# Sexism Identification in Social Networks using a Multi-Task Learning System

Flor Miriam Plaza-del-Arco, M. Dolores Molina-González, L. Alfonso Ureña-López, and M. Teresa Martín-Valdivia

Department of Computer Science, Advanced Studies Center in ICT (CEATIC)
Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
{fmplaza, mdmolina, laurena, maite}@ujaen.es

**Abstract.** This paper describes the participation of SINAI-TL team at sEXism Identification in Social neTworks shared task at IberLEF 2021. In order to accomplish the task, we follow a Multi-Task Learning approach where multiple tasks related to sexism identification are learned in parallel while using a shared representation. Specifically, we test the performance of the combination of different tasks related to sentiment analysis and offensive language detection. Our team ranked second in subtask 1 and third in subtask 2, achieving 78% and 56.67% of accuracy, respectively, among the participants.

**Keywords:** Multi-Task Learning · BERT · Sentiment Analysis · Offensive Language.

## 1 Introduction

Sexism is any discrimination against people on the basis of sex (or, as it is currently expressed, on the basis of gender). Sexism against women is a cultural component, historically widespread, whose principle is the supremacy of men over women in different areas of life, such as in the workplace, politics, society, the family and even in advertising.

We find sexism in daily conversation, in the disregard for opinions expressed by women, in statements loaded with discriminatory ideology, even embedded in hundreds of sayings and fixed expressions. This discrimination against women in society is still deeply rooted in communication, both oral and written, and it is increasingly reproduced on the Internet. Detecting online sexism may be difficult, as it may be expressed in very different forms, but it is necessary in order to design new equality policies, as well as to encourage better behaviour in society.

*IberLEF 2021, September 2021, Málaga, Spain.*

Many academic events and shared tasks took place in the last years related to misogyny identification [11, 10] or related to Hate Speech (HS) detection against immigrants and women (HatEval) [4]. Few works have presented sexism detection and, in particular, they addressed sexism as the detection of hate speech against women. But sexism comprises any form of oppression or prejudice against women and therefore may be hostile (as in the case of misogyny) or subtle. Thus, sexism includes misogyny but is not limited to it [17].

In this paper, we present the system we developed as part of our participation for the sEXism Identification in Social neTworks shared task [17] at IberLEF 2021 [15] in both subtasks. The first subtask consists of classifying whether or not a given text (tweet or gab) is sexist (i.e., it is sexist itself, describes a sexist situation or criticizes a sexist behaviour). Once a message has been classified as sexist, the second subtask aims to categorize the message according to the five type of sexism (ideological and inequality, stereotyping and dominance, objectification, sexual violence, and misogyny and non-sexual violence).

In order to accomplish the EXIST shared task, we propose a Multi-Task Learning system (MTL) that leverages affective and offensive knowledge 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 data used in our experiments. In Section 3, we present the proposed system for addressing the task. In Section 4 and 5, we describe the experiment setup and results, respectively. Finally, the conclusion and future work is presented in Section 6.

## 2 Corpora

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 [17] at IberLEF 2021 [15]. The EXIST dataset incorporates any type of sexist expression or related phenomena, including descriptive or reported assertions where the sexist message is a report or a description of a sexist behaviour. Popular expressions and terms, such as terms 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 various Twitter accounts, and analysed and filtered by two gender experts, Trinidad Donoso and Miriam Comet [19]. The final set contains more than 200 expressions that can be used in gendered contexts. Using the final set of sexism terms (94 seeds for Spanish and 91 seeds for English), tweets were extracted in both languages (over 800,000 tweets were downloaded). As a result, the collected dataset has 4,500 tweets per language for the training set and 2,000 tweets per language for the test set. Final labels of tweets were selected according to the majority vote between five crowdsourcing annotators, who followed the guidelines developed by Trinidad and Miriam, but tweets with 3 to 2 votes were manually reviewed by two people with more than two years of experience analyzing sexist content in social networks. Final EXIST dataset consists of 6,977 tweets for training and 3,386 tweets for testing.

Moreover, we used in our experiments other corpora corresponding to tasks that could be related to sexism identification from Twitter including polarity classification (InterTASS), emotion classification (EmoEvent) HS identification (HatEval), and aggressiveness detection (MEX-A3T). The datasets are described below:

- **International TASS Corpus (InterTASS)** was released in 2017 [14] with Spanish tweets and updated in 2018 with texts written in three different variants of Spanish from Spain, Costa Rica and Peru [13]. In 2019, InterTASS was enlarged with new texts written in two new Spanish variants: Uruguayan and Mexican [9] and finally, it was completed with Chilean-Spanish Tweets in 2020 [12]. 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** [3] 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, 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.
- **HatEval** [4], the HS dataset used in this paper, was provided by the organizers in SemEval 2019 Task 5. The task consisted in detecting hateful content in Twitter posts, against two targets: women and immigrants. For the creation of the corpus, the data was collected using a different time frame. The majority of tweets against women were derived from an earlier collection made in the context of two earlier challenges on misogynistic speech identification, whose collection phase began on July 2017 and ended on November 2017 [11, 10]. The remaining tweets were collected from July to September 2018. The dataset contains tweets composed of an identifier, the text of the tweet and the mark of HS, which is 0 if the text is not hateful and 1 if the text is hateful speech against women or immigrants.
- **MEX-A3T** [2]. It was provided by the organizers in IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets [1]. They built a corpus of tweets to detect aggressiveness from Mexican accounts collected from August to November of 2017. They selected a set of terms that served as seeds for extracting the tweets. They used both words classified as vulgar and non-colloquial in the Dictionary of Mexicanisms . The hashtags were related to sexism, homophobia, politics and discrimination. They used Mexico City as the center and extracted all tweets that were within a radius of 500 km. Finally, the collected tweets were labeled by two people. The dataset contains tweets composed of an identifier, the text of the tweet, and the mark of aggressiveness, being 0 if the tweet is not-aggressive and 1 if the tweet is aggressive.

## 3 System overview

In this section, we describe the systems developed for the sEXism Identification in Social neTworks shared task at IberLEF 2021.

We propose a Multi-Task Learning (MTL) system using the well-known Transformer-based model BERT which has been proven to be very successful in many natural language processing tasks. In the MTL model we integrate knowledge from different tasks related to sexism identification.

In the MTL scenario, the goal is to learn multiple tasks simultaneously instead of learning them separately in order to improve performance on each task [6]. These tasks are usually related, although they may have different data or features. By sharing representations across related tasks, we can allow our model to better generalize to our original task. In this study, we used tasks related to the target task sexism identification. These tasks include offensive language detection, polarity classification, and emotion classification, sharing the same data source: Twitter. The reason for incorporating polarity and emotion information to detect sexism is that these tasks are usually emotional and expresses a negative emotion and polarity towards the recipient.

To develop the MTL system, we follow the most widely used technique to MTL in neural networks introduced by [6], the hard parameter sharing approach. It consists of a single encoder that is shared and updated between all tasks, while keeping a few task-specific layers to specialize in each task [18].

The general architecture of the MTL model is shown in Figure 1. The shared layers are based on BERT [8]. Following [8], in the first step, all the inputs are converted to WordPieces [20], two additional tokens are added at the start ([CLS]) and end ([SEP]) of the input sequence, respectively. In the shared layers, the BERT model first converts the input sequence to a sequence of embedding vectors. This semantic representation is shared across all tasks. Then, on top of the shared BERT layers, the task-specific output heads are created for each task, and task heads are attached to a common sentence encoder. Finally, the layers are fine-tuned according to the given set of downstream tasks.

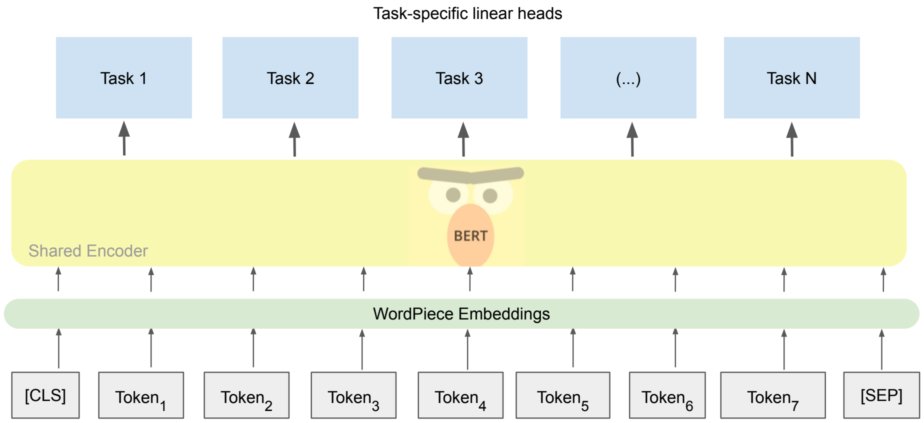
## 4 Experimental setup

### 4.1 Dataset preprocessing

We perform a Twitter-specific data cleaning before including the texts in the models. The following practices to prepare the text for deep learning experiments have been carried out using the ekphrasis module [5]:

- URLs, emails, users' mentions, percentages, monetary amounts, time and date expressions, and phone numbers are normalized.
- Hashtags are unpacked and split to their constituent words.
- Elongated words and repeated characters in words are annotated and reduced.
- Emojis are converted to their alias.

**Fig. 1.** Proposed MTL system for the EXIST task.



### 4.2 System settings

All the models were implemented using PyTorch, a high-performance deep learning library [16] based on the Torch library. The experiments were run on a single Tesla-V100 32 GB GPU with 192 GB of RAM.

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

Regarding our participation, we submitted three runs using the proposed MTL-based system. The details of the modules and the differences of the three settings we presented are described below.

- **Run 1**. In this setting, our goal is to leverage sentiment analysis to aid in the classification of sexism texts. Our assumption is that sexism texts are associated with a negative polarity, then the knowledge share can help to detect easily sexism texts. To this end, we train the MTL model at the same time on the polarity classification and the sexism identification tasks. For the first task, we use the InterTASS dataset. Finally, we obtain the evaluation on the sexism corpora test set.
- **Run 2**. In this setting, our goal is to leverage emotion analysis to aid in the classification of sexism texts. Our assumption is that negative emotions such as *anger*, *fear*, *sadness* and *disgust* could be related to sexism texts while positive emotions are not. For the first task, we use the EmoEvent dataset. Finally, we obtain the evaluation on the sexism corpora test set.
- **Run 3**. In this setting, we train the model on the offensive language identification and the sexism identification tasks. Our assumption is that sexism identification is associated with offensive language and sometimes with hate speech, then the knowledge share during training among these tasks can benefit to the task of sexism identification. For the first task, we use two datasets (HatEval and MEX-A3T). Finally, we obtain the evaluation on the sexism corpora test set.

As the EXIST dataset is composed of English and Spanish texts, while training the MTL system we use two models based on BERT, the BERT base model (cased) trained on English texts and the BETO model [7] trained on Spanish texts. For the first substask (sexism identification) we employ the following hyperparameters: learning rate as 4e-05, batch size as 8, dropout probability as 0.01, the optimization algorithm Adamw, and maximum epoch as 2, while for the second subtask (sexism categorization) the batch size was set to 16 and the number of epochs to 3.

## 5 Results

In this section we present the results obtained by the different runs we have explored in both subtasks of the competition. In order to evaluate them we use the official competition metrics for subtask 1 and subtask 2, accuracy and macro-average F-measure, respectively. Besides, other measures employed in classification tasks including Precision (P) and Recall (R) are computed.

The results of our participation in the EXIST task during the evaluation phase are shown in Table 1 (subtask 1) and Table 3 (subtask 2). In particular, we list the performance of the three runs submitted using the MTL model along with the combination of different tasks as explained in Section 4.2.

If we analyze the results of our 3 runs in subtask 1 and 2, the best result is achieved by the combination of sexism identification and polarity classification tasks, following by run 2, which combines sexism identification and offensive language detection. In subtask 2, it is well noticeable that the run 3 (emotion classification along sexism identification) significantly decreases compared to subtask 1. A possible reason could be that subtask 2 aims to classify 5 different categories that are not significantly associated with emotions, whereas the transfer knowledge of polarity classification and detection of offensive language helps to identify the different categories.

Finally, our results in the competition for both subtasks among the participants (Table 2 and Table 4) show the success of our proposed model achieving the second place in the ranking for the first subtask and the third place for the second subtask. The representations computed by the encoder embed the affective knowledge allows the MTL model to identify sexism more accurately by leveraging the affective nature of the instance.

**Table 1.** Results in Subtask 1 on the test set of EXIST shared task.

| **Run** | **Acc** | **Precision** | **Recall** | **F-measure** |
|---|---|---|---|---|
| 1 | 0.7800 | 0.7796 | 0.7800 | 0.7797 |
| 2 | 0.7766 | 0.7761 | 0.7760 | 0.7761 |
| 3 | 0.7770 | 0.7779 | 0.7751 | 0.7757 |

**Table 2.** Ranking of participants' systems in subtask 1 of EXIST shared task.

| Ranking | Team | Acc |
|---|---|---|
| 1 | AI-UPV_1 | 0.7900 |
| **2** | **SINAI_TL_1** | **0.7800** |
| **3** | **SINAI_TL_3** | **0.7770** |
| **4** | **SINAI_TL_2** | **0.7766** |
| 31 | task1_CIC_1 | 0.7278 |
| 66 | Majority Class | 0.5222 |

**Table 3.** Results in subtask 2 on the test set of EXIST shared task.

| Run | Acc | Precision | Recall | F-measure |
|---|---|---|---|---|
| 1 | 0.6527 | 0.5848 | 0.5527 | 0.5667 |
| 2 | 0.6049 | 0.621 | 0.4082 | 0.4549 |
| 3 | 0.6497 | 0.5774 | 0.5518 | 0.5632 |

**Table 4.** Ranking of participants' systems in subtask 2 of EXIST shared task.

| Ranking | Team | F1 |
|---|---|---|
| 1 | task2_AI-UPV_1 | 0.5787 |
| 2 | task2_LHZ_1 | 0.5706 |
| **3** | **task2_SINAI_TL_1** | **0.5667** |
| **4** | **task2_SINAI_TL_3** | **0.5632** |
| **41** | **task2_SINAI_TL_2** | **0.4549** |
| 62 | Majority Class | 0.1078 |

## 6 Conclusion

This paper presents the participation of the SINAI-TL research group at sEXism Identification in Social neTworks shared task at IberLEF 2021. Our proposal explores how transferred knowledge from tasks related to sexism identification (polarity classification, emotion classification and offensive language detection) may help in a text classification task like EXIST. Experiments conducted show the efficacy of our proposed approach in achieving convincing performance in both subtasks. In particular, polarity classification help the MTL model to classify sexism more accurately by leveraging on the affective knowledge. Finally, as future work we plan to develop a complex model that incorporates other related tasks, such as irony or sarcasm detection, that could be beneficial for sexism identification.

## Acknowledgement

This work has been partially supported by a grant from European Regional Development Fund (FEDER), LIVING-LANG project [RTI2018-094653-B-C21], and Ministry of Science, Innovation and Universities (scholarship [FPI-PRE2019-089310]) from the Spanish Government.

## References

1. Álvarez-Carmona, M.Á., Guzmán-Falcón, E., Montes-y-Gómez, M., Escalante, H.J., Villaseñor-Pineda, L., Reyes-Meza, V., Rico-Sulayes, A.: Overview of MEX-A3T at ibereval 2018: Authorship and aggressiveness analysis in mexican spanish tweets. In: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018. CEUR Workshop Proceedings, vol. 2150, pp. 74–96. CEUR-WS.org (2018)
2. Aragón, M.E., Jarquín-Vásquez, H.J., Montes-y-Gómez, M., Escalante, H.J., Villaseñor-Pineda, L., Gómez-Adorno, H., Posadas-Durán, J.P., Bel-Enguix, G.: Overview of MEX-A3T at IberLEF 2020: Fake News and Aggressiveness Analysis in Mexican Spanish. 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. CEUR Workshop Proceedings, vol. 2664, pp. 222–235. CEUR-WS.org (2020)
3. Plaza-del Arco, F.M., Strapparava, C., Ureña-López, L.A., Martín-Valdivia, M.: EmoEvent: A multilingual emotion corpus based on different events. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 1492–1498. European Language Resources Association, Marseille, France (May 2020)
4. Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F.M., Rosso, P., Sanguinetti, M.: SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In: Proceedings of the 13th International Workshop on Semantic Evaluation. pp. 54–63. Association for Computational Linguistics, Minneapolis, Minnesota, USA (Jun 2019). https://doi.org/10.18653/v1/S19-2007
5. Baziotis, C., Pelekis, N., Doulkeridis, C.: Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 747–754. Association for Computational Linguistics, Vancouver, Canada (August 2017)
6. Caruana, R.: Multitask learning. Machine learning **28**(1), 41–75 (1997)
7. Cañete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., Pérez, J.: Spanish pre-trained bert model and evaluation data. In: PML4DC at ICLR 2020 (2020)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2018)
9. Díaz-Galiano, M.C., García-Vega, M., Casasola, E., Chiruzzo, L., García-Cumbreras, M.Á., Martínez-Cámara, E., Moctezuma, D., Montejo-Ráez, A., Sobrevilla-Cabezudo, M.A., Sadit-Tellez, E., et al.: Overview of TASS 2019: One More Further for the Global Spanish Sentiment Analysis Corpus. In: IberLEF@ SEPLN. pp. 550–560 (2019)
10. Fersini, E., Nozza, D., Rosso, P.: Overview of the evalita 2018 task on automatic misogyny identification (ami). EVALITA Evaluation of NLP and Speech Tools for Italian **12**, 59 (2018)

11. Fersini, E., Rosso, P., Anzovino, M.: Overview of the task on automatic misogyny identification at ibereval 2018. In: Rosso, P., Gonzalo, J., Martínez, R., Montalvo, S., de Albornoz, J.C. (eds.) Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018). CEUR Workshop Proceedings, vol. 2150, pp. 214–228. CEUR-WS.org (2018)
12. García-Vega, M., Díaz-Galiano, M.C., García-Cumbreras, M.Á., Plaza-del-Arco, F.M., Montejo-Ráez, A., Jiménez-Zafra, S.M., Martínez-Cámara, E., 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. CEUR Workshop Proceedings, vol. 2664, pp. 163–170. CEUR-WS.org (2020)
13. Martínez-Cámara, E., Almeida-Cruz, Y., Díaz-Galiano, M.C., Estévez-Velarde, S., García-Cumbreras, M.Á., García-Vega, M., Gutiérrez, Y., Montejo-Ráez, A., Montoyo, A., Muñoz, R., Piad-Morffis, A., Villena-Román, J.: Overview of TASS 2018: Opinions, health and emotions. In: Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN, TASS@SEPLN 2018. CEUR Workshop Proceedings, vol. 2172, pp. 13–27. CEUR-WS.org (2018)
14. Martínez-Cámara, E., Díaz-Galiano, M.C., García-Cumbreras, M.A., García-Vega, M., Villena-Román, J.: Overview of TASS 2017. Proceedings of TASS pp. 13–21 (2017)
15. Montes, M., Rosso, P., Gonzalo, J., Aragón, E., Agerri, R., Álvarez Carmona, M., Álvarez Mellado, E., Carrillo-de Albornoz, J., Chiruzzo, L., Freitas, L., Gómez Adorno, H., Gutiérrez, Y., Jiménez-Zafra, S.M., Lima, S., Plaza-del-Arco, F.M., Taulé, M. (eds.): Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) (2021)
16. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in neural information processing systems. pp. 8026–8037 (2019)
17. Rodríguez-Sánchez, F., de Albornoz, J.C., Plaza, L., Gonzalo, J., Rosso, P., Comet, M., Donoso, T.: Overview of exist 2021: sexism identification in social networks. Procesamiento del Lenguaje Natural **67**(0) (2021)
18. Ruder, S.: Neural transfer learning for natural language processing. Ph.D. thesis, NUI Galway (2019)
19. Vázquez, T.D., Catalán, Á.R.: Violencias de género en entornos virtuales. Ediciones Octaedro (2018)
20. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al.: Google's neural machine translation system: Bridging the gap between human and machine translation. CoRR **abs/1609.08144** (2016)