Combining Transformer-Based Models with Traditional Machine Learning Approaches for Sexism Identification in Social Networks at EXIST 2021

Ezequiel Lopez-Lopez, Jorge Carrillo-de-Albornoz, and Laura Plaza

NLP & IR Group, UNED, 28040 Madrid, Spain
ezequiel.lopez@invi.uned.es
{jcarrillo, lplaza}@lsi.uned.es

Abstract. Sexism in social networks has been increasingly present since their adoption among the connected population, partially due to the disinhibition of individuals when expressing their opinions in social networks and the lack of regulation nor control. Sexism permeates many layers of our society such as media, education or the work environment, and involves many aspects of our society such as gender equality, social respect, gender discrimantion, human rights, etc. In order to fight against this phenomenon, the first step is detecting it properly and in all its varieties. For that purpose, we present our candidate systems for participation in the EXIST challenge (task 1) at IberLEF 2021 for sexism identification. We explore an approach based on the use of pre-trained transformers, such as BERT and roBERTa, in similar tasks and compare them to traditional Machine Learning approaches, such as SVM, Logistic Regression, SGD-based classifier and XGBoost. We achieve our best results of F1=72.4% for the multi-language binary classification task for our system combining transformer-based and traditional models with a majority vote mechanism, positioned #32 in the challenge's ranking, with a F1 difference of 6% against the best result.

Keywords: EXIST · sexism detection · natural language processing · social networks · transformers · IberLEF · BERT · roBERTa

1 Introduction

Sexism [17] is defined as the prejudice or discrimination based on sex or gender, especially against women and girls. It is manifested in many aspects of daily life such as education, work or media and constitutes a source of inequality, verbal and physical aggression, discrimination and other toxic behaviors. The high rate of use of Internet and social networks in the last decade has increased

IberLEF 2021, September 2021, Málaga, Spain.

the presence of sexism in society, due to the visibility of the publications, the connectivity of communities online and the dis-inhibition effect [12]. Efforts to control and regulate [8] these and other toxic behaviors like hate-speech have been made lately, but it is still an ongoing process and new approaches and technologies are needed.

One of the extreme form of sexism is misogyny, the hatred of women. Societies with high levels of misogyny present high rates of domestic violence, rape and commodification of women and their bodies, and they are seen as property or second-class citizens [17]. Although, not every form of sexism is misogynistic [9]. We may find numerous subtle behaviors that establish stereotypes and roles about men and women, which are not easy to detect even for humans.

The automatic identification of sexism in a broad sense may help to create, design and promote the evolution of new equality policies, as well as encourage more adequate behaviors in our society.

The aim of our work is to develop a multi-language binary classification system for sexism in the context of our participation in the EXIST 2021 challenge (task 1). In the scope of the challenge, participants are asked to classify *tweets* and *gabs* (in English and Spanish languages) as *sexist* or *non-sexist*. For such purpose, we have designed several binary classification systems, combining state-of-the-art transformer-based models with traditional ones for both languages, Spanish and English, and benchmarked them to select the best candidates for our participation in the challenge. We found our best performing system (referred as *Task1_MessGroupELL_3*, ranked as #32 in the challenge results) to be our proposal of majority vote of two traditional models and a transformer-based model. These results, compared to the obtained with only traditional approaches or only transformer-based approaches, showed that different features are being complementary captured by different systems, and that using pre-trained transformers models fine-tuned only with the EXIST dataset do not achieve considerable improvements compared to traditional Machine Learning approaches.

2 Resources

Our proposal is focused on addressing the task separately by language. This is based on the fact that the state-of-the-art of techniques and available resources and datasets for this matter are quite different for English and Spanish. Our assumption is that we could achieve good performance results if we attempt to use these elements from the state-of-the-art separately. Additionally, we also consider that sexism could have strong language-dependent (and culture-dependent) characteristics that might be undermined considering a single multi-language approach and better explored separately.

Our purpose is to explore a data-driven approach more than a feature-driven one. In this direction, we attempt to use existing related datasets to support and enrich our systems, alongside with pre-trained transformers that can be trained further in our task, but than can already give us an initial advantage. The comparison of performances between traditional methods and transformers

is especially interesting within our goals. We assume that our approaches can be further extended and improved combining our proposal with a feature-driven one, due to the specific linguistic characteristics of sexist language.

For both languages, we have considered a set of common traditional Machine Learning models based on tf-idf characteristics for benchmarking purposes.

2.1 English Language

For the case of the English subset, we explore pre-trained transformers in other tasks and datasets. We have considered the following pre-trained transformer models to train in our task:

1. **BERT** [6]: pre-trained in a dataset consisting of 11,038 unpublished books and English Wikipedia. We consider this model as neutral one, to compare to our other more specifically trained models.
2. **Twitter-roBERTa-base**: roBERTa-base model trained on 58M tweets, described and evaluated in [2]. Due to the differences between language in Twitter and more traditional sources such as Wikipedia, we consider this model the best starting point to train in our task.
3. **Twitter-roBERTa-base for Hate Speech Detection** [3]: based on the previous roBERTa model and fine-tuned for hate speech with the *TweetEval* benchmark [2].
4. **Twitter-roBERTa-base for Offensive Language Identification** [13]: based on the previous roBERTa model and fine-tuned for offensive language with the *TweetEval* benchmark [2].

The models (3) and (4) correspond to two different tasks (Hate-Speech Detection and Offensive-Language Identification) that we consider interesting and potentially relevant for the detection of sexism or, at least, some aspects of it. As mentioned, sexism is not found as a single unified behavior, but as a wide range of behaviors that might include from a subtle joke to a very aggressive comment.

We regard sexism as partially containing both aspects, hate-speech and offensive-language characteristics, and that the previous training of these models could add a cross-task knowledge to our training. Alternatively, we might expect that both pre-trained models are downstream models of the original model (2) and could limit the performance of a new training, compared to the use of model (2) instead.

2.2 Spanish Language

For Spanish language, we combine our available training data from the EXIST Task in Spanish, with the MeTwo dataset presented in [9]. This dataset has been developed under the same theoretical framework for sexism and uses a similar class scheme: *sexist, non-sexist, dubious*, from which we ignore the third one. In order to preserve the original proportions of our two class in the original

EXIST training dataset, we add a balanced extraction of MeTwo (1152 *sexist* and 1152 *non-sexist* instances). Our reference model for Spanish is BETO [4] a cased BERT-based model, pre-trained in a large Spanish corpus.

3 Combining Transformer based models with traditional Machine Learning approaches

3.1 Preprocessing

The texts are processed for both languages in the same manner, using the spaCy [16] tooling, filtering stop-words and punctuation, with tokenization and lemmatization in lower-case and replacing numbers, currencies, urls, emails and mentions (Twitter) by the keys NUMBER, CURRENCY, URL, EMAIL and MENTION, since we consider these as relevant features for our purpose.

3.2 EXIST dataset

We have split the EXIST [10] training dataset by language and created the extended Spanish dataset:

- English - *non-sexist*: 1800 instances, *sexist*: 1636 instances.
- Spanish - *non-sexist*: 1800 instances, *sexist*: 1741 instances.
- Spanish extended (EXIST + MeTwo) - *non-sexist*: 2952 instances, *sexist*: 2893 instances.

The training and validation sub-datasets have been extracted from the original non-extended datasets, using the holdout method with an 80%-20% split for training and validation, randomly selected and preserving the original class proportions. In the case of Spanish extended dataset, the MeTwo instances have been added only to the training dataset for comparability purposes.

3.3 Baselines

We consider a Support Vector Machine (SVM) implementation as baseline and complementary traditional Machine Learning models for benchmarking purposes, all using TF-IDF features. These models include Logistic Regression (LogReg), XGBoost [5] and Stochastic Gradient Descent (SGD) based classifier [14] modified with a Huber-Loss function, more tolerant to outliers.

3.4 Transformer-based systems.

We have used our training datasets to fine-tune the pre-trained models mentioned in Section 2.1 for English and in Section 2.2 for Spanish. We have performed the training on GPU, with a configuration of 10 train epochs at learning rate $2 \cdot 10^{-5}$, cross-entropy loss function and early stopping at 5 epochs, to prevent overfitting.

3.5 Candidate systems for EXIST 2021 (task 1): binary classification of sexism

English Language. Considering the results presented in Table 1 for English language on the validation dataset, we can observe that transformer-based approaches perform at least 2% better than traditional approaches in macro-average and up to 5% better comparing the best two performances of transformer-based (Roberta) and traditional models (SGD or XGBoost). For the class *sexist*, the difference is ever more pronounced, with almost 10% improvement in some cases.

Table 1. Results obtained for binary classification of sexism for the selected models in English.

	sexist			non-sexist			**macro-avg**		
	F1	R	P	F1	R	P	F1	R	P
SVM	.69	.66	.71	.73	.75	.71	.71	.71	.71
LogReg	.68	.66	.70	.72	.74	.71	.70	.70	.70
SGD	.70	.66	**.74**	.75	**.78**	.72	.72	.72	.73
XGBoost	.70	.68	.72	.74	.76	.72	.72	.72	.72
BERT	.75	.81	.70	.73	.68	.80	.74	.75	.75
Roberta	**.77**	**.83**	.72	**.76**	.71	**.82**	**.77**	**.77**	**.77**
Roberta hate-speech	.75	.77	.72	.75	.73	.78	.75	.75	.75
Roberta offensive	.74	.79	.70	.74	.69	.79	.74	.74	.74
Vote(Roberta, SGD, XGBoost)	.73	.73	.74	.76	.76	.76	.75	.75	.75

It is noticeable that the pre-trained models on other tasks such as *hate-speech* and *offensive-language* don't provide us a clear advantage against the original Roberta. Although the results are quite similar and comparable, it seems plausible that we decrease generalizability using both cross-tasks approaches in the current configuration. This might be due to the heterogeneous composition of the EXIST dataset regarding the different forms of sexism (misogyny, offensive-language, subtle sexism, etc.) that are not necessarily mutually inclusive.

Regarding the performance of traditional models, we can observe their difficulties to identify correctly the *sexist* class, compared to the *non-sexist* against the performance of transformer-based models.

Based on these results, we select Roberta as our one of our candidates for English language for its higher values in almost every aspect, compared to the other candidates.

Nevertheless, the good performance of traditional models to detect the *non-sexist* class and the risk of overfitting while performing fine-tuning in our transformer-based models, make us consider a combination of both approaches (Roberta and traditional models) using a majority vote mechanism as a good candidate to submit, due to its potential better generalizing capabilities. We can observe that, in

macro-average, this approach does not achieve better results than other options, but it offers a considerable well-balanced trade-off for both classes in precision and recall.

Spanish Language Reviewing the performance obtained for Spanish Language in Table 2 on the validation dataset, we can observe that the BERT-based approaches provide more positive overall results compared to traditional approaches. However, the baseline SVM provides the best F1 and recall for the *non-sexist* class in exchange for very poor results on F1 and recall for the *sexist* class. It seems obvious that our best candidates for the task are the BERT-based options. Specifically, we choose the (+MeTwo) BERT approach, since the overall results are better and the per-class trade-offs precision/recall are more balanced than in the original BERT approach. In addition, the inclusion of the MeTwo instances in the training can improve the generalizability of our system.

Table 2. Results obtained for binary classification of sexism for the selected models in Spanish.

	sexist			non-sexist			**macro-avg**		
	F1	R	P	F1	R	P	F1	R	P
SVM	.66	.63	.70	**.72**	**.76**	.69	.70	.69	.70
(+MeTwo) SVM	.68	.67	.69	.71	.72	.69	.69	.69	.69
LogReg	.69	.69	.68	.71	.70	.72	.70	.70	.70
(+MeTwo) LogReg	.71	.71	**.71**	**.72**	.72	.72	.71	.71	.71
SGD	.67	.66	.69	**.72**	.73	.70	.70	.70	.70
(+MeTwo) SGD	.69	.68	.70	.71	.72	.70	.70	.70	.70
XGBoost	.67	.66	.69	.71	.72	.70	.70	.69	.69
(+MeTwo) XGBoost	.70	.73	.67	.69	.66	.72	.69	.69	.69
BERT	**.73**	**.78**	.69	.71	.66	**.76**	**.72**	**.72**	.72
(+MeTwo) BERT	**.73**	.77	.70	.71	.68	.75	**.72**	**.72**	**.73**

3.6 Final system combination for test set

For the prediction of the EXIST's test dataset, we have split the instances by language, preprocessed them with their respective language-dependent preprocessing steps and evaluated them through the selected systems.

As specified by the challenge's organizers, we can submit three runs of our candidate systems on the test dataset. Our strategy to select the candidate configurations consists on using the best Spanish candidate and different options for English to observe how the contributions of one or another candidate impact the performance. The selected configurations are described in Table 3.

We evaluate our run #1 as our transfomer-based approach representative, with the best candidates from both languages. Secondly, we evaluate #2 as a

baseline approach, combining one of the best results from the traditional models, XGBoost, prioritizing a balanced precision/recall trade-off for the *sexist* class. Finally, we combine our best transformer approach with the two best traditional approaches (SGD and XGBoost) through a majority vote, expected to provide a good balance for the *non-sexist* and capture information not captured by the transformer approach.

Table 3. Selected configurations for task evaluation.

Run	Configuration	Description
#1	Roberta(EN), (+MeTwo) BERT(ES)	Roberta pre-trained on tweets and trained on the English subset of EXIST training dataset and BERT pre-trained transformer trained on the Spanish subset from EXIST training dataset and the MeTwo dataset balanced for the classes (*sexist*, *non-sexist*).
#2	XGBoost(EN), (+MeTwo) BERT(ES)	XGBoost based on tf-idf features for English and BERT pre-trained transformer trained on the Spanish subset from EXIST training dataset and the MeTwo dataset balanced for the classes (*sexist*, *non-sexist*).
#3	Majority Vote [SGD(EN) & XGBoost(EN) & Roberta(EN)], (+MeTwo) BERT(ES)	Majority vote of SGD, XGBoost (run #2) and Roberta (run #1) for English and BERT pre-trained transformer trained on the Spanish subset from EXIST training dataset and the MeTwo dataset balanced for the classes (*sexist*, *non-sexist*).

4 Results

The results of the evaluation are shown in Table 4 and present, as expected, the best candidate to be #3 (the majority vote between the best transformer approach and traditional approaches), indicating different and complementary feature learning capabilities. The least performing candidate is, as expected, the traditional approach #2, offering similar but slightly less optimal results than those obtained in the validation dataset. And finally, our intermediate results comes from #1 the combination of our best performing approaches for both languages. As anticipated, the use of our transformers alone in our current configuration has been proven not to be completely effective on its own.

Table 4. Results obtained in the EXIST 2021 challenge (task 1) for our runs (team:Task1_MessGroupELL, positions #39, #33 and #32 in the ranking, respectively for the runs #1, #3 and #2). EXIST challenge's best result and baseline (SVM TF-IDF) are also shown at the upper half of the table, for comparison.

		English			Spanish			Both		
Ranking	Run	F1	R	P	F1	R	P	F1	R	P
#52	baseline	.6886	.6918	.6934	.6766	.6853	.6972	.6832	.6888	.6943
#1	**best**	**.7657**	**.7654**	**.7666**	**.7944**	**.7958**	**.7960**	**.7802**	**.7806**	**.7801**
#33	#1	.7253	.7255	**.7372**	.7144	.7186	.7233	.7225	.7224	.7224
#39	#2	.7070	.7087	.7078	.7144	.7186	.7233	.7109	.7137	.7154
#32	#3	**.7313**	**.7313**	.7313	.7144	.7186	.7233	**.7237**	**.7253**	**.7254**

We have presented as candidates for the challenge three systems, described in Table 3, representing a *baseline system* (run #1), a *pure transformer-based system* (run #2) and a *voted traditional/transformer-based system* (run #3) which have resulted in $F1 = .7237$ for the best performing case (position #32 in the challenge's ranking): our run #3, i.e, the majority vote of two traditional models (XGBoost and SGD) and a transformer-based model (Roberta pre-trained on English tweets). Comparing to the best result in the challenge $F1 = .7802$, we observe a difference of 6% (only 3% for English language), which we consider a very positive outcome. On the other hand, our performance in Spanish language has been shown relatively low compared to the best result in the challenge (8% lower).

With a very similar performance but slightly lower, we find our run #1 (transformer-based) in the next position of the ranking with $F1 = .7225$. This 1% difference in F1-performance relies on a more balanced precision/recall trade-off for the English language examples, showing that the majority vote helps to capture more relevant cases without losing much precision.

Finally, our lowest performing system, run #2 (#39 of the ranking) has shown acceptable results considering its position in the ranking, and regarding that we have used it as a baseline for English language and the same configuration for Spanish language to evaluate differences between our other two candidates. These results are fundamentally based on the Spanish language performance, and show that it has performed considerably well compared to other candidates in the ranking.

In every case, we have presented a transformer-based model for the Spanish language, since our experiments have shown a noticeable difference between this approach and traditional models (Table 2, especially for the *sexist* class, in which traditional approaches present significantly poorer results. For that matter, it is worth noticing that the use of the existing dataset MeTwo for sexism identification for Spanish language has been proven to be an improvement in the performance of every approach for the Spanish language (Table 2).

However, we observe a gap between English and Spanish languages performances for our best result compared to the challenge's top result, that might be due to the fact that we have fine-tuned a BERT model pre-trained in Spanish texts [15], but not specialized in social media as it is our pre-trained model for the English language. We expect that that similar voting mechanism approaches can be followed for the Spanish case, since the *non-sexist* class seems to be better detected by traditional approaches.

5 Conclusions

In this work, we have explored the capabilities of the use of mono-lingual pre-trained transformers as a basis for training a binary classification task on sexism, in the scope of our participation in the EXIST (task 1) challenge of IberLEF [18]. We have explored different state-of-the-art options for both English and Spanish languages, separately, allowing us to test different modular configurations for the task. Different pre-trained transformers have been tested, contrasting the performance of those trained in different, but related, tasks to our current task. This has been shown not to be as well-performing as base pre-trained transformers such as Twitter-roBERTa-base [2]. At the same time, those transformers pre-trained on Twitter corpora have shown as better performers than those trained in traditional or wide-range corpora like the base BERT [6] transformer model.

The results of the evaluation have shown that even well-performing transformers models can be complemented by traditional approaches, in this case, through a simple majority vote. In this case, the traditional approaches can achieve in overall a better recall regarding the secondary class (*non-sexist*), whereas transformer-based approaches achieve better precision for the same class.

The presented proposal and methodology have been proven to be adequate to produce generalizable systems for this task, considering the similarity between performance seen in the training/validation workflow and performance on evaluation, with differences for F1 of 4% for English and 1% for Spanish.

Acknowledgments

This work was supported by the Spanish Ministry of Science and Innovation under Project Misinformation and Miscommunication in Social Media (PGC2018-096212-B-C32).

References

1. Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. Automatic identification and classification of misogynistic language on twitter. In *International Conference on Applications of Natural Language to Information Systems*, pages 57–64. Springer, 2018.

2. Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*, 2020.
3. Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 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*, pages 54–63, Minneapolis, Minnesota, USA, 2019. Association for Computational Linguistics.
4. José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*, 2020.
5. Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4), 2015.
6. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.
7. Simona Frenda, Bilal Ghanem, Manuel Montes-y Gómez, and Paolo Rosso. Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *Journal of Intelligent & Fuzzy Systems*, 36(5):4743–4752, 2019.
8. Dave Gershgorn and Mike Murphy. Facebook is hiring more people to moderate content than twitter has at its entire company, 2017.
9. Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, and Laura Plaza. Automatic classification of sexism in social networks: An empirical study on twitter data. *IEEE Access*, 8:219563–219576, 2020.
10. Francisco Rodríguez-Sánchez, Jorge Carrillo de Albornoz, Laura Plaza, Julio Gonzalo, Paolo Rosso, Miriam Comet, and Trinidad Donoso. Overview of exist 2021: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 67(0), 2021.
11. Zeerak Waseem. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142, 2016.
12. Michelle F Wright, Bridgette D Harper, and Sebastian Wachs. The associations between cyberbullying and callous-unemotional traits among adolescents: The moderating effect of online disinhibition. *Personality and Individual Differences*, 140:41–45, 2019.
13. Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, 2019.
14. Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, page 116, 2004.
15. José Cañete. Compilation of Large Spanish Unannotated Corpora, https://doi.org/10.5281/zenodo.3247731, May, 2019.
16. Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python, 2020.
17. Masequesmay, Gina. "Sexism". In *Encyclopedia Britannica*. 28 May. 2020, https://www.britannica.com/topic/sexism. Accessed 3 June 2021.

18. Manuel Montes, Paolo Rosso, Julio Gonzalo, Ezra Aragón, Rodrigo Agerri, Miguel Ángel Álvarez-Carmona, Elena Álvarez Mellado, Jorge Carrillo-de-Albornoz, Luis Chiruzzo, Larissa Freitas, Helena Gómez Adorno, Yoan Gutiérrez, Salud María Jiménez Zafra, Salvador Lima, Flor Miriam Plaza-de-Arco and Mariona Taulé (eds.): *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), CEUR Workshop Proceedings,* 2021.