additionally, the sense of humor varies across different cultural and geographical groups. For instance, someone disinterested in political issues may find it difficult to understand political jokes. People with different background knowledge will react differently to the same joke. This variability makes it challenging for NLP researchers to detect humorous content[5].

Humor emotion analysis is an intriguing area of study as it reveals alternative ways of expressing human emotions. When people convey various emotions through their words and actions, it is often not straightforward but filled with humorous elements. This is where humor emotion analysis becomes valuable. Previous research has primarily focused on categorizing emotions as positive, negative, or neutral [6]. However, now we aim to delve deeper into the meanings behind humorous emotions in text. Such research not only helps us better understand the diversity of human emotional expression but also provides useful insights for the development of natural language processing and emotional intelligence.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

*Corresponding author.

†These authors contributed equally.

shwu@cyut.edu.tw (S. Wu); s11227615@gm.cyut.edu.tw (Y. Huang); s10927116@gm.cyut.edu.tw (T. Lau)

0000-0002-1769-0613 (S. Wu); 0009-0002-7904-2758 (T. Lau)

Thus, humor emotion analysis is not merely a study of textual emotions but an adventurous journey into the nature of human humor. This exploration will help us comprehensively understand the psychological mechanisms behind human speech and behavior, while also bringing more enjoyment and challenges to our technological advancements.

In this study, we employ RoBERTa, GPT-4, and Llama 3-8B for humor classification. As a result, Llama 3-8B performed the best, achieving an accuracy of 89.68%.

2. Related Work

2.1. Large language models

Large language models (LLMs), such as GPT-4 [7] and Llama 3 [8], have garnered attention due to their outstanding performance on various tasks. These models possess a vast number of parameters and can adapt to new tasks without additional training, a capability known as "in-context learning." [9] Recently, the emergence of ChatGPT, particularly its basis on GPT-3.5 [9] and further refinement through reinforcement learning from human feedback, has drawn significant attention[10, 11].

2.2. Prompt Engineering

Prompt engineering plays a significant role in the fields of artificial intelligence and machine learning[12]. It acts as a communication bridge, especially when using large language models like GPT-3 or GPT-4. We perform fine-tuning [13] to achieve better results, aiming for more accurate and targeted outputs [14, 15]. This concept is crucial in natural language processing (NLP) as it directly impacts the model's performance and output quality. The basic idea of prompt engineering is to guide the model to provide the desired information or execute complex specific tasks through carefully designed prompts [16, 17]. Without clear instructions, the model might generate inaccurate or completely irrelevant responses. We can enhance the accuracy of prompts through several known practices, such as precise instructions, role assignment, give example(one-shot, few-shot)[9], iterative refinement and Chain of Thought (CoT)[18].

3. Dataset

The training dataset for this study was provided by the JOKER organizer and consists of a total of 1,742 entries. The humorous content is categorized into six types: IR (irony) with 210 entries, SC (sarcasm) with 356 entries, EX (exaggeration) with 125 entries, AID (incongruity-absurdity) with 231 entries, SD (self-deprecating) with 169 entries, and WS (wit-surprise) with 651 entries. The distribution of the training data is shown in Figure 1. The test set comprises a total of 722 entries, as illustrated in Figure 2. The results were evaluated by the JOKER organizer.

4. Method

4.1. Deep Learning Models

4.1.1. RoBERTa

We utilize the enhanced BERT[19] model, RoBERTa[20], as our baseline. BERT, which stands for Bidirectional Encoder Representations from Transformers[19], was originally introduced by Google as an encoder-only transformer[21]-based model for natural language processing (NLP) tasks. BERT is pre-trained using the Masked Language Model (MLM) and Next Sentence Prediction (NSP) techniques. Unlike word2vec[22] and GloVe[23], which do not consider context, BERT leverages contextual information during inference, leading to superior performance[19]. In the RoBERTa paper, they mentioned that the BERT model was significantly undertrained[20]. To address this, they implemented several modifications: using larger batches, training the model for a longer duration, dropping the NSP training,

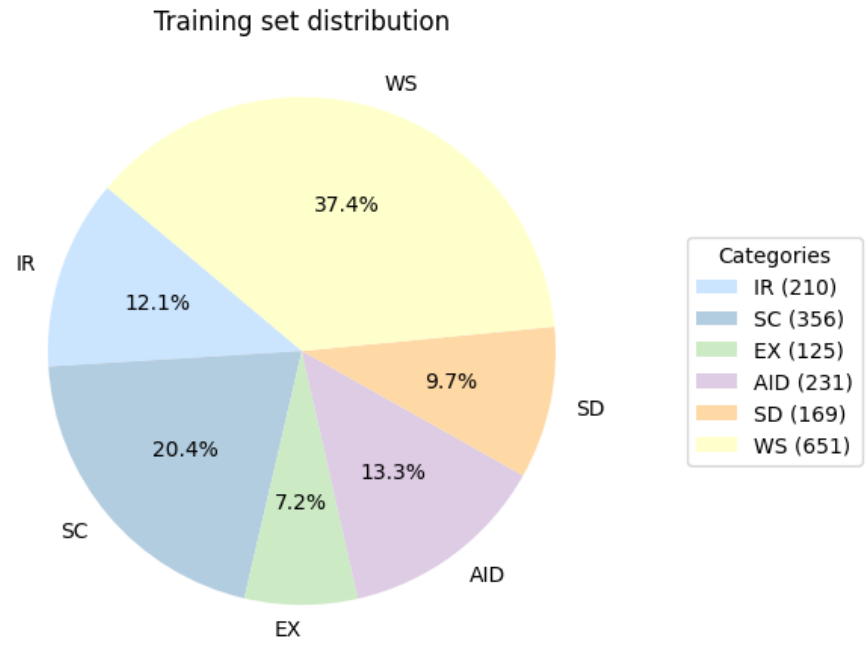

Figure 1: Training set distribution

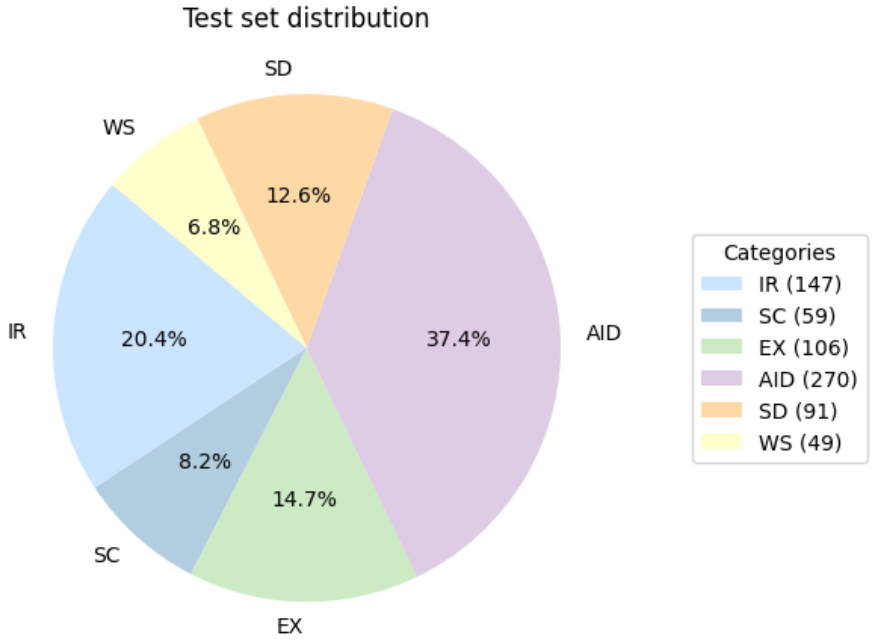

Figure 2: Test set distribution

training on longer sequences, and dynamically changing the masking pattern applied to the training data[20]. For the RoBERTa baseline model, we achieve an accuracy of 72.49%.

4.1.2. GPT-4

In this study, we utilized the GPT-4.0 model with zero-shot prompting and Chain-of-Thought (CoT) prompting to assist with the task. GPT-4, developed by OpenAI, is an advanced natural language processing model built upon its predecessor, GPT-3, with a significantly increased parameter count. This enhancement facilitates a deeper understanding and generation of complex sentence structures,

enabling more nuanced responses and better handling of contextual language features such as irony and humor. GPT-4 is trained using autoregressive language modeling on a diverse dataset, allowing it to perform exceptionally across various NLP tasks like translation, summarization, and question answering.[24]

The model operates by first converting input text into tokens, which are processed by transformer layers using attention mechanisms to evaluate relevance and context. These mechanisms generate intermediate representations of data, which are then decoded into human-readable text. GPT-4 incorporates a randomness function influenced by temperature settings and top-k sampling, which dictate the randomness and determinism of the output, thus enhancing the model's ability to produce contextually appropriate content. This process represents a significant evolution in language model capabilities, setting new benchmarks in language understanding and generation.[24]

4.1.3. Llama 3

Large Language Models (LLMs) are highly capable AI assistants that excel in complex reasoning tasks. They enable interaction with humans through intuitive chat interfaces, leading to rapid and widespread adoption among the general public.[25] Many different LLMs are publicly available, such as GPT-4[7], Mistral 7B[26], Gemma 7B[27], and the LLM we utilize in this study, Llama 3.

Llama 3[8] is an open-source LLM utilizing the Transformer[21] architecture, developed by Meta. The Llama3 model is available in configurations with 8 billion and 70 billion parameters. Llama3 models have achieved state-of-the-art (SOTA) performance across a broad range of tasks due to extensive pre-training on over 15 trillion data tokens, making it the best-performing open-source model. In this study, we fine-tuned the Llama 3-8B model on a single GPU, utilizing 4-bit quantization with QLoRa[28] to reduce GPU RAM usage during training with unsloth[29]. As a result, the model achieved 89.68% accuracy.

5. System Development

5.1. Environment

In our experiment, we utilized a GPU, NVIDIA GeForce RTX 3090 with 24GB of memory. The versions of all packages employed in the experiment will be thoroughly delineated in Table 1.

Table 1
Packages Version

Package	Version
Python	3.10.14
Pytorch	2.2.2
CUDA Toolkit	12.1
CUDA	8.6
Unsloth	2024.4

5.2. RoBERTa

To fine-tune the RoBERTa model, we use 80% of the dataset as the training set and 20% as the test set. The hyperparameters we used for fine-tuning are shown in Table 2.

5.3. GPT 4.0

To evaluate the GPT-4 model, we use the entire dataset for self-testing. We found that direct classification did not yield satisfactory results, so we grouped similar types into broader categories before conducting finer classifications. First, we grouped IR and SC into Category C. Then, we divided the remaining five

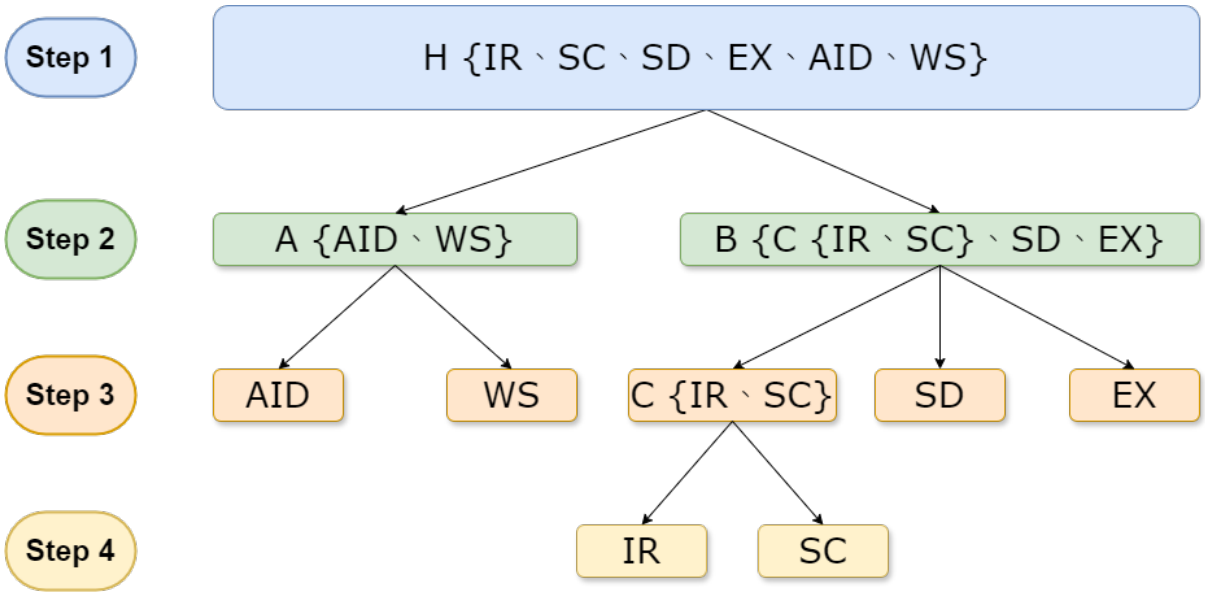

Figure 3: Process of GPT-4 classification with clustering.

types (C, SD, EX, AID, WS) into two categories: Category A (AID, WS) and Category B (C, SD, EX). We then performed a binary classification within Category A to distinguish between AID and WS. For Category B, we conducted a three-way classification to separate C, SD, and EX. Finally, we performed a binary classification within Category C to differentiate between IR and SC. This approach allowed us to consolidate the results for all six types. The flowchart displayed in figure 3 illustrates this classification process. You can check the prompts we applied for each step in the appendix Table 13.14 .

5.3.1. Prompt Design

First, we assign the model a specialized role to enhance its performance in handling complex tasks within a specific domain. Next, we utilize chain-of-thought (CoT) prompting to reduce model hallucinations and increase the probability of generating reasonable responses. We specify the task clearly, provide category names and definitions, and set output constraints. For example, we limit the output to no more than three tokens and restrict the model from producing responses outside the given requirements or repeating the question.

Table 2
Hyperparameters for RoBERTa fine-tuning

Hyperparameters	Value
BERT Model	base
Epochs	5
Batch Size	4
Optimizer	Adam
Learning Rate	1e-5

5.4. Llama 3

To fine-tune the Llama 3-8B model, we use 80% of the dataset as the training set and 20% as the test set. The hyperparameters we used for fine-tuning are shown in Table 4.

Table 3
Stanford Alpaca Format

Model	Format	Prompt
Llama 3-8B	Alpaca Format	Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request. ### Instruction: {} ### Input: {} ### Response: {}

5.4.1. Prompt Design

To fine-tune Llama3, we utilize the Stanford Alpaca Format[30]. The Alpaca format is shown in Table 3. For the instruction, we first tell the model what to do: "Classify the following text into one of the classes." Then, we provide the six classes for classification with explanations: irony, sarcasm, exaggeration, incongruity-absurdity, self-deprecating humor, and wit-surprise. We simply utilize the explanations provided in the official JOKER guideline document here. Based on results from RoBERTa and GPT-4, we discovered that the model struggled to accurately classify irony and sarcasm. Therefore, we added the sequence: "You ought to focus more on classifying irony and sarcasm." Finally, we applied Chain of Thought (CoT) prompting[18] by adding the sequence: "Let's think step by step.".

Meanwhile, the sequence following "### Input:" denotes the text in need of classification, while "### Response:" following with one of six classes: irony, sarcasm, exaggeration, incongruity-absurdity, self-deprecating humor, and wit-surprise. During evaluation, we employ the same prompting technique. The only difference is that we refrain from adding any text after "### Response:" to allow the model to generate the response. The prompt elements are shown in Table 5.

Table 4
Hyperparameters for Llama 3 fine-tuning

Hyperparameters	Value
Parameters	8 billion
Epochs	6
Gradient Accumulation	4
Optimizer	AdamW 8bit
Learning Rate	2e-4
QLoRa	4-bit quantization

6. Experiment Result

6.1. Self-Test Result

Table 6 evaluates the performance of each model. The RoBERTa model achieved an accuracy of 71.63%, 0.64 Macro Average Precision (MAP), 0.65 Macro Average Recall (MAR) and 0.64 Macro Average F1-Score (MA-F1), serving as the baseline. The GPT-4 model chieved an accuracy of 36.24%, 0.37 MAP, 0.34 MAR and 0.34 MA-F1, representing a significant drop compared to the baseline model. The GPT-4 model with Clustering . chieved an accuracy of 38.23%, 0.39 MAP, 0.40 MAR and 0.34 MA-F1, also representing a significant drop compared to the baseline model. However, with clustering, the model performs slightly better. The Llama 3-8B model achieved an accuracy of 89.68%, 0.89 MAP, 0.87 MAR and 0.88 MA-F1, representing a 18.05%, 0.25, 0.23, 0.24 increase compared to the baseline model.

Table 5
Llama 3-8B Prompt

Model	Prompt Element	Prompt
Llama 3-8B	Instruction	Classify the following text into one of the classes. Here are the six types of classes: Irony - Irony relies on a gap between the literal meaning and the intended meaning, creating a humorous twist or reversal. Sarcasm - Sarcasm involves using irony to mock, criticize, or convey contempt. Exaggeration - Exaggeration involves magnifying or overstating something beyond its normal or realistic proportions. Incongruity-Absurdity - Incongruity refers to unexpected or contradictory elements that are combined in a humorous way, and Absurdity involves presenting situations, events, or ideas that are inherently illogical, irrational, or nonsensical. Self-deprecating - Self-deprecating humor involves making fun of oneself or highlighting one's own flaws, weaknesses, or embarrassing situations in a lighthearted manner. Wit-Surprise - Wit refers to clever, quick, and intelligent humor, and Surprise in humor involves introducing unexpected elements, twists, or punchlines that catch the audience off guard. You ought to focus more on classifying irony and sarcasm. Let's think step by step.
	Input	{ text in need of classification from dataset. }
	Response	{ one of six classes from dataset: irony, sarcasm, exaggeration, incongruity-absurdity, self-deprecating humor, and wit-surprise. }

*During the evaluation, leave the "Response" empty.

Table 6
Models Performance of Self-Testing.

Run	Model	Accuracy (%)	MAP	MAR	MA-F1
1	Llama 3-8B	**89.68**	**0.89**	**0.87**	**0.88**
2	GPT-4 with Clustering	38.23	0.39	0.40	0.34
3	RoBERTa	71.63	0.64	0.65	0.64
-	GPT-4	36.24	0.37	0.34	0.34

*MAP: Macro Average Precision
*MAR: Macro Average Recall
*MA-F1: Macro Average F1 Score

6.2. Official Result

All of the models were evaluated by the JOKER organizer [31]. Table 10 presents the official results of each model. The RoBERTa model achieved an accuracy of 18.56%, 0.19 Macro Average Precision (MAP), 0.24 Macro Average Recall (MAR), and 0.21 Macro Average F1-Score (MA-F1). The RoBERTa model showed a significant drop in accuracy compared to our self-test results due to a mistake in our code. The data uploaded for the official result was fine-tuned on extra data containing IR and SC, leading to lower performance than expected. The GPT-4 model with clustering achieved an accuracy of 35.53%, 0.39 MAP, 0.40 MAR, and 0.34 MA-F1. The GPT-4 model with clustering produced results similar to our self-testing.

The Llama 3-8B model used for evaluation is the same model fine-tuned with 80% of the dataset. It

Table 7
Precision, Recall and F1-Score of each class and model. (Self-Testing)

Model	Class	Precision	Recall	F1-Score
Llama 3-8B	IR	0.86	0.84	0.85
	SC	0.86	0.89	0.88
	EX	0.85	0.85	0.85
	AID	0.90	0.79	0.84
	SD	0.96	0.92	0.94
	WS	0.92	0.96	0.94
GPT-4 with Clustering	IR	0.27	0.13	0.18
	SC	0.63	0.13	0.21
	EX	0.18	0.64	0.28
	AID	0.22	0.24	0.23
	SD	0.50	0.74	0.60
	WS	0.53	0.51	0.52
RoBERTa	IR	0.47	0.33	0.38
	SC	0.63	0.61	0.62
	EX	0.52	0.52	0.52
	AID	0.75	0.78	0.76
	SD	0.60	0.76	0.68
	WS	0.87	0.90	0.88

Table 8
Accuracy of GPT-4 with Clustering for each step.

Model	Step	Classes	Accuracy (%)
GPT-4 with Clustering	2	**A** (AID, WS), **B** (C(IR, SC), SD, EX)	78.18
	3	**AID**, **WS**	50.34
	3	**C**(IR, SC), **SD**, **EX**	32.10
	4	**IR**, **SC**	41.95

Table 9
Precision, Recall and F1-Score of each class of Llama 3-8B. (Offical Result)

Model	Class	Precision ↑	Recall ↑	F1-Score ↑
Llama 3-8B	IR	0.63(-0.23)	0.60(-0.24)	0.62(-0.23)
	SC	0.67(-0.19)	0.68(-0.21)	0.67(-0.21)
	EX	0.52(-0.33)	0.41(-0.44)	0.46(-0.39)
	AID	0.86(-0.04)	0.88**(+0.09)**	0.87**(+0.03)**
	SD	0.70(-0.26)	0.69(-0.23)	0.70(-0.24)
	WS	0.44(-0.48)	0.63(-0.33)	0.52(-0.42)

*Blue words represent the differences compared to self-testing.

achieved an accuracy of 69.78%, 0.64 MAP, 0.65 MAR, and 0.64 MA-F1. The Llama 3-8B model exhibited a significant drop in accuracy compared to our self-test results, potentially due to differences in the data distribution between the training and test sets, as shown in Figure 4. However, as seen in Table 9, the model performed exceptionally well on the class AID, even with a small amount of training data. It appears that AID has distinctive features that the model can learn effectively. The model likely overfitted to the training set, impairing its performance on the test set. Balancing the data in the training set may help improve the model's robustness. From the official results, it is evident that the Mistral-7B model performed the best overall in humor classification, achieving an accuracy of 76%, from team ORPAILLEUR[31].

Figure 4: Comparison of Training set and Test set Distributions

Table 10
Models Performance of Official Result.

Run	Model	Accuracy ↑(%)	MAP ↑	MAR ↑	MA-F1 ↑
1	Llama 3-8B	**69.78**(-19.90)	**0.64**(-0.25)	**0.65**(-0.22)	**0.64**(-0.24)
2	GPT-4 with Clustering	35.53(-02.70)	0.39(-0.00)	0.40(-0.00)	0.34(-0.00)
3	RoBERTa	18.56(-53.07)	0.19(-0.45)	0.24(-0.41)	0.21(-0.43)

*MAP: Macro Average Precision
*MAR: Macro Average Recall
*MA-F1: Macro Average F1 Score
*Blue words represent the differences compared to self-testing.

7. Discussion & Error Analysis

7.1. Discussion

From the confusion matrix of RoBERTa and GPT-4, Figure 7 and Figure 6, it is evident that the models struggled to accurately classify between the categories AID and WS, as well as IR and SC. One reason for this difficulty is the existence of two distinct types of irony: verbal irony and situational irony. Verbal irony, often referred to as sarcasm, implies that IR includes SC[32]. Another reason is that identifying sarcasm in a sentence often requires contextual information[32]. Meanwhile, Llama 3-8B demonstrated significantly better performance in the areas where RoBERTa and GPT-4 exhibited weaknesses, as shown in Figure 8. GPT-4 with clustering shows a slight improvement compared to the vanilla GPT-4 model, from Figure 6.

From Table 7, it is not hard to discover that fine-tuning LLMs is an effective method for humor classification. Llama 3 has significantly better performance compared to GPT-4, with substantial improvements in precision, recall, and F1-score for each class. Although non-tuned LLMs have great general performance, they might not excel in specialized tasks. Even if fine-tuning LLMs is not available, using smaller models like RoBERTa can also achieve acceptable performance.

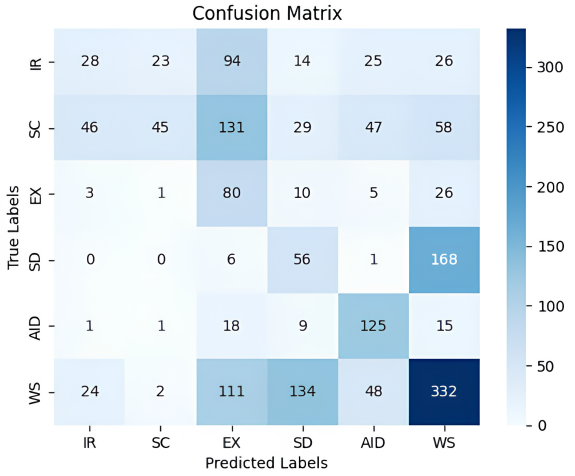

Figure 5: Confusion matrix of GPT4 with Clustering self-testing.

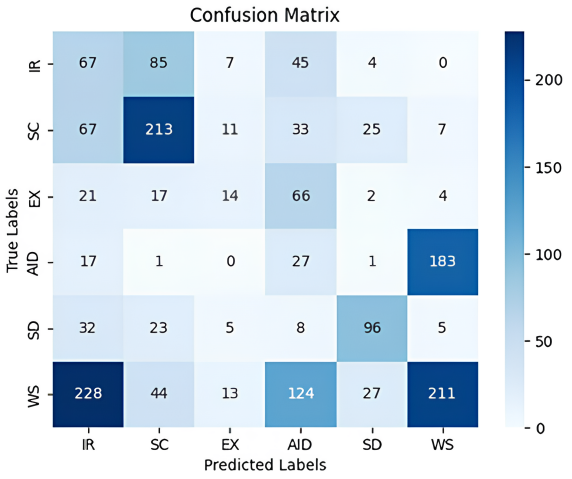

Figure 6: Confusion matrix of GPT4 self-testing.

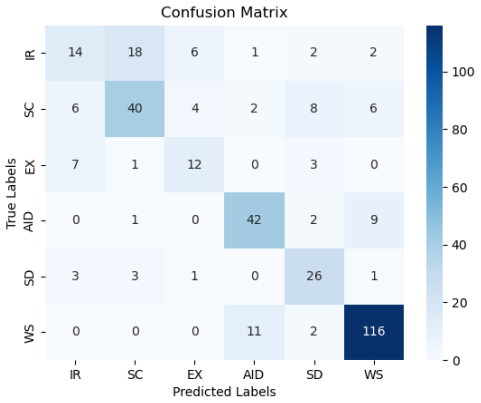

Figure 7: Confusion matrix of RoBERTa self-testing.

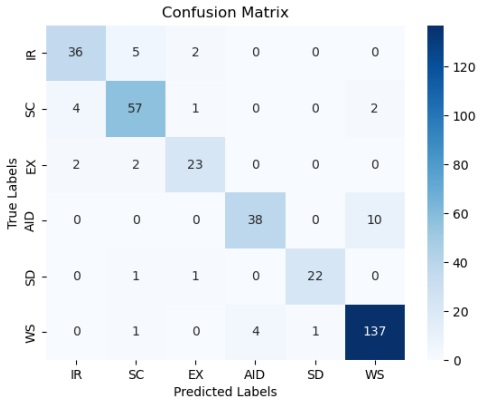

Figure 8: Confusion matrix of Llama 3-8B self-testing.

7.2. Error Analysis

7.2.1. Llama 3

The model may occasionally produce unexpected responses, which can be attributed to the pre-training data. For instance, if the input text is: "When negotiating whether to share your french fries, you have quite a few bargaining chips.", the model might respond with: "lunch." In self-testing, 12 samples produced unexpected outputs. Additional examples are provided in Table 11.

This is the limitation of fine-tuning generative models such as LLMs. When employing the BERT model for classification, the [CLS] token is inputted into the Multilayer Perceptron (MLP) [33]. The model ensures the absence of unexpected output by maintaining a fixed output layer size and employing the softmax function[34] to determine the probability of each output.

We take an additional step to test those errors with ten more opportunities. Some of these errors can be classified into one of the six classes. For instance, consider the input text 1: input text 1: "No longer a female as I refuse to wear heels ever again" Llama 3-8B give an unexpected responce: "twitter", but 1 out of 10 times, it give a responce "sarcasm". The same phenomenon occurred with input text 2: "The leopard tried creeping up on the tigers using its camouflage but it was seen.", which received a "wit-surprise" response 1 out of 10 times. Additionally, input text 8, "Doppelherz. The power of the

Table 11
Samples with unexpected responce.

Error Sample	Input text	Response
1	No longer a female as I refuse to wear heels ever again.	twitter
2	The leopard tried creeping up on the tigers using its camouflage but it was seen.	leopard
3	Time is important to fullbacks. They are always rushing.	football
4	Your road to driving success.	infographic
5	When negotiating whether to share your french fries, you have quite a few bargaining chips.	lunch
6	Nothing can compare to picnicking on a French hillside and savoring the bries.	food
7	Waiters are good at multiplication because they know their tables.	table
8	Doppelherz. The power of the two hearts.	wikipedia
9	Let yourself be transported with every bite. Introducing our new Destination Series.	campaign
10	Looking for a delicious way to stay cool now that it's heating up?	ice
11	Most Manchester United fans will only drink tea because they have all the cups.	manchester
12	Where PR = Public Reactions.	hashtag

two hearts," elicited a "wit-surprise" response 8 out of 10 times. These examples are shown in Table 12. Meanwhile, other input texts remained unchanged.

Table 12
Samples response with one out of six classes in ten trials.

Error Sample	Input text	Response
1	No longer a female as I refuse to wear heels ever again.	sarcasm
2	The leopard tried creeping up on the tigers using its camouflage but it was seen.	wit-surprise
8	Doppelherz. The power of the two hearts.	wit-surprise

8. Conclusion & Future Work

8.1. Conclusion

In this study, we conducted humor classification using deep learning models(RoBERTa), including LLMs such as Llama 3-8B and GPT-4. The best performing model was Llama 3-8B, achieving an accuracy of 89.68% in self-testing and 69.78% in offical result through fine-tuning and prompt engineering. We also analyzed some unexpected responses from the LLMs to understand why they occurred.

In summary, we found that fine-tuning LLMs can be very effective for humor classification. Additionally, we discovered that clustering similar classes allows LLMs to achieve better performance.

8.2. Future Work

For future work, we can observe through the confusion matrix that EX and SD can be grouped into a single category. This approach may improve overall accuracy. Additionally, the LLM could first score the humor type present in the sentences and then classify based on a set threshold. Furthermore, the clustering method can be applied to Llama 3-8B, which might also result in better performance.

Acknowledgments

This study was supported by the National Science and Technology Council under the grant number NSTC 113-2221-E-324-009.

References

[1] L. Ermakova, A.-G. Bosser, T. Miller, T. Thomas, V. M. P. Preciado, G. Sidorov, A. Jatowt, Clef 2024 joker lab: Automatic humour analysis, in: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2024, pp. 36–43.

[2] Z. Li, J. Liu, Y. Wang, Performance analysis on deep learning models in humor detection task, in: 2022 International Conference on Machine Learning and Knowledge Engineering (MLKE), IEEE, 2022. URL: http://dx.doi.org/10.1109/MLKE55170.2022.00023. doi:10.1109/mlke55170.2022.00023.

[3] P. Liang, Discourse analysis on humor, in: 2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), 2011, pp. 5002–5005. doi:10.1109/AIMSEC.2011.6011180.

[4] P. Liang, Discourse analysis on humor, in: 2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), IEEE, 2011. URL: http://dx.doi.org/10.1109/AIMSEC.2011.6011180. doi:10.1109/aimsec.2011.6011180.

[5] Y. Guo, L. Kong, Classification and regression combined model on accessing humor score with explanatory feature, in: 2022 International Conference on Machine Learning and Knowledge Engineering (MLKE), IEEE, 2022. URL: http://dx.doi.org/10.1109/MLKE55170.2022.00050. doi:10.1109/mlke55170.2022.00050.

[6] H. A. Sayyed, S. Rushikesh Sugave, S. Paygude, B. N Jazdale, Study and analysis of emotion classification on textual data, in: 2021 6th International Conference on Communication and Electronics Systems (ICCES), 2021, pp. 1128–1132. doi:10.1109/ICCES51350.2021.9489204.

[7] OpenAI, Gpt-4 technical report, 2024. arXiv:2303.08774.

[8] Meta, Introducing Meta Llama 3: The most capable openly available LLM to date — ai.meta.com, https://ai.meta.com/blog/meta-llama-3/, 2024. [Accessed 29-05-2024].

[9] OpenAI, Language models are few-shot learners, 2020. arXiv:2005.14165.

[10] S. Pitis, M. R. Zhang, A. Wang, J. Ba, Boosted prompt ensembles for large language models, 2023. arXiv:2304.05970.

[11] Q. Guo, R. Wang, J. Guo, B. Li, K. Song, X. Tan, G. Liu, J. Bian, Y. Yang, Connecting large language models with evolutionary algorithms yields powerful prompt optimizers, 2024. arXiv:2309.08532.

[12] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, A. Chadha, A systematic survey of prompt engineering in large language models: Techniques and applications, 2024. arXiv:2402.07927.

[13] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, 2021. arXiv:2107.13586.

[14] Q. Ye, M. Axmed, R. Pryzant, F. Khani, Prompt engineering a prompt engineer, 2024. arXiv:2311.05661.

[15] S. Ekin, Prompt engineering for chatgpt: A quick guide to techniques, tips, and best practices (2023). URL: http://dx.doi.org/10.36227/techrxiv.22683919. doi:10.36227/techrxiv.22683919.

[16] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, D. C. Schmidt, A prompt pattern catalog to enhance prompt engineering with chatgpt, 2023. arXiv:2302.11382.

[17] X. Amatriain, Prompt design and engineering: Introduction and advanced methods, 2024. arXiv:2401.14423.

[18] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, 2023. arXiv:2201.11903.
[19] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.
[20] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. arXiv:1907.11692.
[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2023. arXiv:1706.03762.
[22] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, 2013. arXiv:1301.3781.
[23] J. Pennington, R. Socher, C. Manning, GloVe: Global vectors for word representation, in: A. Moschitti, B. Pang, W. Daelemans (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1532–1543. URL: https://aclanthology.org/D14-1162. doi:10.3115/v1/D14-1162.
[24] B. Chen, Z. Zhang, N. Langrené, S. Zhu, Unleashing the potential of prompt engineering: a comprehensive review, 2024. arXiv:2310.14735.
[25] H. T. et al., Llama 2: Open foundation and fine-tuned chat models, 2023. arXiv:2307.09288.
[26] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. arXiv:2310.06825.
[27] G. Team, Gemma: Open models based on gemini research and technology, 2024. arXiv:arXiv:2403.08295.
[28] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: Efficient finetuning of quantized llms, 2023. arXiv:2305.14314.
[29] D. Han, M. Han, H. H. Nguyen, Qubitium, Y. Belkada, Z, unslothai/unsloth, 2024. URL: https://github.com/unslothai/unsloth.
[30] R. Taori*, I. Gulrajani*, T. Zhang*, Y. Dubois*, X. Li*, C. Guestrin, P. Liang, T. B. Hashimoto, Alpaca: A strong, replicable instruction-following model, https://crfm.stanford.edu/2023/03/13/alpaca.html, 2021. [Accessed 29-05-2024].
[31] L. Ermakova, A.-G. Bosser, T. Miller, V. M. P. Preciado, G. Sidorov, A. Jatowt, Overview of the clef 2024 joker track automatic humour analysis, 2024.
[32] E. Filatova, Irony and sarcasm: Corpus generation and analysis using crowdsourcing, in: N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, 2012, pp. 392–398. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/661_Paper.pdf.
[33] M.-C. Popescu, V. Balas, L. Perescu-Popescu, N. Mastorakis, Multilayer perceptron and neural networks, WSEAS Transactions on Circuits and Systems 8 (2009).
[34] J. Bridle, Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters, in: D. Touretzky (Ed.), Advances in Neural Information Processing Systems, volume 2, Morgan-Kaufmann, 1989. URL: https://proceedings.neurips.cc/paper_files/paper/1989/file/0336dcbab05b9d5ad24f4333c7658a0e-Paper.pdf.

A. Appendix

A.1. Our fine-tuned Llama 3-8B model for Humor Classification

The fine-tuned Llama 3-8B model is available on Hugging Face.

- Hugging Face

A.2. Prompt of GPT-4 with clustering for each step.

Table 13
Prompt of GPT-4 with clustering for each step.

Model	Step	Clases	Prompt
GPT-4 with Clustering	2	**A**, **B**	As a Humor Master, your task is to identify the type of humor from the following two categories . Take it step by step. This is a multi-category classification task. The aim is to automatically classify text according to the following classes: A,B. There are two humor types. Here are the two type of humour: A: These genres are primarily based on unexpected elements or clever twists for humorous effect. B: Usually involves exaggerating or distorting reality, or achieving humorous effects by teasing oneself or others. ###Limit number of words: no more than 3 tokens### ###Please answer directly without restating the question### ###Instructions: For each question, respond using only one of the following abbreviations :A,B. Do not reply with answers other than A,B.###
	3	**AID**, **WS**	As a Humor Master, your task is to identify the type of humor from the following two categories . Take it step by step. This is a multi classification task. The aim is to automatically classify text according to the following classes: WS,AID. There are two humor types. Here are the two type of humour: WS:Includes humor that uses intelligence and wit to elicit laughter through clever language or thought patterns. This type of humor may involve puns, quips, or logical deductions, allowing people to appreciate the author's intelligence and creativity. AID:Includes humor that utilizes elements that defy common sense or logic, or combines unrelated things to create a sense of absurdity or incongruity. This type of humor often surprises and confuses people because it goes against our expectations. ###Limit number of words: no more than 3 tokens### ###Please answer directly without restating the question### ###Instructions: For each question, respond using only one of the following abbreviations:WS,AID. Do not reply with answers other than WS,AID.###

Table 14
Prompt of GPT-4 with clustering for each step.

Model	Step	Clases	Prompt
GPT-4 with Clustering	3	**C**, **SD**, **EX**	As a Humor Master, your task is to identify the type of humor from the following three categories. Take it step by step. This is a multi classification task. The aim is to automatically classify text according to the following classes: IR,EX,SD. There are three humor types. Here are the three type of humour: IR:Includes irony, which relies on the gap between literal meaning and actual intent to create humor, and sarcasm, which is used specifically to mock, criticize, or express contempt. SD:Covers self-deprecating humor that amuses audiences by highlighting personal flaws, weaknesses, or embarrassing situations in a light-hearted way. EX:Involves exaggerating something, exaggerating certain features beyond normal or realistic proportions to create a humorous effect. ###Limit number of words: no more than 3 tokens### ###Please answer directly without restating the question### ###Instructions: For each question, respond using only one of the following abbreviations:IR,SD,EX. Do not reply with answers other than IR,SD,EX.###
	4	**IR**, **SC**	As a Humor Master, your task is to identify the type of humor from the following two categories . Take it step by step. This is a multi classification task. The aim is to automatically classify text according to the following classes: IR,SC. There are two humor types. Here are the two type of humour: IR:Focuses on exploiting the discrepancy between literal meaning and actual intent to create humor, often by reversing or twisting expectations. SC:Focus on the use of irony to ridicule, criticize, or express contempt, often with a certain sharpness or criticalness. ###Limit number of words: no more than 3 tokens### ###Please answer directly without restating the question### ###Instructions: For each question, respond using only one of the following abbreviations:IR,SC. Do not reply with answers other than IR,SC.###