

UMUTeam at Huhu 2023: Detecting Prejudices in Humour Using Ensemble Learning and Knowledge Integration

José Antonio García-Díaz¹, Rafael Valencia-García¹

¹Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Spain

Abstract

These notes detail the participation of the UMUTeam in the Huhu shared task, organised in the IberLEF 2023 workshop. This shared task is concerning the usage of humour for masking prejudice towards minority groups. The organisers propose three challenges: the first one for detecting hurtful humour, which is a binary task; the second one for detecting the targets of the prejudices with a multi-label classification; and the last one is to calculate the intensity of the prejudice. Our team participate in all tasks evaluating several ensemble learning strategies to combine several sentence embeddings extracted from several Large Language Models. Although our results are competitive with our custom validation split, they do not outperform the baselines proposed by the organisers with the test split. After the analysis of the golden labels released by the organisers, we detect an error when incorporating the test split. Accordingly, we re-run our experiments and we obtain the real results by our team. We would have achieved the first position for Task 1 (macro F1-score of 85.5%), the second position in Task 2 (macro F1-score of 78.6%), and the third position in Task 3 (RMSE of 0.878).

Keywords

Humour Analysis, Feature Engineering, Transformers, Knowledge Integration, Ensemble learning, Natural Language Processing

1. Introduction

According to the Oxford dictionary, prejudice is an unreasonable dislike of or preference for a person or group, especially when this prejudice is based on some demographic trait such as their race or sex or psychographic traits such as religion [1]. Therefore, prejudice is related to stereotyping, as they are beliefs about the traits of some social group based on pre-judgements which emphasise negative aspects of others with the aim of presenting the other as different.

Social networks are a common mean of spreading prejudice. Although prejudice is banned from these social environments, often these messages are disguised as humorous messages to dismiss the moral judgement.


In this edition of IberLEF [2], the HUrTful HUmour, Detection of humour spreading prejudice in Twitter (Huhu) shared task is proposed [3]. The participants are challenged to solve three tasks. The first one, hurtful humour detection, is a binary classification task in which the

IberLEF 2023, September 2023, Jaen, Spain

✉ joseantonio.garcia8@um.es (J. A. García-Díaz); valencia@um.es (R. Valencia-García)

🆔 0000-0002-3651-2660 (J. A. García-Díaz); 0000-0003-2457-1791 (R. Valencia-García)

© 2023 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)

participants may identify prejudicial tweets which intent is to cause humour. The systems are ranked employing the F1-measure over the positive class. The second task, prejudice target detection, consists of a multi-label classification to determine which minority group is attacked. This task is measured using the macro-F1 score. The third task consists of measure the degree of prejudice prediction on a continuous scale (1 to 5) that indicates how prejudicial is the message is on average. This task is measured using the Root Mean Squared Error (RMSE).

It is worth mentioning that this is not the first shared task concerning offensive language in humour proposed in IberLEF. In 2021, our team participated in the HAHA shared task [4], in which we achieved the 1st position in the humour rating funniness score prediction and the 3rd position for target classification [5]. Another similar task but organised in SemEval was HaHackathon [6], in which our team also participated [7].

2. Dataset

The organisers of HUUH provided to the participants a dataset for all tasks. However, no details concerning how the dataset is collected or annotated were given to the participants. Please, refer to the overview for more information of the dataset [3]. The participants were limited to the usage of this data. That is, no external data was not allowed in HUUH 2023. We strict to this limitation, except in the case that we use pre-trained LLMs.

From the training dataset, we extract a custom validation split in a ratio of 80%-20% using stratified sampling to keep the ration between prejudicial and non-prejudicial documents.

After the competition, the organisers release the golden labels. Therefore, the full dataset statistics are shown in Table 1 for Task 1 and Table 1 for Task 2. We can observe that both Tasks are imbalanced with more tweets labelled as non-prejudicial in Task 1 and more tweets that expresses prejudices towards women than any other group in Task 2. Concerning the regression task, we observe that the average prejudice score is 2.971, with a standard deviation of 0.838 for the full dataset.

Table 1
Dataset statistics for the Task 1

label	train	val	test	total
non-prejudicial	1441	361	522	2324
prejudicial	695	174	256	1125
total	2136	535	778	3449

Table 2
Dataset statistics for the Task 2

label	train	val	test	total
gordofobia	166	48	55	269
inmigrant race	530	134	178	842
lgbtiq	477	130	250	857
woman	1048	244	688	1980
total	2221	556	1171	3948

3. Methodology

In a nutshell, our methodology follows the basic pipeline of a custom training of machine-learning classifier, in which we apply a basic data cleaning stage, a feature extraction stage, and

the training and validation of several classifiers. The most advanced of our work is the usage of ensemble learning and knowledge integration to combine the features from several LLMs and linguistic features. The stages of our pipeline are described below.

3.1. Data cleaning

During the Data Cleaning stage, we remove from the dataset hyperlinks and mentions as well as other jargon proper from social networks. We also replaced numbers with fixed tokens and remove expressive lengthening from words that use this technique for emphasis. However, as we extract some linguistic clues from the text, we keep the original version of the texts as some features, such as grammatical and orthographic errors require from this original input.

3.2. Feature extraction

For the feature extract stage, we extract Linguistic Features (LFs) using UMUTextStats [8] and sentence embeddings for several LLMs after fine-tuning.

There are 365 linguistic features organised in the following linguistic categories: (1) Phonetics, (2) Morphosyntax, (3) Correction and style, (4) Semantics, (5) Pragmatics, (6) Stylometry, (7) Lexis, (8) Psycho-linguistic processes, (9) Register, and (10) Social media.

The evaluated LLMs include a mix of Spanish and multilingual models: (1) BETO [9], (2) ALBETO [10], (3) BERTIN [11], (4) XLM [12], (5) DilstilBETO [10], (6) MarIA [13], (7) multilingual BERT [14], (8) multilingual DeBERTA [15], and (9) TwHIN [16].

3.3. Large Language Model Fine-tuning

As we want to combine different LLMs, it is more flexible a fixed representation for each text. For this, we fine-tuned all LLMs and then obtain the fixed value of the first classification token [17], extracting the [CLS] token. Therefore, each document is encoded as a unique vector of length 768.

To fine-tune the LLM for the three tasks, we perform an hyperparameter-tuning process. Due to time and memory constraints, we train each LLM under 10 configurations. The parameters evaluated are the learning rate, including strategies to adjust it, such as warm-up steps and the weight decay, the number of epochs, and two batch sizes.

The configuration of each LLM for each Task is showed in Table 3. We did not observe relevant differences in these models, except that BETO requires only one epoch for Task 1 and 2, but 5 (the maximum evaluated) for Task 3.

3.4. Hyperparameter optimisation

We use the classification token for the development of several neural networks using Keras and one special multi-input neural network to combine all LLMs and the LFs using Knowledge Integration. Accordingly, the result models are deep-neural networks that uses the fixed tokens of each document as input. We also train another network for the LFs, as baseline.

The result of this process is depicted in Table 4. Most of the resulting neural networks are very simple. This fact is not surprisingly, as the vectors are already fine-tuning for each label.

In fact, the learning rate for all models for Task 3 is the same as well as the shape of the neural networks for Task 2 and 3, resulting in shallow neural networks with one or two hidden layers. The unique exception for this is XLM in Task 2, with 5 hidden layers.

The last stage in our pipeline is the developing of ensembles as another mean of combine the strengths of the LLM and the LFs. Note that we evaluate different strategies for Task 3 and for Tasks 1 and 2, because Task 3 is a regression task. The evaluated strategies are the (1) mode of predictions, the (2) mean of the probabilities, the (3) highest probability, and (4) a weighted mean, based on the results of the custom validation split.

Table 3
Best hyperparameters for each LLM and Task

	learning rate	epochs	batch size	warm-up steeps	weight decay
Task 1					
albet0	1.9e-05	3	16	500	0.21
bertin	4.3e-05	4	8	250	0.0023
bet0	3.5e-05	1	16	0	0.16
distilbet0	3.3e-05	3	8	0	0.1
maria	2.9e-05	2	8	1000	0.26
mbert	3.6e-05	3	16	250	0.23
mdeberta	3.5e-05	4	16	0	0.12
twhin	1.5e-05	5	16	250	0.29
xlm	4.8e-05	3	16	0	0.022
Task 2					
albet0	3.4e-05	4	8	500	0.076
bertin	3e-05	2	8	0	0.26
bet0	4.7e-05	1	8	250	0.026
distilbet0	4.8e-05	4	16	0	0.24
maria	2.4e-05	3	16	0	0.15
mbert	2e-05	3	16	250	0.049
mdeberta	3.8e-05	4	16	0	0.24
twhin	3.8e-05	2	16	0	0.22
xlm	4.2e-05	5	16	0	0.0091
Task 3					
albet0	3.8e-05	3	8	500	0.099
bertin	3.8e-05	5	8	0	0.27
bet0	4.5e-05	5	8	500	0.19
distilbet0	3.3e-05	3	16	250	0.074
maria	4.4e-05	5	16	250	0.23
mbert	1.1e-05	3	8	500	0.12
mdeberta	3.1e-05	5	8	250	0.23
twhin	3.8e-05	4	8	1000	0.25
xlm	2.9e-05	3	8	0	0.28

Table 4

Results of the hyperparameter optimisation stage using Keras of the LFs (LF), each LLM and the multi-input neural network using Knowledge Integration (KI).

feature set	shape	layers	neurons	dropout	lr	batch size	activation
Task 1							
lf	lfunnel	6	128	0.1	0.01	128	sigmoid
albeto	brick	1	2	0.2	0.01	256	tanh
bertin	brick	4	48	False	0.01	256	tanh
beto	brick	2	128	0.3	0.01	128	sigmoid
distilbeto	brick	2	256	0.3	0.01	512	relu
maria	lfunnel	5	64	0.2	0.001	128	sigmoid
mbert	brick	2	8	0.1	0.01	256	tanh
mdeberta	diamond	3	2	0.3	0.001	512	selu
twhin	rhombus	8	16	False	0.001	512	tanh
xlm	3angle	5	512	0.1	0.01	128	tanh
ki	brick	2	128	False	0.01	256	sigmoid
Task 2							
lf	brick	1	256	False	0.001	256	linear
albeto	brick	1	256	False	0.01	512	tanh
bertin	brick	2	64	0.3	0.01	128	tanh
beto	brick	2	16	0.1	0.001	256	tanh
distilbeto	brick	1	128	False	0.001	512	sigmoid
maria	brick	2	48	0.2	0.01	256	relu
mbert	brick	2	16	0.2	0.001	128	sigmoid
mdeberta	brick	2	512	0.3	0.01	256	linear
twhin	brick	2	128	False	0.001	128	sigmoid
xlm	brick	5	256	0.1	0.001	256	selu
ki	brick	2	95	0.2	0.01	128	relu
Task 3							
lf	brick	2	128	0.1	0.001	64	sigmoid
albeto	brick	2	64	False	0.001	32	sigmoid
bertin	brick	2	4	False	0.001	64	sigmoid
beto	brick	1	512	0.1	0.001	64	sigmoid
distilbeto	brick	1	512	False	0.001	64	sigmoid
maria	brick	2	128	False	0.001	64	sigmoid
mbert	brick	2	256	False	0.001	64	sigmoid
mdeberta	brick	1	256	0.1	0.001	64	sigmoid
twhin	brick	1	512	0.1	0.001	64	sigmoid
xlm	brick	2	64	False	0.001	64	tanh
ki	brick	1	8	False	0.001	32	sigmoid

4. Results and discussion

In this section, we report our results with our custom validation split and the official leader boards for each task.

4.1. Custom validation

Table 5 contains the results for Task 1 (left) and 2 (right). In Task 1, the best result is achieved combining all features using ensemble learning with a weighted mode both for precision, recall, and F1-score, but all the models are very competitive, both in terms of precision and recall. For Task 2, the best results are also very competitive in all tasks. Besides, we check the classification reports of all models, to find if we were suffering overfitting, but we could not find any problem, thinking this task was somehow trivial.

Table 5

Results with custom validation for Task 1 and 2 using the macro-average F1-score

experiment	precision	recall	f1-score	experiment	precision	recall	f1-score
LF	77.817	78.451	78.110	LF	68.252	75.085	71.374
ALBETO	84.286	80.733	82.102	ALBETO	94.103	91.116	92.555
BERTIN	81.227	81.411	81.317	BERTIN	95.958	91.524	93.611
BETO	84.144	83.073	83.566	BETO	95.082	91.154	93.037
DISTILBETO	82.897	81.647	82.212	DISTILBETO	95.873	92.786	94.278
MARIA	84.262	81.754	82.795	MARIA	92.951	92.760	92.816
MBERT	81.450	79.933	80.599	MBERT	94.457	87.703	90.806
MDEBERTA	82.776	82.678	82.727	MDEBERTA	94.497	89.337	91.713
TWHIN	83.448	82.509	82.945	TWHIN	93.728	90.490	92.033
XLM	82.061	79.476	80.522	XLM	93.632	92.565	93.067
KI	85.011	82.318	83.428	KI	95.797	93.762	94.736
EL (HIGHEST)	82.199	81.082	81.591	EL (HIGHEST)	68.372	97.639	80.396
EL (MEAN)	84.204	81.318	82.483	EL (MEAN)	96.611	91.334	93.859
EL (MODE)	85.472	82.744	83.870	EL (MODE)	97.070	91.334	94.062
EL (WEIGHTED)	85.692	83.468	84.419	EL (WEIGHTED)	97.133	93.583	95.288

For Task 3, about prediction of prejudice, the results are reported in Task 6, showing regression metrics: Explained Variance (EV), Root Mean Squared Logarithmic Error (RMSLE), Pearson R (Pearson R), R Square (R2), Mean Average Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The best results are achieved also with the ensemble learning strategy that averages the predictions of each feature set, except for the RMSLE metric.

Concerning the LFs and its interpretability, we calculate the Information Gain for Tasks 1 and 2 with the Linguistic Features (see Figure 1). Concerning Task 1, we observe that linguistic features regarding the usage of hashtags and open questions are more related with documents containing prejudice. This finding suggests that prejudice can be expressed not with direct sentences but with questions. So, prejudice is expressed more indirectly in a way that seems less questionable. Concerning Task 2, we observed a strong presence of offensive speech in

tweets expressing gordofobia, demonyms towards immigrants, topics related to sex correlated with the collective , and lexis concerning family expressing prejudice towards women.

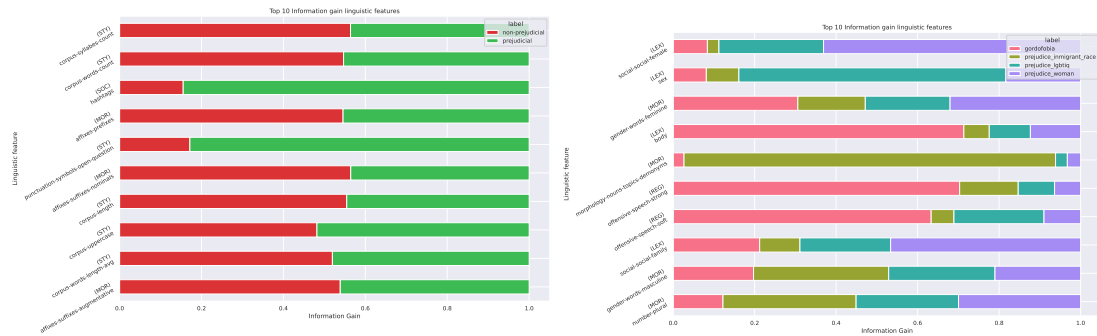


Figure 1: Information gain of linguistic features for Task 1 (left) and Task 2 (right)

4.2. Official results

In this section, we report our results with the official leader board in Table 7 for Task 1, Table 8 for Task 2, and Table 9 for Task 3. For each table, we include the top winning teams, the baselines, and the results of our two runs. These runs are based in the best results achieved during the custom validation split, that are the ensemble based on the weighted mode for Task 1 and the Knowledge Integration for Task 1; the ensemble based on averaging predictions and Knowledge Integration for Task 2; and the Knowledge Integration and the ensemble based on the weighted mode for Task 3.

As it can be observed, our results are limited. They do not outperform any of the baselines and they are far from the results achieved in the custom validation split. After reviewing our pipeline, we discover that we commit an human error when incorporating the test split to

Table 6

Results with custom validation for Task 3

experiment	EV	RMSLE	PEARSONR	R2	MAE	MSE	RMSE
LF	-0.004	0.044	0.319	-0.013	0.640	0.642	0.801
MBERT	-0.091	0.051	0.410	-0.117	0.654	0.708	0.842
DISTILBETO	0.151	0.040	0.505	0.122	0.572	0.557	0.746
XLM	0.050	0.044	0.445	0.009	0.618	0.629	0.793
ALBETO	-0.071	0.049	0.387	-0.076	0.657	0.682	0.826
MARIA	0.315	0.031	0.572	0.310	0.521	0.438	0.661
TWHIN	0.215	0.036	0.507	0.201	0.556	0.507	0.712
BETO	0.283	0.032	0.543	0.280	0.526	0.456	0.676
MDEBERTA	0.212	0.035	0.499	0.212	0.560	0.500	0.707
BERTIN	0.275	0.033	0.535	0.275	0.524	0.460	0.678
KI	0.323	0.031	0.570	0.323	0.511	0.429	0.655
EL (MEAN)	0.336	0.031	0.580	0.334	0.507	0.422	0.650

the files generated during custom validation, because we decided to shuffle the dataset as we observed that the provided dataset has the labels ordered and we want to prevent that the neural networks have the same labels in each batch. We re-run our experiments and we observed the real scores of our system. We decided to incorporate them to the Tables 7, 8, and 9.

Table 7

Official leader-board for Task 1. We include our official real F1-score and the F1-score without errors

position	team	run	Official F1-score	Real F1-score
1	RETUYT-INCO	1	0.820	-
2	BERT 4EVER	2	0.799	-
3	BERT 4EVER	1	0.798	-
BASELINE	BLOOM-1b1	-	0.789	-
BASELINE	BETO	-	0.759	-
BASELINE	SVM-3gram-char	-	0.679	-
BASELINE	AllTrue	-	0.492	-
47	UMUTEAM	2	0.448	0.852
48	UMUTEAM	1	0.443	0.855
54	JPK	1	0.273	-
58	AstonNLP	2	0.116	-

Table 8

Official leader-board for Task 2. We include our official real F1-score and the F1-score without errors

position	team	run	Official F1-score	Real F1-score
1	JUJUNLP	1	0.796	-
2	JOE	1	0.783	-
3	RATOLINS	1	0.778	-
BASELINE	BETO	-	0.760	-
BASELINE	SVM-3gram-char	-	0.603	-
BASELINE	AllTrue	-	0.482	-
38	UMUTEAM	2	0.427	0.786
41	UMUTEAM	1	0.413	0.769
49	cocalao	1	0.109	-

As it can be observed (see Table 7), our proposal for Task 1 would achieved the first position in the ranking with an F1-score of 0.855 using ensemble learning and 0.852 with ensemble learning using a weighted mode. In case of Task 2, we would achieve the second position in the ranking (see Table 8) with an F1-score of 0.786 using Knowledge Integration and our first run based on ensemble learning using weighted mode would have outperformed all the baselines proposed. Finally, concerning Task 3 (see Table 9), our second run based on Knowledge Integration would have achieved position 3 in the ranking but with a result below the baseline based on BETO.

Table 9

Official leader-board for Task 3. We include our official real F1-score and the F1-score without errors

position	team	run	Official RMSE	Real RMSE *
1	M&C	2	0.855	-
2	JOHuhuligans	2	0.875	-
BASELINE	BETO	-	0.874	-
3	MosquitosBiased	1	0.881	-
BASELINE	SVM-3gram-char	-	0.907	-
BASELINE	BLOOM-1b1	-	0.915	-
39	UMUTEAM	2	1.090	0.878
41	UMUTEAM	1	1.144	0.883
46	JPK	1	106.218	-
48	TeamVicente	1	-1	-

5. Conclusions

In this work we have described the participation of the UMUTeam in the HUH2023 shared task, concerning prejudice in humour. We have participated in all tasks, and we have achieved competitive results in all tasks using our custom validation split but very limited results in the official rankings. The cause of these limited results is because we commit a mistake when combining the dataset with the test labels to get the final predictions, and that the data ended up out of order. As the organisers of the task release the golden labels, we have re-run our experiments and include our real results for all tasks, achieving the first position for Task 1, the second position for Task 2, and third position for Task 3.

Acknowledgments

This work is part of the research projects AIInFunds (PDC2021-121112-I00) and LT-SWM (TED2021-131167B-I00) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. This work is also part of the research project LaTe4PSP (PID2019-107652RB-I00/AEI/ 10.13039/501100011033) funded by MCIN/AEI/10.13039/501100011033.

References

- [1] Oxford Learner’s Dictionary, Definition of prejudice noun from the oxford advanced learner’s dictionary, 2023. https://www.oxfordlearnersdictionaries.com/definition/english/prejudice_1 Last Accessed: 2023-06-21.
- [2] S. M. Jiménez-Zafra, F. Rangel, M. Montes-y Gómez, Overview of IberLEF 2023: Natural Language Processing Challenges for Spanish and other Iberian Languages, *Procesamiento del Lenguaje Natural* 71 (2023).
- [3] R. Labadie-Tamayo, B. Chulvi, P. Rosso, Everybody hurts, sometimes. overview of hurtful

- humour at iberlef 2023: Detection of humour spreading prejudice in twitter, in: *Procesamiento del Lenguaje Natural (SEPLN)*, volume 71, 2023.
- [4] L. Chiruzzo, S. Castro, S. Góngora, A. Rosá, J. Meaney, R. Mihalcea, Overview of haha at iberlef 2021: Detecting, rating and analyzing humor in spanish, *Procesamiento del Lenguaje Natural* 67 (2021) 257–268.
- [5] J. A. García-Díaz, R. Valencia-García, Umuteam at haha 2021: Linguistic features and transformers for analysing spanish humor. the what, the how, and to whom, in: *Proceedings of the Iberian Languages Evaluation Forum (Iber-LEF 2021)*, CEUR Workshop Proceedings, Málaga, Spain, volume 2943, 2021, pp. 829–836.
- [6] J. A. Meaney, S. Wilson, L. Chiruzzo, A. Lopez, W. Magdy, SemEval 2021 task 7: HaHackathon, detecting and rating humor and offense, in: *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, Association for Computational Linguistics, Online, 2021, pp. 105–119. URL: <https://aclanthology.org/2021.semeval-1.9>. doi:10.18653/v1/2021.semeval-1.9.
- [7] J. A. García-Díaz, R. Valencia-García, Umuteam at semeval-2021 task 7: Detecting and rating humor and offense with linguistic features and word embeddings, in: *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, 2021, pp. 1096–1101.
- [8] J. A. García-Díaz, P. J. Vivancos-Vicente, A. Almela, R. Valencia-García, Umutextstats: A linguistic feature extraction tool for spanish, in: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 6035–6044.
- [9] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: *PML4DC at ICLR 2020*, 2020, pp. 1–10.
- [10] J. Cañete, S. Donoso, F. Bravo-Marquez, A. Carvallo, V. Araujo, ALBETO and DistilBETO: Lightweight spanish language models, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, European Language Resources Association, 2022, pp. 4291–4298. URL: <https://aclanthology.org/2022.lrec-1.457>.
- [11] J. de la Rosa, E. G. Ponferrada, M. Romero, P. Villegas, P. González de Prado Salas, M. Grandury, BERTIN: efficient pre-training of a spanish language model using perplexity sampling, *Proces. del Leng. Natural* 68 (2022) 13–23. URL: <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6403>.
- [12] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: D. Jurafsky, J. Chai, N. Schluter, J. R. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Association for Computational Linguistics, 2020, pp. 8440–8451. URL: <https://doi.org/10.18653/v1/2020.acl-main.747>. doi:10.18653/v1/2020.acl-main.747.
- [13] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. R. Penagos, A. Gonzalez-Agirre, M. Villegas, MarIA: Spanish language models, *Proces. del Leng. Natural* 68 (2022) 39–60. URL: <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6405>.
- [14] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional

transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. URL: <https://doi.org/10.18653/v1/n19-1423>. doi:10.18653/v1/n19-1423.

- [15] P. He, J. Gao, W. Chen, DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing, CoRR abs/2111.09543 (2021). URL: <https://arxiv.org/abs/2111.09543>. arXiv:2111.09543.
- [16] X. Zhang, Y. Malkov, O. Florez, S. Park, B. McWilliams, J. Han, A. El-Kishky, TwHIN-BERT: A socially-enriched pre-trained language model for multilingual tweet representations, CoRR abs/2209.07562 (2022). URL: <https://doi.org/10.48550/arXiv.2209.07562>. doi:10.48550/arXiv.2209.07562. arXiv:2209.07562.
- [17] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), 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, Hong Kong, China, November 3-7, 2019, Association for Computational Linguistics, 2019, pp. 3980–3990. URL: <https://doi.org/10.18653/v1/D19-1410>. doi:10.18653/v1/D19-1410.