CISUC at IDPT2021: Traditional and Deep Learning for Irony Detection in Portuguese

Hugo Gonçalo Oliveira[0000−0002−5779−8645], José Pereira[0000−0001−8649−9006], and Guilherme Cruz

CISUC, Department of Informatics Engineering,
University of Coimbra, Coimbra, Portugal
hroliv@dei.uc.pt, {jose,gjcruz}@student.dei.uc.pt

Abstract. These notes describe the participation of the CISUC team in the IDPT 2021 shared task. Irony detection was tackled as a text classification task, where both traditional and transformer-based (BERT) approaches were explored. The former performed ok, but not everything went well, and the results achieved by BERT were not evaluated, due to an issue with our official submissions. Nevertheless, we still discuss some of the options taken, identify important features, and present validation results in the training data.

Keywords: Irony Detection · Portuguese · Text Classification · Transformers · Logistic Regression

1 Introduction

Irony is a rhetorical device where interpretation should not be literal [18], because its meaning diverges significantly from, and is often the opposite [7], of the intended meaning. Irony detection is a subtask of Natural Language Processing aiming at the automatic classification of texts as ironic or not, and is extremely relevant for tasks like Sentiment Analysis and Opinion Mining [14]. But irony detection can be challenging, even for humans, who often rely on visual clues, like facial expression or tone [7], for recognising irony. This is especially true when irony is expressed through text only, despite studies on identifying textual clues for irony detection [1].

Irony detection has been tackled by several Natural Language Processing (NLP) researchers, who adopted different approaches. In 2018, there was a SemEval task on *Irony Detection in English Tweets* [18] that covered the binary classification of tweets as ironic or not. Best systems adopted a deep learning approach, e.g., a densely LSTM neural network, based on pre-trained static word embeddings, with syntactic and sentiment features [19]. But there were also more traditional approaches, e.g., an ensemble classifier with Logistic

IberLEF 2021, September 2021, Málaga, Spain.

Regression (LR) and a Support Vector Machine (SVM), considering pretrained word and emoji embeddings, as well as handcrafted sentiment and word-based features [13]. Since then, as it happened for other NLP tasks, pre-trained language models such as BERT [3], or variations like RoBERTa [8], were exploited for obtaining contextualized embeddings, which can be combined with a classifier, e.g, a recurrent Convolutional Neural Network with a LSTM layer [11].

This paper describes the participation of a team from the Center of Informatics and Systems of the University of Coimbra (CISUC) in the *Irony Detection in Portuguese* (IDPT) task [2], included in the 2021 edition of the *Iberian Languages Evaluation Forum* (IberLEF). This was the first time we tackled irony detection, but our interest follows previous work on text classification of Portuguese text, specifically emotions [4] and humour [6].

Given that annotated datasets were made available by IDPT's organisation, one with tweets and another with news, we tackled IDPT with a supervised machine learning approach. Classifiers were learned from the training data and used for classifying the test data, then submitted to be evaluated. Yet, our first step was to look at the data, in order to increase our sensibility with this domain. In the process, we noted some patterns and learned about the sources of the training data, which lead to a data cleaning process, described in Section 2. Following this, we decided to explore both traditional text classification approaches as well as a more recent deep learning approach. The former required us to set some parameters, including the number of features, but it also enabled us to analyse and learn about the most important features for irony, at least in the provided datasets. For this approach, Section 3 provides some insights on the previous process, including the inclusion of lexicon features. The same section describes the deep learning approach, based on the popular transformer-based architecture BERT [3]. For both approaches, we present the results of validation in the training data.

Before concluding, Section 4 has a brief discussion on the official results of the selected classifiers in IDPT. Unfortunately, due to due to our own mistake in the submission process, classifications by the BERT classifiers were not properly evaluated, which made it impossible for us to know their real performance in the test dataset. On the other hand, the performance of the traditional approach, based on Logistic Regression, was good enough for an approach that could be seen as a baseline. This conclusion is mostly based on the results in the news dataset. From the performance in the tweets dataset it is hard to make conclusions. Even though the majority of tweets was automatically classified as ironic, according to the evaluation metrics, more than a half was not. Apparently, this issue was common to all participants.

2 Data

Our starting point was the data provided by IDPT's organisation, namely 15,213 tweets and 18,495 news documents, labelled as ironic (1) or non-ironic (0), which we used for training our models. Test data comprised 300 unlabelled tweets and

300 news documents. For evaluation purposes, test data had to be submitted with automatically-assigned labels.

While analysing the aforementioned datasets, we immediately noticed what could be a discrepancy between training and test data for tweets. Unlike the test data, training data contained little to no emojis, hashtags (#), user mentions (@), URLs, as well as no line breaks. Having in mind that this could have a negative impact on the classification task, and that some of those features could be relevant for irony detection, we tried to understand the differences.

During this process, we learned about the criteria adopted in the creation of the training data, after reading some of the references provided by the organisation [5, 15]. Specifically, we found the dataset created in the scope of da Silva's BSc thesis [15], which seemed to cover most of the tweets of the training data. However, this dataset was available[1] in a slightly different format, where some of aforementioned missing items were either directly present in the textual content or could be recovered from additional properties.

One of such properties was the tweet ID, which enabled us to retrieve most of the original tweets through Twitter's API. With this, we confirmed the hashtag-based criteria adopted for automatically-labelling the dataset:

- Ironic tweets were those containing the hashtags `#ironia` or `#sarcasmo`;
- Non-Ironic tweets were those containing `#economia`, `#politica` or `#educação`.

Based on da Silva's thesis, we made our own pre-processing of the dataset, which included: the complete removal of all five hashtags above; the normalisation of user mentions and URLs, respectively replaced by `@user` and `@link`[2]. Table 1 illustrates this with tweets in the training dataset, provided by the organisation, the original tweets as published on Twitter, and the result after our pre-processing. Differences towards the provided datasets was the inclusion of emojis, the complete removal of hashtags used for non-ironic tweets (e.g., `#economia`), as well as the normalisations.

3 Approaches

This section describes both approaches adopted in our participation in the IDPT task, a traditional machine learning approach, which could be seen as a baseline, and a deep learning approach based on BERT. Moreover, for each approach, validation results are presented and, for the traditional approach, we take a look at important features considered for detecting irony.

[1] https://github.com/fabio-ricardo/deteccao-ironia

[2] We later noticed that using the '@' character was not the best option, because some tokenizers split it from the following word. However, the impact for Portuguese should still be minimal, because these words are in English.

Training	Bravo
Original	Bravo 👏 👏 @JuanManSantos @PoliciaColombia @ELTIEMPO @elheraldoco @RevistaSemana @NoticiasRCN @NoticiasCaracol #ironia
Pre-processed	Bravo 👏 👏 @user @user @user @user @user @user @user @link
Training	5 MOTIVOS PARA AMAR O BRASIL: Old Corrupção Brasil
Original	5 MOTIVOS PARA AMAR O BRASIL: http://ow.ly/9vZX307MOdv #Old #Corrupção #Ironia #Brasil
Pre-processed	5 MOTIVOS PARA AMAR O BRASIL: @link #Old #Corrupção #Brasil
Training	economia KKKKKKKKKKKKK to rindo mas de nervoso rs
Original	#economia KKKKKKKKKKKKK to rindo mas de nervoso rs
Pre-processed	KKKKKKKKKKKKK to rindo mas de nervoso rs

Table 1. Example of tweets pre-processing.

3.1 Traditional Approach

Different traditional machine learning classifiers, implemented in the Python library scikit-learn [10], were trained and validated in different splits of the training data. For this purpose, documents were represented by TF-IDF vectors, also resorting to scikit-learn's *TfidfTransformer*. Portuguese stopwords in the NLTK list were ignored in this process and different parameters were tested, namely the n-gram range, maximum document frequency, minimum document frequency, and maximum number of features. While experimenting in the training datasets, we decided on setting:

- N-gram range to 1 (unigrams), as we saw no improvements with bigrams;
- The maximum document frequency to 0.5, meaning that tokens occurring in more than half of the documents in the collection were ignored, for not being discriminant enough;
- The minimum document frequency to 3, meaning that tokens occurring in only one or two documents were ignored, for not being frequent enough.

We also tested different values for the maximum number of features.

Cross-Validation Tables 2 and 3 report on the performance of three different classifiers in a 10-fold cross-validation, respectively in the tweets and news training datasets, using different numbers of features (500, 1,500 and 5,000). The three classifiers used were Logistic Regression (LR), Naive Bayes (NB), and Random Forest (RF), all white-box, and the metrics considered were: Balanced Accuracy (BAcc), for being the official measure of IDPT; Precision, Recall, and F1 score (F1).

Achieved performances are interesting for a baseline. As expected, performance is slightly higher for news, which should be more formal, than for tweets, where several conventions are broken. But top F1 scores of 89% and 97% may actually suggest that irony detection is not that hard, especially in formal text.

Classifier	Features	BAcc	Precision	Recall	F1
LR	500	0.82±0.03	0.92±0.01	0.92±0.01	0.92±0.01
	1500	0.84±0.02	0.94±0.01	0.94±0.01	0.94±0.01
	5000	0.80±0.03	0.93±0.01	0.93±0.01	0.93±0.01
NB	500	0.84±0.03	0.93±0.01	0.93±0.01	0.93±0.01
	1500	0.88±0.02	**0.95**±0.01	**0.95**±0.01	**0.95**±0.01
	5000	**0.89**±0.02	**0.95**±0.01	**0.95**±0.01	**0.95**±0.01
RF	500	0.84±0.02	0.92±0.01	0.92±0.01	0.92±0.01
	1500	0.87±0.02	0.93±0.01	0.93±0.01	0.93±0.01
	5000	0.87±0.02	0.94±0.01	0.94±0.01	0.94±0.01

Table 2. Performance in 10-fold cross validation in the Tweets training set.

Classifier	Features	BAcc	Precision	Recall	F1
LR	500	0.95±0.00	0.95±0.00	0.95±0.00	0.95±0.00
	1500	0.96±0.00	**0.97**±0.00	**0.97**±0.00	**0.97**±0.00
	5000	**0.97**±0.00	**0.97**±0.00	**0.97**±0.00	**0.97**±0.00
NB	500	0.92±0.01	0.92±0.01	0.92±0.01	0.92±0.01
	1500	0.94±0.01	0.94±0.01	0.94±0.01	0.94±0.01
	5000	0.96±0.00	0.96±0.00	0.96±0.00	0.96±0.00
RF	500	0.94±0.01	0.94±0.01	0.94±0.01	0.94±0.01
	1500	0.95±0.01	0.95±0.01	0.95±0.01	0.95±0.01
	5000	0.96±0.01	0.96±0.00	0.96±0.00	0.96±0.00

Table 3. Performance in 10-fold cross validation in the News training set.

However, these performances are achieved with 5,000 features, which probably leads to models that are over-fitted to the training dataset. Therefore, having in mind that the documents in the test datasets were extracted for a different time period than the training, and there could be significant vocabulary differences (e.g., due to different trending topics), we decided to consider not more than 1,500 features for our submission. Validation performances suggest that a lower number of features has a bigger impact for NB and that, on the other hand, LR is less affected. In fact, for tweets, LR performs better with 1,500 than with 5,000 features. Adding to the simplicity of LR and to its best performance in the news dataset, we decided to use LR in our official IDPT runs.

Lexicon-based features We further decided to explore additional features that we thought could be useful for irony detection, namely:

- Concreteness and imageability scores, obtained from the Minho Word Pool norms [16], where 3,800 Portuguese words have averages of such properties, from 1 to 7, assigned by several judges;
- Sentiment and emotion features, acquired from the NRC Emotion lexicon [9], where such features (0 or 1) are assigned to 14,182 English words through crowdsourcing, then translated to other languages, including Portuguese.

This resulted in ten extra features, averaged for each document: Concreteness, Imageability, Positive, Negative, Anger, Anticipation, Disgust, Fear, Joy, Trust. Our intuition is that these could complement the TF-IDF features, because, indirectly, they end up covering a larger vocabulary, more focused and independent of the training data, and may thus lead to less over-fitting. Since the entries in the previous lexicons are all lemmas, for computing these features, documents were first lemmatized, using the Portuguese models of the Stanza [12] package.

Table 4 shows the performance of the LR classifier using only the extra features or adding them to the 1,500 TF-IDF features. When used alone, their impact is irrelevant for the Tweets, but they seem to make a difference for the News. Alone, they achieve a F1 of 0.71, but when together with TF-IDF, F1 drops by 1 point.

Classifier	Features	BAcc	Precision	Recall	F1
Tweets	Extra	0.50±0.00	0.82±0.00	0.82±0.00	0.82±0.00
	TF-IDF+Extra	0.82±0.02	0.92±0.01	0.92±0.01	0.92±0.01
News	Extra	0.68±0.01	0.71±0.01	0.71±0.01	0.71±0.01
	TF-IDF+Extra	0.95±0.00	0.96±0.00	0.96±0.00	0.96±0.00

Table 4. Performance of LR in 10-fold cross validation in the training sets.

Our option for including these features anyway is further supported by an analysis of their importance coefficient in an LR classifier that learned from them only, and of their values in documents of different classes, especially in the News dataset. For instance, ironic news express slightly more joy, negativity, disgust and anticipation, and are also more imagetic.

Feature Importance After training a LR classifier, each feature has an importance coefficient, which can be useful for interpretation. Tables 5 show the most important features when the previous classifier is trained in the training datasets, as well as the number of documents where they occur and the proportion classified as ironic. Some interesting insights can be observed. For instance, most tweets with user mentions (normalised as '@user') are ironic, and so are more "extreme" tweets that use words like 'adoro', 'tudo' or 'nada'. As for the news, many relevant features for irony are names of politicians, suggesting that they are common targets of irony, or were during the time-span the data was collected. Other features include words that typically appear before a citation, namely 'disse' and 'explicou'.

3.2 Transformer-Based Approach

BERT [3] is a transformer-based model widely used in Natural Language Processing since its release, by Google. It is pretrained in two general language tasks, masked language modelling and next sentence prediction, but can be fine-tuned for other tasks, including text classification, which is our case.

Feature	Docs	Ironic
user	3,118	91.4%
pra	617	1.1%
ironía	261	100%
tá	277	99.6%
adoro	157	100%
tudo	231	97.4%
nada	227	99.6%
pq	210	100%

Feature	Docs	Ironic
disse	3,522	87.7%
temer	936	95.3%
dilma	1,334	95.6%
explicou	948	98.3%
cunha	460	95.0%
aécio	535	97.2%
agora	1,637	78.6%
hoje	1,887	78.7%

Table 5. Most relevant features in the Tweets (left) and News (right) training datasets.

Fine-tuning Our starting point was BERTimbau [17], i.e., *BERT Base Portuguese Cased* (BERT-PT), a model with 110M parameters, pretrained by Neuralmind, exclusively for (Brazilian) Portuguese. In order to fine-tune this model for irony classification, we used the BertForSequenceClassification class of the Transformers library[3], which adds a classification head on top of BERT. Parameters for this model were empirically selected, namely: batch size of 16 for the tweets and 8 for the news, due to memory limitations; and Adam optimizer[4] for being the common option, with lr=2e−5 and eps=1e−8.

Text Size We quickly came across a limitation on the text size, i.e., some documents were longer than the maximum number of tokens that BERT could handle (510 word pieces, plus the initial `[CLS]` and the final `[SEP]` tokens). Figures 1 show the distribution of documents according to the number of tokens in both training sets. As expected, this is much more frequent in the news dataset, as news articles tend to be longer than tweets. Still, after careful analysis and deliberation, we assumed that the proportion of documents that exceeded the limit of tokens was insignificant and deemed that their absence would not produce a noticeable change in the model's overall performance. This left us with two choices: remove the longer documents from the dataset or truncate them. We chose the latter for several reasons. The first 510 tokens of each document would still be relevant for irony detection and, this way, the classifier would learn from all data. Moreover, documents in the test dataset could also exceed this limit and we could not simply remove them.

Validation results In order to select the aforementioned parameters, the BERT-based classifier was validated in the training dataset. For this, we used 60% of the data for training, 10% for validation and 30% for test. Table 6 summarises the performance of the best models of each kind.

Validation performances achieved with BERT are very high in both datasets and outperform the already high F1 of the best traditional approaches, confirming that BERT is a very powerful model. Additional experiments were performed

[3] See https://huggingface.co/transformers/model_doc/bert.html
[4] See https://huggingface.co/transformers/main_classes/optimizer_schedules.html

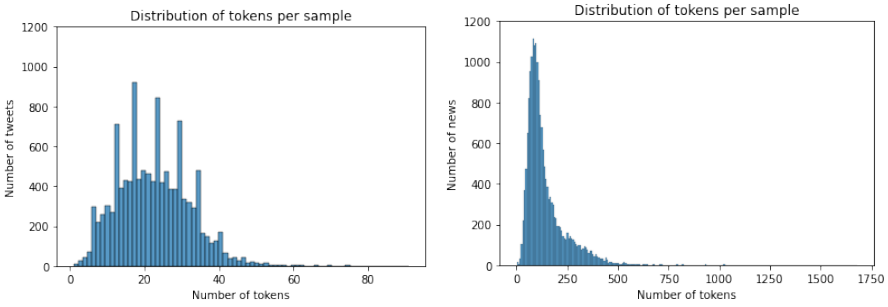

Fig. 1. Histograms with number of tokens in the tweets (left) and news (right) training datasets.

Data	BAcc	Precision	Recall	F1
Tweets	0.97	0.99	1.00	0.99
News	0.98	0.99	0.99	0.99

Table 6. Performance of BERT-based classifiers in the training sets.

with balanced versions of the datasets, obtained with undersampling, but they did not lead to further improvements. We should, nevertheless, recall that these results can be too over-fitted to the training data.

4 Results

Despite all the experiments performed with BERT and our positive expectations regarding their high performance, due to a mistake in our submission[5], it is impossible for us to take any conclusion on the real performance of the BERT-based models and on its comparison to other approaches, including our traditional approach. At least until the labels of the test data are not revealed.

As for the traditional approach, official results are in Table 7. We recall that the model used is based in LR, with 1,500 token features plus 10 lexicon-based features. It achieved sixth position overall in the tweets dataset, but it was not much different from other participants, nor from our BERT submissions where the labels were shuffled. In fact, when analysing our labels, we note that the majority of the tweets were classified as ironic, which results in a recall close to 100%. However, precision is lower than 50%, suggesting that about half of the tweets were not ironic and should have not been classified as so.

While the test labels are not revealed, it is not possible to make an error analysis. However, we believe that performances in the tweet dataset were harmed

[5] More precisely, our script had a switch for shuffling the data to label, which was on when the test data was classified, meaning that the submitted labels were not in the expected order for the official evaluation.

by the criteria adopted in the creation of the data, which we suspect to have diverged between training and test. For instance, in da Silva's thesis [15], one of the criteria for labelling tweets as ironic was the presence of the hashtag `#ironia`, i.e., all the tweets using this hashtag were considered to be positive examples of ironic tweets, and could be included in the training data as such. Yet, when we search Twitter for the tweets of IDPT test data, all of them use the previous hashtag, even if, according to the results, more than a half was not labelled as ironic. This was probably the result of manual analysis, and should definitely be more accurate than da Silva's criteria, which would automatically label them as ironic. However, this also means that the provided training data was misleading and, as we have seen, classifiers trained on such data are not apt for correctly labelling IDPT's tweet test data.

Run	BAcc	Precision	Recall	F1
tweets_3	0.502	0.412	0.992	0.581
news_3	0.802	0.681	0.852	0.757

Table 7. Official results of the traditional approach in the test datasets.

On the news data, our traditional approach achieved the eighth position overall, with a BAcc that was 12 points below the best run. The name of the team that submitted the three best runs (TeamBERT4EVER) suggests that they used BERT, which confirms that irony detection is one more task where BERT is currently the way to achieve the state-of-the-art. Nevertheless, we highlight that, in the traditional approach, a white-box model was used, which enabled us to learn a bit more about how irony is expressed in the datasets, e.g., important features, most of which we would not immediately associate with irony.

5 Conclusion

We have described the participation of the CISUC team in the IDPT 2021 shared task. Despite our issues with the BERT models, the balance is still positive. This participation lead to the application of known approaches to a new challenge, it made us think about the relevance of irony, and taught us a little bit about the way it is expressed in Portuguese.

We tackled this challenge as a text classification task and explored both traditional and deep learning approaches. As expected, deep learning seems to be the best path to achieve top performances, and BERT is definitely a solid model for attempting at state-of-the-art results. Still, sometimes, learning about language and how it works is at least as important as achieving the best performances, and white-box models are much more accessible for this purpose. Unfortunately, while the labels of the test data are not revealed, we cannot compare the performance of the latter with our BERT-based approaches in a real scenario, and thus not analyse the trade-off. The same happens to the comparison with the

runs of other teams. In the news test set, our LR-based approach was ranked eight, and it will definitely be interesting to learn about the other approaches, once the proceedings of IDPT are published.

Now that we had our first contact with this topic, there are plenty of ideas for future work. A possible direction would be studying to what extent it is possible to learn a general classifier of irony, not suitable for a specific type of text or time-span. We did train a BERT-based model on both training datasets (tweets and news), but it was one of the corrupt submissions. A train-validation-test in the previous dataset lead to a surprisingly high performance, i.e., comparable to the performance of the type-specific models. But stronger conclusions can only be taken once we actually test the model in the test data. Still, more than learning from two (or more) different types of text, a general classifier would also have to be trained in texts published in different time-spans. The latter are relevant, because classifiers will learn from the used vocabulary and, during specific time periods, some entities (e.g., politicians, athletes, organisations) can be or become a preferred target of irony, thus skewing the model's evaluation of the words associated with these entities.

Another interesting direction would be a deeper analysis of the actual impact of different features, not only tokens, but also n-grams, case (upper or lower), punctuation, emojis and lexicon-based features, among others. Besides possibly improving the traditional approaches, some of those features could also be appended to the inputs of the BERT-based classifiers.

Finally, one could exploit other available corpora for irony detection, possibly starting with a corpus of humour [6], which typically resorts to irony. More data could also be retrieved from Twitter, possibly relying on additional heuristics for self-labelling (e.g., specific emojis). However, given the specificities of irony, quality of data is especially important. Therefore, any automatically-created dataset should be manually revised.

Acknowledgements

This work is partially funded by national funds through the FCT – Foundation for Science and Technology, I.P., within the scope of the project CISUC – UID/CEC/00326/2020 and by European Social Fund, through the Regional Operational Program Centro 2020.

References

1. Carvalho, P., Sarmento, L., Silva, M.J., De Oliveira, E.: Clues for detecting irony in user-generated contents: oh...!! it's" so easy";-. In: Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion. pp. 53–56 (2009)
2. Corrêa, U.B., dos Santos, L.P., Coelho, L., de Freitas, L.A.: Overview of the IDPT task on Irony Detection in Portuguese at IberLEF 2021. Procesamiento del Lenguaje Natural **67** (2021)

3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Procs. of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics (Jun 2019)
4. Duarte, L., Macedo, L., Gonçalo Oliveira, H.: Exploring emojis for emotion recognition in Portuguese text. In: Proceedings of 19th EPIA Conference on Artificial Intelligence, EPIA 2019, Vila Real, Portugal, September 3-6, 2019, Proceedings, Part I. LNCS/LNAI, vol. 11805, pp. 719–730. Springer (September 2019)
5. de Freitas, L.A., Vanin, A.A., Hogetop, D.N., Bochernitsan, M.N., Vieira, R.: Pathways for irony detection in tweets. In: Proceedings of the 29th Annual ACM Symposium on Applied Computing. pp. 628–633 (2014)
6. Gonçalo Oliveira, H., Clemêncio, A., Alves, A.: Corpora and baselines for humour recognition in Portuguese. In: Proceedings of 12th International Conference on Language Resources and Evaluation. pp. 1278-–1285. LREC 2020, ELRA, Marseille, France (2020)
7. Grice, H.P.: Logic and conversation. In: Speech acts, pp. 41–58. Brill (1975)
8. Liu, B.: Sentiment analysis and opinion mining. Synthesis lectures on human language technologies **5**(1), 1–167 (2012)
9. Mohammad, S.M., Turney, P.D.: Crowdsourcing a word-emotion association lexicon **29**(3), 436–465 (2013)
10. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
11. Potamias, R.A., Siolas, G., Stafylopatis, A.G.: A transformer-based approach to irony and sarcasm detection. Neural Computing and Applications **32**(23), 17309–17320 (2020)
12. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D.: Stanza: A Python natural language processing toolkit for many human languages. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (2020)
13. Rohanian, O., Taslimipoor, S., Evans, R., Mitkov, R.: WLV at SemEval-2018 task 3: Dissecting tweets in search of irony. In: Proceedings of 12th International Workshop on Semantic Evaluation. pp. 553–559. Association for Computational Linguistics, New Orleans, Louisiana (2018)
14. Sarmento, L., Carvalho, P., Silva, M.J., De Oliveira, E.: Automatic creation of a reference corpus for political opinion mining in user-generated content. In: Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion. pp. 29–36 (2009)
15. da Silva, F.R.A.: Detecção de ironia e sarcasmo em língua portuguesa: Uma abordagem utilizando deep learning. Tech. rep., Universidade Federal de Mato Grosso (2018)
16. Soares, A.P., Costa, A.S., Machado, J., Comesaña, M., Oliveira, H.M.: The Minho Word Pool: Norms for imageability, concreteness, and subjective frequency for 3,800 Portuguese words. Behavior Research Methods **49**(3), 1065—1081 (2017)
17. Souza, F., Nogueira, R., Lotufo, R.: BERTimbau: Pretrained BERT models for Brazilian Portuguese. In: Proceedings of the Brazilian Conference on Intelligent Systems (BRACIS 2020). LNCS, vol. 12319, pp. 403–417. Springer, Cham (2020)

18. Van Hee, C., Lefever, E., Hoste, V.: SemEval-2018 task 3: Irony detection in English tweets. In: Proceedings of The 12th International Workshop on Semantic Evaluation. pp. 39–50 (2018)
19. Wu, C., Wu, F., Wu, S., Liu, J., Yuan, Z., Huang, Y.: THU_NGN at SemEval-2018 task 3: Tweet irony detection with densely connected LSTM and multi-task learning. In: Proceedings of 12th International Workshop on Semantic Evaluation. pp. 51–56. Association for Computational Linguistics, New Orleans, Louisiana (2018)