How effective are Quantum Entropy Metrics for Detecting Humor?

Anonymous ACL submission

Abstract

Having ways to measure humor in a text can help provide systems with real-time feedback to enhance their outputs. Thus, we evaluated two Quantum Entropy-based scores as humor quality metrics by comparing their behavior across various corpora in different languages. Results showed significant differences between humorous and non-humorous instances in certain corpora, but an analysis of effect sizes implied minimal practical significance. The metrics' behavior also varies depending on the dataset and language being evaluated. Experiments using a shared multilingual vector space led to more consistent scoring, but also reduced the metric's ability to differentiate humor and non-humor. Results suggest that despite the potential, the effectiveness of these metrics was inconsistent, indicating a need for further research on humor quality measurements.

1 Introduction

Humor is a fundamental aspect of human daily interaction, and it plays a crucial role in various social contexts, including entertainment, coping, and bonding (Chiaro, 2018, p.9–10). However, the subjective and multifaceted nature of humor poses significant challenges for its automated evaluation.

Computational approaches to humor evaluation are valuable for various applications, including humor generation and conversational agents, as they provide real-time feedback on system performance, which could be used as fitness functions (Winters and Delobelle, 2021) or refining criteria (Madaan et al., 2023). In this paper, we delve into the domain of automatic humor quality assessment, exploring the viability of two metrics proposed by Liu and Hou (2023), QE-Uncertainty (QE-U) and QE-Incongruity (QE-I), based on the incongruity theory of humor (Rutter, 1997, p. 16–21), a much-studied approach to interpreting and identifying humor in the context of computational systems. We focus our analysis on such metrics as they are the most recent features that require little to no manual annotation.

To evaluate such appropriateness, we investigate how these measurements behave across six different corpora spanning three languages, extending the original work by Liu and Hou (2023) who evaluated in a single corpus from SemEval 2017 (Miller et al., 2017; Xie et al., 2021). By discerning how these scores differentiate between humorous and non-humorous texts, we also want to test whether their values are consistent across corpora.

Initial results showed discrepancies between classes in some corpora, but effect sizes revealed that these metrics are not good enough to separate humorous and non-humorous texts. We also observed that the values vary according to the corpus and language used, which motivated experiments with multilingual vectors. These showed a more consistent scoring, but with less discerning ability. Despite promising, the effectiveness of the target metrics was inconsistent, opening new paths for future research on humor quality assessment and incongruity-based scores.

The paper is organized as follows. In section 2, we present an overview of incongruity-based humor metrics. A more detailed explanation of Quantum Entropy metrics is in section 3. Our experiments are described in section 4 and their results are reported in section 5. We end with some conclusions and further research paths in section 6. Finally, limitations and ethical considerations are in sections 7 and 8, respectively.

2 Related Work

Since the 2000s, research on Humor Recognition has taken advantage of hand-crafted feature sets (Mihalcea and Strapparava, 2005; Gonçalo Oliveira et al., 2020), including various stylistic characteristics (e.g. alliteration), linguistic resources (e.g. WordNet, sentiment dictionaries), and content features (e.g. frequency counts). More recently, some authors proposed more complex metrics that intend to capture intricate textual relations using language modeling (Kao et al., 2016; He et al., 2019; Xie et al., 2021) and word embeddings (Liu and Hou, 2023). Most of the modern measures, described below, are based on the Incongruity Theory, which states that the humorous effect of a text arises from creating an expectation on the hearer and subsequently subverting it with a logically incongruous conclusion. (Rutter, 1997). For instance, in "Of all the things I lost, I miss my mind the most" (Mihalcea and Pulman, 2007), the author creates an image of a person losing physical objects to break this expectation with the abstract concept of "losing one's mind".

Kao et al. (2016) model the humor effect of punning jokes as the probabilistic difference of seeing the pun word (an ambiguous incongruous term) versus the expected alternative word. Similarly, He et al. (2019) propose a probabilistic metric that considers both the likelihood of seeing the pun word within its local context and the probability of the alternative in the entire text.

Later, Xie et al. (2021) used GPT-2 as a Language Model to compute metrics of Uncertainty and Surprise, by leveraging the probability of generating a joke's punchline given its setup and the likelihood of a model producing such text. Subsequently, (Liu and Hou, 2023) proposed the QE-U and QE-I scores, rooted in Quantum Mechanics principles, especially the von Neumann Entropy (Von Neumann, 1996). Below, these metrics are explained in detail.

3 Quantum Entropy metrics

Liu and Hou (2023) introduce quantum entropy metrics for humor assessment using the concept of the density matrix ρ_S to represent the semantic superposition state of a sentence $S = \{w_1, w_2, \cdots, w_n\}$ based on its word embeddings $\{\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_n\}$. The density matrix is given by the average outer product of all normalized word embeddings, as shown in Equation 1.

$$\rho_S = \frac{1}{n}\sum_{i=0}^{n}\left(\frac{\mathbf{w}_i}{\|\mathbf{w}_i\|} \otimes \frac{\mathbf{w}_i}{\|\mathbf{w}_i\|}\right) \tag{1}$$

From this representation, two metrics are derived using von Neumann entropy calculation to quantify the randomness or uncertainty in the textual representation ρ_S. Additionally, Liu and Hou (2023) assume a setup-punchline structure for input texts, where every joke $\mathcal{J} = \{\mathcal{S}|\mathcal{P}\}$ consists of a setup ($\mathcal{S}$) and a punchline ($\mathcal{P}$).

QE-Uncertainty The Quantum Entropy Uncertainty $U(\mathcal{J})$ of a text $\mathcal{J}$ is defined as the entropy value of the setup ($\mathcal{S}$), expressed in Equation 2.

$$U(\mathcal{J}) = -\operatorname{Tr}(\rho_{\mathcal{S}} \log \rho_{\mathcal{S}}) \tag{2}$$

This metric measures how uncertain (or ambiguous) the meaning of the setup is. As argued by the authors, humorous texts should have more open-ended setup phrases, to leave space for the different incongruous interpretations to emerge.

QE-Incongruity The Quantum Entropy Incongruity $I(\mathcal{J})$ measures how much information about the punchline is already present in the setup, as defined in Equation 3.

$$I(\mathcal{J}) = \overbrace{-\operatorname{Tr}(\rho_{\mathcal{P}}\rho_{\mathcal{S}} \log \rho_{\mathcal{P}}\rho_{\mathcal{S}})}^{\text{Entropy of punchline given setup}} + \underbrace{\operatorname{Tr}(\rho_{\mathcal{S}} \log \rho_{\mathcal{S}})}_{\text{Entropy of setup}} \tag{3}$$

4 Experimental Setup

The quantum entropy metrics have been implemented using the Pytorch library in Python[1,2].

4.1 Corpora

To evaluate the applicability of the metrics across different languages and text types, we selected various corpora used in humor recognition research: (i) SemEval 2017 Task 7 (English) (Miller et al., 2017); (ii) Humicroedit (English) (Hossain et al., 2019); (iii) JOKER CLEF 2023 (English, French, Spanish); (iv) HAHA@IberLEF 2019 (Spanish) (Chiruzzo et al., 2019); (v) HAHA@IberLEF 2021 (Spanish) (Chiruzzo et al., 2021); (vi) HUHU@IberLEF 2023 (Spanish) (Rosso, 2023); (vii) Clemêncio (Portuguese) (Gonçalo Oliveira et al., 2020); (viii) Puntuguese (Portuguese)[3].

As a preprocessing step, we followed the heuristics of Xie et al. (2021) for extracting the setup and punchline from each joke. Texts without two

[1]The code, all results, and visualization scripts will be made publicly available.

[2]We adopted the convention that $0 \log 0 = 0$, as it does not affect matrix trace calculations.

[3]Citation omitted due to reviewing purposes.

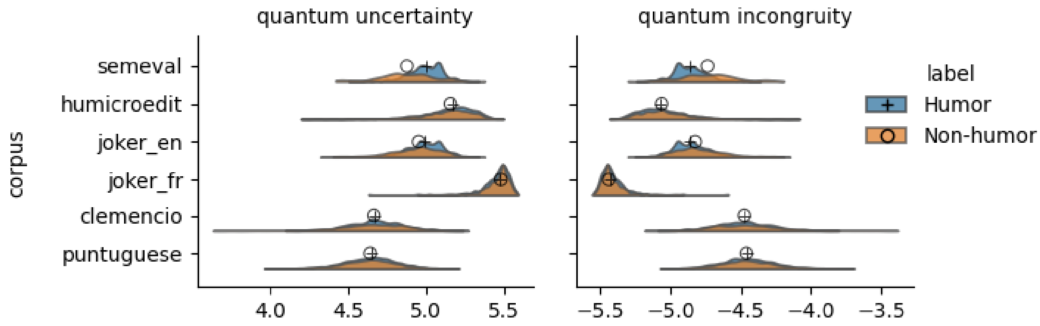

Figure 1: Distributions and medians of Quantum Entropy scores using monolingual (GloVe) word embeddings.

sentences (i.e. the setup followed by the punchline) are excluded. To avoid leaving out too much of the corpora, we included commas and semicolons as possible sentence boundaries. For completeness, resulting dataset sizes are in Appendix A.

4.2 Word embeddings

We used GloVe embeddings (Pennington et al., 2014; Hartmann et al., 2017) with 300 dimensions because not all languages had 50-dimension vectors available (as in Liu and Hou (2023)). Due to the unavailability of monolingual GloVe models for Spanish[4], we only included this language in further experiments using XLM-RoBERTa (XLMR) multilingual embeddings (Conneau et al., 2020). More details on the resources used can be found in Appendix B.

4.3 Statistical Hypothesis Testing

We used non-parametric statistical hypothesis tests for comparing the distributions of humorous and non-humorous texts. As two corpora (Humicroedit and Puntuguese) were created through an editing process, their samples are paired. Therefore, we used the Wilcoxon test (Corder and Foreman, 2011, p. 39) for both. For all remaining corpora, we used the Mann-Whitney U test (Corder and Foreman, 2011, p. 70).

Finally, as the metrics values are expected to be larger for positive instances, we use the alternative hypothesis that the value distribution for humorous texts is greater than that of the non-humorous.

5 Results

The scores distributions, computed with monolingual GloVe embeddings, are in Figure 1. When observing the values for QE-U computed with monolingual embeddings, the majority of corpora exhibit higher values for humorous instances when compared to non-humorous ones. However, this distinction is less pronounced for JOKER-FR and Clemêncio.

Statistical tests also reveal significant differences between humorous and non-humorous instances across all corpora but JOKER-FR and Clemêncio, where p-values are 0.8991 and 0.1694, respectively. Specifically, SemEval and JOKER-EN show substantial effect sizes with Cohen's d measures of 0.92 and 0.41, respectively, while the effect sizes for other datasets are close to zero, suggesting minimal practical significance.

Analyzing the QE-I scores in Figure 1, for some corpora (SemEval, Humicroedit, JOKER-EN), the observed behavior contradicts the expectation that values for humorous instances are higher. Although JOKER-FR exhibits a slightly higher median, the differences in distributions do not reach statistical significance (p-value 0.0751). Examining effect sizes, SemEval and JOKER-EN demonstrate larger practical differences, with Cohen's d values of -0.88 and -0.4, respectively, contrary to the alternative hypothesis that humorous values are higher.

When comparing to the results of Liu and Hou (2023), we observe that their metrics work fairly well with the corpus they evaluated in (SemEval); however, the scenarios are different for other corpora and languages.

5.1 Multilingual experiments

Figure 1 shows that, in monolingual experiments, distribution ranges and shapes vary considerably across corpora, posing challenges for using these metrics as evaluation criteria. For example, JOKER-FR distributions are more concentrated toward extreme values, whereas those in Clemên-

[4]We acknowledge the existence of https://github.com/dccuchile/spanish-word-embeddings; however, the link provided is broken and authors did not answer to our contacts.

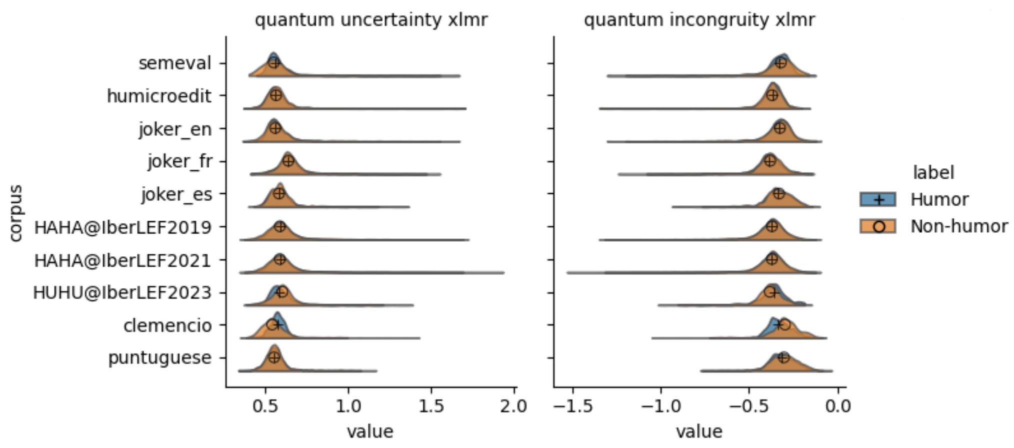

Figure 2: Distributions and medians of Quantum Entropy scores using multilingual (XLMR) word embeddings.

cio and Puntuguese are generally more spread out. This variation could be from substantial differences across the corpora or disparities in monolingual word embedding spaces, leading to notable discrepancies in the final score values. To address this, we conducted the same experiments using a shared multilingual vector space obtained with XLMR (Conneau et al., 2020).

The multilingual experiments for QE-U, depicted in Figure 2, show that using a multilingual embedding space leads to more similar distributions in terms of range and shape. However, while statistically significant differences were observed in some corpora — SemEval (p-value 0.0054), HAHA@IberLEF2019 (0.0004), HAHA@IberLEF2021 (0.0027), Clemêncio (1.7×10^{-16}) — their effect sizes were consistently weak (0.03, 0.04, 0.03, 0.13, respectively). This suggests that, while the use of a shared multilingual embedding model results in more consistent scoring across corpora, it also diminishes the metric's ability to discern subtleties between humor and non-humor instances.

The QE-I score (see Quantum Incongruity XLMR in Figure 2) exhibits a pattern somewhat mirroring that of QE-U. While the latter extends toward positive values, QE-I leans toward the negative. However, both metrics share the problem that they bring the distributions of both classes too close. This is evident in Humicroedit and HUHU@IberLEF2023, which were the only corpora to show significant differences. Yet, Cohen's d effect sizes for these corpora (0.04 and 0.34, respectively) suggest that these differences are either negligible or too small to effectively distinguish the distributions.

6 Conclusion

This work examined the suitability of QE-Uncertainty (QE-U) and QE-Incongruity (QE-I) (Liu and Hou, 2023) as metrics for humor quality assessment across different datasets in four languages: English, French, Spanish, and Portuguese.

The first results, computed using monolingual GloVe embeddings, showed visible discrepancies between classes in most corpora. However, this distinction was less pronounced in others, suggesting that the metrics' effectiveness may vary across different datasets and languages. Statistical tests revealed significant differences between humorous and non-humorous instances in most corpora, but the effect sizes suggested minimal practical significance. Concerning specifically the QE-I score, results contradicted the expectation that humorous instances would have higher values.

The distribution ranges and shapes varied considerably across corpora, posing challenges for using the metrics as evaluation criteria. To better investigate these observations, we conducted the same experiments using a shared multilingual vector space. This resulted in more consistent scoring across corpora at the cost of a lower ability to discern subtleties between humor and non-humor instances.

In conclusion, while QE-U and QE-I show potential as metrics for humor quality, their effectiveness is inconsistent across different corpora and embedding models. Further research is still needed to refine these metrics and explore other potential indicators of humor quality, e.g. the scores proposed by Kao et al. (2016), He et al. (2019), Xie et al. (2021), and others.

7 Limitations

As mentioned in section 1, humor is complex, subjective, and multifaceted phenomenon. We acknowledge that the chosen features cover only a specific aspect of humor: semantic incongruity, without necessarily extending the analysis with extra-linguistic information or cultural contexts.

We also recognize that this analysis is limited to a specific pair of measurements and could be extended to other proposed scores from the literature, such as the ones mentioned in section 2. We note that there are other theories of humor, such as Superiority theory, Relief theory, Social theories, and others (Rutter, 1997); however, we only focused on the Incongruity theory, which provides a one-sided view when analyzing humor.

Finally, we also believe that the lack of monolingual experiments with corpora in Spanish is a downside of this work. We deliberately decided to not use other models available for this language (e.g. word2vec) to ensure that experiments are consistent for all languages.

8 Ethics Statement

We believe that systems capable of dealing with and producing humor can foster unity and ease communication tensions. However, we recognize that humor has been historically used in a derogatory or offensive manner to belittle or discriminate against individuals or groups (Bemiller and Schneider, 2010).

Therefore, the scientific community must not consider it acceptable to automatically generate jokes that incite violence, hatred, or prejudice, including but not limited to racial, gender, and sexual stereotypes, xenophobia, and other forms of discrimination. In this context, we find it vital to mention that one of the corpus used, Clemêncio (Gonçalo Oliveira et al., 2020), is known to have texts labeled as jokes that negatively portray various groups, such as black people, Jewish people, and blonde women; it also touches sensitive topics like suicide and pedophilia (Inácio et al., 2023). Similarly, HUHU@IberLEF 2023 (Rosso, 2023) must contain various texts with the same kind of content, as it was created for a shared task about hurtful humor. We do not know if such an analysis or discussion exists for the other corpora considered in this work.

Acknowledgements

We would like to thank fellow members of the community who kindly replied to our e-mails asking for their datasets and codes.

References

Michelle L. Bemiller and Rachel Zimmer Schneider. 2010. IT'S NOT JUST A JOKE. *Sociological Spectrum*, 30(4):459–479.

Delia Chiaro. 2018. *The Language of Jokes in the Digital Age: Viral Humour*. Routledge, Milton Park, Abingdon, Oxon ; New York, NY.

Luis Chiruzzo, Santiago Castro, Mathias Etcheverry, Diego Garat, Juan José Prada, and Aiala Rosá. 2019. Overview of the HAHA Task: Humor Analysis based on Human Annotation at IberEval 2019. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*, pages 132–144, Bilbao. CEUR-WS.org.

Luis Chiruzzo, Santiago Castro, Santiago Góngora, Aiala Rosá, J. A. Meaney, and Rada Mihalcea. 2021. Overview of HAHA at IberLEF 2021: Detecting, Rating and Analyzing Humor in Spanish. *Procesamiento del Lenguaje Natural*, 67:257–268.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Gregory W Corder and Dale I Foreman. 2011. *Nonparametric Statistics: A Step-by-Step Approach*. Wiley.

Hugo Gonçalo Oliveira, André Clemêncio, and Ana Alves. 2020. Corpora and baselines for humour recognition in Portuguese. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1278–1285, Marseille, France. European Language Resources Association.

Nathan Hartmann, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jéssica Rodrigues, and Sandra Aluísio. 2017. Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, pages 122–131, Uberlândia.

He He, Nanyun Peng, and Percy Liang. 2019. Pun Generation with Surprise. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 1734–1744, Minneapolis. Association for Computational Linguistics.

Nabil Hossain, John Krumm, and Michael Gamon. 2019. "President Vows to Cut Hair": Dataset and Analysis of Creative Text Editing for Humorous Headlines. In *Proceedings of the 2019 Conference of the North*, pages 133–142, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcio Inácio, Gabriela Wick-pedro, and Hugo Gonçalo Oliveira. 2023. What do humor classifiers learn? An attempt to explain humor recognition models. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 88–98, Dubrovnik, Croatia. Association for Computational Linguistics.

Justine T. Kao, Roger Levy, and Noah D. Goodman. 2016. A Computational Model of Linguistic Humor in Puns. *Cognitive Science*, 40(5):1270–1285.

Yang Liu and Yuexian Hou. 2023. Mining Effective Features Using Quantum Entropy for Humor Recognition. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2048–2053, Dubrovnik, Croatia. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-Refine: Iterative Refinement with Self-Feedback.

Rada Mihalcea and Stephen Pulman. 2007. Characterizing Humour: An Exploration of Features in Humorous Texts. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 4394, pages 337–347. Springer Berlin Heidelberg, Berlin, Heidelberg.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 531–538, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Tristan Miller, Christian Hempelmann, and Iryna 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, Vancouver, Canada. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Roberto Labadie Tamayo y Berta Chulvi y Paolo Rosso. 2023. Everybody hurts, sometimes overview of HUrtful HUmour at IberLEF 2023: Detection of humour spreading prejudice in twitter. *Procesamiento del Lenguaje Natural*, 71(0):383–395.

Jason Rutter. 1997. *Stand-up as Interaction : Performance and Audience in Comedy Venues*. Ph.D. thesis, University of Salford, Salford, UK.

John Von Neumann. 1996. *Mathematical Foundations of Quantum Mechanics*. Princeton Landmarks in Mathematics and Physics. Princeton University Press, Princeton Chichester.

Thomas Winters and Pieter Delobelle. 2021. Survival of the Wittiest: Evolving Satire with Language Models. In *Proceedings of the Twelfth International Conference on Computational Creativity*, pages 82–86, Mexico City, Mexico. Association for Computational Creativity (ACC).

Yubo Xie, Junze Li, and Pearl Pu. 2021. Uncertainty and Surprisal Jointly Deliver the Punchline: Exploiting Incongruity-Based Features for Humor Recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 33–39, Online. Association for Computational Linguistics.

A Corpora sizes

As mentioned in subsection 4.1, we conducted a preprocessing step on the corpora to filter out jokes with more than two sentences. The corpora size of the corpora before and after preprocessing can be seen in Table 1.

Corpus	Before		After	
	H	NH	H	NH
SemEval	1,615	645	688	183
Humicroedit	15,095	15,095	3,211	3,211
JOKER-EN	3,084	2,208	1,348	815
JOKER-ES	868	1,131	474	648
JOKER-FR	1,998	2,001	1,029	1,040
HAHA2019	11,595	18,405	3,001	5,123
HAHA2021	14,595	21,405	3,831	5,897
HUHU2023	869	1,802	255	486
Clemêncio	1,400	1,400	622	620
Puntuguese	2,850	2,850	2,488	2,522

Table 1: Number of instances of each class: Humor (H) and Non-humor (NH) for all the studied corpora before and after filtering.

B Resources

Throughout this research, we used multiple resources, especially corpora and embedding models. This is a brief overview of how to locate each resource used.

B.1 Corpora

SemEval 2017 Task 7 https://alt.qcri.org/semeval2017/task7/index.php?id=results;

Humicroedit https://www.cs.rochester.edu/u/nhossain/humicroedit.html;

JOKER CLEF 2023 https://www.joker-project.com/clef-2023/;

HAHA@IberLEF 2019 https://www.fing.edu.uy/inco/grupos/pln/haha/2019/;

HAHA@IberLEF 2021 https://www.fing.edu.uy/inco/grupos/pln/haha/;

HUHU@IberLEF 2023 https://zenodo.org/records/7967255;

Clemêncio https://github.com/NLP-CISUC/Recognizing-Humor-in-Portuguese;

Puntuguese https://anonymous.4open.science/r/Puntuguese-7B67/README.md.

Some corpora, namely JOKER CLEF 2023 and HAHA@IberLEF 2021, are hosted in outdated systems that do not support new registrations or do not have direct links for download. In such cases, we contacted the original authors who promptly granted us access to the data.

B.2 GloVe models

English https://nlp.stanford.edu/projects/glove/

Spanish https://github.com/dccuchile/spanish-word-embeddings

French https://github.com/Ismailhachimi/French-Word-Embeddings

Portuguese http://nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc

As mentioned in subsection 4.2, despite having a repository with Spanish embeddings, the link for downloading specifically the GloVe model is unavailable. We included the repository only for completeness, as we were not able to reach the authors and get a copy of the file needed.