Procesamiento del Lenguaje Natural, Revista nº 67, septiembre de 2021, pp. 257-268 recibido 04-07-2021 revisado 11-07-2021 aceptado 15-07-2021

Overview of HAHA at IberLEF 2021: Detecting, Rating and Analyzing Humor in Spanish

Overview de HAHA en IberLEF 2021: Detección, Valoración y Análisis de Humor en Español

Luis Chiruzzo[1], Santiago Castro[2], Santiago Góngora[1], Aiala Rosá[1], J. A. Meaney[3], Rada Mihalcea[2],
[1]Universidad de la República, Montevideo, Uruguay
{luischir, sgongora, aialar}@fing.edu.uy
[2]University of Michigan, Anne Arbor, USA
{sacastro, mihalcea}@umich.edu
[3]School of Informatics, The University of Edinburgh, Edinburgh, UK
jameaney@ed.ac.uk

Abstract: We present the results of HAHA at IberLEF 2021: Humor Analysis based on Human Annotation. This year's edition of the competition includes the two classic tasks of humor detection and rating, plus two novel tasks of humor logic mechanism and target classification. We describe the corpus created for the challenge, the competition phases, the submitted systems and the main results obtained.
Keywords: Computational humor, Spanish, Humor mechanism, Humor target.

Resumen: Presentamos los resultados de HAHA en IberLEF 2021: Humor Analysis based on Human Annotation. La edición de la competencia de este año incluye las dos tareas clásicas de detección y valoración de humor, más dos tareas nuevas de clasificación de mecanismo y objeto de humor. Describimos la creación del corpus, las fases de la competencia, los sistemas enviados y los principales resultados obtenidos.
Palabras clave: Humor computacional, Español, Mecanismo de humor, Objeto de humor.

1 Introduction

American author E. B. White once said: *"Explaining a joke is like dissecting a frog. You understand it better but the frog dies in the process."* It is generally agreed upon that analyzing humor is a difficult endeavor that removes all the amusement of the activity. However, we believe focusing on humor analysis is important and it is one way of linking current work on computational humor with a more theoretical background.

The field of computational humor has had a surge in recent years, as can be seen by the growing number of shared tasks related to the subject that have been organized. Most of the time these tasks focus on humor detection, and on occasions also humor rating, but a deeper analysis of the way humor works and the topics it deals with continues to be largely unexplored (with some exceptions). Our objective with the HAHA task is to go further in the direction of analyzing humor structure and content, while at the same time continuing to explore the more established tasks of humor detection and rating.

1.1 Background

The study of humor from a computational and machine learning perspective is relatively new. Some noticeable previous works include (Mihalcea and Strapparava, 2005; Sjöbergh and Araki, 2007; Castro et al., 2016), but a characterization of humor that allows its automatic recognition and generation is far from being specified. Figurative language, and in particular humor, has been a productive area of research as regards shared tasks for several years. SemEval-2015 Task 11 (Ghosh et al., 2015) focused on the challenging aspects posed by figurative language, such as metaphors and irony. SemEval-2017 Task 6 (Potash, Romanov, and Rumshisky, 2017) presented humorous tweets submitted to a comedy program, and asked competitors to predict the ranking that the comedy program's audience and producers gave the tweets. The previous two editions of the HAHA: Humor Analysis based on Human Annotation task, at IberEVAL 2018 (Castro,

ISSN 1135-5948. DOI 10.26342/2021-67-22

Chiruzzo, and Rosá, 2018; Castro et al., 2018) and IberLEF 2019 (Chiruzzo et al., 2019), consisted of two subtasks: Humor Detection and Funniness Score Prediction. SemEval-2020 Task 7 (Hossain et al., 2020) proposed a task of humor rating in which participants had to predict how humorous an edited headline was and to predict which of two edits to the same headline was funnier. More recently, SemEval-2021 Task 7, called Hahackaton (Meaney, 2020; Meaney et al., 2021), combined humor detection with offense detection, proposing the same subtasks as in HAHA 2018 and 2019, and adding two additional tasks: Offense Score Prediction and Controversial Humor Classification.

Besides these competitions focusing on the more classical tasks of humor detection and rating, SemEval 2017 (Miller, Hempelmann, and Gurevych, 2017) took another approach trying to analyze one particular (and very common) class of jokes: puns. The first task in this competition was the more usual approach of detecting if a text in English contains a pun, in the second task the participants had to detect exactly which word of the text is the pun, and the third task implied detecting what are the different senses the pun word can be interpreted as. We believe focusing on this type of analysis is a promising way of moving forward in the field of computational humor, but in our new tasks, instead of trying to explore one type of humor mechanism in depth, we take a broader approach to detecting a larger set of humor mechanisms, as well as exploring the most common targets associated to jokes.

1.2 New Tasks

Mirroring the growing interest in computational humor generally, the HAHA task has attracted more participants with each iteration. Three research groups participated in two tasks during the first edition. Interest rose sharply in the second edition of the task, with 18 participants. However, the performance achieved by systems in these first and second editions was still far from human-level for humor detection. For this reason, in this third edition we included the same two tasks of humor detection and rating from the previous editions, with some minor changes. Firstly, in the dataset used for the previous edition, there were about 38.7 % of humorous tweets, but this time the new test set was created with the aim of keeping it as balanced as possible between the humorous and non-humorous classes. Secondly, we endeavoured to include annotators from more diverse backgrounds (see Section 2.1).

We also aimed to advance the field of computational humor by adding two new tasks which are directly inspired by one of the most well-known and comprehensive theories of humor, the General Theory of Verbal Humor (GTVH) (Attardo and Raskin, 1991). This theory claims there are six Knowledge Resources (KRs) used in jokes, which characterize the type of humor contained therein. In particular, we propose to focus on these two:

Logic Mechanism (LM) contemplates how the joke works, what are the means by which it conveys humor (e.g., analogy, exaggeration, wordplay).

Target (TA) identifies if somebody is being laughed at (the butt of the joke) and who that entity is, which relates to the content of the text.

Note that this is not the only possible categorization. (Tsakona, 2009) presents some practical examples of this theory, (Attardo, Hempelmann, and Di Maio, 2002) presents a deeper categorization, while (Reyes et al., 2009) describes another possible way of organizing jokes in a taxonomy, (Berger, 2017) describes a comprehensive list of 45 mechanisms that are used to convey humor in jokes, and (Buijzen and Valkenburg, 2004) follows the same path for analyzing audiovisual humor used in television commercials. We used these ideas as a starting point for our categories' definition and, as we will see in Section 2.2, we adapted them to the types of mechanisms we found in our dataset.

2 Corpus

For the 2019 edition of this task, we built a corpus of 30,000 crowd-annotated tweets (Chiruzzo et al., 2019; Chiruzzo, Castro, and Rosá, 2020). The tweets are labeled to indicate whether they are humorous or not, and each humorous tweet is also annotated with a funniness score, a number between 1 and 5. All tweets considered humorous have at least five annotations, while all tweets considered non-humorous have at least three negative annotations. This corpus was split between 24,000 tweets for training and 6,000 for test. This year's edition of the corpus has

36,000 tweets in total: the 2019 training and test sets were used this year as training and development sets, and we created a new test set of 6,000 tweets. We also annotated a subset of 20 % of each partition with information about humor mechanism and targets.

2.1 New Test Set

The new test corpus was developed by crowdsourcing in a similar way to the corpus used in previous editions, using the *clasificahumor*[1] tool. In previous editions of the competition, we relied on volunteers for the annotation, but this year we aimed to improve this situation by sourcing annotators from the Prolific[2] crowd-sourcing platform and paying them accordingly.

As in previous editions, we searched Twitter for accounts that regularly posted jokes in Spanish. We downloaded 15,655 tweets from 18 humorous accounts. We then performed a check to remove near-duplicates as in (Chiruzzo, Castro, and Rosá, 2020): we calculated the Jaccard coefficient for every pair of tweets and built clusters of tweets with a similarity score of more than 0.5, then we manually inspected every cluster and tagged them as near-duplicates or not. From the near-duplicate clusters, only one tweet was kept and the rest were discarded. We also repeated this analysis on the 2019 corpus to avoid including tweets that were too similar to the previous ones. After this pruning, 13,032 tweets were left in the new collection. For the non-humorous texts, we also downloaded a random collection of 11,353 tweets in Spanish.

The aim was to create a test set of 6,000 tweets, expecting 5 annotations for each of them, with as much balance as possible between the humorous and non-humorous categories. The dataset was annotated in six rounds, each consisting of 1,000 tweets annotated by 26 Spanish-speaking annotators hired through Prolific. Afterward, there was a final smaller round to annotate some tweets that obtained fewer annotations. Each annotator labeled 200 tweets, selecting if each tweet is humorous, and if so how funny it is on a scale from 1 to 5 (see Fig. 1). The task took on average 30 minutes and the annotators were paid USD 5,00 (10,00 USD/hour).

Each batch contained 1,000 tweets drawn from the humorous accounts and the random tweets collection. However, the number of tweets selected from each collection varied between batches to keep the collection as balanced as possible. After each round, we calculated how many tweets were labeled as humorous, and adapted the proportion of tweets between collections in the next round accordingly.

Figure 1: Screenshot of the web tool used for the annotation.

In each batch, ten hand-picked tweets were taken from the 2019 corpus for spam checking (five humorous tweets and five non-humorous tweets). These tweets were mandatory for all annotators. They were intended to check if the users had understood the task and to gauge their attention level. If a user failed to label more than 60 % of these tweets correctly, their annotations were discarded. Fortunately, only a handful of annotators failed this quality check and they were promptly replaced with other annotators.

The result of the annotation process is a corpus of 6,000 tweets with exactly 50 % of the tweets classified as humorous. All of the humorous tweets have five annotations each (at least three positive ones, with their corresponding humor rating), while all of the non-humorous tweets have at least three negative annotations.

Around 170 annotators from different Spanish-speaking countries took part in the process. The countries with the most annotators were Mexico (72), Chile (45), and Spain (36); but there were also annotators from Argentina, Bolivia, Colombia, and Venezuela, among others (see Fig. 2). The agreement between raters for humorous/non-humorous classification measured with Krippendorff's

[1] https://www.clasificahumor.com/
[2] https://www.prolific.co/

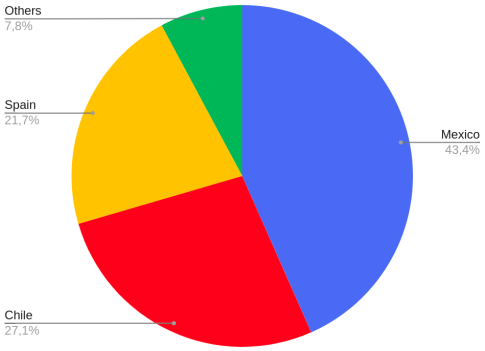

Figure 2: Country of origin of the annotators from Prolific. Our only filter for annotators was that they spoke Spanish as a first language.

alpha was 0.604, similar to the 0.605 score reported in previous editions (Chiruzzo, Castro, and Rosá, 2020).

In contrast, the agreement for humor rating measured as Krippendorff's alpha was considerably lower than in previous tasks. We obtained a 0.085 score (calculated as interval type of measurement), while it was 0.224 for HAHA 2019 (Chiruzzo et al., 2019) and 0.124 for Hahackaton (Meaney et al., 2021). We believe one reason this could be happening is the geographic diversity of the annotators sourced from Prolific, which might have had the effect of increasing the diversity of opinions on the highly subjective matter of humor rating. Due to the way we promoted the web annotation tool in previous editions, the distribution of annotators could have been biased with a large number of annotators from Uruguay and fewer annotators from the rest of the Spanish-speaking countries. This over-representation of annotators from one country, sharing a common background, might have produced more homogeneous opinions on some jokes. To a lesser extent, something similar might have happened in Hahackaton: although the annotators were carefully selected to cover many age groups, they were all from the United States.

2.2 Corpus for Logic Mechanism and Target

The annotation of the corpus for the new tasks was more complex as we were dealing with uncharted territory. There were three rounds of annotation: first an exploratory phase for finding categories, then the bulk of the annotation process, and finally a refinement phase.

Initial definition of categories: We aimed to define a suitable set of categories that would be comprehensive enough to categorize the texts for our dataset, but keeping in mind they were going to be used in the context of a machine learning competition. We started by sampling a set of 200 tweets from the 2019 corpus and having them annotated by four annotators (organizers of the competition) using the humor categories from (Berger, 2017). These are 45 categories grouped in four superclasses: *logic* (e.g. absurd, analogy, mistakes), *language* (including exaggeration, wordplay or puns, sarcasm), *identity* (e.g. embarrassment, parody, unmasking) and *action* (e.g. slapstick). From the beginning it was clear that this number of categories would be too large, and many of them were not found in our dataset (for example, the *action* categories made no sense for the verbal humor in tweets) so the objective was to narrow it down to a manageable number of labels, and also to detect different categories that were not in this original set but could be present in the corpus. The annotators were also asked to identify the individuals/groups which were the target of the jokes in the tweets. We discussed and iterated over the annotations obtained in this first approach until we reached an initial set of categories to use, which included 12 categories for the mechanism (see Section 3.3 for the definitions) and 22 categories for the target. The target categories were organized in a tree with 12 superclasses, so new labels could be added as appropriate leaves of the tree.

Annotation of the corpus: Using this initial set, we selected annotators to process the corpus. Eleven annotators participated in the annotation process for the training and development sets, they were all Computer Science students that had taken at least one course in NLP, and they were compensated with course credit. Each annotator labeled a total of 600 tweets, 500 of these were unique to the annotator, while 100 were shared with another annotator in order to calculate the inter-annotator agreement. Their instructions were to use the 12 categories for mechanism, but add new categories to the targets tree as necessary, as we knew the corpus could contain many more targets.

Refinement: Once the annotation process was finished, a total of 58 target categories

Round 1		Round 2		Round 3
Categories	Subcategories	Categories	Subcategories	Categories
age	children (2), teens, elderly (1)	age	children (41), elderly (52), teens (25)	age (121)
body shaming (9)		body shaming (215)		body shaming (224)
mothers-in-law (1)		family	aunts (3), couples (87), ex (25), fathers (6), grandmothers (4), husbands (1), mothers-in-law (52), mothers (45), orphans (1), widows (1), wives (9)	family/relationships (234)
gender	men (3), women (15), homosexuals (1)	gender	homosexuals (50), men (102), transgender (5), women (329), others (1)	lgbt (57) men (105) women (345)
health	alcoholics (4), illness (1), mental illness (2)	health	addictions (23), alcoholics (77), disability (12), illness (55), mental illness (21)	health (70) substance use (104)
origin	ethnicity (4), race, immigrants (2)	origin	ethnicity (64), immigrants (6), race (17)	ethnicity/origin (93)
professions	doctors (1), footballers (3), musicians (3), politicians (3), other professions (3)	professions	actors (7), bankers (2), boxers (3), builders (6), doctors (49), engineers (6), entertainment figures (2), footballers (47), lawyers (13), musicians (52), nuns (2), politicians (37), sex workers (4), teachers (11), other professions (76)	professions (328)
religion	jewish, jehovah witness	religion	atheists (1), christians (42), jehovah witness (9), jewish (2), others (2)	religion (56)
self-deprecating (20)		self-deprecating (237) self-flattering (3)		self-deprecating (257)
sexual aggressors	paedophiles (1), rapists (1)	sexual aggressors	paedophiles (3), rapists (9), others (3)	sexual aggressors (18)
social status	poor (3), rich	social status	poor (60), rich (5)	social status (68)
		organizations (77)		technology (63)
hipsters (2)		hipsters (11)		
		ideology	communism (1)	

Table 1: Target categories (and subcategories) found in each round of annotation: in the first round of 200 tweets we found 22 categories, in the second round of 6,000 tweets we found 58 categories, in the final annotation round we unified and simplified classes to get to 15 categories. The numbers in parenthesis represent the number of instances found for each category.

were found in the corpus. Two more annotators (organizers of the competition) went through the annotations collecting and unifying all the targets into a tree of categories. We then analyzed the nodes of the tree looking for a set of categories that was manageable but also contained the most representative ones. We ended up settling on a collection of 15 categories.

Table 1 shows a summary of the categories and number of tweets found for each one of them in the three rounds of annotation.

The final step was annotating a subset of the new test set. Two annotators (from the organizing team) took part in this, annotating 650 tweets each (100 tweets were shared for calculating inter-tagger agreement). In this case, we considered the categories for mechanism and target as fixed.

The average Cohen's kappa achieved for inter-annotator agreement of mechanism annotations in the training and dev sets was 0.365. This agreement was a little better for the test set, with 0.449. We believe the higher inter-annotator agreement in test than in train and dev is due to the expertise of the annotators, but in any case we can affirm that it is a very difficult task.

On the other hand, we calculated agreement for the annotations of targets (a multi-

class task) as the F1 score between annotators (taking one annotator as gold and the other one as the candidate). This agreement is 0.375 on average for the train and dev sets, and slightly lower for the test set: 0.350. This way of calculating the agreement does not take into account the large number of labels there are, but at least it may give an idea of how difficult the task is even for humans.

2.3 Composition of the Corpus

Table 2 shows a summary of the composition of the corpus split in train, development, and test sets. We include the number of tweets that are labeled with each of the mechanism and target categories as well.

	Training	Dev	Test
Tweets	24000	6000	6000
Humorous	9253	2342	3000
Mechanism and target labeled	4800	1200	1200
Having at least one target	1629	399	400
Mechanism labels			
absurd	566	142	136
analogy	319	84	53
embarrassment	301	72	28
exaggeration	476	103	75
insults	146	40	21
irony	371	90	100
misunderstanding	416	100	94
parody	255	59	65
reference	578	121	85
stereotype	230	68	35
unmasking	441	130	69
wordplay	701	191	439
Target labels			
age	105	16	15
body shaming	181	43	28
ethnicity/origin	69	24	41
family/relationships	177	57	55
health	58	12	24
lgbt	40	17	13
men	92	13	23
professions	263	65	63
religion	45	11	6
self-deprecating	212	45	36
sexual aggressors	13	5	8
social status	52	16	8
substance use	83	21	15
technology	51	12	10
women	287	58	74

Table 2: Composition of the corpus.

3 Tasks

The HAHA 2021 competition consisted of four tasks. Two of them were analogous to the tasks proposed in HAHA 2018 and 2019, and we proposed two novel tasks for this iteration. We also created new baselines for the first two tasks that are stronger than the ones used in previous editions, aiming to increase the challenge.

3.1 Humor Detection

Given a tweet, the task of humor detection is to determine if its content is humorous or not (intended humor by the author; i.e. a joke). The main metric for measuring performance for this task is the F1 score of the 'humorous' class.

In previous years we used a simple random baseline for this task, which was a very weak baseline meant to encourage participation in the task. This year we used a slightly stronger baseline, but still one of the simplest machine learning methods: we trained a Naïve Bayes classifier with TF-IDF features. This method achieves an F1 of 0.6493 on the dev set, and 0.6619 F1 on the test set.

3.2 Humor Rating

The humor rating task is to predict a funniness score value for a tweet on a 5-star ranking, assuming it is humorous. The performance of this task is measured using the root mean squared error (RMSE) of the humor rating.

In previous years the baseline for this task assigned the average rating found in the training corpus to all tweets. This year we trained a SVM regression model with TF-IDF features – arguably the strongest of the baselines we used for this competition, beating the top scores achieved in previous editions. This method achieves 0.6532 RMSE on the dev set, and 0.6704 RMSE on the test set.

3.3 Humor Logic Mechanism Classification

For a humorous tweet, this task is to predict the mechanism by which the tweet conveys humor from a predefined set of classes. In this task, only one class per tweet is allowed. The possible categories for this task are the following:

- **Absurd**: Humor comes from a logical inconsistency in the reasoning.

- **Analogy**: It is a comparison between dissimilar elements.
- **Embarrassment**: In the punchline one of the participants shames or embarrasses another one.
- **Exaggeration**: There is a situation or comparison that is exaggerated.
- **Insults**: There are insults to the characters in the joke or to real life people.
- **Irony**: They say something but mean the opposite, or they describe a contradictory situation.
- **Misunderstanding**: Humor comes from a participant understanding a question or a situation wrong.
- **Parody**: The text is similar to another known text or work (for example a song, a saying, or a movie dialog) but it is modified to make it humorous.
- **Reference**: It describes a real life situation, generally mundane, that the reader might relate to or not, but when the reader does identify with the situation it results in a humorous effect[3].
- **Stereotype**: Humor comes from using a social group, ethnicity or profession to remark on a stereotypical characteristic.
- **Unmasking**: Humor comes from a character acting in a certain way and later showing that their intentions or characteristics were different than initially thought.
- **Wordplay**: Uses word ambiguity, made up words or combinations of words to give a humorous sense.

The main metric for measuring performance in this task is the macro-averaged F1 score.

The baseline for this task is also a Naïve Bayes model trained with TF-IDF features. This method obtains a 0.1038 F1 score for the dev set, and 0.1001 for the test set.

[3]This is the only mechanism category that does not correspond to at least one of the categories from (Berger, 2017), but it is a particular type of humorous text that is very common in the dataset.

3.4 Humor Target Classification

For a humorous tweet, the target classification task consists in predicting the target of the joke based on its content (what/who it is making fun of) from a predefined set of classes. In this case, there may be many classes associated to a tweet, and also tweets that do not belong to any of the categories (it is a multi-label classification). In this case, each tweet can be labeled with zero or more of the following categories: **age**, **body shaming**, **ethnicity/origin**, **family/relationships**, **health**, **LGBT+**, **men**, **professions**, **religion**, **self-deprecating**, **sexual aggressors**, **social status**, **substance use**, **technology**, and **women**. This task might be related to other important NLP tasks such as detection of offensive content or hate speech.

To measure performance in this task, we consider the labels as pairs (tweet id, category), and calculate the macro-averaged F1 score of finding those exact pairs.

The baseline of this task is more elaborate. We first experimented with using different Naïve Bayes models for each target, but they could capture absolutely none of the targets in the corpus. We thus devised another method: assigning the label X to a tweet if it contains one of the *top words* for label X in the training corpus. The collection of *top words* was created by selecting the 50th to 60th most frequent words for the label (thus discarding the words that were too common, and the ones that were too rare). This method obtained 0.0595 F1 score on the dev set, and 0.0527 on the test set.

4 Competition

The competition ran between March 18 and June 10, 2021 on the CodaLab[4] platform. During that time, a total of 74 users registered to participate, and 18 of those users submitted at least one system for the development or the evaluation phase.

4.1 Phases

The competition consisted of three phases:

Development phase: from April 8 to May 26. At the beginning of this phase, we released the training and development sets. Participants could train their systems and compare their results for the development set. Each

[4]https://competitions.codalab.org/competitions/30090

participant could submit up to 200 systems. There were 276 submissions.

Evaluation phase: from May 27 to June 9. At the beginning of this phase, we released the test set. Participants could run their already trained systems on the test tweets and submit their results. Each participant could submit only up to ten submissions. There were 140 submissions.

Post-Evaluation phase: from June 10 onward. This phase started after the competition ended so anyone can officially benchmark their system. It could be used in the future to advance the state of the art in these tasks or to test alternative methods that the users could not send to the evaluation phase, although the results obtained in this phase would not be part of the official results of the competition.

4.2 Systems Descriptions

Almost all the systems submitted to the competition used neural networks for their solutions, in most cases based on pre-trained neural language models such as BERT (Devlin et al., 2018), GPT-2 (Radford et al., 2019), or BETO (Cañete et al., 2020), a BERT-based model trained entirely with Spanish texts. Some teams also trained other types of models (for example SVM or Decision Trees) in order to make comparisons, but none of the participants that sent their system descriptions submitted any of these models. In what follows, we give a brief description of each system:

Jocoso (Grover and Goel, 2021), user **TanishqGoel**, experimented with an ensemble of multiple transformer architectures, fine-tuned on the humor dataset. They reached very good results in all four tasks, including the best result for Task 1.

icc (García Subies, Betancur Sánchez, and Vaca, 2021) performed a fine-tuning of BETO distinctly for each task, adjusting some hyperparameters for Task 1 also, such as learning rate, batch size and dropout rate. To preprocess the data they normalize every URL, username and laugh using a unique token in each case ("[URL]", "[USER]" and "haha").

ColBERT (Annamoradnejad and Zoghi, 2021), user **moradnejad**, presented an adaptation of the ColBERT model to Spanish using BETO, feeding a neural network which has two parallel paths: One path models each sentence separately, and the other path models the whole text, capturing the incoherence of the final line (punchline) with respect to the previous ones.

BERT4EVER (Wang et al., 2021), user **Neakail**, used a model based on BERT, continuing its pre-training with the training data from the task. For Tasks 3 and 4, they used a pseudo-labeling technique, using the model to predict the categories of the unlabeled tweets and keeping the ones with high confidence, creating 1940 more silver-standard examples. Training a model with this new data obtained the best results for Tasks 3 and 4 in the competition.

RoMa (Rodriguez, Ortega-Bueno, and Rosso, 2021), user **MJason**, presented a neural network approach, combining Siamese Networks, to obtain a representation for each tweet, and Reinforcement Learning, for clustering tweets based on the learned representation.

UMUTeam (García-Díaz and Valencia-García, 2021), user **JAGD**, approached the four tasks using linguistic features and transformers. They use both the preprocessed data and the original one in order to obtain sentence embeddings, using fastText vectors and Spanish BERT, and linguistic features, such as writing style or misspellings, using UMUTextStats. They got the best result for Task 2 in this competition.

skblaz used a model called autoBOT, an autoML technique for text which combines different feature spaces (Škrlj et al., 2021). The system was run for 8 hours, the default neurosymbolic model was used. They report this was one of the first non-English attempts with autoBOT.

kuiyongyi (Kui, 2021) built a system based on Multilingual BERT and LSTM models for Tasks 1, 3 and 4, and a GPT-2 based model for Task 2.

N&&N (Alsalman and Ennab, 2021), user **sarasmadi**, also used an adaptation of the ColBERT model to create embeddings,

Team	Username	Task 1 Score	Task 2 Score	Task 3 Score	Task 4 Score
Jocoso	TanishqGoel	**0.8850** (1)	0.6296 (3)	0.2916 (2)	0.3578 (2)
icc	icc	0.8716 (2)	0.6853 (9)	0.2522 (3)	0.3110 (4)
ColBERT	moradnejad	0.8696 (3)	0.6246 (2)	0.2060 (7)	0.3099 (5)
kuiyongyi	kuiyongyi	0.8681 (4)	0.6797 (8)	0.2187 (5)	0.2836 (6)
noda risa	jgcarrasco	0.8654 (5)	-	-	-
BERT4EVER	Neakail	0.8645 (6)	0.6587 (4)	**0.3396** (1)	**0.4228** (1)
RoMa	Mjason	0.8583 (7)	1.1975 (11)	-	-
UMUTeam	JAGD	0.8544 (8)	**0.6226** (1)	0.2087 (6)	0.3225 (3)
skblaz	skblaz	0.8156 (9)	0.6668 (6)	0.2355 (4)	0.2295 (7)
humBERTor	sgp55	0.8115 (10)	-	-	-
RoBERToCarlos	antoniorv6	0.7961 (11)	0.8602 (10)	0.0128 (10)	0.0000 (9)
N&&N	sarasmadi	0.7693 (12)	-	0.0404 (9)	-
TECHSSN	ayushnanda14	0.7679 (13)	0.6639 (5)	-	-
KdeHumor	kdehumor	0.7441 (14)	1.5164 (12)	-	-
baseline		*0.6619* (15)	*0.6704* (7)	*0.1001* (8)	*0.0527* (8)

Table 3: Best result for each task for all teams in the competition. The numbers in parenthesis indicate the position of the team with respect of the other participants in that task.

and used these embeddings as hidden layers in a neural network.

TECHSSN (Nanda, Singh, and Gupta, 2021), user **ayushnanda14**, created a model based on a fine-tuning of BERT adapted to Tasks 1 and 2. They use the BERT encoding of the whole text and also individual sentences, extracting features from all of them for the final classification.

KdeHumor (Miraj and Aono, 2021) used a neural network approach for Tasks 1 and 2. The network has three layers: an embeddings layer that uses pretrained Spanish word embeddings, a multi kernel CNN layer, and a BiLSTM layer.

Besides these submissions, there were three more teams that participated in the competition (they are located around the middle of the table) but did not send any description of their system. The teams **noda risa**, **humBERTor** and **roBERTocarlos** did not send any description of their systems, but are still included in Table 3.

5 Results

Table 3 shows the results for all the submitting teams, including the top result for each task for each team on the test set. The best system obtained 88.5 % F1 for humor detection, a great improvement over the best result in 2019 (82.1 % F1) and 2018 (79.7 % F1). However, we must take in consideration that the test sets for the three editions were different, so they are not directly comparable. The same happens for humor rating: this year's top system got 0.6226 RMSE for Task 1, while in 2019 the best system achieved 0.736, and in 2018 the best system achieved 0.9784. The numbers for this task seem to be improving as well, again with the caveat that the test sets were different.

The humor mechanism and humor target classification tasks were new this year, and in this case the best system got 33.96 % macro-averaged F1 for the mechanism and 42.28 % for the targets. Even though these are harder tasks, we consider many systems performed better than we expected, beating the baselines by a large margin. Although there is still considerable room for improvement, we find these initial results encouraging to keep advancing in this direction and pushing the limits in humor analysis.

6 Conclusions

We presented the third edition of the HAHA task at IberLEF, including two new subtasks – humor mechanism and humor target classification – in addition to the two tasks already present in previous editions – humor detection and humor rating. For this year's edition, the existing dataset was extended with a new test set, and also a subset was enriched with annotations for the new challenges.

Fourteen teams participated in the task, most of them used neural networks based on pre-trained neural language models.

Regarding Tasks 1 and 2, some participating teams achieved better results than in previous editions, reaching 88.5 % F1 in task 1 and 0.6226 RMSE in Task 2.

For Tasks 3 and 4, even if the results were not very high, most of the teams were able to improve over the proposed baselines. We consider that these are encouraging results, and we believe that with a larger corpus the learning process could be improved.

The labeled datasets compiled for this challenge are publicly available[5]. We hope these datasets and the insights from the current evaluation will encourage more research on the challenging tasks of humor detection, rating, and analysis.

References

Alsalman, N. and N. Ennab. 2021. N&&N at HAHA@IberLEF2021: Determining the Mechanism of Spanish Tweets using ColBERT. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings, Málaga, Spain, 9. CEUR-WS.

Annamoradnejad, I. and G. Zoghi. 2021. ColBERT at HAHA 2021: Parallel Neural Networks for Rating Humor in Spanish Tweets. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings, Málaga, Spain, 9. CEUR-WS.

Attardo, S., C. F. Hempelmann, and S. Di Maio. 2002. Script oppositions and logical mechanisms: Modeling incongruities and their resolutions. *Humor*, 15(1):3–46.

Attardo, S. and V. Raskin. 1991. Script theory revis(it)ed: Joke similarity and joke representation model. *Humor: International Journal of Humor Research*.

Berger, A. A. 2017. *An anatomy of humor*. Routledge.

Buijzen, M. and P. M. Valkenburg. 2004. Developing a typology of humor in audiovisual media. *Media psychology*, 6(2):147–167.

Castro, S., L. Chiruzzo, and A. Rosá. 2018. Overview of the HAHA Task: Humor Analysis based on Human Annotation at IberEval 2018. In *CEUR Workshop Proceedings*, volume 2150, pages 187–194.

Castro, S., L. Chiruzzo, A. Rosá, D. Garat, and G. Moncecchi. 2018. A crowd-annotated Spanish corpus for humor analysis. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 7–11, Melbourne, Australia, July. Association for Computational Linguistics.

Castro, S., M. Cubero, D. Garat, and G. Moncecchi. 2016. Is this a joke? detecting humor in spanish tweets. In *Ibero-American Conference on Artificial Intelligence*, pages 139–150. Springer.

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Chiruzzo, L., S. Castro, M. Etcheverry, D. Garat, J. J. Prada, and A. Rosá. 2019. Overview of HAHA at IberLEF 2019: Humor Analysis based on Human Annotation. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*, CEUR Workshop Proceedings, Bilbao, Spain, September. CEUR-WS.

Chiruzzo, L., S. Castro, and A. Rosá. 2020. HAHA 2019 Dataset: A Corpus for Humor Analysis in Spanish. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5106–5112.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

García-Díaz, J. A. and R. Valencia-García. 2021. UMUTeam at HAHA 2021: Linguistic Features and Transformers for Analysing Spanish Humor. The What, the How, and to Whom. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings, Málaga, Spain, 9. CEUR-WS.

García Subies, G., D. Betancur Sánchez, and A. Vaca. 2021. BERT and SHAP for Humor Analysis based on Human Annotation. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings, Málaga, Spain, 9. CEUR-WS.

[5]https://github.com/pln-fing-udelar/pln-inco-resources/tree/master/humor/haha2021

Ghosh, A., G. Li, T. Veale, P. Rosso, E. Shutova, J. Barnden, and A. Reyes. 2015. Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 470–478.

Grover, K. and T. Goel. 2021. HAHA@IberLEF2021: Humor Analysis using Ensembles of Simple Transformers. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings, Málaga, Spain, 9. CEUR-WS.

Hossain, N., J. Krumm, M. Gamon, and H. Kautz. 2020. Semeval-2020 task 7: Assessing humor in edited news headlines. *arXiv preprint arXiv:2008.00304*.

Kui, Y. 2021. Applying Pre-trained Model and Fine-tune to Conduct Humor Analysis on Spanish Tweets. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings, Málaga, Spain, 9. CEUR-WS.

Meaney, J. 2020. Crossing the line: Where do demographic variables fit into humor detection? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 176–181.

Meaney, J., S. Wilson, L. Chiruzzo, A. Lopez, and W. Magdy. 2021. SemEval 2021 Task 7: HaHackathon, Detecting and Rating Humor and Offense. In *15th International Workshop on Semantic Evaluation*.

Mihalcea, R. and C. Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 531–538, Stroudsburg, PA, USA. Association for Computational Linguistics.

Miller, T., C. F. Hempelmann, and I. Gurevych. 2017. SemEval-2017 task 7: Detection and Interpretation of English Puns. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 58–68.

Miraj, R. and M. Aono. 2021. Humor Detection in Spanish Tweets Using Neural Network. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings, Málaga, Spain, 9. CEUR-WS.

Nanda, A., A. P. Singh, and A. Gupta. 2021. TECHSSN at HAHA @ IberLEF 2021: Humor Detection and Funniness Score Prediction using Deep Learning Techniques. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings, Málaga, Spain, 9. CEUR-WS.

Potash, P., A. Romanov, and A. Rumshisky. 2017. Semeval-2017 task 6:# hashtagwars: Learning a sense of humor. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 49–57.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Reyes, A., P. Rosso, A. Martí, and M. Taulé. 2009. Características y rasgos afectivos del humor: Un estudio de reconocimiento automático del humor en textos escolares en catalán. *Procesamiento del lenguaje natural*, 43:235–243.

Rodriguez, M., R. Ortega-Bueno, and P. Rosso. 2021. RoMa at HAHA-2021: Deep Reinforcement Learning to Improve a Transformed-based Model for Humor Detection. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings, Málaga, Spain, 9. CEUR-WS.

Sjöbergh, J. and K. Araki. 2007. Recognizing humor without recognizing meaning. In F. Masulli, S. Mitra, and G. Pasi, editors, *WILF*, volume 4578 of *Lecture Notes in Computer Science*, pages 469–476. Springer.

Škrlj, B., M. Martinc, N. Lavrač, and S. Pollak. 2021. autobot: evolving neuro-symbolic representations for explainable low resource text classification. *Machine Learning*, 110(5):989–1028.

Tsakona, V. 2009. Language and image interaction in cartoons: Towards a multimodal theory of humor. *Journal of Pragmatics*, 41(6):1171–1188.

Wang, L., X. Lin, N. Lin, Y. Fu, K. Wu, and J. Wu. 2021. Humor Analysis in Spanish Tweets with Multiple Strategies. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings, Málaga, Spain, 9. CEUR-WS.