Assessing Wordplay-Pun classification from JOKER dataset with pretrained BERT humorous models

Victor Manuel Palma Preciado[a], Grigori Sidorov[a] and Carolina Palma Preciado[a]

[a] *Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Av. Juan de Dios Batiz, s/n, 07320, Mexico City, Mexico*

Abstract

Humor is one of the most subjective matters of human behavior since it includes a wide range of variables: sentiments, wordplay, double meanings structurally or phonetic, all of this within the construction of written humor. It is important to assess the humor from a different point of view since this variability tends to provide insight into the true structure or the main core of the humoristic dilemma, as we know the range of humor is so diverse that it presents a high skilled problem even on the simplest tasks. Pre-trained base Bert and DistilBert models trained with a humorous one-liners dataset were used, these trained models were tested with a merged dataset from JOKER from data of tasks 1 and task 3, the collected data was trimmed from duplicated records and special characters to create a final dataset with 3,601 humorous sentences. Under this experiment we try to see if our models were able to detect a different humor from the initial type with which they were trained, it was noted that both methods are able to successfully classify another type of humor. On the one hand, it was expected that the pre-trained models would be able to classify at least a portion of the humor in the data set, the results obtained were much better than anticipated, obtaining 95.64% for BERT and 92.58% for DistilBERT, the models were really able to identify humor, an analysis of the worst and best cases were taken into account.

Keywords 1

Humor identification, Transformers, Humourism, Classifiers

1. Introduction

As we know, humor has a high written complexity, in addition to its different formats and interpretations, which cause quite a big challenge in the field of NLP, as much as it is in the tasks of classification, interpretation, and translation. In previous work under the classification of written humor, the results were quite good for the set of One liners, which had 3 types of short jokes: the riddle type, the differences type, and a short sentence with a single delivery. These One liners, generally considered as humor, were used to train the BERT-like models and in turn, Emebbedigs such as ELMO and USE with simple networks, although these were surpassed by the BERT-like models.

Furthermore, we do not know if these models were really able to identify humor as a general concept or only the structure of humor contained in the data set, this leads to the question of whether the capacity of these models is extended to another type of data with humor, in the style of [1][2][3] that present a high level of typification as does the data set for the JOKER[4] tasks.

In this case, we are interested in knowing if the previously trained model has the ability to recognize the humor found in the data set provided in the JOKER[4] tasks and, therefore, check if this type of model is capable of recognizing another humor, in addition to the approach and vision of what was not classified as humor despite being so.

This work intends to use the data set of tasks 1 and 3 of JOKER[4] with a preprocessing step, joining them to have a larger data set, in order to have a better representation of humor in its different forms.

[1] CLEF 2022 – Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

EMAIL: victorpapre@gmail.com (A. 1); sidorov@cic.ipn.mx (A. 2); c.palma.p0@gmail.com (A. 3)
ORCID: 0000-0003-3901-3522 (A. 2);

All this with the intention to check if our methods have a little more validity in terms of the humor type or in certain cases the same humor, therefore, said training could corroborate this type of classification in a certain way.

2. Implementation

As one of the first steps in the development of the classification, we opted to perform a preprocessing step in which we remove the links, special characters, duplicates, and a very short superficial manual review was made to ensure a certain homogeneity of the conjunction of both sets of data, belonging to tasks 1 and 3. It was chosen a certain portion of the data with the tags [het, hom, pun], with both portions of test and train.

Of a total of 12,540 humorous texts that we initially had for classification, after trimming we were left with 3,639 texts before doing the manual review, to finally end up with a set of 3,601 registers. Subsequently, the models saved in their Tensorflow format were recovered to be used in the classification, using a wrapper called Ktrain[5] to easily load our BERT-like models, which in turn will help us determine if the model perceives the probability of any of the 2 events in the classification between humor and non-humor. With the aim of discovering, in the same way, some structure within the new data set that does allow us to enrich the knowledge that we have of our two models about the capabilities and limitations to see if these coincide with the weaknesses present in the original model.

Once the model was loaded and as part of our classification, the percentage of confidence of belonging to one of the 2 categories (humor-nohumor) was obtained, which were passed through BERT-humor[6] and DistilBERT-humor[6] models, with which we classify the entire data set. As we know in general, the Tranformers scheme models and specifically the BERT-like ones tend to behave favorably with a sufficient amount of information, in most classification tasks, question-answering, ranking explainability, among others. On the other hand, the ELI5 library was used to obtain a weighted attention on the items to be classified, since it allows us to better visualize the structure we want to study.

3. Results

After performing the prediction process under the two pre-trained models proposed for the task of identifying humorous text (BERT and DistilBERT), it was found that the BERT model obtained better results since it identified 95.64% of the data set as humor, which means that of the 3,601 sentences in the data set obtained from JOKER, it correctly detected 3,444 records. For its part, DistilBERT although it did not achieve the same performance, obtained good results since 3,334 texts were correctly identified, thus reaching 92.59%. Table 1 shows the performance of the two models, and although both managed to distinguish the greatest amount of humorous text, there is a small difference of 110 incorrectly classified records between the two.

Table 1
BERT-like models performance

Models	Performance	Correctly detected	Incorrectly detected
BERT	95.64%	3,444	157
DistilBERT	92.58%	3,334	267

For the performance of the models employed, which identify whether a text is humorous or not, a function that predicts the class to which they belong was applied and the confidence probability was also calculated for each record to assess how sure the model is that a text really belongs to the humor class. Even though the functions used to calculate the confidence yield values with six tenths, it was decided to make five ranges of 0.2 to present a more visual representation of the obtained data, these results are shown in Figure 1.

The confidence extracted from the predictions made by BERT reflects that for most of the texts the model classified them as humor with certainty greater than 0.8, although for 105 texts they were identified as humor with a score less than 0.2 which implies that the text is not a humorous one, these of cases were the minority. For the ranges between 0.2 to 0.8, lower records were found among the three, which indicates that the model predicts mostly with high confidence.

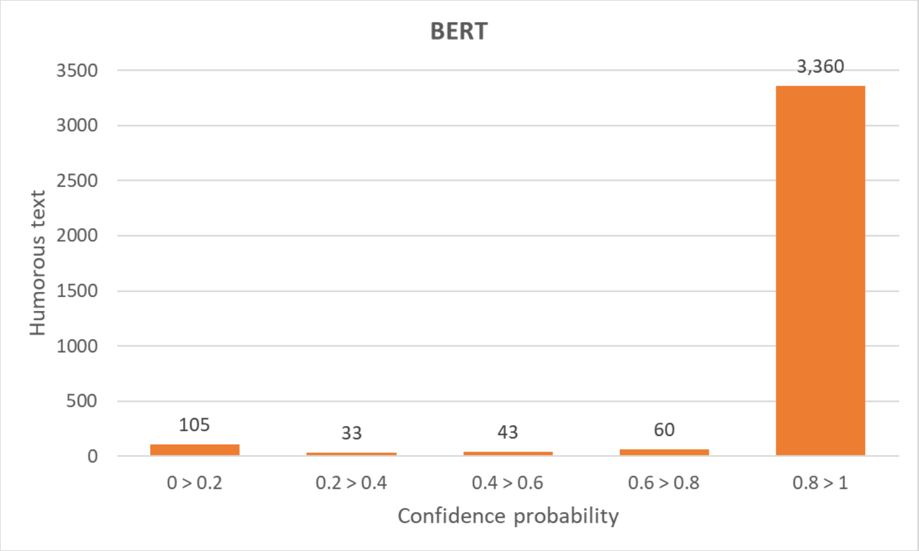

Figure 1: BERT humorous probability classification

In the case of the DistilBERT model, the results obtained are similar to those found with BERT, since the largest amount of data had high confidence as 3,134 texts reached a value greater than 0.8. On the other hand, the other columns with ranges from 0.2 to 0.8 had a greater amount of data in comparison with BERT, but even so, these represent a smaller group; in comparison, both models present results alike (Figure 2). It is worth mentioning that the models recognize as humor texts reach a score above 0.5.

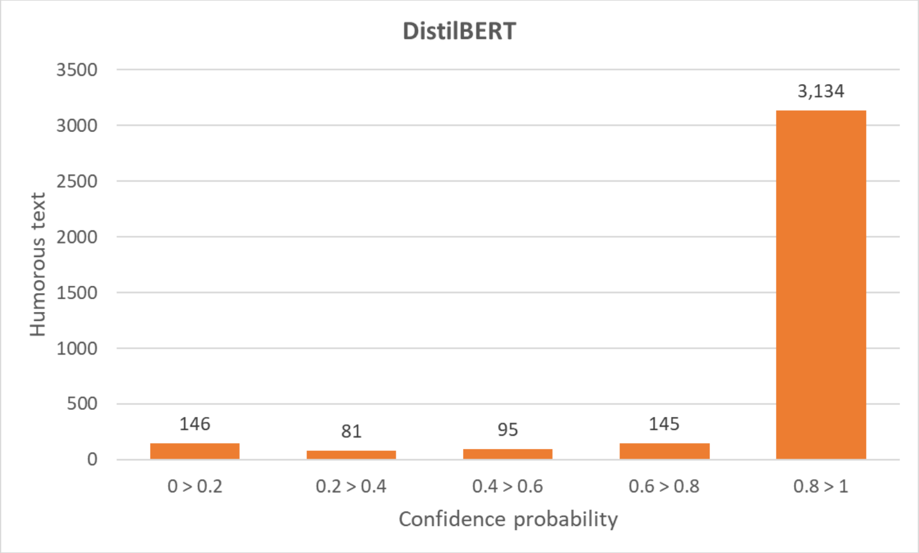

Figure 2: DistilBERT humorous probability classification

Once the analysis of the values obtained from the probabilities of each model was carried out, the humorous texts that obtained the best and worst scores were identified in order to visualize which types of writings the models manage to distinguish in a better way compared to the others. Table 2 presents the best five humorous texts detected by BERT, it is found that puns (PUN) are the ones that obtained the best probability. It can also be observed that the probability score obtained in confidence is high since it almost reaches 1.

Table 2
Top BERT positive cases

Tag	Humorous text	Probability
PUN	Why did the pig leave the party early? Because everyone thought he was a boar!	0.9999985
PUN	Why are ghosts bad liars? Because you can see right through them!	0.9999982
HOM	Why don't people like to talk to garbage men? They mostly talk trash.	0.9999981
PUN	Why don't sharks eat clowns? Because they taste funny.	0.9999980
PUN	Why did the student eat his homework? Because the teacher told him it was a piece of cake!	0.9999980

Likewise, the results of Table 3, which shows the results of the DistilBERT model, are similar to those described above since they are also mostly puns and texts with the tag HOM, with the difference that DistilBERT has among its best scores text the identification tag HET. For both cases, the best scores are riddle texts which have a question-and-answer format.

Table 3
Top DistilBERT positive cases

Tag	Humorous text	Probability
PUN	What's the best fruit for avoiding scurvy? Naval oranges, of course.	0.9996407
PUN	What does an angry pepper do? It gets jalapenos face.	0.9996404
PUN	What do you call a duck that gets all A's? A wise quacker.	0.9996402
HOM	There was an eye doctor who wanted to re-locate but couldn't find a job because he didn't have enough contacts.	0.9996402
HET	What is the best store to be in during an earthquake? A stationery store.	0.9996401

On the other hand, for the case of the humorous texts with the worst probabilities presented in Table 4, the BERT model detected mostly HET and very few HOM and wordplay (puns). As can be seen, the scores obtained were low since they tend to 0, this may be due to the structure of this type of example since in comparison with the best results no riddles are found in the worst rated.

Table 4
Top BERT negative cases

Tag	Humorous text	Probability
HET	Opportunities take "now" for an answer	0.0000654
HET	Podiatrist malpractice: Callous neglect	0.0001483
HOM	Bill Gates took advantage of his Windows of opportunity	0.0001519
HET	Exposure to the son may prevent burning	0.0002049
PUN	Can honeybee abuse lead to a sting operation?	0.0002278

In the same way, this phenomenon occurs for the DistilBERT model since among the humorous text with the worst score there are two texts that also appear in results achieved by BERT, such as: *Opportunities take "now" for an answer* and *Exposure to the son may prevent burning*, this indicates that both models are perceiving the same text as not humorous, which points to a similar detection structure.

Table 5
Top DistilBERT negative cases

Tag	Humorous text	Probability
HET	Exposure to the son may prevent burning	0.0007237
HET	Opportunities take "now" for an answer	0.0007271
HET	Exposure to the son prevents burning	0.0007380
HOM	Could modern submarines be the wave of the future?	0.0007441
HET	A budget helps us live below our yearnings	0.0007591

3.1. ELI5 prospection

The ELI5 library was used to obtain a certain resemblance to where our BERT-like models classify the humorous point of attention, given that it manages a joint probability, we can observe that in general the humorous sentences that fared better were those of the riddle type, which consists of a question and an answer as the humorous delivery is usually made.

Bert - humorous texts with the best scores:

Why did the pig leave the party early? because everyone thought he was a boar!

Why are ghosts bad liars? because you can see right through them!

DistilBERT - humorous texts with the best scores:

What's the best fruit for avoiding scurvy? Naval oranges, of course.

What does an angry pepper do? It gets jalapenos face.

As we can see above, humorous question-and-answer statements had a fairly good rating, since in general this type of text was successfully evaluated. On the other hand, the probability colorations that ELI5 marks make sense since they start with the WH-questions, which is one of the main ways of gathering information, and since in a riddle it is not interesting to reveal a small amount of information, it perfectly fulfills the scheme of attention to delivery after the question taking a darker coloration at the end of the sentence for both the question and answer.

Bert - humorous texts with the worst scores:

Opportunities take "now" for an answer

Podiatrist malpractice: Callous neglect

DistilBERT - humorous texts with the worst scores:

Exposure to the son may prevent burning

Opportunities take "now" for an answer

On the other hand, the sentences that did worse turned out to have a certain pattern, since if we try to look closely these sentences could very well allude to the title of a review or article, given their content between serious and humorous seemed to confuse the decision of the model, when it comes to classification. It seems that the model, given the coloration, focuses on nouns such as *Opportunities*, *Malpractice* and *Exposure* to classify the text as something that does not contain humor.

4. Conclusion

The pre-trained models used in this work that addresses the subject of humorous text identification, managed to detect the majority of the data set obtained from JOKER as humor with a good outcome for both models. It should be noted that it seems that the opposite or negative part (the data set previously used to train the models) strongly affects the result, that is, the non-humorous part of the data set.

The paradigm of the models based on Transformers tend to classify humorous texts well, in this case, BERT and DistilBERT manage to classify humorous texts with high probabilities. A tendency was also found for the models to identify with better scores the riddles that were found in the data set as puns, as opposed to the other types of examples. Therefore, it is obtained that the structure of the humorous text strongly influences its identification.

On the one hand, it is not surprising that it behaved favorably with the JOKER data set, since the humor contained in it in certain aspects becomes very similar to a certain extent to the humor with which these models were trained, it should be noted that Certain curiosities were found within the best and worst classified elements, certain patterns that have a lot to do with the counterpart of the humor with which it was trained, giving a view of some weaknesses and strengths of the models that will be explained later.

5. Acknowledgements

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20220852 and 20220859 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

1. References

[1] Mihalcea Rada and Strapparava Carlo. "Making computers laugh: investigations in automatic humor recognition." In Conference on Human Language Technology and Empirical Methods in Natural Language Processing, (2005): 531–538.

[2] Miller, K.E.L. "The Unuttered Punch Line: Pragmatic Incongruity and the Parsing of 'What's the Difference' Jokes." (3 December 2009).

[3] Orion Weller and Kevin Seppi. "The rJokes Dataset: a Large Scale Humor Collection." In Proceedings of the 12th Language Resources and Evaluation Conference, (2020): 6136–6141, Marseille, France. European Language Resources Association.

[4] Ermakova, L., Miller, T., Puchalski, O., Regattin, F., Mathurin, É., Araújo, S., Bosser, A.-G., Borg, C., Bokiniec, M., Corre, G. L., Jeanjean, B., Hannachi, R., Mallia, Ġ., Matas, G., & Saki, M. (2022). CLEF Workshop JOKER: Automatic Wordplay and Humour Translation. In M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, & V. Setty (Eds.), Advances in Information Retrieval (Vol. 13186, pp. 355–363). Springer International Publishing.

[5] Maiya, A.S. "Ktrain: A Low-Code Library for Augmented Machine Learning" (2020). ArXiv, abs/2004.10703.

[6] V. M. Palma, Automatic Detection of Jokes in Texts, Master’s thesis, Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Ciudad de México, México, 2021.