UMUTeam at HAHA 2021: Linguistic Features and Transformers for Analysing Spanish Humor. The What, the How, and to Whom

José Antonio García-Díaz[1][0000−0002−3651−2660] and
Rafael Valencia-García[1][0000−0003−2457−1791]

Facultad de Informática, Universidad de Murcia,
Campus de Espinardo, 30100, Spain
{joseantonio.garcia8,valencia}@um.es

Abstract. Giving computers basic notions of humour can result in better and more emphatic user interfaces that are perceived more natural. In addition, hate-speech systems that understand humor are more reliable, as humor can be used as a Troy Horse to introduce oppressive-speech passing them off as harmless jokes. Understanding humor, however, is challenging because it is subjective as well as cultural and background dependant. Sharpen forms of humor, moreover, rely on figurative language, in which words loss their literally meaning. Therefore, humor detection has been proposed as shared-task in workshops in the last years. In this paper we describe our participation in the HAHA'2021 shared task, regarding fine-grain humor identification in Spanish. Our proposals to solve these subtasks are grounded on the combination of linguistic features and transformers. We achieved the 1st position for humor rating Funniness Score Prediction with a RMSE of 0.6226, the 8th position for humor classification subtask with an 85.44 F1-score of humours category, and the 7th and the 3rd position for the subtasks of humor mechanism and target classification with an macro averaged F1-score of 20.31 and 32.25 respectively.

Keywords: humor Detection · Feature Engineering · Natural Language Processing.

1 Introduction

According to the Merriam Webster dictionary, humor can be defined as a *mental faculty of discovering, expressing, or appreciating the ludicrous or absurdly incongruous*. Therefore, humor requires sharp mental abilities that are developed starting in the childhood [13]. In addition to social and cognitive skills, other

IberLEF 2021, September 2021, Málaga, Spain.

factors that make identifying humor challenging include subjectivity, as what it is fun for some people may not be fun for others, and cultural background, that influences how humor is perceived and expressed [6]. Moreover, in written communication humor can appear in many forms: from light-heart humor, such as anecdotes told with wit, to bitter forms such as sarcasm. To make humor identification even more difficult, humorists play with figurative language, making use of complex linguistic phenomena such as sharp analogies or double entendres. Figurative language, and its relationship with humor, has been a productive area of research as regards tasks such as satire identification [11, 12].

In this manuscript we describe our participation in the shared task HAHA - Humor Analysis based on Human Annotation (HAHA 2021) [2] from IberLEF (Iberian Languages Evaluation Forum) [9]. From a computational linguistic point of view, humor have been studied in recent years [8], but these studies mainly focused on their identification. As far as we know, the most complete work regarding humor identification and categorisation was the shared task HaHaCkathon 2021 [7], focused on English.

2 Corpus

The evaluation campaign of HAHA proposed four subtasks to the participants regarding humor detection and categorisation. These subtasks are the following: (1) Humor Detection, whose objective is to determine if a text is a joke or not. This subtask is ranked with F1 score of the *humorous* class. The ratio between non-humor and humor was near to 1.5:1, which indicates a strong imbalance among the dataset; (2) Funniness Score Prediction, whose objective is to predict a funniness score value for a text in a 5-star ranking, assuming the document is a joke. This subtask is measured with root mean squared error (RMSE). We observe than the average of the score predictions is 2.046, the mode is 2, and the standard deviation is 0.649, which indicates consensus among the annotators; (3) Humor Mechanism Classification, a multi-classification problem to determine the mechanisms by which a text conveys humor; and (4) Humor Target Classification, a multi-label classification task in which we were requested to predict the targets of the joke.

For the first two subtasks, the organisers provided a gold-standard corpus divided into three splits: (1) training, composed by 24,000 documents; (2) development, composed by 6,000 documents; and (3) evaluation, composed of 6,000 documents. According to the description of the task, the annotators of the corpus used a voting scheme in which they could indicate if the document is humorous or not and, if they answered affordability, how funny it was in a five-star ranking. At the time of writing this working notes, no information regarding the annotation process for subtasks 3 and 4 were provided.

We calculate the correlation of the humor mechanism and humor target[1] and we observe that there is a strong correlation between (1) *embarrassment* and *bodying shaming*, (2) *reference* and *self-deprecating*, (3) *stereotype* and *women*, and

[1] Not included due to page limit, but it is available in the repository

(4) *word play* and *professions*. It can also be observed that some mechanisms work well regardless of the topic of the joke, whether through absurdity, exaggerations, or analogies. Other mechanisms for making humor, on the other hand, aims for specific objectives. We can observe this for humor that consists on embarrassment of body and, in a minor degree, to age and familiar relationships. In addition, targets related to ethnicity seem to be the ones with less variety of mechanisms, being the majority those related to stereotyping and word play.

3 Methodology

Our methodology is grounded on the combination of linguistic features and transformers by mean of neural networks. Our pipeline includes a pre-processing stage for cleaning the texts, a feature extraction for the linguistic features and the transformers, a selection stage for selecting the best features and, finally, a hyper-parameter evaluation stage for tuning the models. For text pre-processing we transform the documents into its lowercase form, we remove hyperlinks, emojis, quotations, mentions, and hashtags. We also replace digits with a fixed token, and we fix misspellings and remove word elongations. Next, both the pre-processed and the original version of the text are used to obtain the features. The cleaned version is used for obtaining the sentence embeddings and the majority of the linguistic features. The original version of the documents, on the other hand, are employed for extracting linguistic features related to misspellings, correction and writing style, and the percentage of uppercase letters. Next, we extract sentence embeddings from Spanish fastText [5] (SE) and transformers based on Spanish BERT (BF) [1]. For fastText, we use the their online tool for extracting fixed vectors of 300. For transformers, we fine-tune the model for each subtask, and then obtained a vector of length 768 from the `[CLS]` token.

It is worth mentioning that during our experimentation, we also evaluated Convolutional and Recurrent Neural Networks because they achieved competitive results in classification tasks such as Sentiment Analysis [10]. Our results with a validation split, however, indicate the accuracy and results achieved by convolutional and neural networks were similar to the ones obtained with fixed-vectors, that are several orders of magnitude faster to train.

We use UMUTextStats [3, 4] for extracting a total of 365 features related to stylometry, phonetics, morphosyntax, discourse markers, figurative language, lexical, psycho-linguistic processes, register, and to detect patterns commonly used in social networks. The linguistic features are scaled using a MinMax strategy. Next, we apply a feature selection based on mutual information. As we faced four challenges, we perform this process four times: Mutual Information for subtasks 1, 3, and 4 and univariate linear regression test for subtask 2. Next, we perform a hyper-parameter tuning for evaluating different combinations of neural networks and select the best one for each subtask based on the main metric. We evaluate 125 neural network models for each feature set {LF, SE, and BF} and for the combinations of features {LF, SE, BF} and {LF, BF}. The hyperparameters include the number of hidden layers and their number of

neurons organised in several shapes, the dropout rate, some activation functions, and different learning rates. Neural networks are provided with an early stopping mechanism, a learning rate scheduler, and the initial weights of the neural network in order to reduce the effect of class imbalance. Source code is available at https://github.com/Smolky/haha-2021.

4 Results

The competition was divided into development and evaluation and the organisers provided a split for each stage. However, as the labels with the grounding truth of the development split were not released, we generate a custom validation split dividing the training set in a ratio of 80% for training and the remaining 20% for validating. Table 1 includes the official leader board to compare our results with the rest of the participants. Our team achieved positions 8th, 1st, 6th, and 3rd for subtasks 1, 2, 3 and 4 respectively.

Table 1. Official results and ranking of the HAHA'2021 task for each subtask, ranked, respectively by F1 score for the humorous category (task 1), RMSE (Task 2), and macro F1-score (Task 3 and 4)

Team / User	Subtask 1	Subtask 2	Subtask 3	Subtask 4
Jocoso	88.50 (1)	0.6296 (3)	0.2916 (2)	0.3578 (2)
icc	87.16 (2)	0.6853 (9)	0.2522 (3)	0.3110 (4)
kuiyongyi	87.00 (3)	0.6797 (8)	0.2187 (5)	0.2836 (6)
ColBERT	86.96 (4)	0.6246 (2)	0.2060 (7)	0.3099 (5)
noda risa	86.54 (5)	-	-	-
BERT4EVER	86.45 (6)	0.6587 (4)	0.3396 (1)	0.4228 (1)
Mjason	85.83 (7)	1.1975 (11)	-	-
UMUTeam	**85.44 (8)**	**0.6226 (1)**	**0.2087 (6)**	**0.3225 (3)**
skblaz	81.56 (9)	0.6668 (6)	0.2355 (4)	0.2295 (7)
humBERTor	81.15 (10)	-	-	-
RoBERToCarlos	79.61 (11)	0.8602 (10)	0.0128 (10)	0.0000 (9)
lunna	76.93 (12)	-	0.0404 (9)	-
N&&N	76.93 (12)	-	0.0404 (9)	-
ayushnanda14	76.79 (13)	0.6639 (5)	-	-
Noor	76.03 (14)	-	0.0404 (9)	-
KdeHumor	74.41 (15)	1.5164 (12)	-	-
baseline	66.19 (16)	0.6704 (7)	0.1001 (8)	0.0527 (8)

Regarding the first subtask, humor detection, all participants outperform the baseline (Naive Bayes classifier based on TFIDF features). Our team achieve a F1-score over the humor class of 85.44, reaching position 8th, only a 3.06 below the best result (Jocoso, F1-score of 88.50). For this subtask we train a neural network with LF and BF. Each feature set is connected to a shallow neural network composed of 2 hidden layers and 8 neurons per layer. We use a dropout

of 0.3, a linear activation function, and a learning rate of 0.01. For understanding the relevance of the linguistic features, we obtain the mutual information[2] and observe that the most relevant features are related to number of questions and the length of the corpus. We also identify as relevant the number of personal pronouns and main verbs.

Regarding the second subtask, humor scoring, we achieve 1st position. For this subtask, only 12 of the 16 participants sent runs. Our results, 0.6226 of RMSE are similar to second (ColBERT, 0.6246) and third (Jocoso, 0.6296) best results. For this subtask we combine LF, SE, and BF in a neural network composed by 6 hidden layers in a long funnel shape, with the first hidden layer composed by 47 neurons. We use a dropout of 0.2 and a learning rate of 0.01. The activation function is *tanh*. We achieve bad results in our first and second runs, obtaining an RMSE score of 1.19713 and 1.21332 respectively. To improve our results, we replace the loss function from MSE to RMSE and we retrain the neural network with our custom validation set for 10 more epochs. As a additional strategy, we observe than our neural network sometimes outputs values lower than 0. As we knew the score could not be negative, we convert them to 0. Moreover, in order to understand how LF performs in this subtask, we calculate Mutual Information Gain. Figure 1 contains the values of the 20 top rated LF. It is worth mentioning that subtask 2 contains only the documents rated as humor and, therefore, the LF are used to discern how funny and which agreement have the documents among the annotators. We can observe that there are linguistic features from several categories, such as stylometry (STY), highlight the number of sentences that are questions. Regarding morphosyntax (MOR), there are features related to the (1) usage of interjections, that are used to express emotions; (2) verbs in third person, as many jokes rely on third person to the targets; (3) adverbs and augmentative suffixes to empathise some actions, and (4) proper nouns, that are used to focus the joke on specific persons. Correction and style (ERR) is another relevant category, as we can observe features related to performance and orthographic errors. Intentional errors in texts are used to make fun from persons from a specific location. However, as we do not identify demonyms as a relevant label, we assume that this resource is mainly used for mocking individuals rather than collectives. The usage of hashtags from Social Media category (MED) is also a relevant feature.

For the third subtask, humor mechanism classification, a total of ten users participated. We achieve the 6th position in the official leader board with a macro F1-score of 20.87. We improve the baseline based on Naive Bayes classifier with TF-IDF features and a macro F1-score of 10.01. The best result is for BERTForever, with an macro F1-score of 33.96. Similar to subtask 1, our neural network consists into a shallow neural network that connects separately LF and BF to a hidden layer composed of 16 neurons. We use a dropout of 0.2 and a learning rate of 0.001, with a sigmoid as activation function. To handle class imbalance, we include in the train split 400 documents from the non-humor class.

[2] Not included due to page limit, but it is available in the repository

We calculate the confusion matrix over the official evaluation split[3]. We observe that *analogy*, *misunderstanding*, and *wordplay* are the classes with the most hits. *Embarrassment*, on the other hand, is confused with *unmasking*. A similar problem can be found between *stereotype* and *wordplay*. Another humor mechanism with poor performance is *parody*, wrongly classified as *reference*, *stereotype*, and *wordplay*. *Irony* is worst performing class. Only 9% of ironic documents were successfully classified. Moreover, ironic documents are wrongly classified uniformly among the rest of the humor mechanisms. Irony consists on suddenly shifting the expected outcome of events. This finding suggests that neither linguistic features nor transformers are able to catch words that deviates from their conventional meaning.

Finally, for the fourth subtask, humor target classification, a total of nine participants sent their results. As our team does not have much experience dealing with multi-label classifications, our approach consists into transforming the problem in a multi classification task. We consider this strategy as we observed that not much of the documents were labelled with more that one tag. Despite the simplicity of our approach, we reached 3rd position with an F1-score of 32.25. In this case, our best neural network consist into a shallow neural network that linked separately each feature set (LF, SE, and BF) to hidden layers of 256 neurons. In this case, we use tanh as activation function, a dropout of 0.3, and a learning rate of 0.001.

[3] Not included due to page limit, but it is available in the repository

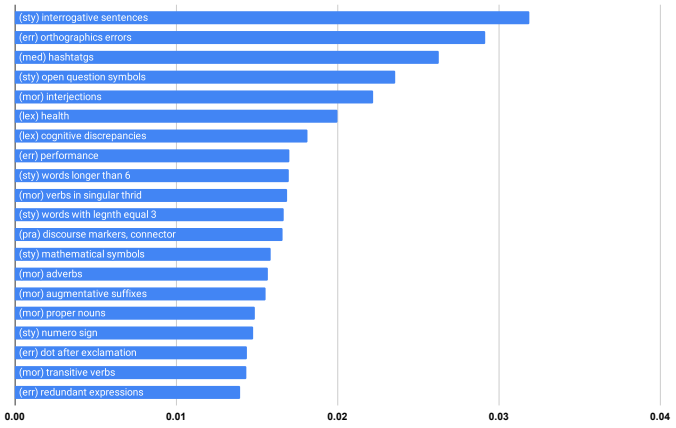

Fig. 1. Mutual Information of linguistic features concerning subtask 2: humor scoring

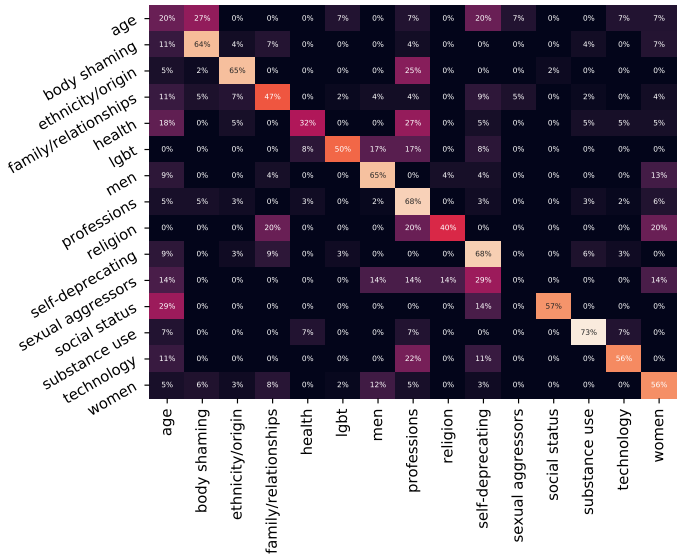

Fig. 2. Confusion matrix of the test split subtask 4: humor target identification

5 Conclusions

This manuscript contains the description of the participation of UMUTeam at HAHA 2021 shared task regarding fine-grain humor identification in Spanish. We are happy with the results achieved in this shared task, as we consider that we are improving our methods and results. As promising research directions, we will improve the linguistic features and we will analyse in which cases the linguistic features and the transformers does not agree in the final prediction, in order to get major insights of the strengths and weaknesses of each method. We also considered interesting to evaluate data augmentation for improving the accuracy of the model. In case of humor this is challenging as it is possible to include more examples during training with translated jokes from other languages. However, as humor is background and cultural dependant, this should be analysed with caution.

6 Acknowledgments

This work was supported by the Spanish National Research Agency (AEI) through project LaTe4PSP (PID2019-107652RB-I00/AEI/10.13039/501100011033). In addition, José Antonio García-Díaz has been supported by Banco Santander and University of Murcia through the industrial doctorate programme.

References

1. Canete, J., Chaperon, G., Fuentes, R., Pérez, J.: Spanish pre-trained bert model and evaluation data. PML4DC at ICLR **2020** (2020)
2. Chiruzzo, L., Castro, S., Góngora, S., Rosá, A., Meaney, J.A., Mihalcea, R.: Overview of HAHA at IberLEF 2021: Detecting, Rating and Analyzing Humor in Spanish. Procesamiento del Lenguaje Natural **67**(0) (2021)
3. García-Díaz, J.A., Cánovas-García, M., Valencia-García, R.: Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in latin america. Future Generation Computer Systems **112**, 614–657 (2020). https://doi.org/10.1016/j.future.2020.06.019
4. García-Díaz, J.A., Cánovas-García, M., Colomo-Palacios, R., Valencia-García, R.: Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings. Future Generation Computer Systems **114**, 506 – 518 (2021). https://doi.org/10.1016/j.future.2020.08.032, http://www.sciencedirect.com/science/article/pii/S0167739X20301928
5. Grave, E., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T.: Learning word vectors for 157 languages. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018) (2018)
6. Jiang, T., Li, H., Hou, Y.: Cultural differences in humor perception, usage, and implications. Frontiers in psychology **10**, 123 (2019)
7. Meaney, J., Wilson, S., Chiruzzo, L., Lopez, A., Magdy, W.: Semeval 2021 task 7: Hahackathon, detecting and rating humor and offense. In: 15th International Workshop on Semantic Evaluation (2021)
8. Mihalcea, R., Strapparava, C.: Making computers laugh: Investigations in automatic humor recognition. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. pp. 531–538 (2005)
9. Montes, M., Rosso, P., Gonzalo, J., Aragón, E., Agerri, R., Álvarez-Carmona, M.Á., Álvarez Mellado, E., Carrillo-de Albornoz, J., Chiruzzo, L., Freitas, L., Gómez Adorno, H., Gutiérrez, Y., Jiménez Zafra, S.M., Lima, S., Plaza-de Arco, F.M., Taulé, M.: Proceedings of the iberian languages evaluation forum (iberlef 2021). In: CEUR workshop (2021)
10. Paredes-Valverde, M.A., Colomo-Palacios, R., Salas-Zárate, M.d.P., Valencia-García, R.: Sentiment analysis in spanish for improvement of products and services: a deep learning approach. Scientific Programming **2017** (2017)
11. del Pilar Salas-Zárate, M., Alor-Hernández, G., Sánchez-Cervantes, J.L., Paredes-Valverde, M.A., García-Alcaraz, J.L., Valencia-García, R.: Review of english literature on figurative language applied to social networks. Knowl. Inf. Syst. **62**(6), 2105–2137 (2020). https://doi.org/10.1007/s10115-019-01425-3, https://doi.org/10.1007/s10115-019-01425-3
12. del Pilar Salas-Zárate, M., Paredes-Valverde, M.A., Rodríguez-García, M.Á., Valencia-García, R., Alor-Hernández, G.: Automatic detection of satire in twitter: A psycholinguistic-based approach. Knowl. Based Syst. **128**, 20–33 (2017). https://doi.org/10.1016/j.knosys.2017.04.009, https://doi.org/10.1016/j.knosys.2017.04.009
13. Semrud-Clikeman, M., Glass, K.: The relation of humor and child development: Social, adaptive, and emotional aspects. Journal of Child Neurology **25**(10), 1248–1260 (2010)