A Crowd-Annotated Spanish Corpus for Humor Analysis

Santiago Castro, Luis Chiruzzo, Aiala Rosá, Diego Garat, Guillermo Moncecchi
Grupo de Procesamiento de Lenguaje Natural
Facultad de Ingeniería
Universidad de la República — Uruguay
{sacastro,luichir,aialar,dgarat,gmonce}@fing.edu.uy

Abstract

Computational Humor involves several tasks, such as humor recognition, humor generation, and humor scoring, for which it is useful to have human-curated data. In this work we present a corpus of 27,000 tweets written in Spanish and crowd-annotated by their humor value and funniness score, with about four annotations per tweet, tagged by 1,300 people over the Internet. It is equally divided between tweets coming from humorous and non-humorous accounts. The inter-annotator agreement Krippendorff's alpha value is 0.5710. The dataset is available for general usage and can serve as a basis for humor detection and as a first step to tackle subjectivity.

1 Introduction

Computational Humor studies humor from a computational perspective, involving several tasks such as humor recognition, which aims to tell if a piece of text is humorous or not; humor generation, with the objective of generating new texts with funny content; and humor scoring, whose goal is to predict how funny a piece of text is.

In order to carry out this kind of tasks through supervised machine learning methods, human-curated data is necessary. Castro et al. (2016) built a humor classifier for Spanish and provided a dataset for humor recognition. However, there are some issues: few annotations per instance, low annotator agreement, and limited variety of sources for the humorous and mostly for the non-humorous tweets (the latter were only about news, inspirational thoughts and curious facts). Up to our knowledge, there is no other dataset to work on humor comprehension in Spanish. Some other authors, such as Mihalcea and Strapparava (2005a,b); Sjöbergh and Araki (2007) have tackled humor recognition in English texts, building their own corpora by downloading *one-liners* (one-sentence jokes) from the Internet, since working with longer texts would involve additional work, such as determining humor scope.

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. Castro et al. (2016) built their corpus based on Twitter, selecting nine humorous accounts and nine non-humorous accounts about news, thoughts and curious facts. Reyes et al. (2013) built a corpus for detecting irony in tweets by searching for several hashtags (i.e., #irony, #humor, #education and #politics), which is also used in Barbieri and Saggion (2014) to train a classifier that detects humor. More recently, Potash et al. (2017) built a corpus based on tweets 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.

In this work we present a crowd-annotated Spanish corpus of tweets tagged with a humor/no humor value and also by a funniness score from one to five. The corpus contains tweets extracted from varied sources and has several annotations per tweet, reaching a high humor inter-annotator agreement.

The contribution of this work is twofold: the dataset is not only useful for building a humor classifier but it also serves to approach subjectivity in humor and funniness. Even though there are not enough annotations per tweet as required to study subjectivity in a genuine way with techniques such as the ones by Geng (2016), the dataset aids as a playground to study the funniness and disagree-

Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media , pages 7–11,
Melbourne, Australia, July 20, 2018.

ment among several people.

This document is organized as follows. Section 2 explains where and how we obtained the data, and Section 3 describes how it was annotated. In Section 4 we present the corpus, and we address the analysis in Section 5. Finally, in Section 6 we present draw the conclusions and present the future work.

2 Extraction

The aim of the extraction and annotation process was to build a corpus of at least 20,000 tweets that was as balanced as possible between the humor and not humor classes. Furthermore, as we intended to have a way of calculating the funniness score of a tweet, we needed to have several votes for the tweets that were considered humorous.

As we wanted to have both humorous and non-humorous tweet samples, we extracted tweets from selected accounts and from realtime samples. For the former, based on Castro et al. (2016), we selected tweets from fifty humorous accounts from Spanish speaking countries, and took a random sample of size 12,000. For the latter, we fetched tweet samples written in Spanish throughout February 2018[1], and from this collection we took another random sample of size 12,000. Note that we preferred to take realtime tweet samples as we did not want to bias by selecting certain negative examples, such as news or inspirational thoughts as in Castro et al. (2016) and Mihalcea and Strapparava (2005b). From both sources we ignored retweets, responses, citations and tweets containing links, as we wanted the text to be self-contained. As expected, both sources contained a mix of humorous and non-humorous tweets. In the case of humorous accounts, this may be due to the fact that many tweets are used to increase the number of followers, expressing an opinion on a current event or supporting some popular cause.

We first aimed to have five votes for each tweet, and to decide which tweet was humorous by simple majority. However, at a certain stage during the annotation process, we noticed that the users were voting too many tweets as non-humorous, and the result was highly unbalanced. Because of this, we made some adjustments in the corpus and the process: as the target was to have five votes for each tweet, we considered that the tweets that already had three non-humorous annotations at this stage should be considered as not humor, then we deprioritized them so the users could focus in annotating the rest of the tweets that were still ambiguous. We also injected 4,500 more tweets randomly extracted only from the humorous accounts. These new tweets were also prioritized since they had less annotations than the rest.

Figure 1: Example of a tweet presented to the annotators. It says: *I hate being bipolar, it's so cool!*. The annotator is asked whether the tweet intends to be humorous. The available options are "Yes", "No" or "Skip". If the annotator selects "Yes", five emoji are shown so the annotator can specify how funny he considers the tweet. The emoji also include labels describing the funniness levels.

3 Annotation

A crowdsourced web annotation task was carried out to tag all tweets.[2] The annotators were shown tweets as in Fig. 1. The tweets were randomly chosen but web session information was kept to avoid showing duplicates. We tried to keep the user interface as 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 as humorous or not, and how funny they are, with as few instructions as possible.

If a person decides that a tweet is humorous, he has to rate it between one to five by using emoji. In this way, the annotator gives more information rather than just stating the tweet is humorous. We also allowed to skip a tweet or click a help button for more information. We consider that explicitly asking the annotator if the text intends to be humorous makes the distinction between the Not Humorous and Not Funny classes less ambiguous,

[1] The language detection feature is provided by the Twitter REST API.

[2] https://clasificahumor.com

which we believe was a problem of (Castro et al., 2016) user interface. Also, we consider our emoji rated funniness score to be clearer for annotators than their stars based rating.

The web page was shared on popular social networks along with some context about the task and the annotation period occurred between March 8th and 27th, 2018. The first tweets shown to every session were the same: three tweets for which we know a clear answer (one of them was humorous and the other two were not). These first tweets ("test tweets") were meant as a way of introducing the user into how the interface works, and also as an initial way for evaluating the quality of the annotations. After the introductory tweets, the rest of the tweets were sampled randomly, starting with the ones with the least number of votes.

4 Corpus

The dataset consists of two CSV files: tweets and annotations. The former contains the identifier and origin (which can be the realtime samples or the selected accounts) for each one of the $27,282$ tweets[3], while the latter contains the tweet identifier, session identifier, date and annotation value for each one of the $117,800$ annotations received during the annotation phase (including the times the skip button was pressed, $2,959$ times). The dataset was released and it is available online.[4]

When compiling the final version of the corpus, we considered the annotations of users that did not answer the first three tweets correctly as having lower quality. These sessions should not be used for training or testing machine learning algorithms. Fortunately, only a small number of annotations had to be discarded because of this reason. The final number of annotations is $107,634$ (not including the times the skip button was pressed), including $3,916$ annotations assigned to the test tweets themselves.

5 Analysis

5.1 Annotation Distribution

Each tweet received 3.8 annotations on average, with a standard deviation of 1.16, not considering the test tweets as they are outliers (they have a large number of annotations). The annotation distribution is shown in Fig. 2. The histogram is highly concentrated: more than 98% of the tweets received between two and six annotations each. Even though the strategy was to show random tweets among the ones with less annotations, note that there are tweets with less than three annotations because some annotations were finally filtered out. At the same time, there are some tweets with more than six annotations because we merged annotations from a few dozen duplicate tweets. Also, note that there is a considerable amount of tweets with at least six annotations $(1,001)$. This subset can be useful to study the different annotator opinions under the same instances.

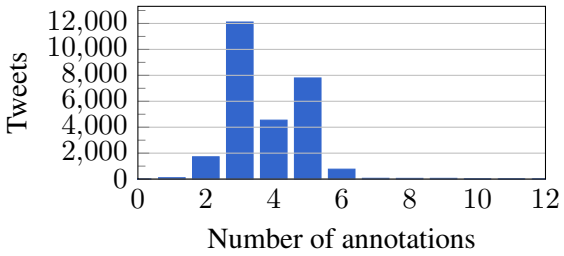

Figure 2: Distribution of tweets by number of annotations. Most tweets have between two and six annotations each.

5.2 Class Distribution

Fig. 3 shows how the classes are distributed between the annotations. Roughly two thirds were assigned to the class Not Humorous, agreeing with the fact that there seem to be more non-humorous tweets from humorous accounts than the other way around. The graph also indicates that there is a bias towards bad jokes in humor, according to the annotators. We use simple majority of votes for categorizing between humorous or not humorous, and weighted average for computing the funniness score only for humorous tweets. The scale goes from one (Not Funny) to five (Excellent). Under this scheme, 27.01% of the tweets are humorous, 70.6% are not-humorous while 2.39% is undecided (2.38% tied and 0.01% no annotations). At the same time, humorous tweets have little funniness overall: the funniness score average is 1.35 and standard deviation 0.85.

5.3 Annotators Distribution

There were $1,271$ annotators who tagged the tweets roughly as follows: two annotators tagged $13,000$ tweets, then one annotated $8,000$, the next eight annotated between one and three thousand,

[3] Tweet text is not included in the corpus due to Twitter Terms and Conditions. They can be obtained from the IDs.

[4] https://pln-fing-udelar.github.io/humor

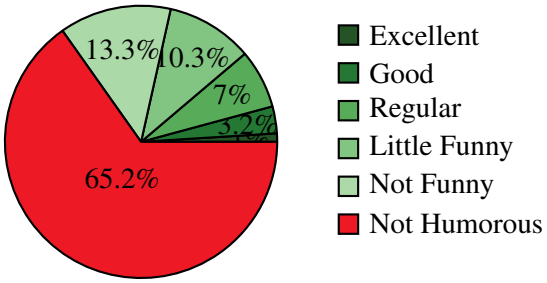

Figure 3: Annotations according to their class.

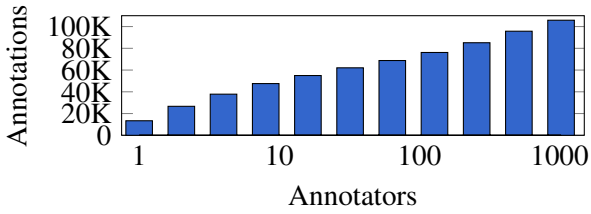

Figure 4: Accumulated distribution of annotations by number of annotators. Notice that the top 100 annotators add up to more than 70,000 annotations.

the next 105 annotated between one hundred and one thousand and the rest annotated less than a hundred, having 32,584 annotations in total (see Fig. 4). The average was 83 tags by annotator, with a standard deviation of 597.

5.4 Annotators Agreement

An important aspect to analyze is to what extent the annotators agree on which tweets are humorous. We used the *alpha* measure from Krippendorff (2012), a generalized version of the *kappa* measure (Cohen, 1960; Fleiss, 1971) that takes in account an arbitray number of raters. The agreement alpha value on humorous versus non-humorous is 0.5710. According to Fleiss (1981), it means that the agreement is somewhat between "moderate" to "substantial", suggesting there is acceptable agreement but the humans cannot completely agree. We believe that the carefully designed user interface impacted in the quality of the annotation, as unlike Castro et al. (2016) this work's annotation web page presented less ambiguity between the class Not Humorous and Not Funny. We clearly outperformed their inter-annotator agreement (which was 0.3654). Additionally, if we consider the whole corpus (including the removed annotations), this figure decreases to 0.5512. This shows that the test tweets were helpful to filter out low quality annotations.

Additionally, we can try to estimate to what extent the annotators agree on the funniness value of the tweets. In this case, disagreement between close values in the scale (e.g. Not Funny and Little Funny) should have less impact than disagreement between values that are further (e.g. Not Funny and Excellent). Following Stevens (1946), in the previous case we were dealing with a *nominal* measure while in this case it is an *ordinal* measure. Alpha considers this into the formula by using a generic distance function between ratings, so we applied it and obtained a value of 0.1625 which is far from good; it is closer to a random annotation. There is a lack of agreement on the funniness. In this case, a machine will not be able to assign a unique value of funniness to a tweet, which makes sense with its subjectivity, albeit other techniques could be used (Geng, 2016). In this case, if we consider the whole dataset, this number decreases to 0.1442.

If we only consider the eleven annotators who tagged more than a thousand times (who tagged 50,939 times in total), the humor and funniness agreement are respectively 0.6345 and 0.2635.

6 Conclusion and Future Work

Our main contribution is a corpus of tweets in Spanish labeled by their humor value and funniness score with respect to a crowd-sourced annotation. The dataset contains 27,282 tweets coming from multiple sources, with 107,634 annotations. The corpus showed high quality because of the significant inter-annotator agreement value.

The dataset serves to build a Spanish humor classifier, but it also serves as a first step to tackle humor and funniness subjectivity. Even though 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.

Future steps involve gathering more annotations per tweet for a considerable amount of tweets, so techniques such as the ones in (Geng, 2016) could be used to study how people perceive the humorous pieces and what subjects and phrases they consider funnier. It would be interesting to consider social strata (e.g. origin, age and gender) when trying to find these patterns. Additionally, a similar dataset could be built for other languages which count with more data to cross over with (such as English) and build a humor classifier exploiting re-

cent Deep Learning techniques based on it.

Acknowledgments

We thank everyone who annotated tweets via the web page. We would not have been able to reach the large number of annotations we have got without their help.

References

Francesco Barbieri and Horacio Saggion. 2014. Automatic detection of irony and humour in twitter. In *ICCC*, pages 155–162.

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

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

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

Joseph L Fleiss. 1981. *Statistical methods for rates and proportions*, 2 edition. John Wiley.

Xin Geng. 2016. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748.

Klaus Krippendorff. 2012. *Content analysis: An introduction to its methodology*. Sage.

Rada Mihalcea and Carlo Strapparava. 2005a. Bootstrapping for fun: Web-based construction of large data sets for humor recognition. In *Proceedings of the Workshop on Negotiation, Behaviour and Language (FINEXIN 2005)*, volume 3814, pages 84–93.

Rada Mihalcea and Carlo Strapparava. 2005b. 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.

Peter Potash, 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)*, pages 49–57.

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

Jonas Sjöbergh and Kenji Araki. 2007. Recognizing humor without recognizing meaning. In *WILF*, volume 4578 of *Lecture Notes in Computer Science*, pages 469–476. Springer.

Stanley Smith Stevens. 1946. On the theory of scales of measurement. *Science*, 103:677–680.