Exploring the Use of Different Linguistic Phenomena for Sexism Identification in Social Networks

Flor Miriam Plaza-del-Arco[†], María-Dolores Molina-González[†], Luis Alfonso Ureña-López[†] and María-Teresa Martín-Valdivia[†]

Department of Computer Science, Advanced Studies Center in ICT (CEATIC)
Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain

Abstract

This paper presents the participation of the SINAI-TL team in the shared task sEXism Identification in Social neTworks at IberLEF 2022 for both English and Spanish. Our goal is to observe which interactional phenomena of sexism expression can help in the detection of this content. To transfer this idea to an automatic Natural Language Processing system, we developed a multi-task learning approach involving different linguistic phenomena, such as emotions, sentiments, sarcasm, insults, constructiveness and targets. We compared the results of this approach with a state-of-the-art benchmark model based on the BERT architecture. Our team ranked fourth in subtask 1 among the 44 runs submitted by the participants, achieving an accuracy score of 78.45%.

Keywords

Multi-task learning, BERT, sexism identification, sentiment analysis, emotion analysis, sarcasm identification, target identification

1. Introduction

Sexism is any discrimination against people based on gender. Sexism against women is a widespread cultural component, whose basis is the superiority of men over women in different sectors of life, such as work, politics, society, family and even advertising.

This problem can be found in different areas such as everyday conversations, statements loaded with discriminatory ideology, contempt for the opinions expressed by women, and even embedded in common sayings and expressions. This discrimination against women in society is still very present in contemporary communication, both written and oral, and is increasingly prevalent on the Internet. Detecting sexism online can be difficult, as it can be expressed in many different ways, but it is necessary in order to design new equality policies, as well as to encourage better behavior in society.

IberLEF 2022, September 2022, A Coruña, Spain.

[†]These authors contributed equally.
fmplaza@ujaen.es (F. M. Plaza-del-Arco); mdmolina@ujaen.es (M. Molina-González); laurena@ujaen.es (L. A. Ureña-López); maite@ujaen.es (M. Martín-Valdivia)
0000-0002-3020-5512 (F. M. Plaza-del-Arco); 0000-0002-8348-7154 (M. Molina-González); 0000-0001-7540-4059 (L. A. Ureña-López); 0000-0002-2874-0401 (M. Martín-Valdivia)

In the Natural Language Processing field, few studies have addressed sexism detection, especially in languages other than English [1, 2, 3]. For this reason, it is important to encourage the NLP community to develop solutions to address this task through, for instance, a series of shared tasks such as the first edition of EXIST [4] and the second edition held this year [5].

In this paper, we present our participation as SINAI-TL team for the sEXism Identification in Social neTworks shared task [5] at IberLEF 2022. In particular, we focus on the first subtask which is a binary classification in which systems have to decide whether or not a given tweet or a Gab post contains sexist expressions or behaviors.

In order to accomplish the EXIST shared task, we aimed to explore whether some linguistic phenomena that might be involved in the expression of sexism could help to address this problem. For this aim, we developed a Multi-Task Learning system (MTL) that leverages affective knowledge (sentiment and emotion) and other linguistic phenomena including sarcasm, constructiveness, insult, and target to detect sexism, using a well-known Transformer-based model.

The rest of the paper is structured as follows. In Section 2 we describe the datasets used in our experiments. In Section 3, we present the proposed system for addressing the task. In Sections 4 and 5, we describe the experimental setup and results, respectively. Finally, the conclusion and future work is presented in Section 6.

2. Datasets

To run our experiments, we used the English and Spanish datasets provided by the organizers of the sEXism Identification in Social neTworks (EXIST) shared task [5] at IberLEF 2022.

The EXIST dataset includes any type of sexist expression or related phenomenon, including descriptive or allegation statements when the sexist message is a denunciation or a statement of sexist behavior. Popular expressions and terms, such as those used in previous approaches to the state of the art, both in English and Spanish, used to undervalue the role of women have been extracted from several Twitter accounts and analyzed and filtered by two gender experts, Trinidad Donoso and Miriam Comet [6]. The final set contains more than 200 expressions that can be used in sexist contexts. In this new edition of EXIST 2022 challenge, the EXIST 2021 dataset is provided as training data. The final EXIST dataset for this edition contains a total of 6,977 tweets for training and 3,386 tweets for testing. For the test set, about 1,058 Twitter tweets were collected and labeled following the procedure used in the EXIST 2021 dataset. Twitter monitoring started on January 1, 2022, and ended on January 31, 2022. This dataset also contains posts from Gab social platform, the organizers retrieved 492 gabs in English and 490 in Spanish. The labeling process was carried out by 6 subject matter experts with several years of experience in gender content analysis considering gender balance, 3 women and 3 men, to avoid gender bias in the labeling process.

In addition, as part of our participation, we used other corpora corresponding to tasks that could be related to sexism identification including polarity classification (TweetEval for English and InterTASS for Spanish), emotion classification (EmoEvent for and Universal Joy in both languages), offensive language identification in English (OLID), detection of toxicity in comments in Spanish (DETOXIS), Constructive Comments Corpus (C3) in English and the

corpus used in Automatic Sarcasm Detection subtask of 2nd FigLang Workshop at ACL 2020. These datasets are described below:

- **TweetEval Corpus.** TweetEval benchmark is the repository for EMNLP 2020 Findings, available at https://github.com/cardiffnlp/tweeteval. This benchmark consisted of seven heterogeneous tasks on Twitter, all framed as multi-class tweet classification. For our experiments in English, we selected the TweetEval dataset used for sentimental analysis in the Twitter subtask [7], whose data is annotated with one of the three following labels: positive, negative and neutral. The training set has 50,333 tweets composed of 19,902 positive, 7,840 negative and 22,591 neutral. The test set has 12,284 tweets composed of 2,375 positive, 3,972 negative and 5,937 neutral.
- **International TASS Corpus (InterTASS)** was released in 2017 [8] with Spanish tweets and updated in 2018 with texts written in three different variants of Spanish from Spain, Costa Rica and Peru [9]. In 2019, InterTASS was enlarged with new texts written in two new Spanish variants: Uruguayan and Mexican [10] and finally, it was completed with Chilean-Spanish Tweets in 2020 [11]. The corpus released in 2019 is the one used in this paper. Each tweet was annotated by at least three annotators with its level of polarity, which could be labeled as positive, negative, neutral and none.
- **EmoEvent** [12] is a multilingual emotion dataset based on events that took place in April 2019. It focuses on tweets in the areas of entertainment, catastrophes, politics, global commemoration and global strikes. For the creation of the corpus, the authors collected Spanish and English tweets from the Twitter platform. Then, each tweet was labeled with one of seven emotions, the six Ekman's basic emotions plus the "neutral or other emotions" label. Focusing on the Spanish language, a total of 8,409 were labeled by three Amazon Mechanical Turkers.
- **Universal Joy.** This dataset was published by Lamprinidis et al. [13] and is composed of over 530k anonymized public Facebook posts in 18 languages, labeled with five basic emotions (anger, anticipation, fear, joy, and sadness). The dataset is a reorganized and cleaned subset of a previously described one that was collected in October 2014 [14]. The authors eliminated all duplicates and classified the language of each instance using three types of methods, and only retained instances where at least two of these methods matched. They manually evaluated 200 randomly selected instances labelled with the code *deu*, *fra*, *eng*, *ita* and *spa*, corresponding to the English, French, German, Italian and Spanish languages, respectively, and found that the average accuracy of their method is 0.97 (±0.04). For our experiments, we used the Spanish and English subsets labeled with emotions.
- **OLID** [15]. The Offensive dataset used in this paper was provided by the organizers of SemEval 2019 Task 6 on Identifying and Categorizing Offensive Language in Social Media. The task was based on a new dataset, the Offensive Language Identification Dataset (OLID). OLID is a large collection of English tweets annotated using a hierarchical three-layer annotation model. It contains 14,100 annotated tweets divided into a training partition of 13,240 tweets and a testing partition of 860 tweets. In OLID was proposed a novel three-level hierarchical annotation schema that encompasses the following three general categories, Offensive Language Detection, Categorization of Offensive Language

and Offensive Language Target Identification. We used this last label about offensive language target identification for our experiments.

- **DETOXIS.** [16] The DETOXIS dataset was compiled from the NewsCom-TOX dataset. This dataset consists of 4,357 comments (approximately) posted in response to different articles extracted from Spanish online newspapers and discussion forums from August 2017 to July 2020. These articles were manually chosen based on their controversial subject matter, their potential toxicity, and the number of comments posted. A keyword-based approach was used to search for articles mainly dealing with immigration. Comments were retrieved in the same order in which they appear on the web timeline. Each comment was labeled into two categories *toxic* and *non-toxic*. In addition, the following characteristics were also annotated: argumentativeness, constructiveness, stance, objective stereotyping, sarcasm, mockery, insult, improper language, aggressiveness and intolerance. All of these categories have a binary classification scheme, except for the level of toxicity.
- **Constructive Comments Corpus (C3)** [17]. The Constructive Comments Corpus (C3) is a subset of comments from the SFU's Opinion and Comments Corpus. This subset is composed of 12,000 news comments in English annotated by crowdworkers as to their constructiveness and characteristics. Among all 12,000 instances, 89.7% instances had a clear consensus among annotators. The corpus is slightly higher in constructive comments (6,516) than in non-constructive comments (5,484).
- **Sarcasm dataset** is available at https://github.com/EducationalTestingService/sarcasm. This corpus consists of 4400 posts extracted from Reddit, and 5,000 tweets for the training set equally balanced between sarcastic and non-sarcastic in both groups, 1,800 posts of Reddit, and 1,800 tweets for the test set equally balanced between sarcastic and non-sarcastic posts. The corpus was used for the shared sarcasm detection task carried out as part of the 2nd Figurative Language Processing Workshop (FigLang 2020) at ACL 2020 [18]. For our experiments, we only used the subsets extracted from Twitter.

3. System overview

In this section, we describe the computational models our SINAI_TL team developed for the sEXism Identification in Social neTworks shared task at IberLEF 2022.

We aim to observe which interactional phenomena of sexism expression can help in the detection of this content. For this aim, we focus on analyzing different linguistic phenomena including sentiments, emotions, sarcasm, irony, the target, insults, and constructiveness. We hypothesize that these related phenomena could help in the detection of sexism. For instance, the expression of sexism could involve negative sentiments and emotions. At the same time, rhetorical figures such as irony and sarcasm are used to mask a hurtful message or to make fun in terms of gender. Most of the time the sexist comment is directed at a person or group of people, therefore, the target to whom it is directed plays an important role in the message. A common element that is often present in the expression of sexism is the use of insults or swear words. Finally, constructive criticism is a respectful judgment of another person to provide help or a positive view of a specific circumstance. This phenomenon occurs in non-sexist messages and can be an indicator to detect these messages.

In order to include this hypothesis in a computational architecture, we propose a model based on a MTL model. This type of system is able to learn multiple tasks simultaneously instead of learning them separately [19]. It generally helps to improve the performance on each task by sharing representations across related tasks. In order to develop this system, we rely on the transformer BERT model [20]. We add as many heads as related tasks to the encoder, one for each task. Then, the layers are fine-tuned according to the given set of downstream tasks. Depending on the task, different classes are taken into account (binary/multiclass classification tasks). During the training process, the objective function weights every task equally. When predicting, for each instance in the EXIST dataset, we assign as many predictions as tasks we are training but the prediction considered for the evaluation is the output of the sexism identification head (sexist, non-sexist).

4. Experimental setup

4.1. Dataset preprocessing

The dataset provided by the organizers contains posts from Twitter and Gab social media platforms. Since the language register used in these platforms is colloquial, we decided to perform some data cleaning steps before including the texts in the model:

- URLs and users' mentions are replaced by the tokens URL and USER, respectively.
- Hashtags are unpacked and split to their constituent words.
- Elongated words and repeated characters in words are reduced.
- Emojis are converted to their alias.

4.2. System settings

All the models have been implemented using PyTorch and HuggingFace libraries [21, 22]. The proposed models have been fine-tuned on a single Tesla-V100 for 2 epochs, with a learning rate of 2e-5 and batch size of 16, the optimization algorithm is Adamw.

As the EXIST dataset is composed of both English and Spanish texts, we split the EXIST dataset into two subsets (EXIST_en) and (EXIST_es). While training the MTL system, we consider each subset separately, thus we develop two different models: one for Spanish and another for English. Regarding the transformer, for EXIST_en subset, we used the BERT model trained on English tweets [23] and for the EXIST_es subset, we opt for BETO [24] [25], a model trained on Spanish texts.

5. Results

During the pre-evaluation phase, we train the model on the training set and then evaluate it on the test set provided by the organizers. For the evaluation phase, we train the model on the training and validation sets, then we evaluate it on the test set.

Table 1
Datasets used for each phenomenon in both English (EN) and Spanish (ES) EXIST subsets.

Phenomenon	EN	ES
Sentiments	TweetEval	InterTASS
Emotions	EmoEvent (EN) & Universal Joy	EmoEvent (ES) & Universal Joy
Sarcasm	Twitter sarcasm	DETOXIS
Target	OLID	DETOXIS
Insults	OLID	DETOXIS
Constructiveness	C3	DETOXIS

Table 2
MTL results for sexist detection on EXIST 2022 dev set (EXIST_es subset). Results in bold show the models that outperform the baseline in terms of F_1 score.

		Macro Average			Class sexist		
Model		P	R	F_1	P	R	F_1
Baseline	BETO	0.7886	0.7889	0.7880	0.8158	0.7649	0.7895
MTL	EXIST_emotion	0.8105	0.8109	**0.8101**	0.8341	0.7925	**0.8128**
	EXIST_sentiment	0.8091	0.8091	**0.8079**	0.8410	0.7774	**0.8080**
	EXIST_sarcasm	0.7994	0.7992	**0.7977**	0.8343	0.7622	**0.7966**
	EXIST_insult	0.7954	0.7956	**0.7944**	0.8249	0.7676	**0.7952**
	EXIST_constructiveness	0.7936	0.7932	**0.7917**	0.8296	0.7542	**0.7901**
	EXIST_target_person	0.7854	0.7846	0.7828	0.8238	0.7409	0.7801

5.1. Pre-evaluation phase

In this phase, we analyze different models and choose the best in terms of performance for the final submission. For this aim, we train our systems on the training set of EXIST 2022 and evaluating them on the validation set. As our hypothesis is that the MTL system trained on related linguistic phenomena to sexism identification helps in the detection of this problem, we decided to compare our results by establish the baseline BERT fine-tuning on the EXIST 2022 corpora. For both English and Spanish subsets, the related tasks (phenomena) we have considered along with the corpora used to train the MTL (see Section 2) are described in Table 1.

The results obtained after fine-tuning the MTL model on each of the related tasks together with the sexism identification task are shown in Tables 2 and 3 in Spanish and English subsets, respectively. These results are reported on the main task of sexism identification. We use the official competition metric macro averaged precision (P), recall (R) and F_1-score as evaluation metrics and further report sexism-specific performance.

For the results on the EXIST_es subset (Table 2) we can observe that all the MTL models except EXIST_target_person surpass the baseline BETO in terms of Macro-F_1 and sexist-F_1 scores. In particular, the setting EXIST_emotion achieved the best performance, followed by EXIST_sentiment and EXIST_sarcasm. The performance of EXIST_emotion increases by 2.21 points Macro-F_1 over the baseline, with macro-P increasing roughly 2.19 points and macro-R

Table 3
MTL results for sexist detection on EXIST 2022 dev set (EXIST_en subset).

Model		Macro Average			Class sexist		
		P	R	F_1	P	R	F_1
Baseline	BERT	0.7964	0.7858	**0.7867**	0.7629	0.8696	0.8128
MTL	EXIST_emotion	0.7928	0.7844	0.7853	0.7658	0.8584	0.8094
	EXIST_sentiment	0.7601	0.7581	0.7586	0.7586	0.7953	0.7766
	EXIST_sarcasm	0.7900	0.7823	0.7832	0.7653	0.8532	0.8069
	EXIST_insult	0.2378	0.5000	0.3223	0	0	0
	EXIST_constructiveness	0.7264	0.7269	0.7259	0.7539	0.7090	0.7308
	EXIST_target	0.7892	0.7851	0.7859	0.7767	0.8351	0.8048

2.02 points. It should be noted that the best model achieved a significant increases by 2.76 points in terms of sexist-recall score.

Regarding the evaluation of the MTL models in the EXIST_en subset, we observe that they do behave differently from the Spanish subset. As can be seen, the baseline BETO is not outperformed by any MTL model. The settings EXIST_emotion and EXIST_target achieve similar performance to the baseline. However, the EXIST_insult and EXIST_constructiveness performance drops considerably.

After comparing the performance of the MTL models on both subsets we can observe that depending on the language considered, the related tasks (phenomena) that help the detection of sexism are different. Therefore, there are two important parameters that have to been carefully analyzed when designing an MTL model for this purpose: the selected datasets and the language.

5.2. Evaluation phase

In this section, we present the results obtained by the different runs we have explored in subtask 1 (sexism identification).

As part of our participation, we present three runs based on the systems reporting the best performance explored during the pre-evaluation phase. Specifically, we chose the three best models for each language (see Tables 2 and 3) and combined them from best to worst performance. The models selected for each language and the differences between the three configurations we presented are described in the following:

- **Run 1**. Baseline (EN) + EXIST_emotion (ES).
- **Run 2**. EXIST_target (EN) + EXIST_sentiment (ES).
- **Run 3**. EXIST_emotion (EN) + EXIST_sarcasm (ES).

In Table 4 can be seen the official results obtained by our SINAI-TL in the different runs for both Spanish and English. With respect to the Spanish language, the three models present an accuracy score similar, with the second sentiment-based MTL model being slightly higher. For the English language, the first model selected for run 1 (Baseline) continues to achieve the best performance. However, the second run that considers the target is significantly lower in

performance compared to the one achieved during the pre-evaluation phase. It can also be seen that the run that considered emotion detection does not improve the baseline.

Table 4
Results in Subtask 1 on the Spanish and English test set of EXIST shared task.

Test set	Run	Acc	Precision	Recall	F-measure
ES	1	0,7500	0,7529	0,7509	0,7497
	2	0,7538	0,7559	0,7546	0,7536
	3	0,7500	0,7529	0,7509	0,7497
EN	1	0,8194	0,8148	0,8206	0,8166
	2	0,6312	0,6180	0,6028	0,6013
	3	0,8194	0,8143	0,8181	0,8158

Finally, Table 5 shows the results obtained by some participants in subtask 1. As we can see, our participation ranks fourth among the participants, with the best run resulting from the combination of the baseline model for English and the EXIST_emotion model for Spanish, followed by runs 3 and 2. Therefore, we consider that the MTL model is a successful system that for the Spanish language provides an improvement over the state-of-the-art BETO. Concerning the English language, we observe that is a challenge to improve the baseline since it already shows values above 80% in the different metrics.

Table 5
Ranking of participants' systems in subtask 1 of EXIST shared task.

Ranking	Team	Acc	F1
1	avacaondata_1	0.7996	0.7978
2	CIMATCOLMEX_1	0.7949	0.7940
3	I2C_1	0.7883	0.7880
4	**SINAI_TL_1**	**0.7845**	**0.7841**
5	**SINAI_TL_3**	**0.7845**	**0.7839**
40	**SINAI_TL_2**	**0.6928**	**0.6882**
44	Majority Class	0.5444	0.3525

6. Conclusions

In this paper, we have presented our participation as SINAI-TL team in the second edition of the task sEXism Identification in Social neTworks at IberLEF 2022. We have explored whether different linguistic phenomena that might be related to sexist expression could help to detect this problem. The experiments have been conducted in two languages: English and Spanish. Our results show that there are two important factors to consider while addressing this task in both languages: the linguistic phenomena considered and the datasets selected. For Spanish, we found that taking into account emotions, sentiments, and sarcasm knowledge helps the detection of sexism. For English, the phenomena studied have not shown any improvement

over the baseline BERT. We consider that this fact could be related mainly to the datasets chosen. Therefore, in future work, we plan to analyze what are the characteristics that should be in line between the datasets considered to study the linguistic phenomena and the sexism dataset, for instance, the source of the text, the number of categories, the number of comments, among others.

7. Acknowledgments

This work has been partially supported by Big Hug project (P20_00956, PAIDI 2020) and WeLee project (1380939, FEDER Andalucía 2014-2020) funded by the Andalusian Regional Government, LIVING-LANG project (RTI2018-094653-B-C21) funded by MCIN/AEI/10.13039/501100011033 and by ERDF A way of making Europe, and the scholarship (FPI-PRE2019-089310) from the Ministry of Science, Innovation, and Universities of the Spanish Government.

References

[1] F.-M. Plaza-Del-Arco, M. D. Molina-González, L. A. Ureña López, M. T. Martín-Valdivia, Detecting Misogyny and Xenophobia in Spanish Tweets Using Language Technologies, ACM Trans. Internet Technol. 20 (2020). URL: https://doi.org/10.1145/3369869. doi:10.1145/3369869.

[2] F. M. Plaza-del-Arco, M. D. Molina-González, L. A. U. López, M. T. Martín-Valdivia, Sexism Identification in Social Networks using a Multi-Task Learning System, 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. 491–499. URL: http://ceur-ws.org/Vol-2943/exist_paper16.pdf.

[3] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, Automatic Classification of Sexism in Social Networks: An Empirical Study on Twitter Data 8 (2020) 219563–219576. doi:10.1109/ACCESS.2020.3042604.

[4] F. Rodríguez-Sánchez, J. C. 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).

[5] 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).

[6] T. D. Vázquez, Á. R. Catalán, Violencias de género en entornos virtuales, Ediciones Octaedro, 2018.

[7] S. Rosenthal, N. Farra, P. Nakov, SemEval-2017 Task 4: Sentiment Analysis in Twitter, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 502–518. URL: https://aclanthology.org/S17-2088. doi:10.18653/v1/S17-2088.

[8] E. Martínez-Cámara, M. C. Díaz-Galiano, M. A. García-Cumbreras, M. García-Vega, J. Villena-Román, Overview of TASS 2017, Proceedings of TASS (2017) 13–21.
[9] E. Martínez-Cámara, Y. Almeida-Cruz, M. C. Díaz-Galiano, S. Estévez-Velarde, M. Á. García-Cumbreras, M. García-Vega, Y. Gutiérrez, A. Montejo-Ráez, A. Montoyo, R. Muñoz, A. Piad-Morffis, J. Villena-Román, Overview of TASS 2018: Opinions, health and emotions, in: Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN, TASS@SEPLN 2018, volume 2172 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018, pp. 13–27.
[10] M. C. Díaz-Galiano, M. García-Vega, E. Casasola, L. Chiruzzo, M. Á. García-Cumbreras, E. Martínez-Cámara, D. Moctezuma, A. Montejo-Ráez, M. A. Sobrevilla-Cabezudo, E. Sadit-Tellez, et al., Overview of TASS 2019: One More Further for the Global Spanish Sentiment Analysis Corpus., in: IberLEF@ SEPLN, 2019, pp. 550–560.
[11] M. García-Vega, M. C. Díaz-Galiano, M. Á. García-Cumbreras, F. M. Plaza-del-Arco, A. Montejo-Ráez, S. M. Jiménez-Zafra, E. Martínez-Cámara, et al., Overview of TASS 2020: Introducing Emotion Detection, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain, September 23th, 2020, volume 2664 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 163–170.
[12] F. Plaza-del-Arco, C. Strapparava, L. A. Ureña-López, M. Martín-Valdivia, EmoEvent: A Multilingual Emotion Corpus based on different Events, in: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 1492–1498.
[13] S. Lamprinidis, F. Bianchi, D. Hardt, D. Hovy, Universal Joy: A data set and results for classifying emotions across languages, in: The 16th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2021.
[14] C. J. Zimmerman, M.-K. Stein, D. Hardt, R. Vatrapu, Emergence of Things Felt: Harnessing the Semantic Space of Facebook Feeling Tags., in: ICIS, 2015.
[15] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval), in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019. URL: https://aclanthology.org/S19-2010. doi:10.18653/v1/S19-2010.
[16] 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 67 (2021) 209–221.
[17] V. Kolhatkar, N. Thain, J. Sorensen, L. Dixon, M. Taboada, Classifying constructive comments, CoRR abs/2004.05476 (2020). URL: https://arxiv.org/abs/2004.05476. arXiv:2004.05476.
[18] D. Ghosh, A. Vajpayee, S. Muresan, A Report on the 2020 Sarcasm Detection Shared Task, in: Proceedings of the Second Workshop on Figurative Language Processing, Association for Computational Linguistics, Online, 2020, pp. 1–11. URL: https://aclanthology.org/2020.figlang-1.1. doi:10.18653/v1/2020.figlang-1.1.
[19] R. Caruana, Multitask learning, Machine learning 28 (1997) 41–75.
[20] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional

Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[21] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, in: Advances in neural information processing systems, 2019, pp. 8026–8037.

[22] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, HuggingFace's Transformers: State-of-the-art Natural Language Processing, arXiv (2019). doi:10.48550/arXiv.1910.03771. arXiv:1910.03771.

[23] [online], https://huggingface.co/vinai/bertweet-large, 2020.

[24] 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.

[25] [online], https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased, 2020.