Overview of the HAHA Task: Humor Analysis based on Human Annotation at IberEval 2018

Santiago Castro, Luis Chiruzzo, and Aiala Rosá

Grupo de Procesamiento de Lenguaje Natural
Universidad de la República — Uruguay
{sacastro,luischir,aialar}@fing.edu.uy

Abstract. This paper presents the HAHA task at IberEval 2018, the first challenge on humor appreciation in the Spanish language. The challenge proposes two tasks related to humor in language: automatic detection and automatic rating of humor in Spanish tweets. In this challenge, we used a corpus of 20,000 annotated tweets in Spanish, where each tweet indicates if it is humorous and the humorous ones contain a funniness score. We summarize the submitted systems and present the general results for both tasks.

Keywords: Humor · Computational Humor · Humor Detection · Natural Language Processing

1 Introduction

In this document we present the task Humor Analysis based on Human Annotation (HAHA), which is part of the IberEval 2018 workshop. This task proposes different subtasks related to automatic humor detection in Spanish.

While humor has been historically studied from a psychological [8,4], cognitive [11], linguistic [18,1,20] and even evolutionary standpoint [16], it is an area yet to be explored from a computational perspective.[1] Humor is a complex human activity that involves many cognitive skills such as abstract thinking, language skills and social perception [16]. Although humor is present in all societies, it is at the same time highly personal, context and culture specific. Even speakers of the same language but living in different geographic areas do not share the same perception of humor [19]. This makes it specially difficult to deal with using automatic processes: a characterization of humor that allows its automatic recognition and generation is far from being specified. One of the aims of this task is to gain better insight in what is humorous and what causes laughter.

There exists previous work related to automatic humor detection in English [10,21] and in Spanish [3]. Furthermore, some similar evaluation campaigns have been performed for English: SemEval-2015 Task 11 [5], which proposed to

[1] See [12] for a comprehensive report about Humor and Computational Humor.

work on figurative language, such as metaphors and irony; and SemEval-2017 Task 6 [17], which aimed at predicting the degree of funniness in tweets given a set of tweets issued in response to a TV game show. Both campaigns have focused on the microblogging platform Twitter, which is particularly suited for these tasks due to its public availability and the fact that its users have to communicate using short messages. The HAHA campaign is, as far as we know, the first attempt to make such an evaluation campaign for the Spanish language.

The rest of the paper is structured as follows: Sect. 2 presents the proposed subtasks for this campaign, Sect. 3 describes the corpus used, Sect. 4 summarizes the submissions made by participants and shows the results for each subtask, and finally Sect. 5 gives some conclusions and presents future research directions.

2 Tasks

Two subtasks were proposed for this task: detecting humor in tweets and trying to estimate how funny a tweet is. The participants could submit up to four submissions per subtask.

2.1 Subtask 1: Humor Detection

The aim of this subtask is to tell if a tweet intends to be humorous (if the intention of the author was to be humorous or not). To do this, a set of training tweets annotated with their corresponding humorous class was given to the participants. The performance metrics used for this subtask were F_1 score for the humorous category and accuracy, being the F_1 score the main measure for this subtask.

Two baselines were provided for this subtask, computed over the test data:

baseline1 Decide randomly with a 50% probability whether a tweet is humorous or not. This baseline achieves 0.42 F_1 score and 0.49 accuracy for the humorous class over the test corpus.

baseline2 Select all tweets that start with a dash as humorous. This baseline was defined by inspecting the corpus and noticing that many tweets considered humorous were dialogues with the utterances delimited by dashes. This heuristic has a very high precision, as almost all the dialogues in tweets are jokes, but a very low recall because there are many more kinds of jokes and humorous tweets. The baseline achieves 0.17 F_1 score and 0.66 accuracy for the humorous class over the test corpus.

2.2 Subtask 2: Funniness Score Prediction

The aim of this subtask is to predict how funny an average person would consider a tweet, taking as ground truth the average funniness value of the tweets in a corpus. The funniness score is defined in a 5-point scale ranging from one (not funny) to five (excellent). The results of this subtask were measured using Root Mean Squared Error (RMSE).

We calculated one baseline for this subtask over the test data, that is choosing the value 3 (middle of the scale) for all tweets. The root mean squared error for this baseline over the test data is 1.14.

It is important to notice that the valid tweets for this subtask are only the humorous ones, as we consider that the average funniness score is only well defined for this category. However, as the participants could not know in advance which of the test tweets were humorous, we asked them to rate all of the tweets in the test corpus, and then our measuring process would take in consideration only the ones that belonged to the class "humorous".

3 Corpus

The corpus for the task consists of 20,000 crowd-annotated tweets. It is detailed in [2] and described hereafter. It was built by extracting 16,500 tweets from humorous Twitter accounts in Spanish (accounts that generally post humorous content such as jokes) and a random sample of 12,000 tweets in Spanish. We selected Spanish speaking Twitter accounts from as many Spanish speaking countries as possible so as to have an acceptable mix of Spanish variations. These tweets were crowd-annotated using a web application[2] between March 8th and 27th, 2018, receiving a total amount of 117,800 annotations by 1,271 annotators (including the number of times tweets were skipped). Each annotation indicates if the user considers the tweet as humorous and, in that case, how funny the user considers it. The funniness rating in the web app could be defined by the annotators using emoji and it was later on translated as a score from 1 to 5. Almost all tweets in the corpus contain at least three votes, and almost all the ones considered humorous contain at least five votes. Table 1 shows an example of tweet annotated in the format of the corpus.

Table 1. Example of annotated tweet

Text	— ¿Tienes Wi-Fi? — Claro. — ¿Cuál es la clave? — Tener dinero y pagarlo.
Is humorous	True
Not-Humor votes	1
1-star votes	1
2-star votes	0
3-star votes	1
4-star votes	1
5-star votes	2
Star average	3.6

[2] https://clasificahumor.com

To prepare it for this task, we post-processed the base corpus in the following way. First, we removed all the annotations considered low quality. It was measured using some cherry-picked tweets that were presented to all users at the beginning of their annotation session, to check the annotator level. The tweet count was 107,634 after it (also excluding the skips). Then, we computed the "is humorous" value for each tweet as a simple majority (e. g., if half of the votes are for the humorous class, we consider the tweet humorous). Finally, we calculated the "funniness score" by averaging the scores of all votes, considering only positive votes (votes for the humorous class).

After computing the humorous class, we found that less than 27% of the total number of tweets were considered humorous. To make the classes as balanced as possible, we randomly dropped 57% of the not humorous tweets so as to get a total number of 20,000 tweets in the corpus. In this new corpus, 36.8% of the tweets are considered humorous. Although it is still not a perfectly balanced corpus, it is better suited for training. This 20,000 tweet corpus was randomly split in 80% for training and 20% for testing.

4 Participants and Results

Three teams took part in the challenge: two of them participated in both subtasks, and the third one only in the first subtask. The proposed systems are mainly based on Neural Networks and traditional Machine Learning techniques. Nevertheless, to our surprise, the best performing system is based on an Evolutionary Algorithm. Tables 2 and 3 show the general results for all teams for Subtasks 1 and 2 respectively. The tables also include the proposed baselines for both subtasks. In almost all the cases, the submissions were able to beat the baselines.

Table 2. Scores for Subtask 1

Team	Run	Accuracy	Precision	Recall	F_1
INGEOTEC	run 2	0.8452	0.7796	0.8157	**0.7972**
UO_UPV	run 1	**0.8455**	0.8158	0.7567	0.7851
UO_UPV	run 2	0.8448	0.8322	0.7312	0.7785
ELiRF-UPV	run 1	0.8367	0.8046	0.7426	0.7724
UO_UPV	run 3	0.8397	0.8281	0.7198	0.7702
INGEOTEC	run 1	0.8403	0.8557	0.6877	0.7625
ELiRF-UPV	run 2	0.7552	0.6546	0.7279	0.6893
baseline	baseline1	0.4915	0.3645	0.4886	0.4175
baseline	baseline2	0.6595	0.9392	0.0932	0.1695

The INGEOTEC [14] team presented systems for Subtasks 1 and 2. They tested several classifiers and regressors from scikit-learn [15] (Naïve Bayes, SVM, Nearest Centroid, Kernel Ridge, Ridge, Ada Boost, Decision Trees and ElasticNet), several models and tools developed by them (μTC [23], B4MSA [22] and EvoDAG [6])

Table 3. Scores for Subtask 2

Team	Score
INGEOTEC	**0.9784**
baseline	1.1419
UO_UPV	1.5919

and a model based on FastText [9]. For Subtask 1 they presented models based on the tools μTC (run 1; using Naïve Bayes) and EvoMSA (run 2; that uses EvoDAG, an Evolutionary Algorithm), where the latter is the best classification system. For Subtask 2 their best regression model uses Kernel Ridge.

The UO_UPV [13] team presented systems for both subtasks. For Subtask 1, they created models based on Bi-LSTM [7] neural networks with attention mechanism using word2vec models as input for the network and also a set of linguistically motivated features (stylistic, structural and content, and affective ones). For the first two runs, they combined the output of the Bi-LSTM network with the linguistic features (the second one uses less features), while in their third run they used only the Bi-LSTM part. For Subtask 2, they used the same architecture but modified the last layer so as to minimize the Mean Squared Error over the expected funniness score.

The ELiRF-UPV team participated in Subtask 1 with two systems. They decided to train their systems using character models since upon manual inspection of the corpus they understood humorous tweets tended to include dashes at the beginning of many sentences, not only at the beginning of the tweet. Their first system is a SVM trained using bag of character n-grams of sizes 1 to 8. Their second system is a Convolutional Neural Network whose filter size ranged from 1 to 8 characters.

In Subtask 1, we find 1,642 non-humorous and 733 humorous tweets correctly classified by all the seven submissions, while at the same time there are 46 non-humorous and 118 humorous tweets incorrectly classified by all. The latter set can be considered "hard". Two examples of this hard set are shown in Table 4. By taking a look at the non-humorous subset, there is a great number of tweets that are shaped like dialogues, although in some cases (as in the example) this structure could be used to indicate a list similar to bullet points. This could suggest that all systems in some way learned that dashes indicate humor. The positive subset of hard tweets is more difficult to characterize, but in many cases the tweets involve some not evident world knowledge, such as the one in the example where the ambiguity of "dar rabia" in Spanish is used, as it could be either "get mad" or "get rabies".

Analyzing the votes received for each tweet during the annotation period, we could consider some of the tweets as more "ambiguous" or more "difficult to classify" for a human. For example, the tweets that got five positive (humorous) votes out of five could be considered to be humorous with high confidence, while the ones that received three humorous votes out of five would be more ambiguous. With this information, we can try to analyze the ratio of submissions

Table 4. Examples of "hard" tweets

Text	Is humorous
— Una mujer bella, no necesita maquillaje. — Una mujer sexy, sólo necesita personalidad. — Una mujer feliz, no depende de un hombre.	False
Me da rabia cuando voy caminando por la calle y de pronto me muerde un perro con espuma en el hocico.	True

that correctly predicted the categories of tweets and compare it to this notion of confidence distribution. This comparison, shown in Table 5, suggests that there could be a correlation between the confidence level of a tweet and the proportion of submissions that correctly predict its category, though more tweets and more annotations would be needed in order to confirm this intuition.

Table 5. Proportion of submissions that correctly predicted tweets by number of votes

Category	Votes	Hits
Humorous	3/5	52.25%
	4/5	75.33%
	5/5	85.04%
Not humorous	3/5	68.54%
	4/5	80.83%
	5/5	82.42%

5 Conclusion

We have presented the results of the HAHA Task, the first competition on humor appreciation in the Spanish language. Three teams participated in two subtasks, involving humor classification and funniness scoring. For the former, the seven runs outperformed the two proposed baselines. Interestingly, the best performing system was based on an Evolutionary Algorithm, EvoDAG. This system, presented by the INGEOTEC team, reached 79.72% in F_1 score. For the latter, a regression model based on Kernel Ridge by the same team was the best model, scoring 0.9784 in Root Mean Squared Error and surpassing the proposed baseline for the subtask.

Some directions we would like to explore in the future include the consideration of social strata (e.g. origin, age and gender) of both the Twitter accounts and the annotators in order to understand how this information could affect the detection and rating of humor, and also trying to predict a distribution of votes for a tweet along the rating spectrum, which would need a significantly greater number of votes for each tweet considered.

References

1. Attardo, S., Raskin, V.: Script theory revis(it)ed: Joke similarity and joke representation model. Humor: International Journal of Humor Research (1991)
2. Castro, S., Chiruzzo, L., Rosá, A., Garat, D., Moncecchi, G.: A Crowd-Annotated Spanish Corpus for Humor Analysis. In: Proceedings of SocialNLP 2018, The 6th International Workshop on Natural Language Processing for Social Media (2018)
3. Castro, S., Cubero, M., Garat, D., Moncecchi, G.: Is This a Joke? Detecting Humor in Spanish Tweets. In: Ibero-American Conference on Artificial Intelligence. pp. 139–150. Springer (2016). https://doi.org/10.1007/978-3-319-47955-2_12
4. Freud, S., Strachey, J.: Jokes and Their Relation to the Unconscious. Complete Psychological Works of Sigmund Freud, W. W. Norton & Company (1905)
5. Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Barnden, J., Reyes, A.: Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). pp. 470–478 (2015). https://doi.org/10.18653/v1/s15-2080
6. Graff, M., Tellez, E.S., Miranda-Jiménez, S., Escalante, H.J.: EvoDAG: A semantic Genetic Programming Python library. In: 2016 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC). pp. 1–6 (Nov 2016). https://doi.org/10.1109/ROPEC.2016.7830633
7. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Networks **18**(5-6), 602–610 (2005). https://doi.org/10.1016/j.neunet.2005.06.042
8. Gruner, C.: The Game of Humor: A Comprehensive Theory of Why We Laugh. Transaction Publishers (2000)
9. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. pp. 427–431. Association for Computational Linguistics (April 2017). https://doi.org/10.18653/v1/e17-2068
10. Mihalcea, R., Strapparava, C.: Making Computers Laugh: Investigations in Automatic Humor Recognition. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. pp. 531–538. HLT '05, Association for Computational Linguistics, Stroudsburg, PA, USA (2005). https://doi.org/10.3115/1220575.1220642
11. Minsky, M.: Jokes and the logic of the cognitive unconscious. Springer (1980)
12. Mulder, M.P., Nijholt, A.: Humour research: State of art. Technical Report TR-CTIT-02-34, Centre for Telematics and Information Technology University of Twente, Enschede (September 2002)
13. Ortega-Bueno, R., Muñiz-Cusa, C., Medina, J., Rosso, P.: UO_UPV: Deep Linguistic Humor Detection in Spanish Social Media. In: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018) (2018)
14. Ortiz-Bejar, J., Salgado, V., Graff, M., Moctezuma, D., Miranda-Jiménez, S., Tellez, E.: INGEOTEC at IberEval 2018 Task HaHa: µTC and EvoMSA to Detect and Score Humor in Texts. In: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018) (2018)

15. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
16. Polimeni, J., Reiss, J.P.: The first joke: Exploring the evolutionary origins of humor. Evolutionary Psychology **4**(1), 347–366 (2006). https://doi.org/10.1177/147470490600400129
17. Potash, P., Romanov, A., Rumshisky, A.: SemEval-2017 Task 6:# HashtagWars: Learning a sense of humor. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 49–57 (2017). https://doi.org/10.18653/v1/s17-2004
18. Raskin, V.: Semantic Mechanisms of Humor. Studies in Linguistics and Philosophy, Springer (1985)
19. Reimann, A.: Intercultural communication and the essence of humour. Journal of the Faculty of International Studies **29**(1), 23–34 (2010)
20. Ruch, W., Attardo, S., Raskin, V.: Toward an empirical verification of the general theory of verbal humor. HUMOR: the International Journal of Humor Research (1993)
21. Sjöbergh, J., Araki, K.: Recognizing Humor Without Recognizing Meaning. In: Masulli, F., Mitra, S., Pasi, G. (eds.) WILF. Lecture Notes in Computer Science, vol. 4578, pp. 469–476. Springer (2007). https://doi.org/10.1007/978-3-540-73400-0_59
22. Tellez, E.S., Miranda-Jiménez, S., Graff, M., Moctezuma, D., Suárez, R.R., Siordia, O.S.: A Simple Approach to Multilingual Polarity Classification in Twitter. Pattern Recognition Letters (2017). https://doi.org/10.1016/j.patrec.2017.05.024, http://www.sciencedirect.com/science/article/pii/S0167865517301721
23. Tellez, E.S., Moctezuma, D., Miranda-Jiménez, S., Graff, M.: An automated text categorization framework based on hyperparameter optimization. Knowledge-Based Systems **149**, 110–123 (2018). https://doi.org/10.1016/j.knosys.2018.03.003, https://www.sciencedirect.com/science/article/pii/S0950705118301217