

# UMUTeam at IberLEF-2022 DETESTS task: Feature Engineering for the Identification and Categorization of Racial Stereotypes in Spanish

José Antonio García-Díaz<sup>1</sup>, Salud María Jiménez-Zafra<sup>2</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>Computer Science Department, SINAI, CEATIC, Universidad de Jaén, 23071, Spain

## Abstract

This paper describes the participation of the UMUTeam in the DETESTS shared task organized at IberLEF 2022 within the SEPLN conference. We have addressed the two proposed subtasks. The first one on determining the presence of racial stereotypes in a given sentence (binary task), and the second one on classifying the stereotypes in a given set of categories (multi-label task). The approach presented for both subtasks is based on the combination of linguistics features and Transformers using knowledge integration and ensemble learning strategies. In subtask 1, our team ranked in third position, out of 39 participants, with an F-score of 69.90, while in subtask 2 we placed in second position, out of 5 participants, with an ICM metric of -0.3298.

## Keywords

Natural Language Processing, Racial stereotypes identification and categorization, Feature engineering, Linguistic features, UMUTextStats, Negation processing, Transformers, Knowledge integration, Ensemble learning

## 1. Introduction

Stereotypes are an image or idea commonly accepted by a group or society as immutable [1]. They are a generalized preconception about qualities and features that are attributed to a group of people based on cultural, social and economic elements [2]. They reinforce toxic and hateful speech in social media, so their automatic detection through the development of Natural Language Processing systems is of growing interest.

Some tasks related to the identification of stereotypes about women and immigrants have already been organized, such as the Automatic Misogyny Identification task (AMI) [3, 4], the sEXism Identification in Social neTworks task (EXIST) [5, 6], the HatEval task [7], for the detection of hate speech against immigrants and women, or the HaSpeeDe 2 task [8], on identifying stereotypes about muslims, Roma and immigrants.

With the aim of mitigating hateful content towards immigrants, it has been organized the shared-task *DETESTS, DETECTION and classification of racial STereotypes in Spanish* [9], as part of


*IberLEF 2022, September 2022, A Coruña, Spain.*

✉ joseantonio.garcia8@um.es (J. A. García-Díaz); sjzafra@ujaen.es (S. M. Jiménez-Zafra); valencia@um.es (R. Valencia-García)

🆔 0000-0002-3651-2660 (J. A. García-Díaz); 0000-0003-3274-8825 (S. M. Jiménez-Zafra); 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)

the IberLEF 2022 workshop within the framework of the SEPLN 2022 conference. The organizers proposed two subtasks. The first one is a binary classification task to identify whether a text contains stereotypes or not. The second one is a multi-label classification task to categorize the stereotypes present in the text, if any.

This work presents the participation of the UMUTeam in both subtasks, which is based on the exploration of different strategies to combine Transformers and linguistic features, extracted from our UMUTextStats tool [10, 11] and from our negation processing system [12, 13]. Specifically, we study different combination mechanisms based on knowledge integration and ensemble learning. The rest of the paper is organized as follows. Section 2 presents the task and the dataset provided. Section 3 describes the methodology of our proposed system for addressing subtask 1 and subtask 2. Section 4 shows the results obtained and a discussion thereof. Finally, Section 5 concludes the paper with some findings and future directions.

## 2. Task description

The shared task DETESTS 2022, organized at IberLEF workshop, aims to detect and classify stereotypes related to immigration. Specifically, the organizers propose two tasks for dealing with stereotypes in comments posted in Spanish:

- Subtask 1: Determine if a given sentence contains at least one stereotype or none.
- Subtask 2: Detect whether a sentence contains stereotypes or not and, if it does, assign it to one or more of the following categories: 1) ‘victims of xenophobia’, 2) ‘suffering victims’, 3) ‘economic resources’, 4) a problem of ‘migration control’, 5) people with ‘cultural and religious differences’, 6) people which takes ‘benefits’ of our social policy, 7) a problem for ‘public health’, 8) a threat to ‘security’, 9) ‘dehumanization’ and 10) ‘other’ types of stereotypes.

The dataset provided is made up of comments from the NewsCom-TOX corpus [14] and the StereoCom corpus, consisting of Spanish comments published in response to articles from online newspapers and discussion forums. The statistics of the dataset, grouped by each subtask, are shown in Table 1. It can be observed that both subtasks are not balanced. In case of subtask 1, the proportion between stereotype and non stereotype is near to 1:3. For the second subtask, the ten traits are also not balanced, being the majority of traits focused on migration, security, and benefits. It is worth mentioning that the organizers provided only training and testing, so we select a custom split for validating. The custom validation split is created using stratified sampling, in order to keep the balance among the labels.

Finally, it is worth mentioning that, in order to evaluate the participants’ systems, the organizers used F1-score and cross-entropy for subtask 1, and hierarchical F-measure, propensity F-measure and ICM metric for subtask 2.

## 3. Methodology

The pipeline used for participating in both subtasks can be described as follows. First, the dataset was divided into training and validation. Second, a cleaned version of the dataset is created.

**Table 1**

Dataset distribution for subtask 1 and 2.

	train	val	test	total
Subtask 1				
non-stereotype	2363	583	-	2946
stereotype	691	180	-	871
Subtask 2				
benefits	161	45	-	206
culture	143	46	-	189
dehumanisation	53	12	-	65
economic	41	14	-	55
health	15	2	-	17
migration	257	64	-	321
others	52	15	-	67
security	210	45	-	255
suffering	51	12	-	63
xenophobia	10	6	-	16

Third, the feature sets are extracted, including linguistic features (LF), non-contextual sentence embeddings from FastText (SE), and contextual embeddings from BERT (BF), and RoBERTa (RF). Fourth, several neural networks are trained using the feature sets separately. Fifth, two strategies for combining the feature sets in a unique system are evaluated: knowledge integration and ensemble learning. Finally, the best runs are obtained using the custom validation split.

The cleaned version of the dataset is obtained by removing punctuation marks, hyperlinks, and emojis. In addition, some misspellings are fixed using the PSpell library [15], using a custom threshold. Besides, acronyms and abbreviations are transcribed by what they stand for and all texts are transformed into their lowercase form. It is worth noting that the cleaned version of the dataset is used to generate the sentence embeddings and to calculate the majority of the linguistic features except the ones based on correction and style.

The LF features are two fold. First, we obtain linguistic features from UMUTextStats [10, 11] and they are expanded with fine-grained negation [12, 13]. The 389 LF from UMUTextStats are organised in the following categories: (1) phonetics, (2) morphosyntax, (3) correction and style, (4) semantics, (5) pragmatics and figurative language [16], (6) stylometry, (7) lexis, (8) psycho linguistic processes, (9), and (10) social media jargon. The fine-grained negation features include simple cues (e.g., “no”/ *not*), continuous cues (e.g. “en mi vida”/ *in my life*) and discontinuous cues (e.g. “ni...ni”/ *neither...nor*).

The non-contextual sentence embeddings from FastText (SE) are obtained with the Spanish model [17]. The contextual sentence embeddings are based on BETO [18] and MarIA [19]. These contextual embeddings are obtained in basis of the value of the [CLS] token. For this, we apply an approach similar as the one described in [20]. Before this, we apply a fine-tuning approach and a hyper-parameter optimization stage using RayTune [21]. In order to maximize the performance of Transformers, we evaluate for both subtasks 10 models with Tree of Parzen

Estimators (TPE) [22], which is based on selecting the next hyper-parameter combination using Bayesian reasoning and the expected improvement. This stage involves the following hyperparameters: the (1) weight decay, (2) the batch size, (3) the warm-up speed, (4) the number of epochs, and (5) the learning rate.

Once the feature sets are obtained, we train several neural network models for each feature set. For this, we evaluate characteristics of the neural networks such as the shape, the number of hidden layers and the number of neurons per layer. The shape of the neural network is the number of neurons per layer. For example, in a brick shape, all hidden layers have the same number of neurons. In a long funnel shape, however, the number of neurons decreases as the depth of the network increases. We only evaluate brick shape for shallow neural networks (with one or two hidden layers) but several shapes are evaluated with networks that have between 3 and 8 hidden layers. The learning rate and dropout are also evaluated. Table 2 depicts the best hyperparameters for subtask 1 and 2 respectively. According to the characteristics of the best neural networks, we can observe that they are more simpler in case of subtask 2 but they require higher dropout ratio.

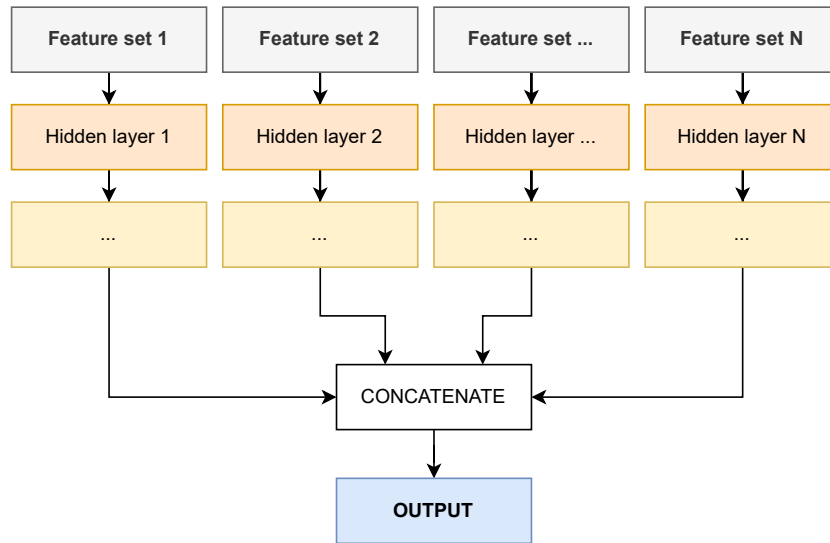
**Table 2**

Best hyper-parameters for subtasks 1 and 2 for each feature set trained separately and combined using knowledge integration.

Feature set	shape	hidden layers	neurons	dropout	lr	activation
Subtask 1						
lf	lfunnel	6	128	.1	0.010	sigmoid
se	brick	1	128	-	0.001	relu
bf	lfunnel	6	512	.3	0.001	selu
rf	brick	3	8	.1	0.010	sigmoid
Knowledge Integration	brick	2	48	.1	0.010	tanh
Subtask 2						
lf	brick	1	256	.3	0.010	relu
se	brick	2	64	.2	0.001	relu
bf	brick	2	37	.3	0.001	tanh
rf	brick	1	128	.3	0.010	relu
Knowledge Integration	brick	1	128	.3	0.010	relu

The final step of our pipeline is the integration of the features in one robust solution. We evaluate two major strategies for this: knowledge integration and ensemble learning. Knowledge integration consists of creating from scratch a neural network that have multiple inputs. Each feature set is fed an input and then connected to their own hidden layers. All these hidden layers are concatenated in new hidden layers and connected to the final layer for the prediction. Figure 1 has a diagram of the basic KI architecture. The ensemble learning strategy, on the other hand, consists of output of the labels based on the outputs of each neural network trained with each feature set separately. For this, we evaluate four strategies: (1) mode, that consists of selecting the label most repeated among the classifiers; (2) soft voting, that is a weighted mode in which the weights are based on the F1-score achieved with the validation split; (3)

averaging probabilities, that consists of averaging the probabilities obtained with each model; and (4) highest probability, that selects the label with highest probability.



**Figure 1:** Knowledge Integration architecture.

## 4. Results

This section describes the systems submitted by our team in each run. It is worth mentioning that each participating team could submit five runs. Moreover, it shows the results obtained in subtask 1 and subtask 2.

### 4.1. Subtask 1

We sent five runs for the first subtask. The results, and a brief description of each one, are depicted in Table 3. The first run, based on knowledge integration, achieved an F-score of 68.284. The second run, which consisted of ensemble learning with soft voting strategy, achieved our best result with an F-score of 69.903. The result achieved with the third run was similar, with an F-score of 69.656 based on ensemble learning averaging the probabilities of each model. The fourth run, consisted of ensemble learning with highest probability, was discarded due to a code error. Finally, the fifth run, consisted in a knowledge integration strategy but removing negation features, which achieved an F-score of 69.066.

The official leaderboard for subtask 1 is depicted in Table 4. We achieved the third position in the ranking with a F-score of 69.90%. Two teams from I2C outperformed our best run with a F-score of 70.42% and 70.05%, respectively. As commented above, we achieve this result using an ensemble learning strategy that combined all the feature sets using a soft voting strategy. The average result of all participants is a F-score of 52.955% with a standard deviation of 0.131. Besides, we outperform the four baselines proposed: a FastText Model using Support Vector

**Table 3**

Individual results for each run of subtask 1.

Run	Description	F-score
2	Ensemble learning (soft voting)	<b>69.903</b>
3	Ensemble learning (average probabilities)	69.659
5	Knowledge Integration (without negation)	69.066
1	Knowledge Integration	68.284
4	Ensemble learning (highest probability)	-

Machines, TF-IDF features using Support Vector Machines, a baseline based in predicting always the Stereotype class, and a random classifier.

**Table 4**

Official leaderboard for subtask 1.

#	Team Name	F-Score	#	Team Name	F-Score
	GoldStandard	1.0000	22	TheATeam	53.05
1	I2C_III	70.42	23	Francesco & Álvaro	52.42
2	I2C	70.05	24	TMNT	52.33
<b>3</b>	<b>UMUTeam</b>	<b>69.90</b>	25	B&C	52.16
4	I2C_II	66.89	26	DATTABAYÖ	52.02
5	Lak_NLP	66.27	27	pink-team	51.22
6	daminci	65.96	28	hectors	50.47
7	Elias-Urios-Alacreu	64.38	29	PPG	50.46
8	Salsa Version	63.87	30	IndecisionTrees	50.45
9	MALNIS	63.82	31	Humble_Team	49.29
10	JPG	63.48		FastText+SVC	48.61
11	Laura y Marta	62.31	32	Jorge Maté - Arturo Serrano	47.97
12	Alejandro & Raquel	62.16		TFIDF+SVC	47.06
13	tulseros	58.93		AllOnes	42.43
14	Roborecrea	58.59	33	Limi	40.57
15	Monty Python	57.23	34	UO	40.34
16	Izarcos	55.94	35	H3Lambda	39.26
17	Uwuntu	55.81	36	Gabriel Gros	26.28
18	Paloma_y_Karina	55.17	37	gfermuo	26.24
19	las_orchateras	54.23		RandomClassifier	22.95
20	Team Rocket	54.05	38	ArnauGarciaCuco_JoseVicenteGrauGil	19.69
21	Carles&Jorge	53.40	39	mcarmaf	18.16

## 4.2. Subtask 2

For the second subtask we also send five runs. The results for each run and a description of each one are shown in Table 5. In this case, the approach that provided the best results was the one based on knowledge integration using all of the features.

Table 6 depicts the official leaderboard for the second subtask. A total of five participants sent

**Table 5**

Individual results for each run of subtask 2.

Run	Description	ICM	Hierarchical F	Propensity F
1	Knowledge Integration	<b>-0.330</b>	<b>0.882</b>	<b>0.872</b>
5	Knowledge Integration without negation	-0.408	0.878	0.868
4	Ensemble learning (highest probability)	-0.429	0.861	0.848
2	Ensemble learning (soft voting)	-0.457	0.875	0.866
3	Ensemble learning (average probabilities)	-0.567	0.870	0.860

their results. We achieved the second position in the official leaderboard with our run based on knowledge integration. Besides, we outperformed the five baselines proposed. These baselines are the same that the ones used for subtask 1 but including a baseline that predicts all zeros.

**Table 6**

Official leaderboard for subtask 2.

#	Team Name	ICM	Hierarchical-F	Propensity-F
	GoldStandard	1.6676	1.0000	1.0000
1	MALNIS	-0.2380	0.8813	0.8717
<b>2</b>	<b>UMUTeam</b>	<b>-0.3298</b>	<b>0.8818</b>	<b>0.8718</b>
3	Elias-Urios-Alacreu	-0.3628	0.8668	0.8554
4	Lak_NLP	-0.4242	0.8606	0.8470
5	tulseros	-0.6433	0.8578	0.8468
	TFIDF+SVC	-0.6954	0.8552	0.8442
	AllZeros	-1.1280	0.8317	0.8215
	FastText+SVC	-1.1348	0.8314	0.8154
	RandomClassifier	-2.0403	0.7493	0.7308
	AllOnes	-36.3162	0.2224	0.1354

## 5. Conclusions

These working notes summarises the participation of the UMUTeam in the DETESTS shared task (IberLEF 2022). We participated in the two challenges proposed, achieving promising results in both. Specifically, we ranked the 3/39 in the subtask 1, a binary classification task in which we achieved an F-score of 69.90, and 2/5 in the subtask 2, a multi-label classification task in which we achieved an ICM metric of -0.3298. For both subtasks, we evaluate different feature combination strategies for improving the reliability of the models obtained with the feature sets evaluated separately. These feature sets are based on linguistic features, fine-grained negation features, and three different types of embeddings.

As future work we are planning to improve our pipeline, by implementing nested cross-validation instead of using stratified sampling, as we suspect that our best models are biased to our custom validation split. We will also explore the effect of highly skewed models towards certain labels. Besides, we are planning to use the linguistic and negation features (as they are

**Table 7**

Macro precision, recall, and f1-score for the second subtask with our custom validation split.

label	precision	recall	f1-score
benefits	78.378	64.444	70.732
culture	75.000	39.130	51.429
dehumanisation	0.000	0.000	0.000
economic	100.000	42.857	60.000
health	100.000	50.000	66.667
migration	78.947	46.875	58.824
others	80.000	26.667	40.000
security	68.421	28.889	40.625
suffering	20.000	8.333	11.765
xenophobia	100.000	16.667	28.571
micro avg	75.182	39.464	51.759
macro avg	70.075	32.386	42.861
weighted avg	71.834	39.464	49.913
samples avg	11.075	9.720	10.097

the ones that provide more interpretability) to analyse the differences among features based on Transformers. Finally, we will analyze why negation features are adding noise to the modelling of the first task according to the results obtained.

## Acknowledgments

This work was supported by Project LaTe4PSP (PID2019-107652RB-I00) funded by MCIN/AEI/10.13039/501100011033, Project AllnFunds (PDC2021-121112-I00) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, Project LIVING-LANG (RTI2018-094653-B-C21) funded by MCIN/AEI/10.13039/501100011033 and by ERDF A way of making Europe, and Big Hug project (P20\_00956, PAIDI 2020) and WeLee project (1380939, FEDER Andalucía 2014-2020) funded by the Andalusian Regional Government. In addition, José Antonio García-Díaz has been supported by Banco Santander and University of Murcia through the industrial doctorate programme, and Salud María Jiménez-Zafra has been partially supported by a grant from Fondo Social Europeo and Administración de la Junta de Andalucía (DOC\_01073).

## References

- [1] REAL ACADEMIA ESPAÑOLA: Diccionario de la lengua española, 23.<sup>a</sup> ed., [versión 23.5 en línea], 2022, URL: <https://dle.rae.es>.
- [2] B. G. Gavaldón, Los estereotipos como factor de socialización en el género, *Comunicar* (1999) 79–88.
- [3] E. Fersini, P. Rosso, M. Anzovino, Overview of the Task on Automatic Misogyny Identification at IberEval 2018., *IberEval@ SEPLN 2150 (2018)* 214–228.



- [4] E. Fersini, D. Nozza, P. Rosso, Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI), *Evalita Evaluation of NLP and Speech Tools for Italian 12* (2018) 59–66.
- [5] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, T. Donoso, Overview of EXIST 2021: Sexism Identification in Social Networks, *Procesamiento del Lenguaje Natural 67* (2021) 195–207.
- [6] F. Rodríguez-Sánchez, J. C. de Albornoz, L. Plaza, A. Mendieta-Aragón, G. Marco-Remón, M. Makeienko, M. Plaza, J. Gonzalo, D. Spina, P. Rosso, Overview of EXIST 2022: sEXism Identification in Social neTworks, *Procesamiento del Lenguaje Natural 69* (2022).
- [7] V. Basile, C. Bosco, E. Fersini, N. Debora, V. Patti, F. M. R. Pardo, P. Rosso, M. Sanguinetti, Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter, in: *13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019*, pp. 54–63.
- [8] S. Manuela, C. Gloria, E. Di Nuovo, S. Frenda, M. A. Stranisci, C. Bosco, C. Tommaso, V. Patti, R. Irene, et al., Haspeede 2@ EVALITA2020: Overview of the EVALITA 2020 Hate Speech Detection Task, *EVALITA 2020 Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian* (2020) 1–9.
- [9] A. Ariza, W. S. Schmeisser-Nieto, M. Nofre, M. Taulé, E. Amigó, B. Chulvi, P. Rosso, Overview of the DETESTS Task at IberLEF-2022: DETECTION and classification of racial STereotypes in Spanish, *Procesamiento del Lenguaje Natural 69* (2022).
- [10] 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.
- [11] 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.
- [12] S. M. Jiménez-Zafra, R. Morante, E. Blanco, M. T. Martín-Valdivia, L. A. Ureña-López, Detecting negation cues and scopes in Spanish, in: *Proceedings of The 12th Language Resources and Evaluation Conference, 2020*, pp. 6902–6911.
- [13] S. M. Jiménez-Zafra, M. Taulé, M. T. Martín-Valdivia, L. A. Ureña-López, M. A. Martí, SFU Review SP-NEG: a Spanish corpus annotated with negation for sentiment analysis. A typology of negation patterns, *Language Resources and Evaluation* 52 (2018) 533–569.
- [14] M. Taulé, A. Ariza, M. Nofre, E. Amigó, P. Rosso, Overview of DETOXIS at IberLEF 2021: DETECTION of TOXicity in comments In Spanish, *Procesamiento del lenguaje natural, 2021*, num. 67, p. 209-221 (2021).
- [15] [Online], 2019, PSPELL library, URL: <http://aspell.net/>.
- [16] M. del Pilar Salas-Zárate, G. Alor-Hernández, J. L. Sánchez-Cervantes, M. A. Paredes-Valverde, J. L. García-Alcaraz, R. Valencia-García, Review of English literature on figurative language applied to social networks, *Knowledge and Information Systems* 62 (2020) 2105–2137.
- [17] É. Grave, P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov, Learning Word Vectors for 157 Languages, in: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018, pp. 3483–3487.
- [18] J. Cañete, G. Chaperon, R. Fuentes, J. Pérez, Spanish pre-trained BERT model and evaluation

- data, PML4DC at ICLR 2020 (2020) 1–10.
- [19] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodríguez-Penagos, A. Gonzalez-Agirre, M. Villegas, MarIA: Spanish Language Models, *Procesamiento del Lenguaje Natural* 68 (2022) 39–60. URL: <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6405>.
  - [20] 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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3982–3992.
  - [21] R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, I. Stoica, Tune: A Research Platform for Distributed Model Selection and Training, *arXiv preprint arXiv:1807.05118* (2018).
  - [22] J. Bergstra, D. Yamins, D. Cox, Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures, in: *International conference on machine learning*, PMLR, 2013, pp. 115–123.

## A. Online Resources

The source code is available via [GitHub](#).