Overview of HAHA at IberLEF 2019:
Humor Analysis based on Human Annotation

Luis Chiruzzo[1], Santiago Castro[21], Mathias Etcheverry[1], Diego Garat[1], Juan José Prada[1], and Aiala Rosá[1]

[1] Universidad de la República, Uruguay
{luischir,mathiase,dgarat,prada,aialar}@fing.edu.uy
[2] University of Michigan, USA
sacastro@umich.edu

Abstract. This paper presents the results of the HAHA task at IberLEF 2019, the second edition of the challenge on automatic humor recognition and analysis in Spanish. The challenge consists of two subtasks related to humor in language: automatic detection and automatic rating of humor in Spanish tweets. This year we used a corpus of 30,000 annotated Spanish tweets labeled as humorous or non-humorous and the humorous ones contain a funniness score. A total of 18 participants submitted their systems obtaining good results overall, we present a summary of their systems and the general results for both subtasks.

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

1 Introduction

This paper describes the results of the second edition of the task Humor Analysis based on Human Annotation (HAHA), part of the IberLEF 2019 workshop.

Despite humor and laughter being universal and fundamental human experiences [32], it has only recently become an active area of research within Machine Learning and Computational Linguistics [37]. Some previous works focus on the computational processing of humor [34,49,10], but a characterization of humor that allows its automatic recognition and generation is far from being specified, even though it has been historically studied from a psychological [19,25], cognitive [36] and linguistic [44,3,46] standpoint. The aim of this task is to gain better insight in what is humorous and what causes laughter, while at the same time fostering the Computational Humor field.

This is the second edition of the HAHA evaluation challenge. In 2018 edition [8] three teams took part in the competition to assess the humor value

and funniness score in a corpus of 20,000 tweets. Although the results for this first edition of the competition were satisfactory, there is room for improvement. There have been similar or related evaluation campaigns in the past, for example: SemEval-2015 Task 11 [21] proposed to work on figurative language, such as metaphors and irony, but focused on Sentiment Analysis. SemEval-2017 Task 6 [43] presented a similar task to this one as well. Additionally, this campaign is related to other evaluation campaigns focused on subjectivity analysis in language such as irony detection [53] [15] and sentiment analysis [33].

In order to address humor in this challenge, we need a working definition of what we will call humor and what could be considered a humorous tweet. In the literature, it is generally accepted that a fundamental part of defining humor is the perception of something being funny [45], which means the opinion of human subjects is essential for determining if something is humorous. However, we must also consider the intent of the author of being humorous or not. In this challenge we define two dimensions: first, we consider a text (tweet) is humorous if the intention of the author was to be funny, as assessed by human judges. Second, we consider how funny a tweet is according to those human judges, but only for the tweets that have already been regarded as attempted humor. These two dimensions are translated into the two subtasks in this challenge.

2 Task description

The following subtasks are proposed for this track:

2.1 Subtask 1: Humor Detection

This subtask has the aim to tell if a tweet attempts 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 (while accuracy is used as another reference).

Two baselines were computed for this subtask over the test data, although finally the first one was published to the participants:

random: Decide randomly with a 50% probability whether a tweet is humorous or not. This baseline achieves 42.0% F_1 score and 50.5% accuracy for the humorous class over the test corpus. This was the only published baseline for Subtask 1.

dash: Select all tweets that start with a dash as humorous (em dash, among many other Unicode variants). This baseline was based on [10], in which the authors found that in Twitter you can get quite decent results given that many tweets considered humorous were dialogues with the utterances delimited by dashes. This heuristic has a high precision (94.5%), as almost all the dialogues in tweets are jokes, but a low recall (16.3%) because there

are more kinds of humorous tweets. The baseline achieves 27.8% F_1 score and 66.9% accuracy for the humorous class over the test corpus.

Note that a majority baseline does not make sense using this evaluation metric because the F_1 score is zero or undefined.

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 the ground truth the average funniness value of the tweets in a corpus. The funniness score is a value from one (attempted humour but not funny) to five (hilarious). This subtask was evaluated using Root Mean Squared Error (RMSE).

We calculated two baselines for this subtask over the test data, but we finally published only one of them:

random: Choose the value 3 (middle of the scale) for all the tweets. The root mean squared error for this baseline over the test data is 2.455. This was the only published baseline for Subtask 2.

average: Choose the average funniness score for the training corpus (2.0464) for all test tweets. The root mean squared error for this baseline over the test data is 1.651.

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 the tweets in the test set, so then the evaluation metric considers only those that truly belong to the humorous class.

3 Corpus

The annotation process for this task followed the same approach as in [9] and [8]. We extracted tweets from specific humorous accounts and random tweets using the Twitter API using Tweepy[3], then we used a web application to crowdsource the labeling of the tweets. Using the app, each annotator has to label a tweet as attempted humor or not attempted humor, and if the annotator chose the former, a score between one and five has to be chosen for the tweet. The main differences between the annotation process this year and last year are the following:

– We extracted the new tweets from the same fifty humorous Twitter accounts we used last year plus all the tweets from thirteen new accounts we found this year from varied Spanish dialects (10,000 new tweets in total).

[3] `https://www.tweepy.org/`

– We extracted 3,000 randomly sampled real-time tweets in Spanish using the Twitter GET statuses/sample endpoint[4] on February 4th and February 7th, 2019.
– The dataset from HAHA 2018 contained some instances of duplicate or near-duplicate tweets (tweets that only differed in a few words and did not change their semantics significantly). We used a semi-automatic process to detect and remove duplicate instances: first we collected all tweet pairs whose Jaccard coefficient was greater than 0.5, then we manually examined those pairs and classified them in equivalence classes, taking only one tweet from each class for the final corpus. 1,278 tweets were removed from last year's corpus.
– Using the web app[5], we crowd-sourced the annotation of all the new tweets and the tweets that had received less than five annotations during the HAHA 2018 annotation process and were considered humorous. The annotation process took part between February and March, 2019. Almost 800 participants took part during the annotation process producing 75,000 votes.

The final corpus consists of 30,000 tweets, where 11,595 (38.7%) are humorous. This is marginally more balanced than that of HAHA 2018, which had 36.8% humorous tweets. This version of the corpus is also cleaner as many near-duplicates have been pruned and we tried to avoid including new ones. We also made sure that all the humorous tweets had at least five votes and all the non-humorous had at least three negative votes.

Text	— Mami, ¿a que no adivinas dónde estoy? — Hijo, ahora no puedo hablar, llámame luego. — No puedo, sólo tengo derecho a una llamada... *— Mommy, can you guess where I am?* *— Son, I can't talk now, call me later.* *— I can't, I'm only entitled to one phone call...*
Is it humorous?	True
Votes: Not humor	1
Votes: 1 star	0
Votes: 2 stars	0
Votes: 3 stars	1
Votes: 4 stars	2
Votes: 5 stars	1
Average Score	4

Table 1. Example instance from the dataset.

[4] https://developer.twitter.com/en/docs/tweets/sample-realtime/api-reference/get-statuses-sample.html
[5] https://clasificahumor.com/

The corpus is divided into 80% training and 20% test. The training set contains both the training and test partitions from last year, and some new tweets to make a total of 24,000 tweets. The new test partition consists entirely of new tweets (6,000). Table 1 shows an example of an instance from the dataset.

4 Systems descriptions

101 teams signed up (asked for the dataset) in the CodaLab competition website[6] but only 18 teams submitted their test predictions at least once. Table 2 lists the submitting teams and related information. We describe hereafter each of their best systems, ordered by their position in F_1 score of the Subtask 1. We do not list the teams jamestjw, vaduvabogdan, Taha, LadyHeidy and jmeaney as they did not submit a paper nor provide information about their models.

Team	CodaLab username	Submissions
acattle	acattle	17
adilism	adilism	8
Amrita_CEN	premjithb	12
Aspie96	Aspie96	13
bfarzin	bfarzin	20
BLAIR_GMU	jimblair	15
garain	garain	10
INGEOTEC	job80	2
jamestjw	jamestjw	14
jmeaney	jmeaney	9
LadyHeidy	LadyHeidy	7
LaSTUS/TALN	abravo	18
OFAI–UKP	dodinh	13
Kevin & Hiromi	kevinb	20
Taha	Taha	12
UO_UPV2	reynier	4
UTMN	zuma	20
vaduvabogdan	vaduvabogdan	10

Table 2. Participant teams ordered alphabetically by team name. The submission count includes both tasks as their trials had to be submitted together by the teams.

adilism [28] used the multilingual cased BERT-base [18] pretrained model along with the fastai library [27]. They first continued its training with the BERT's unsupervised language model objective on the dataset provided for the competition, without the labels. Then they fine-tuned it separately for each task with one-cycle learning-rate style [50] and discriminative fine-tuning [26]. For the Subtask 1, they used a linear layer on top of the output of the last layer for

[6] `https://competitions.codalab.org/competition/22194/`

the [CLS] token with a *tanh* activation linear, then a dropout layer and another linear layer with a binary cross-entropy loss. Apart from this, they used a binarized Multinomial Naïve Bayes as proposed in [55] with unigram and bigram tf-idf features, and combine its predictions with those of the neural network with logistic regression to obtain the final predictions. For the Subtask 2, they changed the BERT model to use the mean-squared loss and combine the predictions with a gradient-boosted tree model from LightGBM [29] instead.

Kevin & Hiromi[7] built an ensemble of 5 models: a forward ULMFiT model [26], a backward ULMFiT model (both with fastai [27]), a pre-trained multimodal cased BERT-base model, a pre-trained multimodal uncased BERT-base model and an SVM with Naive Bayes features from [55] (NBSVM). They combined the predictions with a linear regression model and drew a decision threshold graph to decide win which point the F_1 score is maximized. The first two models were pre-trained (along with a SentencePiece[8] model) on 500,000 new tweets. For the Subtask 2, they use the same ensemble without the NBSVM, and the output is only one score. In the end, they report that they made the model for the Subtask 1 benefit from the model trained for the Subtask 2.

bfarzin [16] trained ULMFiT [26] from scratch using fastai [27] on 475,143 new tweets and tokenizing with Byte Pair Encoding (BPE) [48] (with SentencePiece). Then, they fine-tuned the Language Model on the competition data (without labels) and they fine-tuned each Subtask in a supervised way (separately). They reported that they also have tried with Transformer [54] models and LSTMs instead of QRNNs [5] but found similar performance. The Language Model training was executed with one-cycle learning rate [50] and for the task-specific training they first froze the pre-trained weights for a third of the epochs but then continued the training by fine-tuning them. The best weight initialization were obtained by sampling 20 random seeds. Two linear layers were used as the task-specific layers with ULMFiT. For the Subtask 1 they used cross-entropy loss with label smoothing [42] and they over-sampled the minority class with the Synthetic Minority Oversampling Technique [12]. For the Subtask 2, they used the non-humorous instances as "0" and mean-squared error loss.

INGEOTEC [39] used μTC [52] with sparse and dense word representations. Linear SVM seemed to be the best approach for the Subtask 1 while and SVM regressor was the one for Subtask 2. For text classification, they also explored fastText [4] and flair [1] along with multiple combinations of token embeddings which range from simple characters to BERT [18], as well as EvoMSA [24] and B4MSA [51], but did not obtain an improvement.

BLAIR_GMU [31] used the multilingual cased pre-trained BERT-base [18]. The authors took the last-layer output corresponding to the [CLS] token and added a linear output for classification in the Subtask 1 and use binary cross-entropy loss, while they use mean-squared error for Subtask 2. The authors also

[7] See `http://kevinbird15.com/2019/06/26/High-Level-Haha-Architecture.html` for more information.

[8] `https://github.com/google/sentencepiece`

improved the model for Subtask 1 by considering not only the correct labels but also the output predictions of their model for Subtask 2. The authors do not report whether they use BERT as a feature extraction model or if they fine-tune it.

UO_UPV2 [38] performed lemmatization using FreeLing [40], then used a Spanish word embedding collection developed in-house and hand-crafted features of the type stylistic, structural and content, and affective (including features based on LIWC [41]), to create a vector used as the initial hidden state of a BiGRU [14] neural network with attention followed by three dense layers.

UTMN [23] approached it as a multi-task learning setting with hard parameter sharing [47]: a neural network that processes several types of features in parallel with a common scheme, then concatenates the layers and feeds the outcome to a dense layer. The four concatenated features are: a sentence representation coming from Spanish word embeddings [7] input to a 1D-CNN and a Max Pooling, tf-idf features restricted to 5,000 words plus two dense layers, sentiment and topic modeling features with two dense layers, and some format and other types of hand-crafted features.

LaSTUS/TALN [2] developed a multi-task supervised learning scheme for Humor along with Irony, Sentiment and Aggressiveness using dialect-specific word embeddings, a common BiLSTM layer and two dense layers as classifiers for each task (including both Subtasks).

Aspie96 [22] trained a character-level 1D-CNN with three layers followed by a BiRNN and then a dense layer to output a binary value for the Subtask 1, and a similar approach but with an output value of up to 5 for the Subtask 2.

OFAI–UKP [35] used Gaussian Processes Preference Learning [13], training Gaussian processes using three word representations (Spanish Twitter embeddings [17], the average token frequency in a Wikipedia dump and the word's lemma average polysemy) and several format hand-crafted features.

acattle [11] created a document tensor space for embedding tweets, considering each tweet as a document, and trained Random Trees. They also tried propagating the labels based on the Instance-based Learning technique k-Nearest Neighbors, but this technique did not outperform the first one.

garain [20] used Google Translate for transforming the sentences to English and applied SenticNet5 [6] to get sentiment of the words. The authors transformed the tweets into one-hot vectors and included some manually extracted features of format and sentiment to train a BiLSTM neural network.

premjithb processed the tweets through an embeddings layer and then an LSTM layer for the Subtask 1. For the Subtask 2, they applied doc2vec [30] and used linear regression.

Team	F_1	Precision	Recall	Accuracy
adilism	**82.1**	79.1	**85.2**	**85.5**
Kevin & Hiromi	81.6	80.2	83.1	85.4
bfarzin	81.0	78.2	83.9	84.6
jamestjw	79.8	79.3	80.4	84.2
INGEOTEC	78.8	75.8	81.9	82.8
BLAIR_GMU	78.4	74.5	82.7	82.2
UO_UPV2	77.3	78.0	76.5	82.4
vaduvabogdan	77.2	72.9	82.0	81.1
UTMN	76.0	75.6	76.5	81.2
LaSTUS/TALN	75.9	77.4	74.5	81.6
Taha	75.7	**81.0**	71.1	82.2
LadyHeidy	72.5	74.4	70.8	79.1
Aspie96	71.1	67.8	74.9	76.3
OFAI–UKP	66.0	58.8	75.3	69.8
acattle	64.0	68.3	60.2	73.6
jmeaney	63.6	61.3	66.1	70.5
garain	59.3	49.1	74.8	59.9
Amrita_CEN	49.5	47.8	51.4	59.1
random	44.0	39.4	49.7	50.5
dash	27.8	**94.5**	16.3	66.9

Table 3. Results for the Subtask 1, only the best submission according to F_1 for each team is shown.

5 Results

Table 3 shows the participants results for the Subtask 1. The results are ordered from best to worst in terms of the F_1 score. All participants surpassed the random baseline and also the unpublished baseline that considers the tweets that start with a dash as humorous. The best system was submitted by adilism and achieved an F_1 score of 82.1% and an accuracy of 85.5%. It was followed closely by Kevin & Hiromi (F_1 81.6%) and bfarzin (81.0%).

Table 4 presents the results obtained by the participants in the Subtask 2, which are only thirteen teams, from best to worst ordered by RMSE. It is interesting to observe that the relative order between the participants in this Subtasks is similar to that of the Subtask 1. From this result, we hypothesize that tackling attempted humorousness is highly related to saying how funny a tweet is. All participants surpassed the random baseline, and most of the systems also surpassed the second unpublished baseline of using the average score in the training set. The best system was submitted by adilism and got 0.736 RMSE, followed closely by bfarzin (0.746 RMSE) and Kevin & Hiromi (0.769 RMSE).

By analyzing the team systems, it seems that leveraging the knowledge of other models is what worked the best. Pre-trained language models such as BERT [18] and ULMFiT [26] were used by the best performing systems. However, they need careful manipulation such as with slanted triangular learning rate scheduling or gradual unfreezing [26], as the best teams considered, otherwise

Team	RMSE
adilism	**0.736**
bfarzin	0.746
Kevin & Hiromi	0.769
jamestjw	0.798
INGEOTEC	0.822
BLAIR_GM	0.910
LaSTUS/TALN	0.919
UTMN	0.945
acattle	0.963
Amrita_CEN	1.074
average	1.651
garain	1.653
Aspie96	1.673
OFAI–UKP	1.810
random	2.455

Table 4. Results for the Subtask 2, only the best submission for each team is shown.

the models may incur in catastrophic forgetting (thus not leveraging the existing knowledge) or overfitting. It is also important to test several random seeds to get robust results, as transfer learning based on pre-trained models such as BERT show high variance. Fastai [27] proved to be useful and practical to accomplish it for the best systems. Domain adaptation also seemed to be important, such as continue the language model training with the competition dataset or new tweets, as the pre-trained models are not well-suited to tweets. Apart from this, multi-task learning, which is another way to leverage knowledge, seemed to be useful to many teams based on their results and on what they have reported, including benefiting one subtask from this competition from the other one. To take advantage of multiple techniques, some teams built ensembles (e.g., ensembling neural networks with Naïve Bayes models) that boosted the results according to what they reported. Lastly, we observed that teams signed up regularly during the whole competition timeline, and that their sign up time did not show an clear correlation with their later performance (i.e., teams that started later or download the training data later did not perform worse in general).

6 Conclusions

We presented the HAHA (Humor Analysis based on Human Annotation) task at IberLEF 2019. This automatic humor detection and analysis challenge consists of two subtasks: identifying if a tweet attempts to be humorous or not, and giving a funniness score for the humorous ones. Eighteen participants submitted systems for Subtask 1, the best system achieved 82.1% F_1 for the humorous class and 85.5% accuracy. Thirteen participants submitted systems for Subtask 2, the best system achieved 0.736 in RMSE. All systems surpassed the random

baselines. The top scores in this edition of the competition also beat the top scores achieved last year (79.7% F_1 for Subtask 1 and 0.978 for Subtask 2), although the corpora are different: this year's training set contains all training and test set from last year and some more tweets, and this year's test set is completely new.

Given this year's interest in the task (more than a hundred teams applied to the competition, eighteen teams sent their submissions) and that many of the participants (and potential ones) do not speak Spanish as their main language, it would be interesting to run a challenge similar to this one in other languages, particularly in English. Even for Spanish, given the high variability of language and humor across geography and demographics, it would be interesting to see how this affects the detection and rating of humor. To do this, we would need larger corpora annotated by more people from different linguistic, geographical and social background.

References

1. Akbik, A., Blythe, D., Vollgraf, R.: Contextual string embeddings for sequence labeling. In: COLING 2018, 27th International Conference on Computational Linguistics. pp. 1638–1649 (2018)
2. Altin, L.S.M., Àlex Bravo, Saggion, H.: LaSTUS/TALN at HAHA: Humor Analysis based on Human Annotation. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings, CEUR-WS, Bilbao, Spain (9 2019)
3. Attardo, S., Raskin, V.: Script theory revis(it)ed: Joke similarity and joke representation model. Humor: International Journal of Humor Research (1991)
4. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017). https://doi.org/10.1162/tacl_a_00051, `https://doi.org/10.1162/tacl_a_00051`
5. Bradbury, J., Merity, S., Xiong, C., Socher, R.: Quasi-recurrent neural networks. ArXiv **abs/1611.01576** (2017)
6. Cambria, E., Poria, S., Hazarika, D., Kwok, K.: Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
7. Cardellino, C.: Spanish Billion Words Corpus and Embeddings (March 2016), `https://crscardellino.github.io/SBWCE/`
8. Castro, S., Chiruzzo, L., Rosá, A.: Overview of the HAHA Task: Humor Analysis based on Human Annotation at IberEval 2018. In: CEUR Workshop Proceedings. vol. 2150, pp. 187–194 (2018)
9. Castro, S., Chiruzzo, L., Rosá, A., Garat, D., Moncecchi, G.: A Crowd-Annotated Spanish Corpus for Humor Analysis. In: Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media. pp. 7–11 (2018)
10. 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

11. Cattle, A., Papalexakis, Z.Z.E., Ma, X.: Generating Document Embeddings for Humor Recognition using Tensor Decomposition. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings, CEUR-WS, Bilbao, Spain (9 2019)
12. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
13. Chu, W., Ghahramani, Z.: Preference learning with gaussian processes. In: Proceedings of the 22nd international conference on Machine learning. pp. 137–144. ACM (2005)
14. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
15. Cignarella, A.T., Frenda, S., Basile, V., Bosco, C., Patti, V., Rosso, P., et al.: Overview of the evalita 2018 task on irony detection in italian tweets (ironita). In: Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018). vol. 2263, pp. 1–6. CEUR-WS (2018)
16. Czapla, B.F.P., Howard, J.: Applying a Pre-trained Language Model to Spanish Twitter Humor Prediction. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings, CEUR-WS, Bilbao, Spain (9 2019)
17. Deriu, J., Lucchi, A., De Luca, V., Severyn, A., Müller, S., Cieliebak, M., Hofmann, T., Jaggi, M.: Leveraging large amounts of weakly supervised data for multi-language sentiment classification. In: Proceedings of the 26th international conference on world wide web. pp. 1045–1052. International World Wide Web Conferences Steering Committee (2017)
18. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186 (2019)
19. Freud, S., Strachey, J.: Jokes and Their Relation to the Unconscious. Complete Psychological Works of Sigmund Freud, W. W. Norton & Company (1905)
20. Garain, A.: Humor Analysis based on Human Annotation(HAHA)-2019: Humor Analysis at Tweet Level using Deep Learning. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings, CEUR-WS, Bilbao, Spain (9 2019)
21. 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
22. Giudice, V.: Aspie96 at HAHA (IberLEF 2019): Humor Detection in Spanish Tweets with Character-Level Convolutional RNN. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings, CEUR-WS, Bilbao, Spain (9 2019)
23. Glazkova, A., Ganzherli, N., Mikhalkova, E.: UTMN at HAHA@IberLEF2019: Recognizing Humor in Spanish Tweets using Hard Parameter Sharing for Neural Networks. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings, CEUR-WS, Bilbao, Spain (9 2019)
24. Graff, M., Miranda-Jiménez, S., Tellez, E.S., Moctezuma, D.: Evomsa: A multilingual evolutionary approach for sentiment analysis. arXiv preprint arXiv:1812.02307 (2018)

25. Gruner, C.: The Game of Humor: A Comprehensive Theory of Why We Laugh. Transaction Publishers (2000)
26. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 328–339 (2018)
27. Howard, J., et al.: fastai. `https://github.com/fastai/fastai` (2018)
28. Ismailov, A.: Humor Analysis Based on Human Annotation Challenge at IberLEF 2019: First-place Solution. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings, CEUR-WS, Bilbao, Spain (9 2019)
29. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: Lightgbm: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems. pp. 3146–3154 (2017)
30. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International conference on machine learning. pp. 1188–1196 (2014)
31. Mao, J., Liu, W.: A BERT-based Approach for Automatic Humor Detection and Scoring. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings, CEUR-WS, Bilbao, Spain (9 2019)
32. Martin, R.A., Ford, T.: The psychology of humor: An integrative approach. Academic press (2018)
33. Martínez-Cámara, E., Almeida Cruz, Y., Díaz-Galiano, M.C., Estévez Velarde, S., García-Cumbreras, M.A., García-Vega, M., Gutiérrez Vázquez, Y., Montejo Ráez, A., Montoyo Guijarro, A., Muñoz Guillena, R., Piad Morffis, A., Villena-Román, J.: Overview of TASS 2018: Opinions, health and emotions. In: Martínez-Cámara, E., Almeida Cruz, Y., Díaz-Galiano, M.C., Estévez Velarde, S., García-Cumbreras, M.A., García-Vega, M., Gutiérrez Vázquez, Y., Montejo Ráez, A., Montoyo Guijarro, A., Muñoz Guillena, R., Piad Morffis, A., Villena-Román, J. (eds.) Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018). CEUR Workshop Proceedings, vol. 2172. CEUR-WS, Sevilla, Spain (September 2018)
34. 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
35. Miller, T., Dinh, E.L.D., Simpson, E., Gurevych, I.: OFAI–UKP at HAHA@IberLEF2019: Predicting the Humorousness of Tweets Using Gaussian Process Preference Learning. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings, CEUR-WS, Bilbao, Spain (9 2019)
36. Minsky, M.: Jokes and the logic of the cognitive unconscious. Springer (1980)
37. 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)
38. Ortega-Bueno, R., Rosso, P., Pagola, J.E.M.: UO_UPV2 at HAHA 2019: BiGRU Neural Network Informed with Linguistic Features for Humor Recognition. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings, CEUR-WS, Bilbao, Spain (9 2019)
39. Ortiz-Bejar, J., Tellez, E., Graff, M., Moctezuma, D., Miranda'Jiménez, S.: INGEOTEC at IberLEF 2019 Task HaHa. In: Proceedings of the Iberian Languages

Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings, CEUR-WS, Bilbao, Spain (9 2019)
40. Padró, L., Stanilovsky, E.: Freeling 3.0: Towards wider multilinguality. In: Proceedings of the Language Resources and Evaluation Conference (LREC 2012). ELRA, Istanbul, Turkey (May 2012)
41. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic inquiry and word count: Liwc 2001. Mahway: Lawrence Erlbaum Associates **71**(2001), 2001 (2001)
42. Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., Hinton, G.: Regularizing neural networks by penalizing confident output distributions. arXiv preprint arXiv:1701.06548 (2017)
43. 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
44. Raskin, V.: Semantic Mechanisms of Humor. Studies in Linguistics and Philosophy, Springer (1985)
45. Ruch, W.: Psychology of humor. The primer of humor research **8**, 17–101 (2008)
46. 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)
47. Ruder, S.: An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098 (2017)
48. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909 (2015)
49. 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
50. Smith, L.N.: A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. arXiv preprint arXiv:1803.09820 (2018)
51. 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 **94**, 68–74 (2017)
52. 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)
53. Van Hee, C., Lefever, E., Hoste, V.: Semeval-2018 task 3: Irony detection in english tweets. In: Proceedings of The 12th International Workshop on Semantic Evaluation. pp. 39–50 (2018)
54. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
55. Wang, S., Manning, C.D.: Baselines and bigrams: Simple, good sentiment and topic classification. In: Proceedings of the 50th annual meeting of the association for computational linguistics: Short papers-volume 2. pp. 90–94. Association for Computational Linguistics (2012)