NLPalma Joker 2024: Yet, no Humor with Humorousness - Task 2 Humour Classification According to Genre and Technique*

Victor Manuel Palma-Preciado,[1, 2*,†], Carolina Palma-Preciado[1,†] and Grigori Sidorov[1]

[1] Instituto Politécnico Nacional (IPN), Centro de Investigacion en Computacion (CIC), Mexico City, Mexico
[2] Université de Bretagne Occidentale, HCTI, France

Abstract

The following work aims to describe the team participation in JOKER 2024, which focuses on developing various methods for classifying text that exhibit different techniques and humorous intentions. Understanding such aspects of humor can often be challenging for human beings. By classifying humor into these categories, we aim to establish more robust methods for classification, which can be applied across various fields of study. Current models offer high potential for training and fine-tuning complex tasks like humor classification. This ranges from the traditional use of Convolutional Neural Networks (CNNs) to the widely utilized modern Transformer paradigm BERT-like models. The results were mixed, as different approaches were chosen. It is believed that, given their performance, the models can still be optimized and their accuracy improved. Overall, the results are satisfactory for a first approach using the usual BERT-like model and embeddings such a USE with a CNN.

Keywords

BERT, Natural language processing, Humour classification, Humour, Wordplays, Jokes.

1. Introduction

The main objective of this work is to find robust methods to achieve good results in Task 2 "Classification According to Genre and Technique "of JOKER CLEF 2024[5] for different types of humor. In Task 2, the model must classify sentences containing a wide range of humorous constructs into different classes. This task is based on the English dataset of JOKER 2024 [2]. The primary goal is to accurately perform multiclass classification, automatically categorizing text into the following classes: irony, sarcasm, exaggeration, incongruity-absurdity, self-deprecating, and wit-surprise. The aim is to develop a model capable of clearly identifying these classes.

The study of humor is an underexplored topic, making resources such as corpora and models trained in different kinds of humor difficult to obtain or non-existent. Humor is

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

* Corresponding author.

† These authors contributed equally.

victorpapre@gmail.com (A. 1); cpalmap2020@cic.ipn.mx (A. 2); sidorov@cic.ipn.mx (A. 3)
0000-0001-8711-1106 (A. 1); 0000-0003-3253-4464 (A. 2); 0000-0003-3901-3522 (A. 3)

believed to be imbued with cultural characteristics, increasing the complexity of understanding humorous expressions and making humor subjective and challenging to tackle.

Since humans find it difficult to generalize humor, certain features can help machines understand it in a specific way, as their ability to compute similarities is stronger than that of humans. Consequently, humor detection methods can infer some aspects of humor better than humans, who interpret it subjectively. Access to a sufficiently robust dataset and baseline provides a measure to scale studies in the field of humor [10].

The main objective of this work is to find robust methods to achieve good results in Task 2 for different types of humor. In Task 2, the model must classify sentences containing a wide range of humorous constructs into different classes. This task is based on the English dataset of JOKER 2024 [2]. Where the primary goal is to accurately perform multiclass classification, automatically categorizing text into the following classes: irony, sarcasm, exaggeration, incongruity-absurdity, self-deprecating, and wit-surprise. The aim is to develop a model capable of clearly identifying them.

Adapting various methods and models to achieve the desired results for each class should be the main approach. It is important to note that each class contains a different number of examples. Taking this into account can ease the training process by maintaining balance among the classes. Additionally, we need to consider whether there is sufficient data to train the model. If not, we should increase the data using other methods to ensure better results.

Classifying humorous sentences presents a unique challenge because some sentences unintentionally resemble others, causing confusion. For example, irony and sarcasm can often blur together, as can incongruity-absurdity and exaggeration, making them tricky to interpret even for humans. Therefore, we need a method that excels at differentiating these nuances. Understanding the types of sentences our model needs to discern between each class is crucial, as the similarity between them can hinder and confuse the model, introducing noise that must be addressed.

2. Materials and Methods

This section details the three key stages of the experiments. First, it describes how the dataset for the classification task was composed and process. Next, it outlines the selection process, workflow, and configuration of the models. Finally, it details the resources used.

2.1. Dataset

The provided dataset for Task 2, as previously described, is a multiclassification dataset as it contains text labeled in six different class labels: irony (IR), sarcasm (SC), exaggeration (EX), incongruity-absurdity (AID), self-deprecating (SD), and wit-surprise (WS).

Since the aim is to compare different approaches made by the participants, the organizers have prepared a train and test sub dataset for this purpose. For the training dataset, 1,742 samples were given, and their distribution among the classes can be seen in (Figure 1), where class WS has the most samples and EX has the fewest.

The models were trained on the English corpus consisting wordplays, which was further divided into an internal training and validation set with a 70:30 (1,219:523) for a hold-out stratify validation method. Then each model was then used to evaluate the test set provided by JOKER, comprising 6,642 sentences.

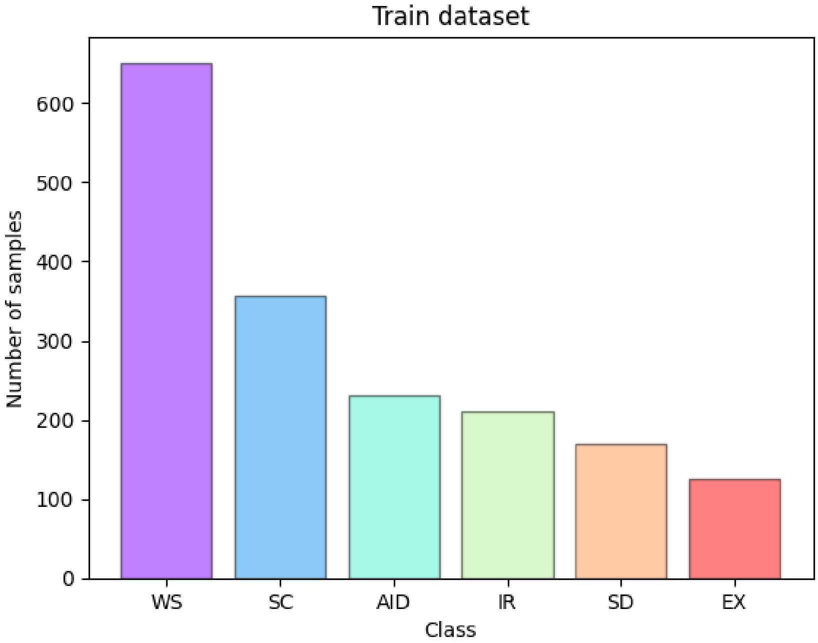


Figure 1: Train dataset frequency by classes.

2.2. Model selection

As part of the model selection different approaches were taken the structure of the treatments is as follows:

- Transformers Paradigm Models (BERT, BERT multilingual, DistilBERT, among others).
- CNN (Convolutional Neural Network) + Universal Sentence Encoder.

Since experimentation can be time-consuming, it is often necessary to repeat them to improve results. Specifically, generating embeddings can take a long time. Therefore, using increasingly powerful computers becomes a necessity to reduce computational times and push the models further, allowing multiple runs of experimentation simultaneously. In this case, the Google Colab Pro environment allows the utilization of GPU resources without runtime limits, facilitating the experimentation stage.

The best yield results of our methods are taken for submission, the firsts treatments were based one of the most know constructs the transformers BERT-like, and the second one based on a quite simple CNN structure paired with a quite powerful representation of embeddings (USE).

A multilingual model such as multilingual BERT, pre-trained in 104 languages, was initially considered for training [8]. However, with the purpose of finding the best model during evaluation, it was decided to keep the model but apply a separate approach. As a result, two models were trained: one specifically for English and one with multilingual capabilities, both utilizing the same BERT architecture [6].

The models were trained with BERT [4] and multilingual BERT [3], loaded from the Hugging Face transformers module with the help of the Ktrain wrapper [1], which facilitated the process of loading and fine-tuning the models quickly and simply.

Since transformers do not require extensive preprocessing [9], the training process was relatively straightforward. However, during the fine-tuning phase, it was necessary to experiment with different parameters to achieve the best results in validation accuracy and loss. The best models were saved in Keras H5 format for future reference. For the training of the BERT-like models, the following parameters were utilized: a batch size of 32, 8 training epochs, and a learning rate of 5e-5.

In the case of the CNN, KERAS was used, as simple and easy form to deploy this structure, for the embedding (USE) due to changes on the Tensorhub platform, Kaggle was used to handle the embeddings obviously with a different checkpoint as the one on Tensorhub, in this case 512 characteristics were taken from the USE model and then utilizing a wide variety of optimizers for the multiclass classification. The creation and approach of the network was influence from the work of [7], of course with our own twist was taken, in this case lower blocks of CNN, with two and three stacked convolutional blocks after a few attempts with different architectures, two blocks of simple connected convolutional network yield the best results, using a learning rate of 3e-1 and 5e-6. We obtained mixed results but in comparation with the BERT-like model, were not as good.

2.3. Resources

Among the resources used to train and evaluate the models, Google's Colab environment was employed. This platform enables Python programming and execution, while also providing easier use of GPU. The use of GPU resources allowed for faster execution in the performance of the described tasks, by enabling multiple simultaneous computations. The server used has the following specifications: GPU NVIDIA-SMI 525.85.12, CUDA v12.0, and 25 of RAM. As part of the resources, BERT-like models that are less demanding in the use of resources, it is necessary to consider the data volume and the desired width of tokenization.

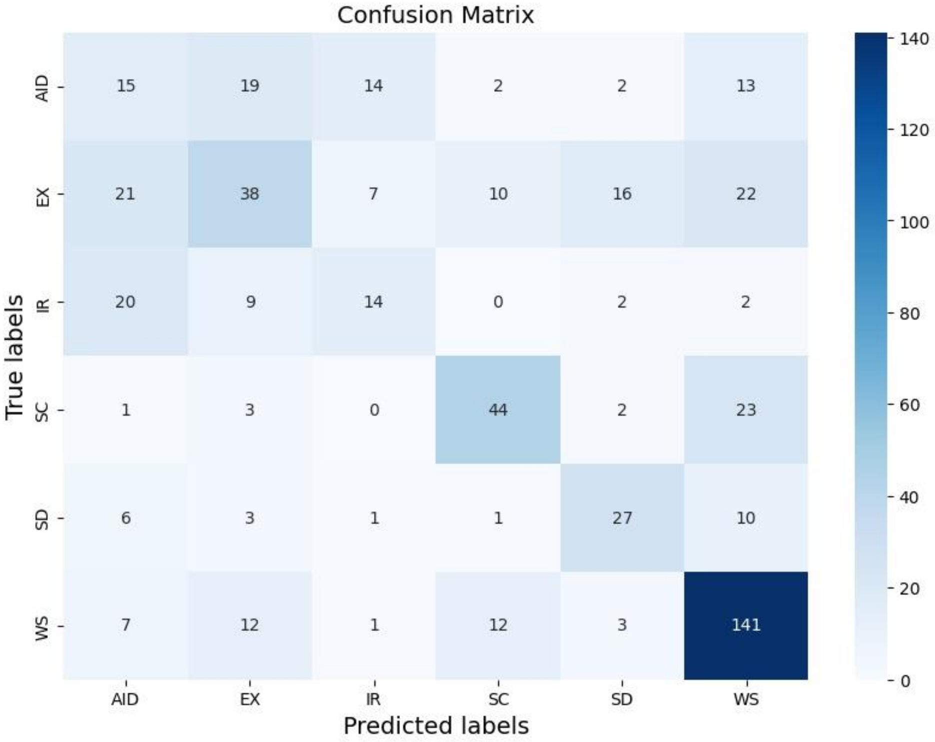
3. Results

This section presents the results of the internal evaluation, which used a 30% split of the training dataset for validation with both the CNN+USE and BERT models. The evaluation of the test dataset is described in the task overview document provided by the organizers.

3.1. CNN+USE

The results of the CNN neural network together with USE for text representations are presented in Figure 2. Compared to the BERT-like model paradigm, the classifier has lower performance but still achieves some acceptable results. The network's overall performance had a weighted average accuracy of 47%, which is quite low in general.

The model struggled to distinguish between incongruity-absurdity with exaggeration and irony. This suggests that the embeddings did not capture the subtle characteristics that differentiate these types of humor.



	AID	EX	IR	SC	SD	WS
AID	15	19	14	2	2	13
EX	21	38	7	10	16	22
IR	20	9	14	0	2	2
SC	1	3	0	44	2	23
SD	6	3	1	1	27	10
WS	7	12	1	12	3	141

Figure 2: Confusion matrix for the CNN+USE model.

3.2. BERT model

In the case of the BERT model, two runs were conducted: one with the base model and another with the multilingual version, both had similar results, with the multilingual model showing only slight improvement.

Figure 3 shows the results obtained, exhibiting behavior very similar to the CNN+USE model. Exaggeration, irony, and sarcasm are the most conflicting classes, so it is expected that these classes have slightly lower performance. The class wit-surprise have, similar results in both models, as it has the best overall performance. The weighted average accuracy of the model was 70.%.

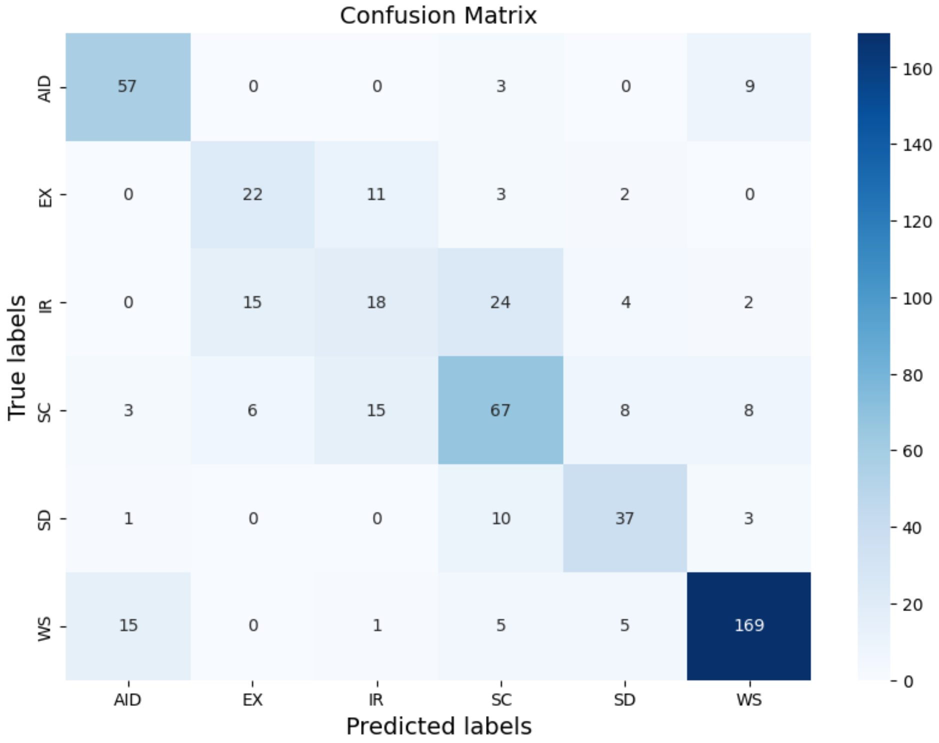


Figure 3: Confusion matrix for the BERT model.

3.3. Best Results

Table 1 showcases the best results achieved with the BERT model. it's noteworthy that the CNN+USE network's performance was inferior to that of BERT-like models and its analysis is not provide in this study. Each class is represented by a sentence that embodies its distinct characteristics. To illustrate our findings, we'll utilize the top result for each one.

As depicted below, the table presents probabilities for each class alongside the corresponding text (joke). It seems that the most effective methods yield results favoring longer, more anecdotal humorous sentences, as all of them have a confidence probability above 90%. This observation is somewhat corroborated when compared with Table 2, where the worst cases are presented. This contrasts with the majority of the poorer results, which still maintain a relatively high probability, typically around 40%.

Table 1

Best classify wordplay by class with BERT

Text	Class	Probability
Yogi had a whiskey, water, and tea drink every night. He was a toddy bear.	IR	0.9944
How did the hipster burn his mouth? He ate his pizza before it was cool.	SC	0.9864
Covid 19 coronavirus: Women are claiming 'boobs get bigger' after having Pfizer jab	EX	0.9577
Someone, please help me! I'm way too young to be this old already.	AID	0.9818
I've always known #Bez is my spirit animal but seeing the mess he makes on confirms it 100% what a legend that man is.	SD	0.9871
it's all about women in stem struggles. what about women in interactive media struggles: i no longer win against my friends in smash because they major in goddamn video game.	WS	0.9732

While some sentences vividly exemplify their class, others pose ambiguity. This variability in humor structures poses a challenge for accurate sentence-joke identification. The model may occasionally struggle to discern highly similar sentences, an issue that appears somewhat neglected.

Table 2

Worst classify wordplay by class with BERT

Text	Class	Probability
I can't believe today is the last day we can be gay.	IR	0.4298
I started a band called 999 megabytes. We haven't gotten a gig yet.	SC	0.5050
Children of Karen's don't get autism because they weren't vaccinated. They do however have hearing problems from listening to their moms scream at managers.	EX	0.4932
I put my phone on vibrate. An hour later, I finally received a text message.	AID	0.4845
I don't know anything about Coronavirus other than if you have it; you get an undeniable urge to go the airport.	SD	0.3680
The satellite went into orbit on January 1st causing a new year's revolution.	WS	0.3714

It appears that sentence length influences classification. Notably, poorly classified instances tend to be concise one-liners. Hence, detailed descriptions and contextual information could prove beneficial for each class. Additionally, BERT seems to utilize question marks and exclamation points for differentiation.

For instance, it's widely recognized that sentences employing exaggerated humor typically feature magnifying adjectives. In the case of irony and sarcasm, the fine line between them is determined by the underlying context. Thus, we posit that considering these nuances could mitigate misclassifications and errors.

4. Conclusions

The results obtained through the proposed classification approaches show promise, yet further refinement could strengthen and improve them. Specifically, in the classification of humorous classes, employing BERT-like models yielded favorable outcomes; however, there remains room for improvement through more effective fine-tuning and exploration of diverse variations in BERT-like architectures. In essence, this study has yielded encouraging findings, demonstrating the potential of transformer-based models in multiclass classification tasks.

Although a detailed performance analysis with evaluation metrics is yet to be provided, the analysis has demonstrated that the model exhibits strong confidence in classifying each category, even when confronted with a heavily imbalanced original dataset. Addressing this imbalance by augmenting the data or increasing the sample size for each class could potentially enhance the model's performance.

In conclusion, while this study primarily utilized deep learning models such as neural networks like CNNs and transformers like BERT, it's worth noting that other architectures remain unexplored and warrant investigation. The field of machine learning continually evolves, and exploring diverse models could lead to further insights and advancements.

References

[1] Arun S. Maiya (2020). ktrain: A Low-Code Library for Augmented Machine Learning. arXiv preprint arXiv:2004.10703. BigScience Workshop. (2022). BLOOM (Revision 4ab0472)

[2] Ermakova, L., Miller, T., Bosser, A-G., Palma, V., Sidorov, G., and Jatowt, A. 2024. CLEF 2024 JOKER Lab: Automatic Humour Analysis. In: Goharian, N., et al. Advances in Information Retrieval. ECIR 2024. Lecture Notes in Computer Science, vol 14613. Springer, Cham. https://doi.org/10.1007/978-3-031-56072-9_5

[3] google-bert/bert-base-multilingual-uncased · Hugging Face. (2001, march 11). https://huggingface.co/google-bert/bert-base-multilingual-uncased

[4] google-bert/bert-base-uncased · Hugging Face. (2001, march 11). https://huggingface.co/google-bert/bert-base-uncased

[5] Liana Ermakova, Anne-Gwenn Bosser, Tristan Miller, Victor Manuel Palma Preciado, Grigori Sidorov, and Adam Jatowt. Overview of JOKER @ CLEF-2024: Automatic humour analysis. In Lorraine Goeuriot, Philippe Mulhem, Georges Quénot, Didier Schwab, Laure

Soulier, Giorgio Maria Di Nunzio, Petra Galuščáková, Alba García Seco de Herrera, Guglielmo Faggioli, and Nicola Ferro, editors, Experimental IR Meets Multilinguality, Multimodality, and Interaction: Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Cham, 2024. Springer.

[6] Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR, abs/1810.04805. Retrieved from http://arxiv.org/abs/1810.04805

[7] Peng-Yu Chen and Von-Wun Soo. 2018. Humor Recognition Using Deep Learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 113–117, New Orleans, Louisiana. Association for Computational Linguistics.

[8] Pires, T., Schlinger, E., & Garrette, D. (2019). How Multilingual is Multilingual BERT? https://doi.org/10.18653/v1/p19-1493

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. En arXiv (Cornell University) (Vol. 30, pp. 5998- 6008). Cornell University. https://arxiv.org/pdf/1706.03762v5

[10] Victor Manuel Palma Preciado, Grigori Sidorov, Carolina Palma Preciado Assessing WordplayPun classification from JOKER dataset with pretrained BERT humorous models, JokeR: Automatic Wordplay and Humour Translation, pages (1828-1833), CLEF (2022) [6] Devlin, J., Chang, M.-W.,