Profiling Irony and Stereotype Spreaders on Twitter: PAN Shared Task (IROSTEREO) 2022

Álvaro Rodríguez Sánchez[1], Martín Barroso Ordóñez[1]

[1]*Student at the Master's Degree in Artificial Intelligence, Pattern Recognition and Digital Imaging, Department of Computer Systems and Computation (DSIC), Polytechnic University of Valencia (UPV), Valencia (Spain)*

Abstract

We present our solution to the problem proposed by IROSTEREO's PAN Shared Task in 2022. It proposes the detection of irony and stereotype spreaders on Twitter. Throughout the memory, we will show how through: the technique based on neural networks for the pre-training of natural language processing such as BERT, the use of sentence embeddings, and two alternatives to put them together; it will be achieved from a set of tweets associated with a set of authors, predict whether an author is ironic or not. Lastly, it will be presented a model that achieves high accuracy and with which finally participated in the competition.

Keywords

Author profiling, Irony detection, Twitter, Natural language processing, Sentence embeddings, BERT.

1. Introduction

In this paper, we present our participation in the task proposed by the PAN in the year 2022 [1] [2]. This task, Profiling Irony and Stereotype Spreaders on Twitter, as well as some others [3], proposes the objective of determining whether an author is ironic or not by re-collecting a series of tweets associated with him/her. This challenge has been addressed by creating several models based on Machine Learning and Natural Language Processing techniques.

For the past few years, Author profiling has become highly relevant due to its potential applications [4], for example, in forensic linguistic studies, marketing analysis and verification of authorship of historical/literary texts. The aim of author profiling is to automatically extract demographic characteristics of the author of a text, such as gender, age, mother tongue or sexual orientation.
Nowadays, there are some Author profiling related to irony, for example, the detection of irony and humor from social networks [5], where a series of components that are the key to achieving their automatic processing are identified; finally, it is possible to relate that humorous texts are certainly positive, while ironic texts tend to be a bit more threatening. Nevertheless, there are sometimes that one of the concepts may have the characteristics of the other; this is why this task is so intriguing.

The following sections present the procedure and strategies that have been carried out and the different experiments that have been performed to reach a final model.

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

alrodsa2@inf.upv.es (R. Sánchez); marbaro1@inf.upv.es (M. B. Ordóñez)

2. Dataset

The data, for creating the first models, is located inside a folder, it contains 420 XML files, each of them referring to different authors; these authors are represented with an id. Each XML file is composed of 200 tweets from the author, these are in the English language. A *truth.txt* file is also provided which contains the tag associated with each author, this tag can be *I* if the author is ironic, or *NI* if the author is not ironic.

After performing the different experiments with the previous dataset, we were provided with a test dataset, which consists of 180 authors with 200 tweets each. Unlike the previous dataset, this one does not contain the truth.txt, since this is the one that will be taken into account to be evaluated by the contest.

3. Proposed solution

It has been proposed to solve this problem by using Machine Learning based models and without the use of a pre-processing, as the tweets are already pre-processed. Firstly, the sentence embedding representation is going to be used through BERT [6], during this process 2 alternatives were chosen:

- Average the 200 tweets of each author just before passing them through BERT, i.e., having applied all the necessary pre-processes to be able to enter them in BERT (this representation is shown in Figure 1 as the Model Input), all the 200 tweets are converted into array (Word2vec), then applied a padding and mask for each tweet to make all the same length and finally average all the 200 tweets into a final average vector. Once this average is obtained, pass it through BERT to subsequently use the result as a representation of that author.
- Pass every single tweet of each author through BERT and average the sentence embeddings obtained from BERT of all the tweets associated with each author to obtain a final sentence embedding per author that describes it.

After obtaining a sentence embedding for each of the authors (with either of the two methods), we will proceed to the training of a classical model. In our case, we have chosen to use the sklearn library, which offers a set of classical models; we will pass these sentence embeddings and the associated author label to these models for training.

4. Experiments

This section shows the results, using accuracy as a metric, of both methods presented in the previous section for different classical classifiers. The classifiers that have been used are the SVM[1], MLP[2], GaussianNB and RandomForest. It is important to note that all experiments have been performed using 10-fold cross-validation.

[1] Support Vector Machine

[2] Multi Layer Perceptron

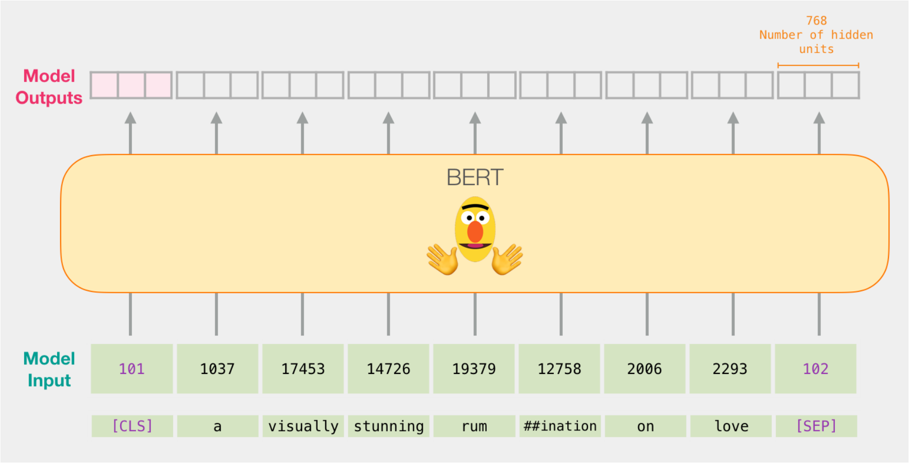

Figure 1: Tweet to sentence embedding conversion process using BERT.

4.1. Pre-BERT average embeddings method

The best result obtained from each of the classifiers and the value of their hyperparameters are shown below.

Classifier	Hyperparameters	Accuracy
SVM	C=0.0001	0.53809
MLP	hl1=128, hl2=64, hl3=32	0.64286
GaussianNB	smoothing=0.0001	0.63095
RandomForest	max_depth=10, n_estimators=100, max_features=50	0.64248

Table 1
Accuracies of the best classifiers in the Pre-BERT method.

As shown in Table 1, the model with the highest accuracy is the MLP model when 3 hidden layers are incorporated between the input and output layers. Despite this, the accuracy value obtained is not entirely promising, since accuracy of **0.64286** in the classification of 2 different classes is not a very optimistic value. In the following method, the strategy implemented in this method will be improved.

4.2. Post average embeddings method

The best result obtained from each of the pre-trained models used, the best classifier obtained with each one, and the value of its hyperparameters are shown below.

As we can see in Table 2 , we got a great improvement in the results achieved by averaging after (and not before) putting the data into the pre-training model. The best performer was given by the Multilingual Universal Sentence Encoder, although the difference is not significant.

Classifier	Hyperparameters	Accuracy
BERT MLP	hl1=32, hl2=32, hl3=16	0.933
Distill BERT SVM	C=21	0.933
MUSE MLP	hl1=4, hl2=64, hl3=32	0.935

Table 2
Accuracies of the best classifiers in the Post-BERT method.

5. Final model and conclusion

Finally, we have to train a final model with all the data with the characteristics that we have seen that have performed best in the experiments. In order not to over-fit we are going to put only a few iterations, 250, since we have observed that if we do more we easily reach 100% accuracy in the training samples, which is not desirable. So we have trained a final MLP with 4, 64, and 32 hidden units for 250 iterations that receive as input the averaged embeddings of a user's tweets that have been taken as input from the MUSE.

In the first submission to the TIRA platform [7] the test dataset predictions were uploaded in the requested XML format; an accuracy of 0.9556 was achieved.

As shown throughout the experiments and results, the use of MUSE together with an MLP has proven to be the most successful case for maximizing accuracy and detecting the presence of irony or non-irony of an author within a set of tweets associated with him/her.

References

[1] O.-B. Reynier, C. Berta, R. Francisco, R. Paolo, F. Elisabetta, Profiling Irony and Stereotype Spreaders on Twitter (IROSTEREO) at PAN 2022, in: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2022.

[2] J. Bevendorff, B. Chulvi, E. Fersini, A. Heini, M. Kestemont, K. Kredens, M. Mayerl, R. Ortega-Bueno, P. Pezik, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, E. Zangerle, Overview of PAN 2022: Authorship Verification, Profiling Irony and Stereotype Spreaders, and Style Change Detection, in: M. D. E. F. S. C. M. G. P. A. H. M. P. G. F. N. F. Alberto Barron-Cedeno, Giovanni Da San Martino (Ed.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Thirteenth International Conference of the CLEF Association (CLEF 2022), volume 13390 of *Lecture Notes in Computer Science*, Springer, ????

[3] I. H. F. L. C. W. Z. A. C. Paolo Rosso, Francisco Rangel, A survey on author profiling, deception, and irony detection for the arabic language, 2018. URL: https://compass.onlinelibrary.wiley.com. doi:10.1111/lnc3.12275.

[4] L. Wanner, J. Soler, Feature engineering for author profiling and identification: on the relevance of syntax and discourse, Universitat Pompeu Fabra. Departament de Tecnologies de la Informació i les Comunicacions (2017). URL: http://hdl.handle.net/10803/404984, applications of Natural Language to Information Systems.

[5] A. Reyes, P. Rosso, D. Buscaldi, From humor recognition to irony detection: The figurative language of social media, Data Knowledge Engineering 74 (2012) 1–12. URL: https://www.sciencedirect.com/science/article/pii/S0169023X12000237. doi:https://doi.org/10.1016/j.datak.2012.02.005, applications of Natural Language to Information Systems.

[6] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. URL: https://arxiv.org/abs/1810.04805. doi:10.48550/ARXIV.1810.04805.

[7] 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.