

# BERT Sentence Embeddings in different Machine Learning and Deep Learning Models for Author Profiling applied to Irony and Stereotype Spreaders on Twitter

Daniel Parres<sup>1,†</sup>, Claudia Gomez<sup>1,†</sup>

<sup>1</sup>Universitat Politècnica de Valencia, Camí de Vera, s/n, 46022 València, Valencia, Spain

## Abstract

Irony detection is an interesting problem with different fields of application, both for managing social media and for studying people's behavior and opinions. This paper focuses on classifying Twitter author profiles as ironic or non-ironic based on their tweets. For this purpose, different types of feature extraction are applied to each author's tweets, from classical techniques to the most recent ones that are part of the state of the art. Furthermore, once the feature extraction has been performed, classical Machine Learning techniques such as SVM, Random Forest, and Logistic Regression are applied, up to more recent methods such as artificial neural networks with Self Attention mechanisms. Finally, a discussion is opened on the methods used and what each technique can contribute to solving this task. As it happens in most of the tasks where Embedding techniques are applied in natural language, new frontiers of study, analysis, and application are opened. Therefore, this study provides different brushstrokes for the development of robust systems for the detection of ironic and non-ironic authors.

## Keywords

Author profiling, Sentence Embeddings, Machine Learning, Deep Learning, BERT, Irony,

## 1. Introduction

This work focuses on profiling authors based on their tweets into ironic or non-ironic, emphasizing authors who employ irony to spread stereotypes. With this objective we use data provided by PAN'22 [1]. This data is composed of 420 Twitter profiles and 200 tweets from each of them; each user is labeled as ironic or non-ironic.

The task of classifying authors as ironic or non-ironic based on their tweets is very interesting due to the current context in which we find ourselves, where anyone has access to social media and the freedom to share and spread their ideas. Because of this, identifying which authors are spreading comments that can be considered harmful from an assertive perspective is important both for managing and administering social media and sociological or psychological studies.

Thanks to the emergence of different Artificial Intelligence techniques and algorithms, this task can be covered by Machine Learning. This work first performs a study of the data, analyzing


---

<sup>†</sup>These authors contributed equally.

✉ dparres@prhlt.upv.es (D. Parres); cgomros@posgrado.upv.es (C. Gomez)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

tweets and preprocessing them. Once the data is studied, different classical techniques and state-of-the-art algorithms such as artificial neural networks are applied. And finally, the results obtained are evaluated and discussed.

## 2. Related Work

Since 2013, there are works focused on the detection of irony, an example is [2] which describes a set of textual characteristics to recognize irony at the linguistic level, focusing on short texts, such as tweets. The proposed model is evaluated in two dimensions: representativeness and relevance.

In [3] encouraging results are demonstrated for deriving pragmatic contextual models for irony detection, which provides the application of a new approach beyond the use of features. While on the other hand, in [4], in a contest to detect ironic tweets in the Arabic language, it is shown that classical feature-based models are superior to neural ones.

The task of author profiling is applied to different areas, such as detecting hate speech [5], gender [6] and age of authors [7], by analyzing their tweets, but this work focuses on detecting whether an author is ironic or not.

In author profiling, different features have been used for years, such as text length, cosine similarity ranking, word retrieval, Okapi BM25 ranking, and NRC emotions, as proposed in [8]. With the approach of using different features, in [9] for irony detection, classical models and Multilayer Perceptron are used, together with statistical techniques such as counting, post-tagger, textual markers, and lexicon-based as wordnet similarity. But the best performing models are SVM, MLP, and Random Forest, while the best performing features are textual markers, sentiment score, and polarity value.

In [10] and [11] the techniques and preprocessing that have obtained the best results in the PAN'19 and 20 contests for gender, bots and hate speech profiling are presented, where the extensive use of BERT word embeddings [12] and their use in neural models are highlighted.

It can be seen from all the literature to date that in author profiling problems, the use of classical versus neural models tends to perform better. Although this is with the help of BERT word embeddings, without preprocessing techniques given that they usually do not improve accuracy rates as discussed in [13] with multiple experiments.

## 3. Methodology

This section is divided into 2 subsections. The first one focuses on the analysis of the data and the study of the balancing of the two classes to know which metric is the most representative for the training of the algorithms, if there are repeated tweets and the most used words in each class.

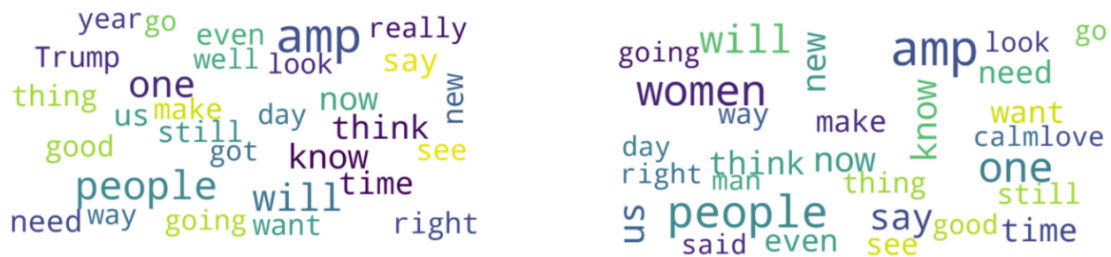
In the second subsection, the treatment of the data and the different models used are presented. To compare the models, 10-Fold Cross-Validation is used, with 90% for training and 10% test for classical models and 80% for training, 10% validation, and 10

### 3.1. Data Analysis

The data provided by PAN'22 is composed of 420 users, with 200 tweets each, where each profile is labeled as ironic or non-ironic. The distribution of the two classes is 210 ironic and 210 non-ironic authors, so we have a balanced problem, and a good metric for the analysis and comparison of models is the accuracy. Another feature of the dataset is found in the tweets where USER, HASHTAG, and URL tags are used to refer to users, hashtags, and URLs within the tweets themselves.

Analyzing the tweets, being a task with 200 tweets for each user, we have a total of 84,000 tweets, where we have 749 repeated tweets. In the case of the repeated tweets, since they are so few in proportion to the data set, they do not negatively affect the performance of the models.

On the other hand, the most used words by class have been analyzed, to observe if there is any pattern that repeats significantly in ironic or non-ironic authors. Eliminating USER, HASHTAG, URL, and stopwords, the texts have been used to construct the word clouds for ironic and non-ironic authors (Figure 1). There is hardly any difference in the words most frequently used for each type of author, although in non-ironic authors the word "women" is quite frequent, while in ironic authors the word "Trump" is one of the most repeated.



**Figure 1:** Wordclouds for ironic (left) and non-ironic (right) user tweets.

### 3.2. Models and Evaluation

Before presenting the different Machine Learning models for the task of author profiling, it is necessary to study which forms of tweet preprocessing best fit the task. Two different approaches have been used, the first one is the use of the classical TF-IDF vectorizer, the second one is the BERT tokenizer, and finally the BERT Sentence Embedding. For comparison, the experiments performed are presented in Table 1, where they have been tested with Support Vector Machines (SVM), Multilayer Perceptron (MLP), and Recurrent Neural Network (RNN).

It should be noted, as mentioned in the Methodology section, that the experiments were performed using 10-Fold Cross-Validation. It can be seen from the results of Table 1, that with the use of BERT Sentence Embedding better results are obtained. Since the model is trained for capturing the relationship between words taking into account the context, and performance is achieved that in comparison to the TF-IDF Vectorizer and BERT Tokenizer techniques is much higher. Therefore in the following experiments it was decided to use BERT Sentence Embedding for feature extraction.

**Table 1**

Accuracy of different feature extraction for tweets.

Model	Accuracy using TF-IDF Vectorizer	Accuracy using BERT Tokenizer	Accuracy using BERT Sentence Embedding
SVM	0.67	0.72	<b>0.93</b>
MLP(5 layers)	0.76	0.79	<b>0.92</b>
RNN(Bidirectional LSTM)	0.82	0.83	<b>0.93</b>

Having studied that BERT Sentence Embedding is the best text representation, it is interesting to analyze if the preprocessing of USER, HASHTAG, and URL tags can provide any improvement compared to no preprocessing. The removal of these terms cause a worsening in the accuracy of 1% in the models of Table 1, which may be because in the cases of ironic users, the use of hashtags is quite widespread and this helps the classifiers to find patterns and improve the understanding of what is irony detection.

Regarding the Machine Learning models used, two different scenarios are presented: on one hand, classical techniques are applied, and on the other one, more advanced techniques such as Artificial Neural Networks.

For the classical models, about 10,000 experiments have been carried out comparing different techniques such as Decision Trees, Logistic Regression, Gaussian Naive Bayes, Multinomial Naive Bayes, Support Vector Machine, Bernoulli Naive Bayes, K-Nearest Neighbors, Logistic Regression with different preprocessing techniques on the BERT Sentence Embedding vector such as Binarizer, Feature Agglomeration, MaxAbsScaler, MinMaxScaler, Normalizer, Principal Component Analysis, RBFSampler, Robust Scaler, StandardScaler, ZeroCount. The most significant results of the different combinations are presented in Table 2, where the best accuracy obtained by far is 0.94 using Logistic Regression with Normalizer L2, Robust Scaler and a Variance Threshold of 0.005.

**Table 2**

Most relevant experiments with classical algorithms.

Model	Accuracy
Decision Tree gini and max-depth=3	0.89
K-NearestNeighbors k=70, weights=distance and power=1	0.75
Gaussian Naive Bayes	0.88
MultinomialNB alpha=1 and fit <sub>prior</sub> = False	0.88
<b>Logistic Regression penalty=l2, C=5 and dual=False</b>	<b>0.94</b>
BernoulliNB alpha=100 and fit <sub>prior</sub> = True	0.90
Support Vector Machine C=10 y kernel=rbf	0.93

Currently, the best performing models are artificial neural networks due to their great generalization capacity and high performance, so a wide family of neural models is applied in this task. Five different types of architectures have been developed, a 5-layer Multilayer Perceptron, a bilinear convolutional neural network, and three different types of bidirectional LSTM RNN, a simple one, one with 1D Convolution mechanisms and another with 1D Convolutions and

Self-Attention.

Table 3 presents the different architectures with their corresponding accuracies. It can be seen that the neural methods obtain more or less the same results among them, surpassing on average the classical methods in Table 2. Despite being better on average than the classical methods, the highest accuracy reached is 0.94 for the classical Logistic Regression algorithm.

**Table 3**

Most relevant experiments with neural algorithms.

Model Architecture	Accuracy
Multilayer Perceptron (5 layers)	0.92
Bilinear CNN (BERT Sentence Embedding to Image)	0.93
Bidirectional LSTM	0.93
Conv. 1D + Bidirectional LSTM	0.93
Bidirectional LSTM + Self Attention + Conv. 1D	0.92

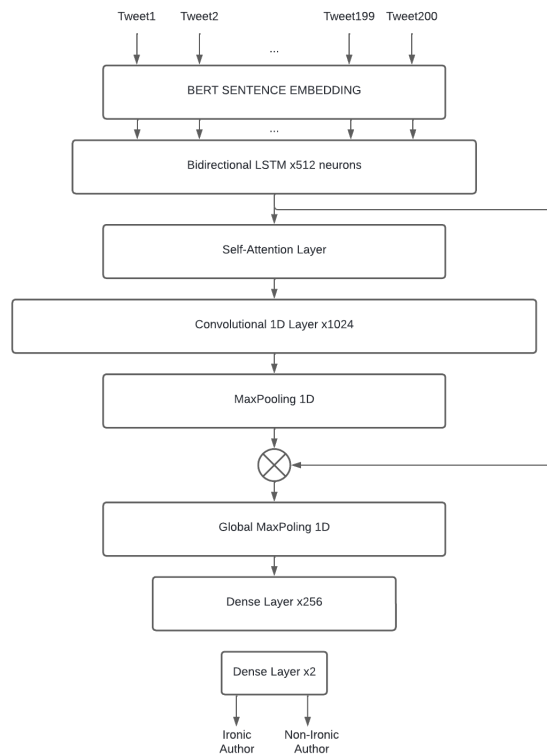
The main idea when training the Multilayer Perceptron or the classical methods is to average the vectors calculated by BERT for each author. That is, as each author has 200 tweets, BERT returns 200 vectors of dimension 784, so an average of these 200 vectors is performed to obtain only one. While in the rest of the neural models, the 200 BERT vectors have been used, an architecture that can perform well for this task it's the Convolutional 1D with Bidirectional LSTM layers, presented in Figure 3, called Iro-Net for simplicity.

Using all BERT vectors is enriching. All state-of-the-art neural models are able to achieve better results due to the large amount of information. Because of this, it has been decided to develop a neural network using convolution mechanisms together with a bidirectional LSTM layer, named Iro-Net and inspired by the model proposed by [14]. Moreover, in the field of Natural Language Processing, models that implement attention mechanisms such as Transformers are the standard. Therefore Iro-Net incorporates a Self-Attention layer after the recurrent LSTM layer. The Iro-Net architecture is shown in Figure 2 and has been designed specifically for this work.

### 3.3. Iro-Net Architecture

The Iro-Net architecture and its corresponding hyperparameters are presented in this section. It should be noted that the parametrization described corresponds to the best model submitted to Tira [15].

This artificial neural network is composed of different layers that perform different functions. The first layer corresponds to the Sentence Embedding of BERT and is in charge of handling the 200 tweets of each user and obtaining an interesting representation of them. A Bidirectional LSTM layer is applied to the Embeddings of the tweets and tries to perform a context-aware representation of the tweets. At the output of the Bidirectional LSTM layer is a Self-Attention layer, in order to discriminate which words or conditions are the most important when detecting irony. The next layer is a 1D Convolutional with a corresponding Max-Pooling layer that performs well for pattern recognition. A Residual connection from the output of the Bidirectional LSTM is added to the Max-Pooling output. Residual connections are widely used in Deep



**Figure 2:** Iro-Net Architecture.

Learning. Specifically, this design achieves a more discriminative representation, because of this it has been considered interesting to apply it. Finally, a Global Max-Pooling layer, a layer of 256 neurons and a Softmax output in charge of calculating the probability of whether the author is ironic or not are added.

Regarding the Hyperparameters used it is interesting to mention the following. All layers use Glorot weight initialization. In Glorot initialization the biases are initialized to 0 and the weights are calculated with the uniform distribution taking into account the size of the previous layer. The layers: Bidirectional LSTM, Max-Pooling and the penultimate layer with 256 neurons use the Dropout technique with values between 0.2 and 0.3. In addition, it is interesting to mention that all layers use the L2 regularizer with a value of 0.00005.

As for the Hyperparameters of the training, the Adam optimizer with a learning rate of 0.001 has been used. During training it is important to use a Learning Rate Annealing technique because changing the learning rate can improve the performance of the model and not stagnate at local minima. The best performing technique is Learning Rate On Plateau with a minimum learning rate value of 0.000001. Finally, the training consists of 100 epochs and a batch size of 16.

## 4. Discussion and Conclusion

As has been studied in the Related Work section, the use of BERT embeddings and the contextual relationship that can be obtained by using them is a key point in any natural language task. In this work, BERT's Sentence Embeddings have been compared against TF-IDF vectorizer and BERT's tokenizer, with Sentence Embedding being the best representation with a significant difference.

Furthermore, it's worth noting that the power of BERT's Sentence Embeddings representation is so powerful, that such a process does not need any kind of preprocessing. This opens new application frontiers, which in the not too distant future will allow Embeddings techniques to be used for other types of problems in the Machine Learning field.

With the reviewed bibliography, the classical models in the task of author profiling seem to have in their majority better performance than the neural ones, but we think that the cause of this is due to the reduced number of data that we have. If in the future we get more Twitter profiles labeled as ironic or non-ironic, it would be interesting to replicate the experiments corresponding to Tables 2 and 3, but for the moment, the best accuracy obtained is 0.94 using BERT's Sentence Embedding and Logistic Regression classifier.

As already mentioned, the detection of ironic profiles is an important task with many applications, both for directing and managing social media, as well as for sociological studies. Therefore, stressing the importance of a necessary continuous and in-depth improvement in this area is necessary.

In addition, another point to highlight about the work done is that despite how mainstream artificial neural network methods have become, for the moment we should not forget the classic algorithms, such as SVM, Decision Trees, Logistic Regression, and so on.

Since the PAN contest for irony detection consists of two phases, in the first phase (early bird) it has been decided to deliver two models, a classical one and a neural one. The classical algorithm delivered is Logistic Regression, with the parameterization presented in Table 2 and the neural network presented is Iro-Net, whose architecture is shown in Figure 3. In the results of the first phase, the best model among the two phases was Iro-Net with 96.11% accuracy, so in the second phase (final submission) the Iro-Net model will be presented as the definitive model.

As future work and possible extensions, using BERT as a pre-trained model and performing fine-tuning with different variations of BERT, such as distilled BERT, could bring improvements in the performance of ironic profile detection models based on tweets.

## References

- [1] O.-B. Reynier, C. Berta, R. Francisco, R. Paolo, F. Elisabetta, Profiling Irony and Stereotype Spreaders on Twitter (IROSTEREO) at PAN 2022, 2022. URL: <https://pan.webis.de/clef22/pan22-web/author-profiling.html>.
- [2] A. Reyes, P. Rosso, T. Veale, A multidimensional approach for detecting irony in twitter, *Language resources and evaluation* 47 (2013) 239–268.
- [3] J. Karoui, F. Benamara, V. Moriceau, N. Aussenac-Gilles, L. H. Belguith, Towards a contex-

- tual pragmatic model to detect irony in tweets, in: 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015), 2015, pp. PP-644.
- [4] B. Ghanem, J. Karoui, F. Benamara, V. Moriceau, P. Rosso, Idat at fire2019: Overview of the track on irony detection in arabic tweets, in: Proceedings of the 11th Forum for Information Retrieval Evaluation, 2019, pp. 10-13.
  - [5] F. Rangel, G. L. De la Peña Sarracén, B. Chulvi, E. Fersini, P. Rosso, Profiling hate speech spreaders on twitter task at pan 2021., in: CLEF (Working Notes), 2021, pp. 1772-1789.
  - [6] F. Rangel, P. Rosso, Pan19 author profiling: Bots and gender profiling, 2019. URL: <https://doi.org/10.5281/zenodo.3692340>. doi:10.5281/zenodo.3692340.
  - [7] F. Rangel, P. Rosso, M. Koppel, E. Stamatatos, G. Inches, Overview of the author profiling task at pan 2013, in: CLEF Conference on Multilingual and Multimodal Information Access Evaluation, CELCT, 2013, pp. 352-365.
  - [8] E. R. Weren, A. U. Kauer, L. Mizusaki, V. P. Moreira, J. P. M. de Oliveira, L. K. Wives, Examining multiple features for author profiling, *Journal of information and data management* 5 (2014) 266-266.
  - [9] E. Gose, *Pattern recognition and image analysis* (1997).
  - [10] F. Rangel, P. Rosso, Overview of the 7th author profiling task at pan 2019: bots and gender profiling in twitter, in: Working Notes Papers of the CLEF 2019 Evaluation Labs Volume 2380 of CEUR Workshop, 2019.
  - [11] F. Rangel, A. Giachanou, B. H. H. Ghanem, P. Rosso, Overview of the 8th author profiling task at pan 2020: Profiling fake news spreaders on twitter, in: CEUR Workshop Proceedings, volume 2696, Sun SITE Central Europe, 2020, pp. 1-18.
  - [12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* (2018).
  - [13] E. Alzahrani, L. Jololian, How different text-preprocessing techniques using the bert model affect the gender profiling of authors, *arXiv preprint arXiv:2109.13890* (2021).
  - [14] M. Polignano, M. d. Gemmis, G. Semeraro, Contextualized bert sentence embeddings for author profiling: The cost of performances, in: *International Conference on Computational Science and Its Applications*, Springer, 2020, pp. 135-149.
  - [15] M. Potthast, T. Gollub, M. Wiegmann, B. Stein, TIRA Integrated Research Architecture, in: N. Ferro, C. Peters (Eds.), *Information Retrieval Evaluation in a Changing World, The Information Retrieval Series*, Springer, Berlin Heidelberg New York, 2019. doi:10.1007/978-3-030-22948-1\\_5.