

# UMUTeam at IROSTEREO: Profiling Irony and Stereotype spreaders on Twitter combining Linguistic Features with Transformers

José Antonio García-Díaz<sup>1</sup>, Miguel Ángel Rodríguez-García<sup>2</sup>,  
Francisco García-Sánchez<sup>1</sup> and Rafael Valencia-García<sup>1</sup>

<sup>1</sup>Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Spain

<sup>2</sup>Departamento de Ciencias de la Computación, Universidad Rey Juan Carlos, 28933 Madrid, Spain

## Abstract

Irony is a curious mode of communication in which the speaker says something that wants the audience to be interpreted oppositely. Its automatic detection is a very challenging task due to its complex interpretation, and it has a significant potential for various applications in text mining. Social Media platforms like Twitter offer a vital chance to analyze this literary technique since users frequently utilize it to give their opinions. In this working note, we describe the contribution designed for the PAN's shared author profiling task and its subtask concerning Stereotype Stance Detection. The former consists in determining whether the authors spread irony and stereotypes and the latter is focused on identifying stereotypes that can hurt vulnerable groups. The organizers provide a set compiled from Twitter to carry out the task. In particular, we have proposed a supervised learning pipeline consisting of a combination of Deep Learning techniques that utilizes context and non-context embeddings to address the binary classification. The resulting system reaches promising results, achieving the fifth-best score in the main task with an accuracy of 96.67%.

## Keywords

Author Profiling, Irony and Stereotypes, Stance detection, Feature Engineering, Deep Learning, Transformers, Natural Language Processing

## 1. Introduction

With the proliferation of social media, irony has made it one of the most literary device utilized in this communication manner [1]. Several definitions have been provided in the literature about irony, but they concurred with the same binary classification, verbal and situational irony. The former has been conceived as the act of using words that mean the opposite of what you think, particularly to make funny [2, 3]. The latter has been defined as a strange or funny situation because things happen in a way that seems to be the opposite of what you expected [4, 5]. Both definitions highlight one of the primary features of this rhetorical device, to make something understandable by expressing the opposite [3]. Such rhetorical complexity

---

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

✉ joseantonio.garcia8@um.es (J. A. García-Díaz); miguel.rodriguez@urjc.es (M. Á. Rodríguez-García); frgarcia@um.es (F. García-Sánchez); valencia@um.es (R. Valencia-García)

>ID 0000-0002-3651-2660 (J. A. García-Díaz); 0000-0001-6244-653 (M. Á. Rodríguez-García); 0000-0003-2667-5359 (F. García-Sánchez); 0000-0003-2457-1791 (R. Valencia-García)



© 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)

makes dialogue that is occasionally arduous to comprehend by humans [6]. This challenge has attracted the research community's attention. In recent years, several approaches have been published addressing the detection of irony in natural language text obtained from different social media sources. In particular, we have focused on identifying whether its author spreads Irony and Stereotypes for the PAN shared challenge [7].

Irony and Stereotype identification is an essential task in social media applications since it enables the identification of online abuse and harassment [8]. The automatic detection in written discourse is a complex task where traditional text mining methods cannot be applied successfully [9]. This conventional method's drawback is that the identification requires semantics that cannot be inferred from word counts computed from document analysis [10]. To overcome this deficiency, more sophisticated Machine Learning methods are applied to solve the problem, but although the obtained results are quite competitive, there is still scope for improvement [11].

In this working note, we have faced the author profiling challenge proposed by constructing a supervised classification pipeline. The method comprises four stages: pre-processing stage to clean the provided dataset; the collecting features a stage, where contextual and non-contextual embeddings were utilized; training stage, where several machine models were used and finally, the evaluation stage, where we evaluated the designed models.

The remainder of this working note is organized as follows: Section 2 provides a brief review of the related work. It examines distinct approaches proposed in the literature that address the challenge thrown. Section 3 specifies the methods developed for addressing the challenge. In Section 4 the results achieved in the challenge are presented. Besides, we report separately our participation in a subtask concerning Stereotype Stance Detection in Section 5. Finally, Section 6 summarizes the findings obtained developing this work, and it also scrutinizes some of the future lines to explore.

## 2. Related work

Due to the complexity of recognizing verbal irony in a natural language test, we can find different approaches, from ones that utilize simple strategies to more complex ones. Barbieri and Saggion in [12] proposed two tree-based classifiers, Random Forest and Decision Tree. They represent each tweet by using the following seven groups of features: i) frequency to analyze the gap between rare and common words utilized by users; ii) written-spoken to capture the users' style; iii) intensity to measure the power of the adverbs and adjectives; iv) structure that analyzes the length, punctuation and emoticons; v) sentiments utilize SentiWordNet to measure the gap between positive and negative terms; vi) synonyms for comparing common vs rare synonyms utilized; and, finally, vii) ambiguity to analyze possible ambiguities. Furthermore, this approach explores the usage of a bag of word representation based on frequency analysis. Anchieta et al. in [13] proposed two differentiated strategies in a more complex way. Firstly, they combined Term Frequency, Inverse Frequency (TF-IDF), and Linear Support Vector Machine (SVM). The former was used to extract the features from the datasets, and the latter was the classifier utilized for the identification task. The classifier was trained by using the Stochastic Gradient Descent (SGD) technique. Secondly, they combine embeddings created by using Distributed Bag of Words Paragraph Vector model and a Multi-Layer Perceptron (MLP) for

tackling the classification task. With a different level of complexity, Wu et al. in [14] proposed the Dense-LSTM model based on a densely connected LSTM network with a multi-task learning strategy. It comprises an embedding layer to convert the inputs tweets into a sequence of dense vectors and four Bi-LSTM layers concatenated with 200-dim hidden states to learn different levels of information simultaneously. Furthermore, they combine two different pre-trained word embeddings that are concatenated and used.

### 3. Methodology

The IROSTEREO challenge consists of a binary classification from an author profiling perspective. The dataset proposed in this task is compiled from Twitter. The training dataset has a total of 420 different users. The users are grouped in those who are irony and stereotype spreaders (I) and those who are not (NI). For each user, there are 200 of their tweets written in English [15]. We separate a small subset from the training dataset to perform a custom validation. The statistics of the dataset are depicted in Table 1.

**Table 1**  
IROSTEREO dataset's users

	train	val	total
I	166	44	210
NI	176	34	210
<b>TOTAL</b>	<b>342</b>	<b>78</b>	<b>420</b>

We followed a typical pipeline of supervised classification for solving the proposed task. We started applying a pre-processing stage of the dataset. Then, we compile the feature sets, train several machine learning models, and evaluate them using a custom validation split.

The pre-processing stage consists in the creation of an alternative version of the documents by encoding them in lowercase, removing mentions, hyperlinks, digits, punctuation, and expressive lengthening. Besides, we expand texting language and fix misspellings. The alternative version is used to extract the majority of the feature sets based on sentence embeddings and linguistic features.

The feature sets involved in our experimentation consist into linguistic features from UMU-TextStats (LF) [16, 17], and three sentence embeddings: non-contextual sentence embeddings from FastText (SE) [18], and two contextual embeddings from BERT (BF) [19] and RoBERTa (RF) [20]. These feature sets were used separately and combined using two approaches. One is based on knowledge integration, and another is based in ensemble learning. For the ensemble learning, we evaluate four strategies: i) soft voting, ii) hard voting, iii) average probabilities, and iv) highest probability. Concerning the hard voting strategy, it is a weighted mode with the weights based on the F1-score results of the custom validation split.

As we deal with author analysis, the results are reported at author level. Nevertheless, some of the described stages of our pipeline are performed at document level. For example, the features are compiled at the document level and then combined by each user to produce a unique vector per user.

For extracting the contextual sentence embeddings from BERT and RoBERTa we fine-tune the models with the IROSTEREO dataset, and then we obtained the value of the [CLS] token [21]. In order to find the best hyperparameters, we trained ten models for BERT and 10 models for RoBERTa. The hyperparameters are i) the weight decay, ii) the batch size, iii) the warm-up speed, iv) the number of epochs, and v) the learning rate. This step is performed using Tree of Parzen Estimators (TPE) [22], which is a method for choosing the hyper-parameters based on Bayesian reasoning and expected improvement.

Next, we train several neural networks for each feature set and for the combination of all feature sets using a knowledge integration strategy. These hyperparameters include the shape of the network, the dropout mechanism, the learning rate and the activation function. Table 2 depicts the best hyperparameters for this task. It can be observed that the majority of best results are obtained with shallow neural networks, with two hidden layers but a large number of neurons. The only exception is SE, which achieved its best result with 7 hidden layers and 27 neurons in a long funnel shape. Besides, all experiments achieved better results with high dropout mechanisms and a learning rate of 0.010 using no activation function (linear). The exception again is SE, which uses a smaller learning rate, a smaller ratio of the dropout and elu as an activation function.

**Table 2**

Best hyper-parameters for each feature set trained separately and combined using knowledge integration.

Feature set	shape	hidden layers	neurons	dropout	lr	activation
LF	brick	2	128	.3	0.010	linear
SE	long funnel	7	27	.1	0.001	elu
BF	brick	2	512	.3	0.010	linear
RF	brick	2	512	.3	0.010	linear
KI	brick	2	512	.3	0.010	linear

## 4. Results and analysis

First, we report the results achieved with our custom validation split. These results include the label's precision, recall, and F1-score, and the macro and weighted average of the whole task. We report the results of the best feature set trained separately in Tables 3, 4, 5, and 6 for LF, SE, BF and RF respectively. The results for the KI strategy in Table 7, and the results for the four strategies using ensemble learning in Tables 8, 9, 10, and 11 for hard voting, soft voting, averaging probabilities and highest probability, respectively.

From the results achieved with the custom validation split, that are reported at the user level, we can assume that determining if a user is an irony and stereotype spreader is somehow a trivial task. It is worth mentioning that these results at the document level will be more limited. The best results are achieved with BERT from the feature sets separately. However, it draws our attention to the limited results achieved with RoBERTa (see Table 6). We observed that all the incorrect predictions are from the *I* label, but the model reports the *NI* label. We also compared the predictions between BF and RF and observed that the BF model outputs probabilities near

100% whereas RF is less accurate.

We can observe that the features based on pure linguistics also achieve similar results to the ones obtained with state-of-the-art embeddings. The LF features include features related to

**Table 3**  
Classification report for LF.

	precision	recall	f1-score
I	94.737	97.297	96.000
NI	97.826	95.745	96.774
macro avg	96.281	96.521	96.387
weighted avg	96.465	96.429	96.433

**Table 5**  
Classification report for BF.

	precision	recall	f1-score
I	97.297	97.297	97.297
NI	97.872	97.872	97.872
macro avg	97.585	97.585	97.585
weighted avg	97.619	97.619	97.619

**Table 7**  
Classification report for KI.

	precision	recall	f1-score
I	97.297	97.297	97.297
NI	97.872	97.872	97.872
macro avg	97.585	97.585	97.585
weighted avg	97.619	97.619	97.619

**Table 9**  
Classification report for EL (soft-voting).

	precision	recall	f1-score	support
I	97.297	97.297	97.297	
NI	97.872	97.872	97.872	
macro avg	97.585	97.585	97.585	
weighted avg	97.619	97.619	97.619	

**Table 11**  
Classification report for EL (highest probability).

	precision	recall	f1-score
I	44.048	100.000	61.157
NI	0.000	0.000	0.000
macro avg	22.024	50.000	30.579
weighted avg	19.402	44.048	26.938

**Table 4**  
Classification report for SE.

	precision	recall	f1-score
I	100.000	94.595	97.222
NI	95.918	100.000	97.917
macro avg	97.959	97.297	97.569
weighted avg	97.716	97.619	97.611

**Table 6**  
Classification report for RF.

	precision	recall	f1-score
I	64.286	48.649	55.385
NI	66.071	78.723	71.845
macro avg	65.179	63.686	63.615
weighted avg	65.285	65.476	64.594

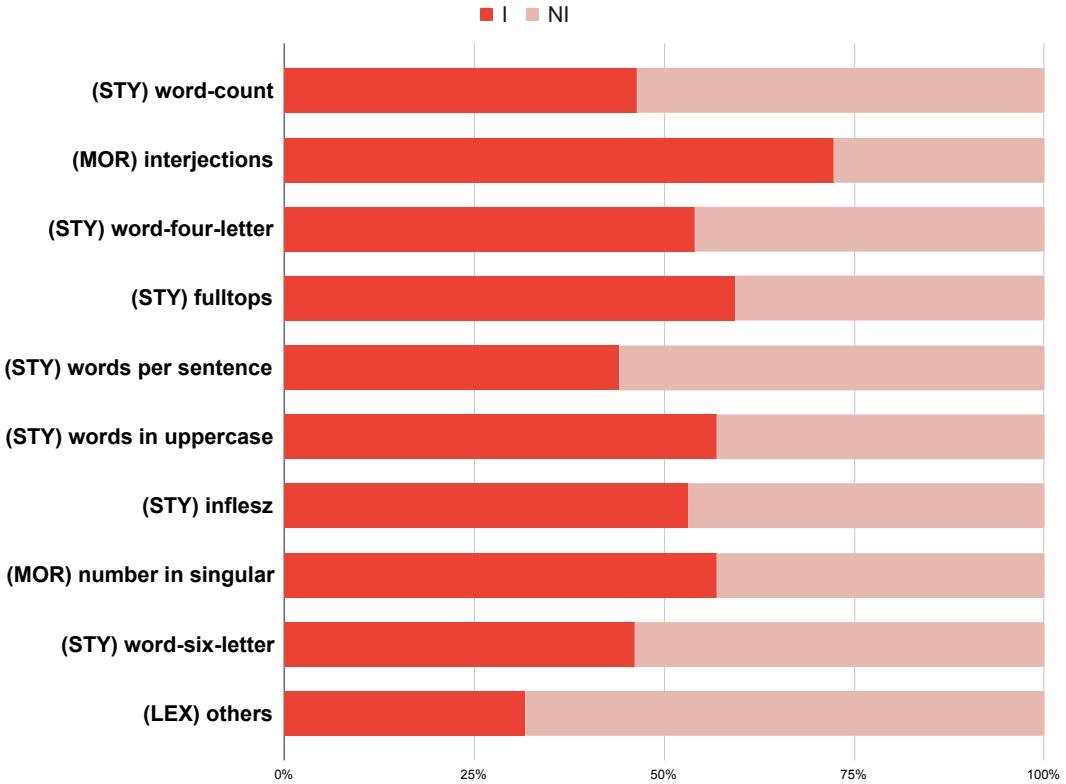
**Table 8**  
Classification report for EL (hard-voting).

	precision	recall	f1-score
I	97.297	97.297	97.297
NI	97.872	97.872	97.872
macro avg	97.585	97.585	97.585
weighted avg	97.619	97.619	97.619

**Table 10**  
Classification report for EL (avg. probabilities).

	precision	recall	f1-score
I	97.222	94.595	95.890
NI	95.833	97.872	96.842
macro avg	96.528	96.233	96.366
weighted avg	96.445	96.429	96.423

stylometry, lexis, social media jargon, and Part-of-Speech features. In order to gain insights concerning the interpretability of the features, we calculate the Information Gain of the linguistic features and we normalize the top-ten that achieved a better coefficient for the *I* and *NI* labels (see Figure 1). It can be observed that the majority of the most discerning features are related to stylometry, including the number of words, the number of words per sentence, the usage of full stops, and some readability formulas. There are two linguistic features concerning morphology: the usage of interjections and the usage of words in singular.



**Figure 1:** Information gain of the ten features with higher information gain

Because of the results achieved, for participating in this shared task, we sent one run based on the Knowledge Integration strategy, achieving the fifth best result an accuracy of 96.67% from a total of 65 participants. We selected the Knowledge Integration strategy over the two ensemble learning strategies that achieved the same results (hard and soft voting) because the Knowledge Integration have reported better results in other shared tasks in the past. Table 12 contains the best results along with the baselines proposed by the organizers. It is worth mentioning that these results were yielded from TIRA [23], an Integrated Research Architecture utilized by IROSTEREO organizers for managing the participants' algorithms executions.

## 5. Stereotype Stance Detection subtask

The organizers of the IROSTEREO shared task proposed a minor challenge that consisted in determining whether the stereotypes are used in favor of the target or against them. For this, they released a training dataset in which 94 authors were tagged *against* and 46 authors were tagged *in favor*.

To solve this challenge, we utilized the same pipeline described for the main challenge. Our results with our custom validation split are promising. We report the Knowledge Integration strategy and the four ensemble learning strategies in Table 13. We achieved a macro F1-score of 82.8753% with the Knowledge Integration strategy and a macro F1-score of 78.5714% with the ensemble learning based on soft-voting.

However, our results with the official leader board were limited. We achieved a macro F1-score of 53.12% (F1 with the *In Favor* label of 25% and F1 of the *Against* label of 81.25%).

## 6. Conclusions and future work

This working note describes the participation of the UMUTeam at IROSTEREO shared task concerning author profiling. This is a binary classification task in which the participants are

**Table 12**

Top results and baselines from the official leader board for the IROSTEREO 2022 shared task, ranked by accuracy

POS	Team	Accuracy
1	wentaoyu	0.9944
2	harshv	0.9778
3	edapal	0.9722
3	ikae	0.9722
5	<b>UMUTEAM</b>	<b>0.9667</b>
5	Enrub	0.9667
	LDSE	0.9389
	RF + char 2-ngrams	0.8610
	LR + word 1-ngrams	0.8490
	LSTM+Bert-encoding	0.6940

**Table 13**

Macro precision, recall and F1-score for the Stance detection subtask using the custom validation split. KI stands for Knowledge Integration and EL for Ensemble Learning

	precision	recall	f1-score
KI	93.478	78.571	82.875
EL - soft-voting	83.182	76.071	78.571
EL - hard-voting	91.667	71.429	75.455
EL - average probabilities	78.804	68.929	71.459
EL - highest probability	37.037	50.000	42.553

challenged to identify which profiles from Twitter are spreaders of Irony and Stereotypes. Our proposal is grounded on the combination of several feature sets based on linguistic features and sentence embeddings. We achieved promising results with our custom validation split and achieved a final accuracy of 96.67% on the official leader board.

One of the limitations of our work is the results achieved with RoBERTa (RF). Although we searched for common errors in our pipeline, we could identify the reason for the limited results. To address this issue, we suggest combining document level analysis with tools such as SHAP [24] in order to find the reason for the wrong predictions. Besides, we obtained wrong predictions with the highest probability strategy (see Table 11) as this ensemble outputs always the I label (100% of accuracy). We suspect this issue is related to an error in code while generating the final report.

As future work we will incorporate cross-validation techniques into our pipeline and data-augmentation techniques to increase our models' generalization.

## Acknowledgments

This work is part of the research project LaTe4PSP (PID2019-107652RB-I00) funded by MCIN/AEI/10.13039/501100011033. This work is also part of the research project PDC2021-121112-I00 funded by MCIN/AEI/10.13039/501100011033, by the European Union NextGenerationEU/PRTR, and by "Programa para la Recualificación del Sistema Universitario Español 2021-2023". In addition, José Antonio García-Díaz is supported by Banco Santander and the University of Murcia through the Doctorado Industrial programme.

## References

- [1] K. Buschmeier, P. Cimiano, R. Klinger, An impact analysis of features in a classification approach to irony detection in product reviews, in: Proceedings of the 5th workshop on computational approaches to subjectivity, sentiment and social media analysis, 2014, pp. 42–49.
- [2] J. A. García-Díaz, R. Valencia-García, Compilation and evaluation of the spanish saticorpus 2021 for satire identification using linguistic features and transformers, *Complex & Intelligent Systems* (2022) 1–14.
- [3] J. Garmendia, *Irony*, Cambridge University Press, 2018.
- [4] J. Hunter, *Evaluating the Circumstances*, John P. Hunter III, 2014. URL: <https://books.google.es/books?id=w7sYBAAAQBAJ>.
- [5] V. P. Maiorana, *Preparation for Critical Instruction: How to Explain Subject Matter While Teaching All Learners to Think, Read, and Write Critically*, Rowman & Littlefield, 2016.
- [6] N. Schwarz, *A Deep Learning Model for Detecting Sarcasm in Written Product Reviews*, Master's thesis, Interactive Media; FH Oberösterreich – Fakultät für Informatik, Kommunikation und Medien, 4232 Hagenberg, Austria, 2019.
- [7] J. Bevendorff, B. Chulvi, E. Fersini, A. Heini, M. Kestemont, K. Kredens, M. Mayerl, R. Ortega-Bueno, P. Pezik, M. Potthast, et al., Overview of pan 2022: Authorship ver-

- ification, profiling irony and stereotype spreaders, style change detection, and trigger detection, in: European Conference on Information Retrieval, Springer, 2022, pp. 331–338.
- [8] A. Chaudhary, S. A. Hayati, N. Otani, A. W. Black, What a sunny day: toward emoji sensitive irony detection, W-NUT 2019 (2019) 212.
  - [9] H. Taslioglu, P. Karagoz, Irony detection on microposts with limited set of features, in: Proceedings of the Symposium on Applied Computing, 2017, pp. 1076–1081.
  - [10] B. C. Wallace, Computational irony: A survey and new perspectives, Artificial intelligence review 43 (2015) 467–483.
  - [11] J. Sánchez-Junquera, P. Rosso, M. Montes, B. Chulvi, et al., Masking and bert-based models for stereotype identification, Procesamiento del Lenguaje Natural 67 (2021) 83–94.
  - [12] F. Barbieri, H. Saggion, Modelling irony in twitter, in: Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics, 2014, pp. 56–64.
  - [13] R. T. Anchieta, F. A. R. Neto, J. C. Marinho, K. V. do Nascimento, R. S. Moura, 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), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, September, 2021, volume 2943 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 917–924.
  - [14] C. Wu, F. Wu, S. Wu, J. Liu, Z. Yuan, Y. Huang, Thu\_ngn at semeval-2018 task 3: Tweet irony detection with densely connected lstm and multi-task learning, in: Proceedings of The 12th International Workshop on Semantic Evaluation, 2018, pp. 51–56.
  - [15] O.-B. Reynier, C. Berta, R. Francisco, R. Paolo, F. Elisabetta, Profiling Irony and Stereotype Spreaders on Twitter (IROSTERO) at PAN 2022, in: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2022.
  - [16] J. A. García-Díaz, R. Colomo-Palacios, R. Valencia-García, Psychographic traits identification based on political ideology: An author analysis study on spanish politicians' tweets posted in 2020, Future Generation Computer Systems 130 (2022) 59–74.
  - [17] J. A. García-Díaz, S. M. Jiménez-Zafra, M. A. García-Cumbreras, R. Valencia-García, Evaluating feature combination strategies for hate-speech detection in spanish using linguistic features and transformers, Complex & Intelligent Systems (2022) 1–22.
  - [18] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov, Learning word vectors for 157 languages, CoRR abs/1802.06893 (2018). URL: <http://arxiv.org/abs/1802.06893>. arXiv:1802.06893.
  - [19] J. D. M.-W. C. Kenton, L. K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, 2019, pp. 4171–4186.
  - [20] 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, CoRR abs/1907.11692 (2019). URL: <http://arxiv.org/abs/1907.11692>. arXiv:1907.11692.
  - [21] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: <https://arxiv.org/abs/1908.10084>.
  - [22] J. Bergstra, D. Yamins, D. Cox, Making a science of model search: Hyperparameter opti-

- mization in hundreds of dimensions for vision architectures, in: International conference on machine learning, PMLR, 2013, pp. 115–123.
- [23] 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.
- [24] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, *Advances in neural information processing systems* 30 (2017).