Attempting to Recognize Humor via One-Class Classification

CISHUHUC@HUHU-IberLEF

Marcio Lima Inácio[1,*], Hugo Gonçalo Oliveira[1]

[1]*University of Coimbra, Centre for Informatics and Systems (CISUC), Polo II, Pinhal de Marrocos, 3030-290 Coimbra, Portugal*

Abstract

Teaching machines to understand and identify humor is an important step to create Natural Language Processing systems that are able to treat challenging linguistic phenomena. Additionally, it might help other researchers on understanding the dynamics of how humor occurs and how it is used to propagate and maintain social stereotypes and prejudices. For that, we participated in the HUHU Shared Task in the first subtask of Humor Detection. Given recent concerns about the trustfulness of humor classifiers, and how they take advantage of the training data, we decided to explore using One-Class Classification methods for this task. Yet, in general, using this kind of approach produced lower scores (0.541 F1) than other fully-supervised methods (0.796 F1). We also showed that training on non-humor examples reached better results than the opposite, however we argue that this idea might not be mostly useful for real-scenario applications and that using better textual representations of verbal humor may overcome this setting.

Keywords

Humor Recognition, One-Class Classification, Novelty Detection, Large Language Models

1. Introduction

Humor is known to be a form of expressing and propagating prejudice and stereotypes toward marginalized groups, being used as a way to legitimize power and oppression of some groups upon others [1]. In this sense, Humor Recognition systems can help us to better filter this kind of content, which can ultimately enable us to understand how it occurs and which are the social dynamics behind their realization.

Additionally, understanding verbal humor requires a deep and complex knowledge of the language [2]. Therefore, creating such systems can help us to develop novel computational methods that can analyze language and its use more deeply. In this context, we decided to participate in the HUHU Shared Task, specifically in Subtask 1: HUrtful HUmor Detection [3].

Recently, we showed that content-based methods for Humor Recognition might not necessarily learn characteristics that are inherently related to the humorous effect but rely on other

IberLEF 2023, September 2023, Jaén, Spain

[*] Corresponding author.

mlinacio@dei.uc.pt (M. L. Inácio); hroliv@dei.uc.pt (H. Gonçalo Oliveira)
https://eden.dei.uc.pt/~mlinacio/ (M. L. Inácio); https://eden.dei.uc.pt/~hroliv/ (H. Gonçalo Oliveira)
0000-0002-0875-4574 (M. L. Inácio); 0000-0002-5779-8645 (H. Gonçalo Oliveira)

CEUR Workshop Proceedings (CEUR-WS.org)

stylistic details that are more related to the author or source [4]. This raised a question if the models are indeed learning the desired task of Humor Recognition. These observations also brought concerns about the difficulties of finding suitable non-humorous examples so that the machine actually learns about the phenomenon rather than resort to unrelated idiosyncrasies of the text.

Given the previous background, we decided to explore the feasibility of using One-Class Classification (also known as Novelty Detection, Outlier Detection, or Out-of-domain Detection) [5, 6] methods for this task, under the assumption that we can model the desired phenomenon by only looking at its positive instances. This kind of model would help us to create Humor Classification systems that better capture humor-related characteristics and that do not rely on negative instances, making corpora creation simpler and less demanding.

One-Class Classification is usually used in binary classification scenarios where examples of one of the classes are scarce or difficult to gather, focusing on identifying if a given example consists of an abnormal pattern compared to the "standard" data presented during training. It is also used as means of cleaning datasets, by identifying outliers that largely differ from the general distribution of the data; for example, Larson et al. [7] applied this approach to curate data for dialog systems. For text classification, this kind of approach has been recently explored specially for fake news detection [6, 8].

Furthermore, as for means of comparison, we also used Jocoso, the winning method of the HAHA 2021 Shared Task on Humor Recognition [9, 10]. In general, we achieved the best F1-Scores (0.796) with our re-implementation of Jocoso. Our approach based on One-Class Classification, however, did not achieve such high scores (0.541).

The remainder of the paper is organized as follows: section 2 shows the methods we considered and implemented for this task, followed by the first experiments used to select the final two approaches for submission in section 3. Our main results are presented in **??**, alongside a general discussion in section 4.

2. Humor Recognition Methods

For the task of Humor Recognition, we followed two main paths: a strong baseline based on the HAHA-winning [9] system, Jocoso [10], and One-Class Classification methods, which are binary classification systems that are trained only on one of the available classes.

2.1. LLMs Ensemble

Essentially, Jocoso [10] is an ensemble of different fine-tuned Large Language Models (LLMs) through a hard-voting process, as can be seen in Figure 1, that describes a general architecture of the system, with which the authors tested different model combinations, achieving their final ensemble with the following methods:

- Multilingual Base Cased BERT (mBERT) [11];
- BETO [12];
- ALBERT [13];
- sBETO [12, 14];

- RoBERTa [15];
- Naive Bayes: this is the only classifier that is not based on an LLM;

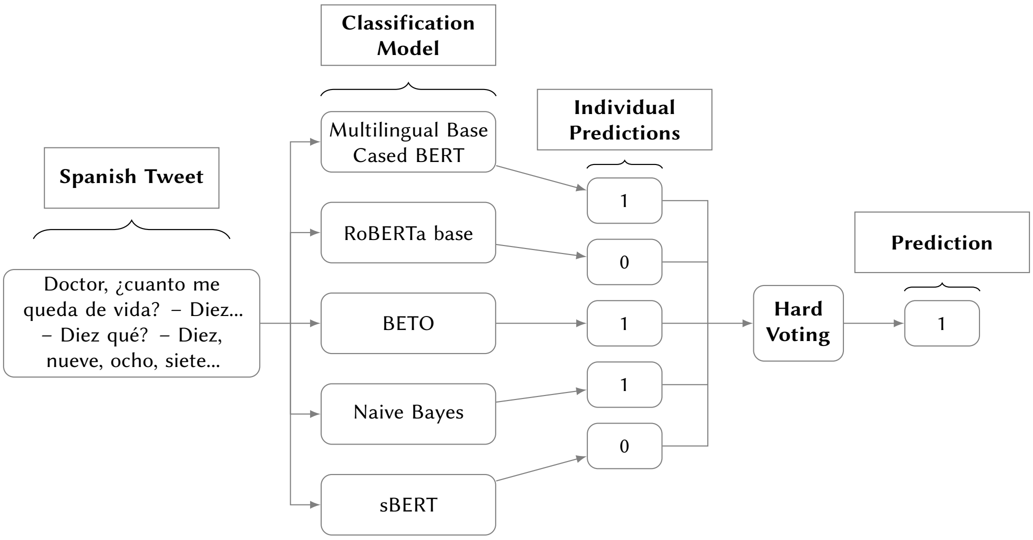

Figure 1: Architecture of the Jocoso system. Source: Adapted from [10].

In our study, we decided to give more attention to monolingual LLMs for Spanish, so that the model has better representations of the language's structures, since it has seen more Spanish data during pre-training and usually has a tokenizer better adapted to this specific language, which can increase performance [16]. We also gathered models with different architectures and ones that use different corpora for their pre-training. The LLMs that we tested are:

- mBERT [11];
- BETO [12];
- ALBETO: an ALBERT model trained on monolingual Spanish data [17];
- RoBERTuito: a follow-up version of sBETO [12, 18, 14];
- RoBERTaBNE: a RoBERTa model trained on data of the National Library of Spain (BNE) [19];

As can be noticed, we did not include the Naive Bayes approach, since we wanted to have a method that is focused entirely on using LLMs. Finally, we will describe in section 3 how we tested different combinations of these models to find the best one, and we reached to the final combination used in our submission: ALBETO, BETO, and RoBERTaBNE.

Besides the hard-voting process, we also compared using a soft-voting ensemble, by using the confidence scores given by each model. In our preliminary experiments, this kind of approach produced comparable results to the hard voting (as will be shown in section 3), but, as they were not largely different, we kept using hard-voting for simplicity and to keep it closer to the original Jocoso approach.

2.2. One-Class Classification

For One-Class Classification, we used two classical approaches: One-Class SVM [20, 21] and Isolation Forest [22]. Additionally, we also decided to explore a distance-based method named DCDistanceOCC [6], previously explored for fake news detection.

As this kind of approach is focused on creating a model based on instances of a single class, we explored two different scenarios: train on humor or train on non-humor instances. At test time, a given example is compared by the model with its modeling of the training class.

Another variation that we can have is regarding how to represent the texts. For this, we took three ways: using TF-IDF counts, using BETO, and using RoBERTuito. We decided to test two Spanish pre-trained LLMs to check if using a model focused on tweets (RoBERTuito) is better than using a general-purpose BERT model (BETO). Since these LLMs were not pre-trained for creating sentence representations, we used mean pooling, i.e. the text is represented as the mean vector across its tokens.

3. Model Selection and Submissions Results

To define the exact methods that were going to be used in the final submissions, we first conducted a preliminary study with the approaches that were presented in section 2. For these first experiments, we split the provided data into provisional training (75%) and validation (25%) sets. The results from this study can be seen in Table 1.

The table shows that the best results were achieved by using the Jocoso-based methods. For the sake of simplicity, we omit every other combination of the models and show only the best results and the combination of all five fine-tuned models. This way, we reached an F1 score of 0.79 on the held-out validation set using both hard and soft voting. As mentioned earlier, for our first submission (run1), we chose to use the hard voting method with the models shown in the table.

Regarding the use of One-Class Classification methods, they did not achieve such high F1 values. In general, both Isolation Forest and DCDistanceOCC classified every – or almost every – instance to the class used during training, meaning that the model they created for the class was too general and did not capture significantly distinct characteristics of humor or non-humor.

On the other hand, One-Class SVM was somewhat able to model the texts; however, the results are still lacking. From its results, we can observe that, despite the low F1 scores, it is preferable to model non-humor – i.e. train on non-humor. This might indicate that humorous language is more diverse and harder to model, but can also be a limitation of the models used to get the textual representations, as they were not necessarily pre-trained with humorous data.

Finally, for our second submission (run2), we used One-Class SVM trained on non-humor. Since we did not notice a difference among the representations used, we kept using RoBERTuito, as it is mostly related to the textual domain of tweets.

For the final submissions, we trained the two selected approaches with the whole training set provided for the task. The final results on the official test set, as shared in the official HUHU website, are shown in Table 2. We can confirm that, as expected from the preliminary experiments (Table 1), using the hard voting method produced the best results, which are even competitive, getting 4th place in the general ranking and outperforming both baselines.

Table 1
Results of the preliminary study.

Method	F1
ALBETO	0.73
mBERT	0.77
BETO	0.77
RoBERTaBNE	0.76
RoBERTuito	0.68
Hard Voting	
ALBETO + BETO + RoBERTaBNE	**0.79**
All five models	0.77
Soft voting	
ALBETO + mBERT + BETO	**0.79**
ALBETO + BETO + RoBERTaBNE	**0.79**
All five models	0.78
One-Class SVM + TF-IDF	
Train on humor	0.38
Train on non-humor	0.53
One-Class SVM + BETO	
Train on humor	0.37
Train on non-humor	0.53
One-Class SVM + RoBERTuito	
Train on humor	0.35
Train on non-humor	0.53
Isolation Forest + TF-IDF	
Train on humor	0.49
Train on non-humor	0.00
Isolation Forest + BETO	
Train on humor	0.49
Train on non-humor	0.03
Isolation Forest + RoBERTuito	
Train on humor	0.49
Train on non-humor	0.03
DCDistanceOCC + TF-IDF	
Train on humor	0.49
Train on non-humor	0.23
DCDistanceOCC + BETO	
Train on humor	0.49
Train on non-humor	0.00
DCDistanceOCC + RoBERTuito	
Train on humor	0.49
Train on non-humor	0.00

Table 2
Final classification results.

Method	F1
Winning method (RETUYT-INCO_run1)	0.820
Baseline (BLOOM-1b1)	0.789
Baseline (BETO)	0.759
Hard Voting (CISHUHUC_run1)	0.796
One-Class SVM + RoBERTuito (CISHUHUC_run2)	0.541

4. Conclusions and Related Work

During the HUHU Shared Task, we focused on the first subtask of HUrtful HUmor Detection [3]. For this, we decided to explore using One-Class Classification methods, which try to perform the binary classification by only looking to examples of one class during training.

Our main results showed that using this kind of approach with general textual representations (TF-IDF and LLMs with mean pooling) did not achieve satisfactory results. Additionally, we found out that, for One-Class SVM, modeling non-humor – and, consequently, considering deviant instances as humor – produced better results than the opposite. Nonetheless, we argue that this approach (training on non-humor) might not be well-suited for applications in real scenarios, as it would consider any kind of creative or non-standard use of language as being humor; to avoid this we would need a largely diverse training set of non-humorous texts, which can become unfeasible. Therefore, we argue that further research on One-Class Classification for Humor Recognition should focus on actually modeling humor.

In this sense, we believe that looking for other kinds of textual representation, such as explicit feature sets or sentential LLMs fine-tuned with humorous data, can be an interesting path to enhance such models [23].

A deeper analysis of the models' performance in different scenarios and domains may also be a fruitful path for future research, which can help us understand if they are actually learning to identify general humor characteristics. This would be interesting to investigate not only for One-Class Classification, but for any kind of approach on Humor Detection.

Acknowledgments

This work was funded by national funds through the FCT – Foundation for Science and Technology, I.P. (grant number UI/BD/153496/2022), within the scope of the project CISUC (UID/CEC/00326/2020) and by the European Social Fund, through the Regional Operational Program Centro 2020.

References

[1] M. L. Bemiller, R. Z. Schneider, IT'S NOT JUST A JOKE, Sociological Spectrum 30 (2010) 459–479. URL: http://www.tandfonline.com/doi/abs/10.1080/02732171003641040. doi:10.1080/02732171003641040.

[2] S. E. O. Tagnin, O humor como quebra da convencionalidade, Revista Brasileira de Linguística Aplicada 5 (2005) 247–257. URL: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1984-63982005000100013&lng=pt&tlng=pt. doi:10.1590/S1984-63982005000100013.

[3] R. Labadie-Tamayo, B. Chulvi, P. Rosso, Everybody Hurts, Sometimes. Overview of HUrtful HUmour at IberLEF 2023: Detection of Humour Spreading Prejudice in Twitter, in: Procesamiento del Lenguaje Natural (SEPLN), volume 71, 2023.

[4] M. Inácio, G. Wick-Pedro, H. Gonçalo Oliveira, 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, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 88–98. URL: https://aclanthology.org/2023.latechclfl-1.10.

[5] D. Miljković, Review of novelty detection methods, in: The 33rd International Convention MIPRO, 2010, pp. 593–598. URL: https://ieeexplore.ieee.org/document/5533467/.

[6] P. Faustini, T. Covoes, Fake News Detection Using One-Class Classification, in: 2019 8th Brazilian Conference on Intelligent Systems (BRACIS), IEEE, Salvador, Brazil, 2019, pp. 592–597. URL: https://ieeexplore.ieee.org/document/8923888/. doi:10.1109/BRACIS.2019.00109.

[7] S. Larson, A. Mahendran, A. Lee, J. K. Kummerfeld, P. Hill, M. A. Laurenzano, J. Hauswald, L. Tang, J. Mars, Outlier Detection for Improved Data Quality and Diversity in Dialog Systems, in: Proceedings of the 2019 Conference of the North, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 517–527. URL: http://aclweb.org/anthology/N19-1051. doi:10.18653/v1/N19-1051.

[8] M. P. S. Gôlo, M. C. De Souza, R. G. Rossi, S. O. Rezende, B. M. Nogueira, R. M. Marcacini, One-class learning for fake news detection through multimodal variational autoencoders, Engineering Applications of Artificial Intelligence 122 (2023) 106088. URL: https://linkinghub.elsevier.com/retrieve/pii/S0952197623002725. doi:10.1016/j.engappai.2023.106088.

[9] L. Chiruzzo, S. Castro, S. Góngora, A. Rosá, J. A. Meaney, R. Mihalcea, Overview of HAHA at IberLEF 2021: Detecting, Rating and Analyzing Humor in Spanish, Procesamiento del Lenguaje Natural 67 (2021) 257–268. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/download/6394/3814.

[10] K. Grover, T. Goel, HAHA@IberLEF2021: Humor Analysis using Ensembles of Simple Transformers, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), CEUR-WS.org, Málaga, 2021, pp. 883–890. URL: http://ceur-ws.org/Vol-2943/haha_paper1.pdf.

[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[12] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained BERT model and evaluation data, in: PML4DC at ICLR 2020, 2020. URL: https://pml4dc.github.io/iclr2020/papers/PML4DC2020_10.pdf.

[13] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, ALBERT: A lite BERT for self-supervised learning of language representations, in: International Conference on Learning Representations, 2020. URL: https://openreview.net/forum?id=H1eA7AEtvS.

[14] J. M. Pérez, J. C. Giudici, F. Luque, Pysentimiento: A python toolkit for sentiment analysis and SocialNLP tasks, 2021. arXiv:2106.09462.

[15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2019. URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[16] P. Rust, J. Pfeiffer, I. Vulić, S. Ruder, I. Gurevych, How Good is Your Tokenizer? On

the Monolingual Performance of Multilingual Language Models, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 3118–3135. URL: https://aclanthology.org/2021.acl-long.243. doi:10.18653/v1/2021.acl-long.243.

[17] J. Cañete, S. Donoso, F. Bravo-Marquez, A. Carvallo, V. Araujo, ALBETO and DistilBETO: Lightweight spanish language models, in: Proceedings of the 13th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 4291–4298. URL: https://aclanthology.org/2022.lrec-1.457.

[18] M. García-Vega, MC. Díaz-Galiano, MA. García-Cumbreras, FMP. Del Arco, A. Montejo-Ráez, SM. Jiménez-Zafra, E. Martínez Cámara, CA. Aguilar, MAS. Cabezudo, L. Chiruzzo, et al., Overview of TASS 2020: Introducing emotion detection, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) Co-Located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain, 2020, pp. 163–170.

[19] A. G. Fandiño, J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, M. Villegas, MarIA: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022). URL: https://upcommons.upc.edu/handle/2117/367156#.YyMTB4X9A-0.mendeley. doi:10.26342/2022-68-3.

[20] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, R. C. Williamson, Estimating the Support of a High-Dimensional Distribution, Neural Computation 13 (2001) 1443–1471. URL: https://direct.mit.edu/neco/article/13/7/1443-1471/6529. doi:10.1162/089976601750264965.

[21] S. Alam, S. K. Sonbhadra, S. Agarwal, P. Nagabhushan, One-class support vector classifiers: A survey, Knowledge-Based Systems 196 (2020) 105754. URL: https://linkinghub.elsevier.com/retrieve/pii/S0950705120301647. doi:10.1016/j.knosys.2020.105754.

[22] F. T. Liu, K. M. Ting, Z.-H. Zhou, Isolation Forest, in: 2008 Eighth IEEE International Conference on Data Mining, IEEE, Pisa, Italy, 2008, pp. 413–422. URL: http://ieeexplore.ieee.org/document/4781136/. doi:10.1109/ICDM.2008.17.

[23] M. Gôlo, M. Caravanti, R. Rossi, S. Rezende, B. Nogueira, R. Marcacini, Learning Textual Representations from Multiple Modalities to Detect Fake News Through One-Class Learning, in: Proceedings of the Brazilian Symposium on Multimedia and the Web, ACM, Belo Horizonte Minas Gerais Brazil, 2021, pp. 197–204. URL: https://dl.acm.org/doi/10.1145/3470482.3479634. doi:10.1145/3470482.3479634.