Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5106–5112
Marseille, 11–16 May 2020

HAHA 2019 Dataset: A Corpus for Humor Analysis in Spanish

Luis Chiruzzo[R], **Santiago Castro**[MR], **Aiala Rosá**[R]
[R]Universidad de la República, Uruguay
[M]University of Michigan, USA
luischir@fing.edu.uy, sacastro@umich.edu, aialar@fing.edu.uy

Abstract

This paper presents the development of a corpus of 30,000 Spanish tweets that were crowd-annotated with humor value and funniness score. The corpus contains approximately 38.6% intended-humor tweets with a 2.04/5 average funniness score. It has been used in an automatic humor recognition and analysis competition, obtaining encouraging results from the participants.
Keywords: Humor, Computational Humor, Humor Detection, Humor Analysis, Natural Language Processing

1. Introduction

Within the subfield of Computational Humor, there have been several works that have built resources to study different forms of Humor in texts (Mihalcea and Strapparava, 2005; Sjöbergh and Araki, 2007; Yang et al., 2015; Potash et al., 2017). For the case of the Spanish language, (Castro et al., 2016) created a first humor dataset, in which several issues were subsequently addressed by (Castro et al., 2018). Both of these works extracted tweets from both live random samples and selected accounts and had them annotated according to intended humor (binary), and a funniness score from one to five for the latter case. The dataset created in (Castro et al., 2018) was used by Castro et al. (2018) in the context of the HAHA 2018 competition for Humor Detection and Funniness Average Prediction.

However, the dataset presented in (Castro et al., 2018) still presents some issues. First, there are many duplicate tweets (around 6%), which only differ in their format or spacing, but the text is essentially the same. This is a potential problem with the quality of the data as they may have different annotations. Second, we believe there is some room for improvement for the inter-annotator agreement, which is 0.57 in Krippendorff's alpha value (Krippendorff, 2012). Finally, there is a class unbalance that should be tackled as it does not represent a sample from the reality (as tweets are picked from different sources) and complicates training and evaluation.

In this work, we build on top of (Castro et al., 2018), tackling the mentioned issues and creating a new dataset. We gather more tweets following a similar crowdsourcing annotation procedure but tackling some of the issues to increase the agreement score and to have a more balanced dataset, and we put it together along with the tweets from their dataset by removing the duplicate tweets and merging the annotations. Additionally, during the dataset annotation, we asked the annotators to tell if they considered the tweet text to be offensive or not. We find interesting to study how humor plays along with hate speech. This dataset was used (Chiruzzo et al., 2019) in the context of the HAHA 2019 competition, hosted at IberLEF, for Humor Detection and Funniness Score Prediction.

2. Related Work

Different authors have constructed datasets for humor recognition in English texts, most of them focusing on recognizing humorous short texts (called one-liners). (Mihalcea and Strapparava, 2005) created a corpus of 16,000 one-liner jokes. (Sjöbergh and Araki, 2007) built their corpus by downloading 6,100 one-liner jokes collected from the Internet. (Yang et al., 2015) also constructed a humor dataset, collecting 2,423 short texts from the site Pun of the Day (`http://www.punoftheday.com`). (van den Beukel and Aroyo, 2018) collected 12,000 humorous one-liners with a web-scraper from five selected jokes web-sites.

The microblogging platform Twitter has been found particularly useful for building humor corpora due to its public availability and the fact that its short messages are suitable for jokes or humorous comments. (Reyes et al., 2013) built a corpus for detecting irony in tweets by searching for several hashtags (i.e., #irony, #humor, #education and #politics). More recently, (Potash et al., 2017) built a tweet corpus that aims to distinguish the degree of funniness, assigning the values 0, 1 or 2 to each tweet. They used the tweet set issued in response to a TV game show, labeling which tweets were considered humorous by the show. The dataset includes 12,734 tweets and was used for the SemEval-2017 Task 6 (#HashtagWars: Learning a Sense of Humor).

For languages other than English, the available resources are scarce. (Khandelwal et al., 2018) created a corpus containing English-Hindi code-mixed tweets, with 1,755 humorous tweets and 1,698 non-humorous tweets. For Spanish, different versions of the corpus presented in this paper have been available (Castro et al., 2016; Castro et al., 2018), in this work we focus on an extension and improvement of this resource.

The main differences between our work and the ones previously discussed are the construction of a resource to study Humor for the Spanish language, the five-point funniness scale used for the annotation, and the crowdsourcing process through which the dataset was annotated so that the humorous nature of each tweet was decided by multiple and varied people.

3. Dataset construction

In this section, we describe the approach we take to defining intended humor and funniness, and the way we built the

dataset which has had two iterations so far and has been used for the 2018 and 2019 editions of the HAHA (Humor Analysis based on Human Annotations) evaluation campaign[1].

3.1. Approach to Humor and Funniness

We define two separate dimensions to conceptualize what we consider humorous. First, the type of humor we are trying to deal with in this work is *intentional humor*, i.e. the author of a text intended to be humorous or to amuse others with it. It clearly exists a relationship between humor and funniness, so in a first approach, following this criteria, we could be tempted to consider a piece of text as humorous if any number of people find it funny. However, funniness is a highly subjective property that varies significantly from person to person and also could vary across time for the same person. It would not be correct to say that the text is not humorous because the recipients of the message did not find it funny, it could as well happen that the message was a joke in bad taste which failed to amuse the recipients, but the intention of being humorous existed nonetheless. That is why we consider the two separate but related dimensions:

- *Humor* is an attribute of a piece of text that refers to the intention of the writer of being humorous.
- *Funniness* is an attribute that refers to the subjective experience of the reader if he or she finds the text amusing.

Of these two dimensions, the former could be considered as more objective and the latter as more subjective. As we show in Section 4.2., this seems to be the case, as the annotation process yielded greater agreement measures for the former than for the latter.

The two dimensions are translated into two different subtasks in the HAHA evaluation campaign. The first task refers to automatically determining if a tweet is humorous or not (a classification problem) and the second one refers to automatically assessing how funny a tweet is (a regression problem). As we will see in Section 4.4., the humor intention of a tweet seems to be much easier to predict than its expected funniness, which could in part happen due to the objectivity or subjectivity intrinsic to these dimensions.

3.2. Annotation interface

The graphical interface[2] presented to the annotators was designed to have these concepts in mind: we want to distinguish between tweets that were intended to be humorous or not humorous, and for the humorous ones we want to know how funny the annotator finds them. We also tried to make the interface as intuitive and engaging as possible, so we could use any number of annotators without prior training and keep them long enough in the platform so we could collect votes for several tweets.

The interface is shown in Figure 1. It displays an example tweet, and the only guiding text in the screen asks the user if the tweet *intends* to be humorous, which corresponds exactly to the first dimension we want to capture. The available options are *"yes"* or *"no"* (which are emphasized), or

Figure 1: Graphical interface used for the annotation process in 2019. The sample tweet says: *"- Boss, you underpaid me this month. - But I overpaid you last month. - Yes, one error is understandable, but two..."* The only difference between the 2018 and the 2019 versions of the tool is that the latter contains the *"Ofensivo"/"Offensive"* checkbox.

"skip" (which is de-emphasized). There is also an *"offensive"* checkbox which will be explained in section 3.4.. If the user chooses the option *"no"*, it will be recorded as a negative vote for that tweet (not humor) and no further questions will be asked. On the other hand, if the user chooses *"yes"*, she is immediately prompted with a list of options for scoring the tweet from 1 to 5. The options are displayed as emojis depicting different states of amusement from *"Nada gracioso"/"Not funny"* to *"¡Buenísimo!"/"Great!"*. The vote is not recorded until the user selects one of the scoring options. This way, we make sure that any vote for a humorous tweet has a corresponding score.

The tweets are presented randomly, but we keep track of an identifier for the session so as not to present the same tweet twice to the same user. During the annotations periods, the page was shared on popular social networks (Facebook and Twitter) to draw as much attention as possible and thus get votes from many different users from different backgrounds. As we will describe in section 3.3., we used some test tweets to try to measure the quality of the annotations in each session.

3.3. Corpus 2018

The first iteration was between February and March 2018 (Castro et al., 2018). In this first version of the corpus, the aim was to collect 20,000 tweets with labels for humorous or non-humorous and a corresponding score, trying to make it as balanced as possible between the humorous and non-humorous classes. We sampled 16,500 tweets from humorous Twitter accounts in Spanish that were found by manual inspection and 12,000 random tweets in Spanish. We tried to find humorous accounts from different Spanish speaking countries (including Spain, Mexico, Uruguay, Colombia, Argentina, and others) so as not to bias the corpus to a single Spanish variant. These tweets were crowd annotated by volunteers using a web tool during March 2018. The annotators had to decide, for each tweet, if it was humorous or not, and in case it was humorous, how funny the annotator considered it on a five-point scale.

All the users were presented with the same three test tweets for which we already knew if they were humorous or not

[1] https://www.fing.edu.uy/inco/grupos/pln/haha/

[2] `http://clasificahumor.com`

(two humorous and a non-humorous one). The purpose of these test tweets was to rule out users that did not understand the premise of the annotation process, we considered the sessions where any of these tweets were mislabeled as invalid sessions and did not use their votes in the final version of the corpus.

First, we aimed at getting at least five votes for each tweet and determine the humorous tweets by simple majority, i.e. the tweets that got at least three humor votes out of five should be considered humorous. We tried to shuffle the tweets presented to the annotators to keep the number of votes for each tweet as close as possible on average, not letting some of the tweets to lag (a notable exception to this are the three test tweets, which received as many votes as sessions). As the voting period proceeded, we realized that the tweets that were already getting three negative votes did not have any possibility of being considered humorous in the final corpus, even if the remaining two votes were of the humorous categories. Because of this, occasionally during the voting period, we manually deprioritized the tweets that got three or more negative votes, to keep in the pool only the tweets that still had a chance of being considered positive. As a result, the corpus contains some tweets that do not have five votes, mainly the non-humorous ones.

Once the voting period ended, we had received 117,800 votes from 1,546 users. We collected all the annotations, discarding the invalid sessions, determined the humorous value by simple majority, and the average score for the humorous tweets. In total, around 26.9% of the tweets were considered humorous. We then randomly discarded non-humorous tweets until getting 20,000 tweets in total, achieving a final proportion of 36.8% humorous tweets in the corpus. This 2018 version of the corpus contains 20,000 tweets where 7,357 are humorous and 12,643 are not, the average funniness score for the humorous tweets is 2.10. The corpus was divided into an 80/20 train-test split and it was used in the HAHA at IberEval 2018 competition (Castro et al., 2018).

3.4. Corpus 2019

The second iteration was done between December 2018 and March 2019. First, we started by analyzing some tweets in the 2018 version of the corpus that we noticed were near-duplicates, i.e. the content was almost the same with a few different words that did not change their semantics. We used a semi-automatic process to find duplicate candidates by collecting all pairs of tweets that had a Jaccard coefficient greater than 0.5. We manually inspected all pairs, clustered them into equivalence classes, and took one example from each class discarding the others from the corpus. As a result, we pruned 1,278 tweets from the corpus, most of them were humorous. Table 1 shows some examples of near-duplicate tweets found in the previous version of the corpus. The most common differences between tweets considered near-duplicates include slight changes in spelling or capitalization, differences in punctuation, repetition of characters and use of hashtags.

The aim for this second version of the corpus was to get 30,000 tweets in total, so we extracted 10,000 new tweets from humorous accounts (the same accounts as in the previous year plus thirteen new accounts), and 3,000 new random Spanish tweets. We used the same web tool for annotating the new tweets with a small modification.

Si tuviera un peso por cada persona que me dice "feo", pues sería pobre porque soy perfecto.	*Si tuviera un peso por cada vez que alguien me dice "feo", sería pobre porque soy perfecto.*
—¿Tienes Wi-Fi? *—Claro.* *—¿Cuál es la clave?* *—Tener dinero y pagarlo.*	*- ¿Tienes wi-fi?* *- Sí* *- ¿Y cuál es la clave?* *- Tener dinero y pagarlo.*
Me encanta encontrar dinero en mi ropa. Es como un regalo para mí de mí.	*Me encanta encontrar el dinero en mi ropa, es como un regalo para mí de mí.*
Cuando te digan ESTUDIA no hagas nada, significa ES-TU-DIA, aprovéchalo.	*Cuando te digan ESTUDIA no hagas nada, significa ES-TU-DIA, aprovechalo! #fb*
¿Cursi yo? ¿Cursi YOO? Cursi el viento..!! ..que acaricia tu cabello, impregnándome de tu aroma, y el dulce terciopelo...	*Cursi yo?? CURSI YOOOOOOO?????... cursi el viento que acaricia tu cabello impregnandome con tu aroma y el dulce terciopelo...*

Table 1: Examples of different types of near-duplicate tweets.

From our experience during the 2018 annotation, we found out that some annotators were still confused between considering a tweet as "non-humorous" or considering it "humorous but not funny". This was more evident for tweets that contained insults or offensive content, on occasions tweets that would be considered a bad taste joke (i.e. humorous but not funny) could be labeled as not humor if they contained insults. To alleviate this situation, we decided to slightly modify the graphical interface by adding a new option to mark a tweet as offensive. This option, as shown in Figure 1, is a checkbox, and its information is saved only after the user chose whether the tweet is humorous or not. The purpose of this new option is twofold: On one hand, it could help us collect information about tweets that are offensive or not offensive to analyze if there is any correlation between offensive content and humor. On the other hand, it would help to make clearer to an annotator that there are tweets that could be offensive or in bad taste but should be marked as humorous nonetheless. We hoped that making this option explicit would help disentangle these possibilities and show that offensiveness and humor are different dimensions.

Between February and March 2019, we received 74,312 votes from 780 users. This time we used two test tweets presented to all users, different from the ones used the previous year but with the same intent of trying to detect invalid sessions. After determining the humorous tweets and their respective scores, we discarded non-humorous tweets until we got the 30,000 tweets we wanted for this version of the corpus, which ended up being slightly more balanced than the 2018 version having 38.6% of humorous tweets. In the 2019 version of the corpus, there are 30,000 tweets where 11,595 are humorous and 18,405 are not, the average fun-

niness score for the humorous tweets is 2.04. The corpus was divided in an 80/20 train-test split with the following criteria: all tweets that had been part of the train and test partitions in the 2018 version of the corpus are part of the training partition in the 2019 corpus, the tweets that were annotated in 2019 would be split between train and test to keep the best possible balance given the number of humorous tweets. In this way, the 2019 test partition contains only tweets that the participants of the previous year had not seen. This corpus was used in the HAHA at IberLEF 2019 competition (Chiruzzo et al., 2019). Refer to the Appendix for more details.

4. Analysis

In this section, we present the composition of the final dataset and an analysis of some aspects of the corpus.

4.1. Dataset information

	Train	Test	Total
Tweets	24,000	6,000	30,000
Non-humorous	14,757	3,658	18,405
Humorous	9,253	2,342	11,595
Average Score	2.04	2.03	2.04
Votes No	59,440	13,605	73,045
Votes 1	19,063	4,818	23,881
Votes 2	14,713	3,777	18,490
Votes 3	10,206	2,649	12,855
Votes 4	4,493	1,122	5,615
Votes 5	1,305	275	1,580

Table 2: Composition of the final corpus for the total count and each class.

Table 2 shows the composition of the corpus, which is provided as two CSV files containing the training data and test data. Each row in the files includes the tweet unique identifier, the text of the tweet, the number of votes for each category (not humor, 1, 2, 3, 4 or 5 stars), and two values that can be calculated from the number of votes: a boolean value indicating if the tweet should be considered humorous or not and a real value indicating the average funniness score (if the tweet is not humorous, this value is `NULL`). Figure 2 shows the general distribution of votes for tweets in the corpus. The corpus contains around 38.6% of humorous tweets (i.e. tweets that receive less negative votes than positive votes), although in total the number of all positive votes is around 46.1%.

4.2. Agreement

	Humor		Funniness	
	2018	2019	2018	2019
All sessions	0.551	0.605	0.144	0.208
Valid sessions	0.571	0.639	0.163	0.224

Table 3: Annotator agreement measured as Krippendorff's alpha for the categorical humor value and the ranged funniness value.

Table 3 shows the agreement of the annotators calculated using Krippendorff's alpha for the 2018 and 2019 versions of the corpus. First of all, the agreement for the humorous/non-humorous classes is above 0.5 in all cases, which indicates a moderate to substantial agreement (Fleiss, 1971). Compare this to the agreement values obtained in (Castro et al., 2016), which reports an agreement of 0.365 for a similar task. The agreement for the funniness score value is considerably lower, which is expected due to the high subjectivity of this measure.

It is also interesting that the agreement increases appreciably in all cases when considering only the valid sessions. This could indicate that the process of presenting test tweets to all users helps ruling out some low-quality annotations. The agreement values for the 2019 annotations have also increased significantly respect to the 2018 corpus.

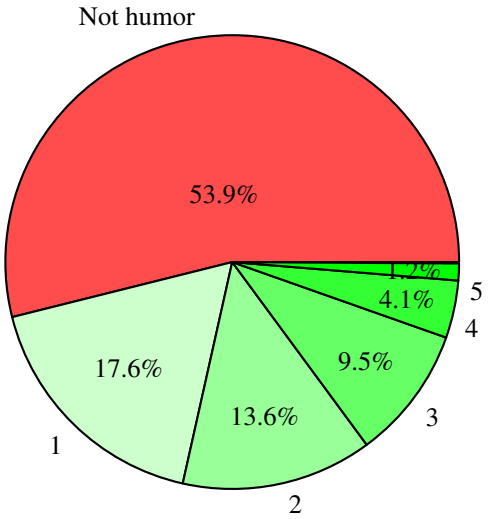

Figure 2: Distribution of votes in the final version of the corpus. The numbers 1 to 5 are the different scores the annotators could assign to the humorous tweets.

4.3. Offensiveness

In total, we received 1,438 votes that were marked as offensive. Although this number is not enough for creating a corpus of offensive tweets (indeed very few tweets were voted as offensive more than once) we found an interesting property of the votes that had the offensive mark. Figure 3 shows the distribution by category of all votes marked as offensive and all votes not marked as offensive. Notice that in the cases were a user marked the tweet as offensive, the most common voted category is "1" (humorous, but with the lowest score). On the other hand, if the vote does not have the offensive mark, the most common category is "x" (not humor). This could indicate that the users that understood the possibility of marking a tweet as offensive, also understood more clearly that it is possible to have a tweet that is both offensive and humorous, while other users opted for marking more tweets as not humorous. Another possibility is that offensive tweets (such as tweets containing insults) have a higher chance of being jokes in bad taste. Further analysis is needed to understand what the case is in our corpus.

Year	System	Task1 Precision	Recall	F1	Acc	Task2 RMSE
2018	INGEOTEC	77.9	81.6	**79.2**	84.5	**0.978**
	UO_UPV	81.6	75.7	78.5	**84.6**	1.592
	ELiRF_UPV	80.5	74.3	77.2	83.7	-
	random baseline	36.5	48.9	41.8	49.2	1.142
	dash baseline	93.9	9.3	16.9	65.9	-
2019	adilism	79.1	85.2	**82.1**	**85.5**	**0.736**
	Kevin & Hiromi	80.2	83.1	81.6	85.4	0.769
	bfarzin	78.2	83.9	81.0	84.6	0.746
	random baseline	39.4	49.7	44.0	50.5	1.651
	dash baseline	94.5	16.3	27.8	66.9	-

Table 4: Performance of the top three teams that took part in the competitions in 2018 and 2019. Task 1 refers to humor identification (classification task) while Task 2 refers to funniness score prediction (regression task).

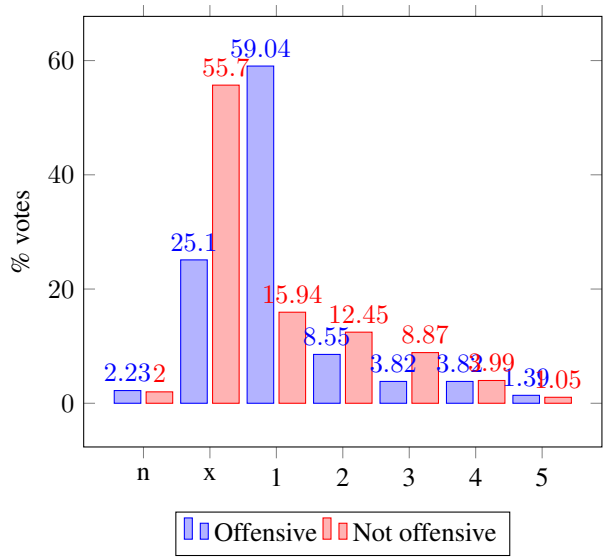

Figure 3: Percentage of votes for each category considering tweets marked as offensive or not marked as offensive. Votes marked as "n" mean the user skipped the tweet, votes marked as "x" mean a tweet that is not humorous, and votes marked as a number mean a tweet that is humorous with that score.

4.4. How baselines have performed on it

Table 4 shows the performance of the top three systems that participated in the competition in 2018 and 2019 together with two baselines. In 2018 the top system used an evolutionary algorithm for training the system (Ortiz-Bejar et al., 2018), while in 2019 the top system performed fine-tuning over a multilingual BERT language model (Ismailov, 2019). For task 1, the random baseline selects the positive class randomly with the probability of the training corpus, while for task 2 it selects always the average score in the training corpus.

The dash baseline, which is only defined for task 1, selects all tweets that start with a dash as humorous. The intuition behind this baseline is that very often the humorous tweets are written in a dialogue format, starting each line with a dash. This baseline has quite a high precision, more than 90% in both versions of the corpus. None of the systems could beat this baseline in terms of precision. On the other hand, the recall of this baseline is very low, because many humorous tweets are not written as dialogues, and that is why its F1 score is not that high.

5. Conclusion

We presented a corpus of Spanish tweets annotated with information about humor: if the tweets are humorous or not, and how funny the humorous tweets are. This information was crowd annotated by users that rated each tweet as non-humorous or humorous, the humorous ones were also annotated with a score in a range from 1 to 5. The corpus contains 30,000 tweets with about 38.6% instances of the humorous class, with an 80/20 train-test split.

This corpus is slightly more balanced than the 2018 version of the corpus (Castro et al., 2018) and is also less noisy because we manually analyzed and resolved all cases of near-duplicated tweets. The annotators also had the option of marking a tweet as offensive. Although the number of votes for offensive tweets is not enough to create a corpus of offensive tweets by itself, the marks in the corpus could help analyze the relationship between humor, funniness, and offensiveness.

As future work, it would be very interesting to generate a similar resource as this one but for other languages, particularly for English.

6. Bibliographical References

Bender, E. M. and Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

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

Chiruzzo, L., Castro, S., Etcheverry, M., Garat, D., Prada, J. J., and Rosá, A. (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.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Ismailov, A. (2019). 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, Bilbao, Spain, 9. CEUR-WS.

Krippendorff, K. (2012). *Content analysis: An introduction to its methodology*. Sage.

Ortiz-Bejar, J., Salgado, V., Graff, M., Moctezuma, D., Miranda-Jiménez, S., and Tellez, E. (2018). INGEOTEC at IberEval 2018 Task HaHa: mu-TC 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)*.

7. Language Resource References

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

Castro, S., Chiruzzo, L., Rosá, A., Garat, D., and Moncecchi, G. (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.

Khandelwal, A., Swami, S., Akhtar, S. S., and Shrivastava, M. (2018). Humor detection in English-Hindi code-mixed social media content : Corpus and baseline system. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Mihalcea, R. and Strapparava, C. (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.

Potash, P., Romanov, A., and Rumshisky, A. (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.

Reyes, A., Rosso, P., and Veale, T. (2013). A multidimensional approach for detecting irony in twitter. *Language resources and evaluation*, 47(1):239–268.

Sjöbergh, J. and Araki, K. (2007). Recognizing humor without recognizing meaning. In Francesco Masulli, et al., editors, *WILF*, volume 4578 of *Lecture Notes in Computer Science*, pages 469–476. Springer.

van den Beukel, S. and Aroyo, L. (2018). Homonym detection for humor recognition in short text. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 286–291, Brussels, Belgium, October. Association for Computational Linguistics.

Yang, D., Lavie, A., Dyer, C., and Hovy, E. (2015). Humor recognition and humor anchor extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2367–2376, Lisbon, Portugal, September. Association for Computational Linguistics.

Appendix

We present the data statement following (Bender and Friedman, 2018). HAHA 2019 is a dataset of 30,000 tweets with annotations for intended humor (binary) and funniness (five-point scale). It can be accessed via `https://www.fing.edu.uy/inco/grupos/pln/haha/`.

A. CURATION RATIONALE Jokes are hard to find online automatically using heuristics. At the same time, finding jokes within in-the-wild long texts can be problematic since you have to account for its boundaries concerning non-humorous content. Thus, we collect jokes from Twitter, supposing a tweet is either completely humorous or not at all. We rely on cherry-picked humorous accounts to source humorous tweets, and randomly sampled tweets for non-humorous content (which we hypothesize are harder to tell apart from jokes compared to specific types of tweets such as headlines or proverbs). We collected the data between December 2018 and February 2019.

Because the data from each source type is not clean, we carried out an online crowd-annotation between February and March 2019, in which any person could enter the web page and annotate tweets voluntarily. We shared this web page with our acquaintances and also on social networks (Facebook and Twitter). We used three tweets that we knew the intended humor answer for spam detection and we used an HTTP cookie with a long expiration time to avoid showing repeated tweets (note a user could eventually see the same tweet twice if entering from different devices). We always showed a random tweet among those that had the minimum annotation count. Finally, because we detected duplicate tweet texts (same content with the same or different format), we merged them along with their annotations.

B. LANGUAGE VARIETY Because we do not target a specific Spanish dialect, for the former we used a varied number of humorous accounts that declared to be from each of the countries in which Spanish is the main language, while for the latter we obtained randomly streamed tweets in Spanish (using Twitter's language detection). It includes dialects of Spanish (es) from the following origins: Argentina (es-AR), Bolivia (es-BO), Chile (es-CL), Colombia (es-CO), Costa Rica (es-CR), Dominican Republic (es-DO), Ecuador (es-EC), El Salvador (es-SV), Guatemala (es-GT), Honduras (es-HN), Mexico (es-MX), Nicaragua (es-NI), Panama (es-PA), Paraguay (es-PY), Peru (es-PE), Puerto Rico (es-PR), Spain (es-ES), and Uruguay (es-UY).

C. SPEAKER DEMOGRAPHIC The only information we know is that they likely speak Spanish.

D. ANNOTATOR DEMOGRAPHIC For privacy and practical reasons, we do not ask annotators for information. However, we have the following information for the annotation period from Google Analytics (February 1st to March 31st, 2019):

- 92% bounce rate. The following data only counts the non-bounced visits.

- 1,222 page views (note that in one page view there can be many annotations).
- 1,083 sessions.
- 793 users (8 users had at least 10 sessions).
- User age: 7.3% 18–24, 46.6% 25–34, 20.65% 35–44, 11.08% 45-54, 9.82% 55-64, and 4.53% 65+.
- 54.9% male and 45.1% female.
- >72% of the users' web browser language was in Spanish, >25% was in English.
- Top 10 countries: 635 users from Uruguay, 38 from Argentina, 31 from Mexico, 17 from the United States, 12 from Spain, 11 from Chile, 11 from the United Kingdom, 7 from China, 6 from Ecuador and 5 from Guatemala.
- Device: 72.51% mobile users, 26.86% desktop, and 0.63% tablet.
- 64% sessions from social networks (60% Facebook and 40% Twitter), 24.5% direct access, 9.2% from organic searches, and 2.3% from other types of referrals.

E. SPEECH SITUATION The tweets (written text) were extracted between December 2018 and February 2019 (however, note some tweets are older). They are publicly-accessible tweets.

F. TEXT CHARACTERISTICS The texts are tweets, which contain up to 280 characters and may include emoji.

G. RECORDING QUALITY N/A.

H. OTHER N/A.

I. PROVENANCE APPENDIX N/A.