Improving Irony Detection by Balancing Methods and Feature Selection

Anthony I. M. Luz[1], **Henrique Santos**[2], **Manoel M. P. Medeiros**[1], **Rafael T. Anchiêta**[1]

[1]Federal Institute of Piauí - Picos (IFPI)
Laboratory of Artificial Intelligence, Robotics, and Automation (LIARA)

[2]Federal University of Piauí - Picos (UFPI)

{marquesanthony62, henriques.santod}@gmail.com

{mmessias,rta}@ifpi.edu.br

Abstract. *Irony is a linguistic phenomenon that can be seen as a funny or strange aspect of a situation that is very different from what is expected, using words that say the opposite of what they really mean, often as a joke, and with a voice that shows that. When it is just text, detecting irony becomes quite challenging. In this paper, we adopt an approach organized into three stages: feature extraction, sampling techniques, and feature selection to detect ironic texts written in the Portuguese language. We evaluate our strategy on the IDPT corpus and achieve* 0.55 *balanced accuracy, outperforming state-of-the-art results. Moreover, we found out that both sampling techniques and feature selection may improve the results.*

1. Introduction

According to the Oxford dictionary[1], irony may be viewed as a funny or strange aspect of a very different situation from what is expected, using words that say the opposite of what really mean, often as a joke, and with a tone of voice that shows this. When dealing only with texts, the task of irony detection becomes even more challenging due to the lack of other elements to identify whether a text is ironic or not. For example, Figure 1 shows an example of ironic text. The irony in this sentence is that a player nicknamed "Valdívia" (possibly after the Chilean player who played for Palmeiras) scored a "brilliant goal" against the Palmeiras team. The irony lies in the fact that a player with the same nickname as a former player from the opposing team scored an extraordinary goal against that team. This can be seen as an unexpected and ironic turn of events. One can see that identifying irony without context may be a difficult task, even for humans.

Jogador do Inter apelidado de Valdívia
faz um golaço contra o Palmeiras

Figure 1. Example of ironic text.

Detecting ironic texts has important implications for Natural Language Processing (NLP) tasks. For example, hate speech detection, toxic language, opinion mining, and others [Reyes et al. 2009].

[1]https://www.oxfordlearnersdictionaries.com/

The usefulness of methods to detect irony has led to a growing interest in the study of irony. Several shared tasks have contributed to the development of methods in different languages. SemEval-2018 (English), IroSvA-2019 (Spanish),WANLP-2021 (Arabic), and IDPT-21 (Portuguese) [Corrêa et al. 2021].

In particular, IDPT-21, which focused on identifying ironic text in Portuguese, the obtained results by the teams are still far from the other shared tasks. While for the other shared tasks, the teams achieved an accuracy greater than 0.7, for the IDPT-21, the best result reached only 0.52. This suggests that there is much room for improvement in this area.

In this paper, we adopted a supervised machine-learning approach to detect ironic texts written in Portuguese. Our strategy is organized into three stages. First, we extracted thirteen features and fed some machine learning algorithms to detect ironic texts. In this stage, we identified the classifier that performs best in this task. Second, we explored sampling techniques in order to improve the obtained results. Finally, we computed the importance of each feature, aiming to find which are the most predictive features.

We evaluated our approach on the IDPT dataset, achieving 0.55 balanced accuracy, outperforming the best team in the IDPT shared task. We found out that both sampling strategies and selecting important features may improve classification results.

The rest of this paper is organized as follows. In Section 2, we present an overview of the teams in the IDPT shared task. In Section 3, we detail the feature extraction process. Section 4 presents an analysis and discussions of the results achieved. Finally, in Section 5, we conclude the paper and outline potential future directions for research.

2. Related Work

There are several developed strategies to detect ironic texts in various languages. Here, we focused on approaches that deal with the Portuguese language.

[Anchiêta et al. 2021] developed an approach based on superficial features as Text Frequency-Inverse Document Frequency (TF-IDF) and fed the Support Vector Machine (SVM) classifier to identify ironic texts. Also, the authors used back-translation as data augmentation to balance the corpus. The method achieved 0.523 balanced accuracy.

[Oliveira et al. 2021] also adopted the TF-IDF weigh scheme as a feature and evaluated three different classifies Logistic Regression, Naive Bayes, and Random Forest. The authors reached 0.511 balanced accuracy with the Naïve Bayes classifier.

[Heinrich et al. 2021] explored different strategies of pre-processing, feature extraction, and ten machine learning algorithms. The best result was reached with the Multilayer Perceptron, a pre-processing stage that focuses on cleaning up the undesirable and irrelevant patterns, and the TF-IDF as a feature. This approach achieved 0.502 balanced accuracy.

3. Feature Extraction

The task of detecting irony is challenging, as irony often happens in different ways and may be related to factors intrinsic to the observer. However, here we assume a much simpler and more objective definition, in view of the definition of the "Houaiss Dictionary", which conceived irony as a "figure of speech through which a different message

is passed - in general contrary – to the literal message, usually with the aim of criticizing or promoting humor". Based on this concept, we extract thirteen linguistic features from [Pedro 2018] to determine whether or not a text is ironic, as depicted in Table 1.

Table 1. Extracted features to detect ironic tweets.

Feature	Example
# Punctuation (1)	"!!??!?!?"
# Abbreviations (1)	"lol", "ah", "eh", "hi"
# Laughing expressions (1)	"kkkk", "rsrsrs", "hahaha"
# Emoticons (1)	":-)", ":-D", ":-P"
# Expressions (5)	"na boa" "só que não" "só que" "sim" "seria"
# Fear text (1)	"medo", "#medo", "medo!"
# Irony hashtags (1)	"#ironia", "#sarcasmo", "#joking", "#kidding"
# Morphosyntactic patterns (2)	"ADJ + ADJ" |"ADV + ADV" "tão + ADJ" |"tão + ADV"

From this table, the '#' symbol indicates the 'number of'. Also, we have five features for the number of expressions and two for morphosyntactic patterns. We use regular expressions to extract the former features, while for the last ones, we use the NLPNet tagger [Fonseca and Rosa 2013].

Based on the extracted features, we trained and assessed some supervised machine learning algorithms, such as K-Nearest Neighbors (KNN), Decision Trees (DT), Support Vector Machine (SVM), and Random Forest (RF) from the scikit-learn package [Pedregosa et al. 2011]. In the following section, we present the results and discussions about our strategy.

4. Results and Discussion

As aforementioned, we evaluated some supervised machine-learning algorithms for the task of detecting ironic tweets in Portuguese. We achieved the best result on the test set using the decision tree classifier[2], reaching 0.505 in the balanced accuracy (bacc) metric. We adopted this metric in order to compare our results with those of the IDPT shared task [Corrêa et al. 2021]. Table 2 shows the comparison of our results with the shared task teams.

We can see that our simple approach achieved good results, our method ranked third out of seven teams. In addition to the balanced accuracy, we analyzed the confusion matrix (Table 3) to identify which class our strategy misclassifies.

As we can see, our method misclassifies the non-ironic class. This is because the corpus is unbalanced, i.e., a few non-ironic tweets concerning ironic ones. To mitigate this problem and inspired by the PiLN team [Anchiêta et al. 2021], which adopts

[2]We use the default parameters for each classifier from the scikit-learn package.

Table 2. Comparison of the obtained results.

Team	Bacc
PiLN [Anchiêta et al. 2021]	0.523
CISUC [Oliveira et al. 2021]	0.511
Our	**0.505**

Table 3. Confusion Matrix.

		Predicted	
		Ironic	**Non-Ironic**
Actual	**Ironic**	123	0
	Non-Ironic	175	2

a back-translation strategy to balance the corpus, we used a simpler approach. We explored both undersampling and oversampling techniques from the Imbalanced-learn toolbox [Lemaître et al. 2017].

Undersampling techniques remove examples from the training set that belong to the majority class with the objective of equalizing the number of examples in each class [Fernández et al. 2018]. As an undersampling technique, we use the RandomUnderSampler that randomly balances the dataset. After balancing the corpus, we achieved 0.53 balanced accuracy on the test set using the decision tree classifier, which outperforms the best team in the IDPT shared task. This result indicates that not much data is needed to improve the results of the shared tasks. Furthermore, this result may also suggest that the corpus may suffer from various noises. Table 4 presents the confusion matrix result after the undersampling technique.

Table 4. Confusion matrix after the undersampling technique.

		Predicted	
		Ironic	**Non-Ironic**
Actual	**Ironic**	51	28
	Non-Ironic	126	95

One can see that the decision tree correctly classified more non-ironic tweets. On the other hand, it misclassified more ironic tweets. We believe it is due to the undersampling, as it removed examples of the ironic class. In this way, we explored the oversampling strategy.

Oversampling techniques create artificial samples of the minority class to equalize the data set [Fernández et al. 2018]. As an oversampling technique, we use the RandomOverSampler, which randomly selects samples from the minority class and duplicates them to balance the corpus. Our result with this technique was a balanced accuracy of 0.55 on the test set using the same classifier as above, which slightly surpasses our previous result (0.53).

Table 5. Confusion matrix after the oversampling technique.

		Predicted	
		Ironic	**Non-Ironic**
Actual	**Ironic**	42	15
	Non-Ironic	135	108

From this table, one can see that the classification of non-ironic tweets has improved. However, as with the undersampling technique, the decision tree misclassified more ironic tweets. We believe it is because of noises in the corpus, as pointed out by [Oliveira et al. 2021]. We will explore this problem in future work. After balancing the corpus, our better result was 0.557 balanced accuracy with oversampling approach, thus ranking first out of the seven teams.

Besides exploring sampling strategies, we investigated which features are most important for the ironic tweets classification task. Thus, we computed the importance of each feature. The importance of a feature is computed as the (normalized) total reduction of the criterion brought by that feature. It is also known as the Gini importance [Pedregosa et al. 2011].

We found out that the most important features are 0 and 2, which are the number of punctuations and the number of laughing expressions, respectively. Based on this feature importance score, we selected the top three features and evaluated the decision tree classifier using only those features, without sampling strategies. The resulting balanced accuracy was 0.557. Interestingly, this result is identical to the balanced accuracy achieved when we used an oversampling technique with all thirteen features. This indicates that these three features are the most influential in predicting ironic and non-ironic tweets and that including additional features does not significantly improve the performance of the model.

Our method is publicly available at `https://github.com/Anth0nYM/irony-detector`.

5. Conclusion

This paper presented an approach to detect ironic tweets written in the Portuguese language. The strategy was based on the extraction of thirteen features to evaluate some machine learning algorithms. The decision classifier obtained the best results among the analyzed classifiers. Moreover, we explored sampling techniques and found out that both undersampling and oversampling improve the results. More than that, we also perform a feature selection to identify which features are most predictive. We found out that with only three features, we achieve the best result, which is the same as the oversampling technique.

For future work, we intend to investigate deep-learning techniques, such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). We also aim to explore data augmentation methods, including automatic text generation, to further enhance the results.

Acknowledgments

The authors are grateful to IFPI for supporting this work.

References

Anchiêta, R. T., Neto, F. A. R., Marinho, J. C., do Nascimento, K. V., and Moura, R. S. (2021). Piln IDPT 2021: Irony detection in portuguese texts with superficial features and embeddings. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021)*, pages 917–924, Málaga, Spain. CEUR-WS.org.

Corrêa, U. B., Coelho, L., Santos, L., and de Freitas, L. A. (2021). Overview of the idpt task on irony detection in portuguese at iberlef 2021. *Procesamiento del Lenguaje Natural*, 67:269–276.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., and Herrera, F. (2018). *Learning from imbalanced data sets*, volume 10. Springer.

Fonseca, E. R. and Rosa, J. L. G. (2013). Mac-morpho revisited: Towards robust part-of-speech tagging. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, pages 98–107, Fortaleza, Brazil. Sociedade Brasileira de Computação.

Heinrich, T., Ceschin, F., and Marchi, F. (2021). Teamufpr at IDPT 2021: Equalizing a strategy using machine learning for two types of data in detecting irony. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021)*, pages 952–932, Málaga, Spain. CEUR-WS.org.

Lemaître, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5.

Oliveira, H. G., Pereira, J., and Cruz, G. (2021). Cisuc at IDPT2021: Traditional and deep learning for irony detection in portuguese. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021)*, pages 898–909, Málaga, Spain. CEUR-WS.org.

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., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pedro, G. W. (2018). Comentcorpus: Identificação e pistas linguísticas para detecção de ironia no português do brasil. Master's thesis, Universidade Federal de São Carlos.

Reyes, A., Rosso, P., and Buscaldi, D. (2009). Humor in the blogosphere: First clues for a verbal humor taxonomy. *Journal of Intelligent Systems*, 18(4):311–332.