# UMUTeam at HOMO-MEX 2023: Fine-tuning Large Language Models integration for solving hate-speech detection in Mexican Spanish

José Antonio García-Díaz[1], Salud María Jiménez-Zafra[2] and Rafael Valencia-García[1]

[1]Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Spain
[2]Computer Science Department, SINAI, CEATIC, Universidad de Jaén, 23071, Spain

Abstract

This work describes the participation of the UMUTeam in the HOMO-MEX shared task at IberLEF 2023, on Hate speech detection in Online Messages directed tOwards the MEXican Spanish speaking LGBTQ+ population. We have addressed the two proposed tasks: Task 1, consisting of identifying the category of hate speech and, Task 2, on determining the types of phobia from a given set of tweets. For both tasks, we have evaluated different approaches based on the combination of sentence embeddings using ensemble learning and knowledge integration. Specifically, the sentence embeddings have been extracted from several Spanish and multilingual Large Language Models after fine-tuning them for each task separately. In total, 11 teams participated in Task 1 and 9 teams in Task 2. The best run sent by our team placed in position 3rd for Task1 and position 8th for Task 2 with an F1-score of 0.842 and a macro-average F1-score of 0.669, respectively, with 0.885 and 0.696 being the results obtained by the teams ranked in 1st position.

**Keywords**

Hate-speech Identification, Feature Engineering, Transformers, Knowledge Integration, Ensemble learning, Natural Language Processing



## 1. Introduction

The LGBT+ community is a vulnerable group that often suffers discrimination [1, 2]. LGBT+ phobia is the hatred or aversion towards people who belong to the LGBT+ community. We speak of LGBT+ phobia when someone uses vexatious expressions, behaves in an aggressive manner, rejects or isolates someone, or prevents someone from a procedure or access to a public service, in an intentional manner, because of the sexual orientation, gender identity or gender expression of that person.

Despite global progress against this form of discrimination, LGBT+ phobia is still a problem. In this context, it is organized the shared task *HOMO-MEX: Hate speech detection in Online Messages directed tOwards the MEXican Spanish speaking LGBTQ+ population* [3], as part of IberLEF 2023, a shared evaluation campaign for Natural Language Processing (NLP) systems in

*IberLEF 2023, September 2023, Jaén, 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)

Spanish and other Iberian languages. The competition was organized through CodaLab and can be accessed at the following link: https://codalab.lisn.upsaclay.fr/competitions/10019.

The objective of the HOMO-MEX shared task is to improve automatic detection systems designed for the classification of hate speech directed towards the LGBT+ community. Specifically, it is proposed two tasks:

- **Task 1: Hate speech detection.** It is a multi-class classification task and consists of, given a tweet, classifying it in one of the following categories:
  - **LGBT+ phobic**, if the tweet contains hate speech directed anyone whose sexual orientation and/or gender identity differs from cis-heterosexuality.
  - **not LGBT+ phobic**, if the tweet does not include hate speech towards the LGBT+ population, but mentions this community.
  - **not LGTB+ related**, if the tweet is not related to the LGBT+ community.
- **Task 2: Fine-grained hate speech detection.** It is a multi-label classification task and consists of, given a tweet that contains LGBT+ phobia, identifying one or more types of phobia present in it:
  - **Lesbophobia**: homophobia explicitly directed at homosexuals who identify as female.
  - **Gayphobia**: homophobia explicitly directed at homosexuals who identify as male.
  - **Biphobia**: hate speech directed against people who are attracted to more than one gender.
  - **Transphobia**: hate speech directed against non-cis-gendered people.
  - **Other LGBT+phobia**: hate speech against other sexual and gender minorities not included in any of the previous categories (e.g "aphobia": hatred received by people who do not feel sexual attraction).

Our team has participated in both tasks, in which we sent a total of 5 runs, based on the combination of sentence embeddings extracted from several Large Language Models (LLMs) after fine-tuning them for each task separately. These LLMs include Spanish models such as BETO [4], MarIA [5], AlBETO and DistilBETO [6], and multilingual models such as multilingual BERT [7], multilingual deBERTA [8], TwHIN [9], and XLM [10]. These features are combined using ensemble learning and knowledge integration. Specifically, the first run is based on knowledge integration that consists of training a multi-input neural network introducing all sentence embeddings at once. The second, third, fourth and fifth run are based on ensemble learning using different heuristic for combining the results. These heuristics are based on the mode of labels, in the highest probability of each class, on averaging the probabilities, and a weighted mode based on the results achieved with a custom validation split.

The rest of the paper is organized as follows. Section 2 presents the details of the dataset provided by the organizers to the participating teams. Subsequently, in Section 3, the methodology followed to carry out the experimentation is described. Next, Section 4 shows the results obtained during the validation and evaluation phases. In addition, a discussion of the results is presented. Finally, Section 5 concludes the paper with the main insights and future directions.

## 2. Dataset

The dataset of this shared task is composed of 12,416 tweets written in Mexican Spanish that have been extracted between 2012 and 2022, out of which 11,000 corresponds to Task 1 and 1,416 to Task 2. At a first stage, training was made available in order to the participants develop their systems. We select a subset of these tweets for custom validation in a ratio of 80-20. Later, test sets were released to participate in both tasks. The distribution of the datasets for *Task 1: Hate speech detection* and *Task 2: Fine-grained hate speech detection* are presented in Table 1 and Table 2, respectively. We can observe that, for Task 1, the majority of the tweets do not include hate speech towards the LGTB+ community, but mention it and, for Task 2, most of the tweets have content about *gayphobia*. Finally mention that the organizers decided not to make the test set public after the end of the competition, so it is not possible to provide statistics on the distribution of the test data, beyond the total of tweets, nor to analyze it.

**Table 1**
Dataset for Task 1: Hate speech detection

| label | train | val | test | total |
|---|---|---|---|---|
| LGBT+ phobic | 689 | 173 | ? | 862 |
| not LGBT+ phobic | 3488 | 872 | ? | 4360 |
| not LGTB+ related | 1422 | 356 | ? | 1778 |
| total | 5599 | 1401 | 4000 | 11000 |

**Table 2**
Dataset for Task 2: Fine-grained hate speech detection

| label | train | val | test | total |
|---|---|---|---|---|
| biphobia | 8 | 2 | ? | 10 |
| gayphobia | 571 | 143 | ? | 714 |
| lesbophobia | 56 | 16 | ? | 72 |
| others | 51 | 13 | ? | 64 |
| transphobia | 65 | 14 | ? | 79 |
| total | 751 | 188 | 477 | 1416 |

Next, we examined the corpus and its correlation with the labels using the UMUTextStats tool [11]. This tool is capable of extracting more than 350 linguistic features related to different linguistic categories such as register, morphosyntax, lexis or stylometric among others. We use these features to measure the information gain concerning the ground labels for Task 1 and Task 2 (see Figure 1). As expected, in Task 1, we found that features correlated with offensive speech are relevant but also the number of orthographic errors and lexis concerning sex. Similarly, in Task 2, lexis related to sex is also relevant but in this case it is the most significant feature, highly correlated with gayphobia and, in a minor degree, with biphobia and lesbophobia. Continuing with Task 2, lexis concerning female social groups is also very present in texts labelled as lesbophobia. In case of offensive speech, it is highly correlated with gayphobia.

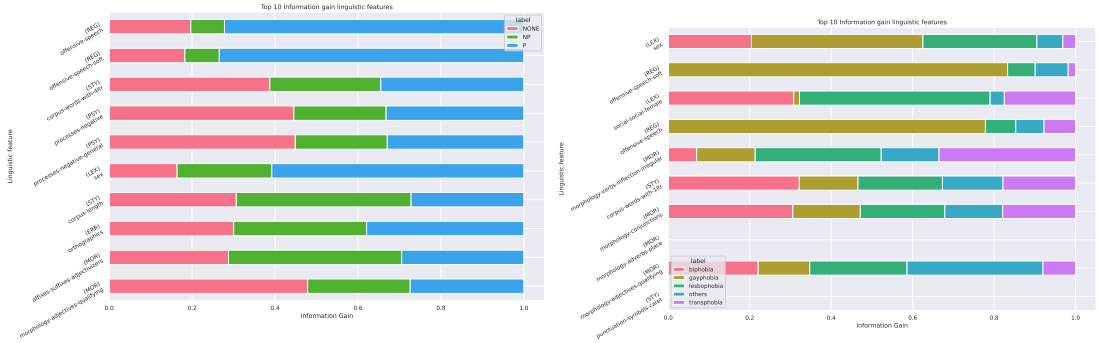


**Figure 1:** Information gain of linguistic features for Task 1 (left) and Task 2 (right) - P: LGBT+ phobic , NP: not LGBT+ phobic, None: not LGTB+ related

## 3. Methodology

In a nutshell, our methodology can be described as follows. First, we apply some basic data-cleaning to the dataset. Second, we fine-tuned each evaluated LLM separately before extracting their sentence embeddings. This fine-tuning process involve the training of 10 different models. Third, we evaluate several neural networks using these sentence embeddings together in a knowledge integration strategy with the objective of finding the best hyperparameters. Besides, we conduct extract hyperparameter optimization stages for each LLM separately to use their outputs in the ensemble learning strategies.

### 3.1. Data-cleaning

We conduct a basic data cleaning process to obtain a more generic model. We remove for the texts argot used in social networks, such as hyperlinks, hashtags, mentions and extra white spaces. We also expand some abbreviations typically used in social networks and the language used in short texts, expanded acronyms, and replace numbers with the token `[NUMBER]`.

### 3.2. Fine-tuning of the LLMs

Once the dataset is cleaned, we fine-tune several LLMs for both tasks separately. We evaluate 10 models per LLM evaluating the following hyperparameters: the learning rate (between 1e-5 and 5e-5 following a uniform distribution), the number of epochs (between 1 and 5), the batch size (8 or 16), the warm-up steps (0, 250, 500 or 1000) and the weight decay (between 0.0 and .3 following a uniform distribution). The models are selected using HyperOptSearch with Tree of Parzen Estimators (TPE) and the ASHA Scheduler with the objective of maximizing the macro weighted f1-score. All this process is conducted using RayTune.

Table 3 depicts the results achieved in this process for Tasks 1 (left) and Task 2 (right). It can be observed than LLMs for Task 1 have much more lower warm-up steps and weight decay except in the case of BERTIN that have the same number of warm-up steps and lower weight decay. Besides, both tasks have achieved better results with smaller batch size (8 vs 16). Finally,

we did not find relevant information concerning the number of training epochs nor the learning rate.

**Table 3**
Hyperparameter tuning of the LLMs for Task 1 and Task 2. The hyperparameters evaluated are the learning rate (lr), the training epochs (epochs), the batch size (bs), the warm-up steps (ws) and the weight decay (wd)

| | Task 1. Hate speech detection | | | | | Task 2. Fine-grained hate speech detection | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **LLM** | **lr** | **epochs** | **bs** | **ws** | **wd** | **lr** | **epochs** | **bs** | **ws** | **wd** |
| AlBETO | 2.2e-05 | 5 | 8 | 250 | 0.26 | 4e-05 | 4 | 16 | 0 | 0.046 |
| BERTIN | 2.7e-05 | 2 | 8 | 250 | 0.048 | 4.1e-05 | 4 | 8 | 250 | 0.21 |
| BETO | 3.9e-05 | 3 | 8 | 500 | 0.17 | 4.4e-05 | 3 | 8 | 0 | 0.096 |
| DistilBETO | 4.9e-05 | 5 | 8 | 1000 | 0.16 | 3.1e-05 | 5 | 8 | 250 | 0.012 |
| MarIA | 4.4e-05 | 3 | 16 | 500 | 0.18 | 4.3e-05 | 4 | 8 | 0 | 0.072 |
| mBERT | 3.1e-05 | 2 | 8 | 500 | 0.23 | 1.7e-05 | 3 | 8 | 250 | 0.11 |
| mDeBerta | 4.9e-05 | 5 | 16 | 500 | 0.28 | 2.3e-05 | 1 | 8 | 250 | 0.077 |
| TwHIN | 1.6e-05 | 5 | 8 | 1000 | 0.17 | 3.2e-05 | 5 | 8 | 500 | 0.27 |
| XLM | 3.9e-05 | 3 | 8 | 1000 | 0.023 | 4.4e-05 | 2 | 8 | 250 | 0.14 |

After this step, we extract the sentence embeddings for the best model of each LLM. It is worth mentioning that we extract the embeddings at sentence level because it allows us to combine them more easily in a new multi-input neural network taking profit of the strengths of each LLM. The sentence embeddings are obtained from the encoding of the classification token, as suggested in [12]. These embeddings are a fixed-length vector of 768.

### 3.3. Feature combination

Once the sentence embeddings are obtained, we evaluate to combine them using a knowledge integration strategy by feeding them in a multi-input neural network. The best configuration of this new neural network is also determined by a hyper optimization stage conducted in Keras. Now that the input are fixed sentence embeddings, we evaluate traditional neural network architectures, in which we assess the number of hidden layers and the number of neurons per layer, the learning rate, the batch size, the dropout mechanism for regularization and the activation function between layers. For Task 1, the best neural network consists of a deep neural network with 8 hidden layers and 16 neurons per layer stacked in a rhombus shape. The network uses no dropout and a learning rate of 0.01. The batch size is 512 and it uses *tanh* as activation function. For Task 2, however, the best results are achieved with a shallow neural network with 2 hidden layers but with 512 neurons per layer and no activation function between the hidden layers. The batch size is 64, the learning 0.01 and a strong dropout mechanism of 512.

## 4. Results and discussion

In this section we report and discuss the results obtained during the validation phase and the official results achieved in the evaluation phase for *Task 1: Hate speech detection* and *Task 2: Fine-grained hate speech detection.*

### 4.1. Results with custom validation

We tested different Spanish and multilingual LLMs. Specifically, different approaches based on sentence embeddings extracted from the LLMs were evaluated after fine-tuning them for each task separately. In addition, different sentence embeddings combination strategies were also evaluated by using knowledge integration and ensemble learning. The Spanish LLMs evaluated were BETO [4], MarIA [5], AlBETO and DistilBETO [6], and multilingual LLMs were BERT [7], MdeBERTA [8], TwHIN [9], and XLM [10]. On the other hand, the knowledge integration (KI) strategy consisted of training a multi-input neural network introducing all sentence embeddings at once and, the ensemble learning approaches tested were based on the highest probability of each class (EL (HIGHEST)), on averaging the probabilities (EL (MEAN)), on the mode of the labels (EL (MODE)), and a weighted mode (EL (WEIGHTED)).

Table 4 and Table 5 present the results obtained with the validation set for Task 1 and Task 2, respectively. As can be seen in Table 4, the best performing strategy in the hate speech detection task was knowledge integration and the best individual model was TWHIN. If we take a look at Table 5, for the fine-grained hate speech detection task, knowledge integration and TWHIN were also the best performing approach and best individual model, respectively.

**Table 4**
Results for Task 1 using the custom validation split

| model | precision | recall | f1-score |
|---|---|---|---|
| ALBETO | 78.447 | 81.132 | 79.673 |
| BERTIN | 77.173 | 80.673 | 78.714 |
| BETO | 82.414 | 81.352 | 81.867 |
| DISTILBETO | 77.978 | 80.496 | 79.133 |
| MARIA | 78.260 | 81.538 | 79.724 |
| MBERT | 78.620 | 79.278 | 78.942 |
| MDEBERTA | 82.223 | 78.517 | 80.176 |
| TWHIN | 81.789 | **83.669** | 82.679 |
| XLM | 77.181 | 79.746 | 78.351 |
| KI | **83.623** | 82.680 | **83.139** |
| EL (HIGHEST) | 39.669 | 48.665 | 15.192 |
| EL (MEAN) | 80.416 | 81.531 | 80.955 |
| EL (MODE) | 81.339 | 80.313 | 80.811 |
| EL (WEIGHTED) | 82.056 | 81.758 | 81.905 |

## 4.2. Official results

This subsection presents the results obtained in the evaluation phase. The organizers selected F1-score to rank the systems performance for Task 1 and they chose the macro-average F1-score for Task 2. Each team could submit a maximum of 5 runs, selecting the best one for ranking. We defined our 5 runs to evaluate the different feature integration strategies implemented. The results for each of the runs are depicted in Table 6 as well as the strategy followed in each of the run. The best result for Task 1 was obtained with the ensemble learning on a weighted mode. For Task 2, the best result was also reached with ensemble learning, but this time with the ensemble based on the mode of the labels. In general, it is observed that the combination strategies evaluated provide similar results, except for the approach of ensemble learning on the highest probability of each class in Task 2, where a notable difference is observed.

For the competition, we selected run 5 for Task 1 and run 2 for Task 2, as they were the ones that provided the best results.

Table 7 shows the official leader-board for Task 1, in which we achieved the 3rd position with a score of 84.21%. The results of our team are highlighted with a gray background.

**Table 5**
Results for Task 2 using the custom validation split

| model | precision | recall | f1-score |
|---|---|---|---|
| ALBETO | 74.897 | 57.449 | 63.176 |
| BERTIN | 42.504 | 63.804 | 49.630 |
| BETO | 85.525 | 66.231 | 72.956 |
| DISTILBETO | 78.490 | 68.490 | 71.176 |
| MARIA | 82.902 | 65.410 | 71.873 |
| MBERT | 67.625 | 53.649 | 58.541 |
| MDEBERTA | 31.694 | 50.288 | 37.780 |
| TWHIN | 86.645 | 66.808 | 74.152 |
| XLM | 51.961 | 29.863 | 33.475 |
| KI | 87.424 | 69.046 | **75.660** |
| EL (HIGHEST) | 33.435 | **93.956** | 45.391 |
| EL (MEAN) | 87.657 | 60.794 | 68.693 |
| EL (MODE) | 88.000 | 60.794 | 68.882 |
| EL (WEIGHTED) | **88.110** | 62.044 | 69.663 |

**Table 6**
Results for Task 1 and Task 2 per run

| run | Task 1 | Task 2 |
|---|---|---|
| 01. Knowledge Integration | 0.833 | 0.654 |
| 02. Ensemble learning (mode) | 0.839 | **0.669** |
| 03. Ensemble learning (highest probability) | 0.821 | 0.492 |
| 04. Ensemble learning (average probabilities) | 0.840 | 0.662 |
| 05. Ensemble learning (weighted mode) | **0.842** | 0.667 |

In Task 2, we achieved more limited results, as it can be observed in Table 8 reaching to position 8 in the ranking with a score of 66.87%. In this case, the results among all participants are more similar, achieving a average results of 67.69 with a standard deviation of 1.17.

## 5. Conclusions

In this working notes we have described our participation in the HOMO-MEX shared task concerning hate-speech identification and categorization in Mexican-Spanish. We are very proud with our participation as we achieved competitive results, reaching the third position in the first task concerning hate-speech identification. In the second task, however, we achieved only the 8th position, but our results are only about a 3% less than the winner. To participate in both tasks, we fine-tuned several Spanish and Multi-lingual LLMs, extracted their sentence embedddings and combined their strengths into a multi-input neural network Knowledge Integration fashion. Besides, we evaluated other integration techniques such as ensemble

**Table 7**
Official leader-board for Task 1

| rank | team | F1-score |
|---|---|---|
| 01 | bayesiano98 | 0.885 |
| 02 | carfer | 0.843 |
| 03 | UMUTeam | **0.842** |
| 04 | homomex | 0.839 |
| 05 | Cordyceps | 0.835 |
| 06 | I2CHuelva | 0.833 |
| 07 | UTB_NLP | 0.821 |
| 08 | INGEOTEC | 0.805 |
| 09 | Habesha | 0.797 |
| 10 | cesar_m | 0.764 |
| 11 | moeintash | 0.733 |

**Table 8**
Official leader-board for Task 2

| rank | team | macro-avg F1-score |
|---|---|---|
| 01 | I2CHuelva | 0.696 |
| 02 | carfer | 0.685 |
| 03 | ErikaRivadeneira | 0.683 |
| 04 | bayesiano98 | 0.681 |
| 05 | Cordyceps | 0.679 |
| 06 | Habesha | 0.673 |
| 07 | HomoMex | 0.670 |
| 08 | UMUTeam | **0.669** |
| 09 | cesar_m | 0.655 |

learning and analyzed the data set using linguistic features.

There is space for improvement in our proposal. First, we observed that most of the errors performed by our systems using a custom validation split are related to words that very tied to Mexican Spanish. In this sense, we need to analyze if these words are recognized in the LLMs and how their embeddings are similar to more generic Spanish words. Second, our research group have already evaluated different datasets in Spanish concerning hate-speech [13]. We will use these models to validate this dataset in order to understand differences between Spanish from Spain and Spanish from Mexico. Third, we did not evaluate data augmentation techniques for solving data-imbalanced nor the integration of the linguistic features in the ensemble or the knowledge integration model.

## 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. It also has been partially supported by Project CONSENSO (PID2021-122263OB-C21), Project MODERATES (TED2021-130145B-I00) and Project Social-Tox (PDC2022-133146-C21) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, Project PRECOM (SUBV-00016) funded by the Ministry of Consumer Affairs of the Spanish Government, Project FedDAP (PID2020-116118GA-I00) supported by MICINN/AEI/10.13039/501100011033 and WeLee project (1380939, FEDER Andalucía 2014-2020) funded by the Andalusian Regional Government. Salud María Jiménez-Zafra has been partially supported by a grant from Fondo Social Europeo and the Administration of the Junta de Andalucía (DOC_01073).

## References


[1] R. Hidalgo Sánchez, et al., El auge de los delitos de odio: la LGTBI-fobia en la actualidad, 2022.

[2] L. S. Casey, S. L. Reisner, M. G. Findling, R. J. Blendon, J. M. Benson, J. M. Sayde, C. Miller, Discrimination in the United States: Experiences of lesbian, gay, bisexual, transgender, and queer Americans, Health services research 54 (2019) 1454–1466.

[3] G. Bel-Enguix, H. Gómez-Adorno, G. Sierra, J. Vásquez, S.-T. Andersen, S.-L. Ojeda-Trueba, Overview of HOMO-MEX at Iberlef 2023: Paraphrase Detection in Spanish Shared Task, Procesamiento del Lenguaje Natural 71 (2023).

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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, arXiv preprint arXiv:2209.07562 (2022).

[10] 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.

[11] 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.

[12] 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.

[13] 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.