A Crowd-Annotated Spanish Corpus for Humor Analysis

Santiago Castro, Luis Chiruzzo, Aiala Rosá, Diego Garat and Guillermo Moncecchi

July 20th, 2018

Grupo de Procesamiento de Lenguaje Natural,
Universidad de la República — Uruguay

Outline

Background

Background i

- **Humor Detection** is about telling if a text is humorous (e. g., a joke).

My grandpa came to America looking for freedom, but it didn't work out, in the next flight my grandma was coming.

IT'S REALLY HOT

Background ii

- Some previous work, such as Barbieri and Saggion (2014), Mihalcea and Strapparava (2005), and Sjöbergh and Araki (2007), created binary Humor Classifiers for short texts written in English.
  - They extracted one-liners from the Internet and from Twitter, such as:

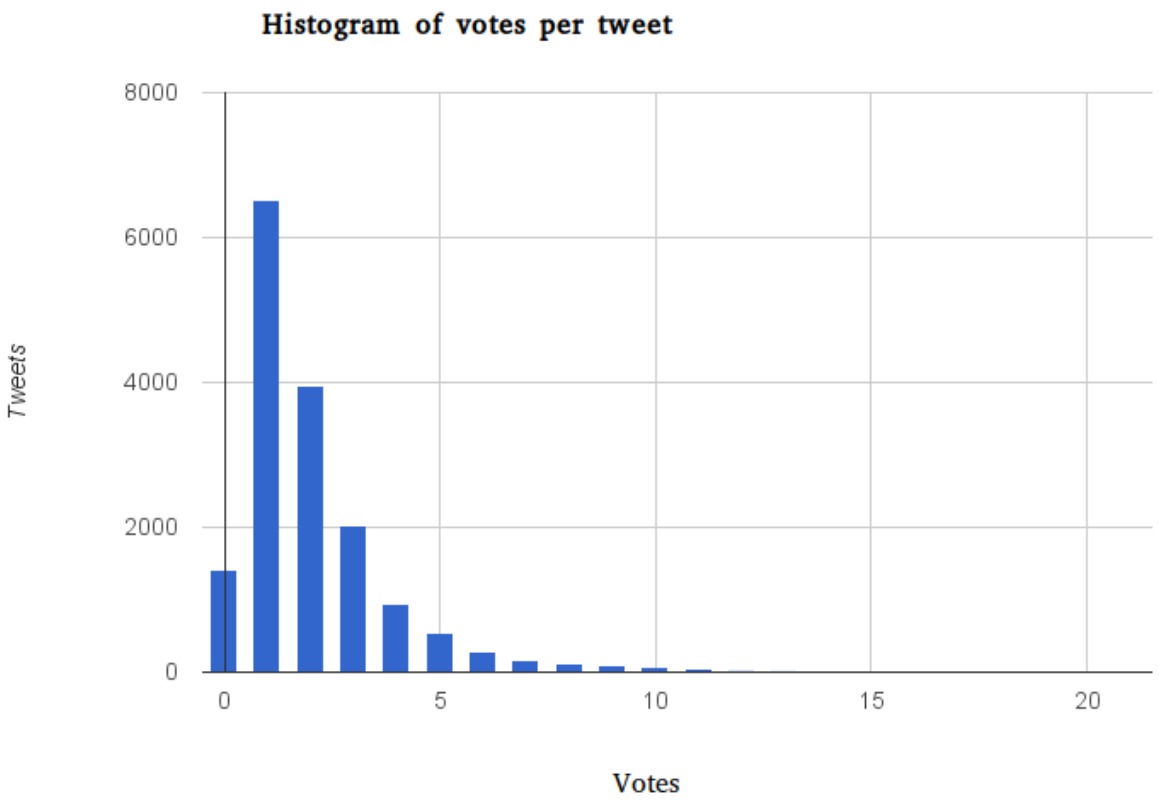
    Beauty is in the eye of the beer holder.

- Castro et al. (2016) worked on Spanish tweets since our group is interested in leveraging tools for Spanish.
  - Back then, we conceived the first and only Spanish dataset to study Humor.

Background iii

- Castro et al. (2016) corpus provided 40k tweets from 18 accounts, with 34k annotations. The annotators decided if the tweets were humorous or not, and if so they rated them from 1 to 5.
- However, the dataset has some issues:
  1. low inter-annotator agreement (Fleiss' $\kappa = 0.3654$)
  2. limited variety of sources (humorous: 9 Twitter accounts, non-humorous: 3 about news accounts, 3 about inspirational thoughts and 3 about curious facts)
  3. very few annotations per tweet (less than 2 in average, around 500 with ≥ 5 annotations)
  4. only 6k were considered humorous by the crowd

Background iv



Related work

Potash, Romanov, and Rumshisky (2017) built a corpus based on tweets in English that aims to distinguish the degree of funniness in a given tweet. They used the tweet set issued in response to a TV game show, labeling which tweets were considered humorous by the show. Used in SemEval 2017 Task 6 — #HashtagWars.

Extraction

Extraction i

1. We wanted to have at least 20k tweets as balanced *as possible*, at least 5 annotations each.
2. We fetched tweets from 50 humorous accounts from Spanish speaking countries, taking 12k at random.
3. We fetched tweet samples written in Spanish throughout February 2018, taking 12k at random.

Extraction ii

4. As expected, both sources contained a mix of humorous and non-humorous tweets.

Annotation

Annotation i

We built a web page, similar to the one used by Castro et al. (2016):

Annotation ii

clasificahumor.com

Annotation iii

- Tweets were randomly shown to annotators, but avoiding duplicates (by using web cookies).
- We wanted UI to be the more intuitive and self-explanatory as possible, trying not to induce any bias on users and letting them come up with their own definition of humor.
- The simple and friendly interface is meant to keep the users engaged and having fun while classifying tweets.

Annotation iv

- People annotated from March 8th to 27th, 2018.
- The first tweets shown to every session were the same: 3 tweets for which we know a clear answer.
- During the annotation process, we added around 4,500 tweets coming from humorous accounts to help the balance.

Dataset

Dataset i

- The dataset consists of two CSV files: **tweets** and **annotations**.

tweet ID	origin
24	humorous account

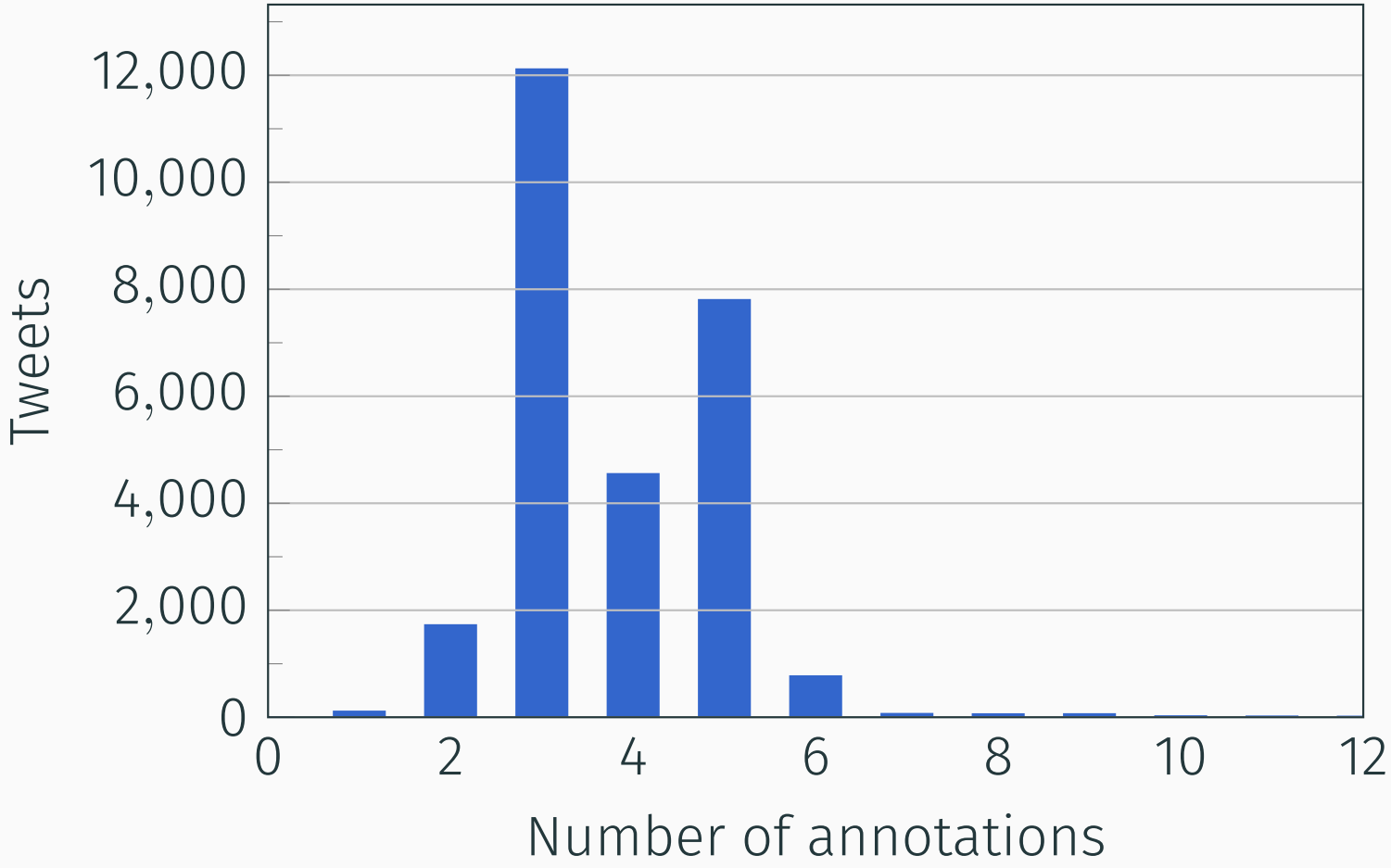
tweet ID	session ID	date	value
24	YOH113F...C4R	2018-03-15 19:30:34	2

Dataset ii

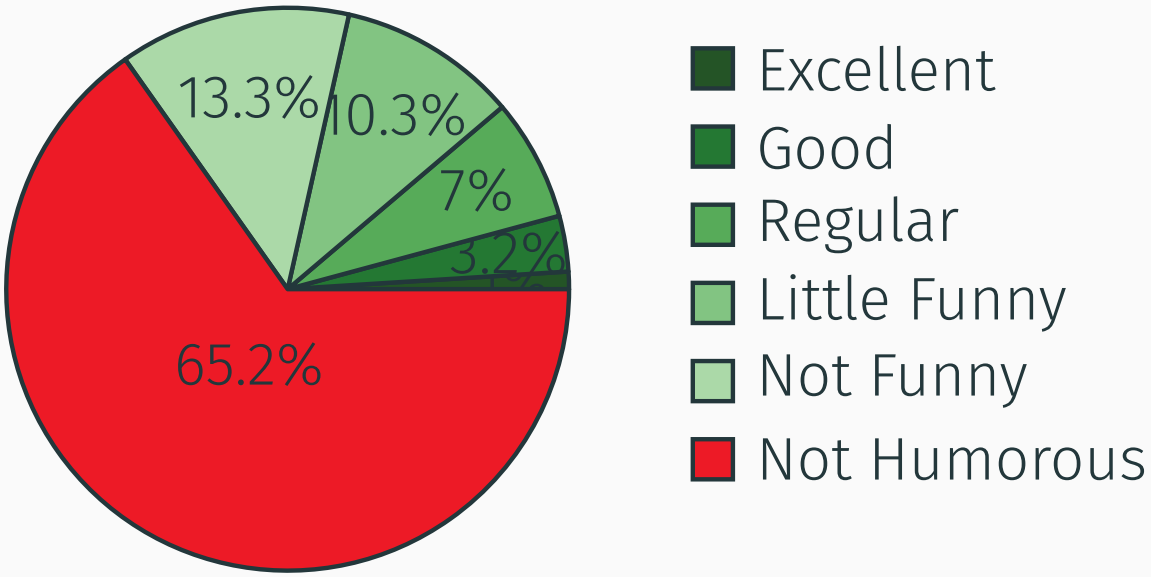
- 27,282 tweets
- 117,800 annotations (including 2,959 skips)
- **107,634** “high quality” annotations (excluding skips)

Analysis

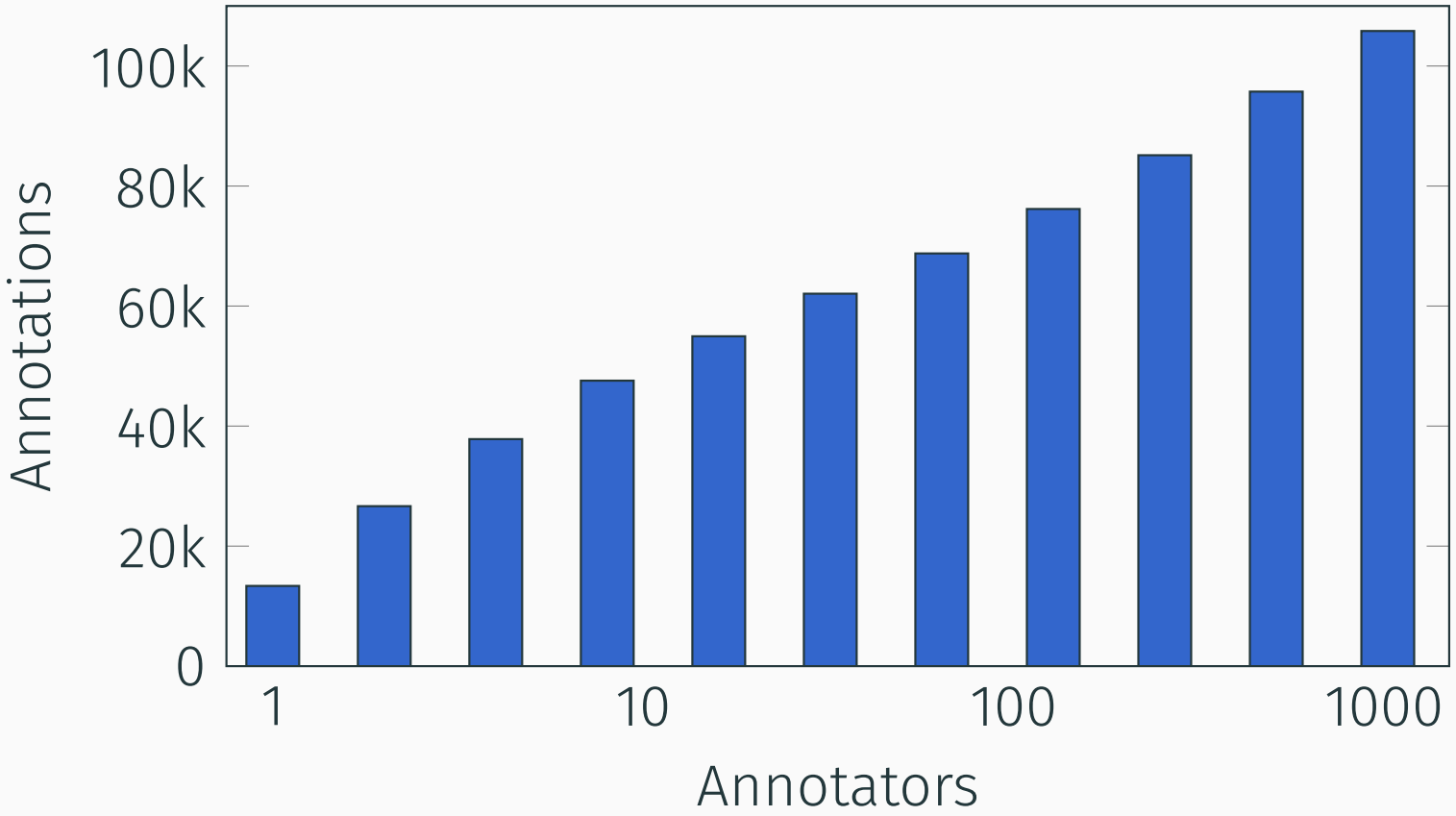
Annotation Distribution



Class Distribution



Annotators Distribution



Agreement

- Krippendorff's $\alpha = 0.5710$ (vs. 0.3654)
- If we include the "low quality", $\alpha = 0.5512$
- Funniness: $\alpha = 0.1625$
- If we only consider the 11 annotators who tagged more than a 1,000 times (who tagged 50,939 times in total), the humor and funniness agreement are respectively 0.6345 and 0.2635.

Conclusion

Conclusion

- We created a better version of a dataset to study Humor in Spanish. 27,282 tweets coming from multiple sources, with 107,634 annotations “high quality” annotations.
- Significant inter-annotator agreement value.
- It is also a first step to study subjectivity. Although more annotations per tweet would be appropriate, there is a subset of a thousand tweets with at least six annotations that could be used to study people’s opinion on the same instances.

HAHA Task

HAHA Task

- An IberEval 2018 task.
- Two subtasks: Humor Classification and Funniness Average Prediction.
- Subset of 20k tweets.
- 3 participants,
- 7 and 2 submissions respectively.

Analysis

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

References

- Barbieri, Francesco and Horacio Saggion (2014). “Automatic Detection of Irony and Humour in Twitter”. In: *ICCC*, pp. 155–162.
- Castro, Santiago et al. (2016). “Is This a Joke? Detecting Humor in Spanish Tweets”. In: *Ibero-American Conference on Artificial Intelligence*. Springer, pp. 139–150. DOI: 10.1007/978-3-319-47955-2_12.

References ii

- Fleiss, Joseph L (1971). "Measuring nominal scale agreement among many raters". In: *Psychological bulletin* 76.5, p. 378. DOI: 10.1037/h0031619.
- Krippendorff, Klaus (2012). *Content analysis: An introduction to its methodology*. Sage. DOI: 10.1111/j.1468-4446.2007.00153_10.x.
- Mihalcea, Rada and Carlo 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. Vancouver, British Columbia, Canada: Association for Computational Linguistics, pp. 531–538. DOI: 10.3115/1220575.1220642.

References iii

- Potash, Peter, Alexey Romanov, and Anna Rumshisky (2017). "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. DOI: 10.18653/v1/s17-2004.
- Sjöbergh, Jonas and Kenji Araki (2007). "Recognizing Humor Without Recognizing Meaning". In: *WILF*. Ed. by Francesco Masulli, Sushmita Mitra, and Gabriella Pasi. Vol. 4578. Lecture Notes in Computer Science. Springer, pp. 469–476. ISBN: 978-3-540-73399-7. DOI: 10.1007/978-3-540-73400-0_59.

Questions?

https://pln-fing-udelar.github.io/humor/