Automatic Sexism Identification Using an Ensemble of Pretrained Transformers

Victoria Pachón Álvarez[1], Jacinto Mata Vázquez[2], Wissam Chibane[3], Juan Luis Domínguez Olmedo[4]

[1 2 4] *University of Huelva, Escuela Técnica Superior de Ingeniería, Huelva, Spain*
[3] *Mouloud Mammeri University, Tizi-Ouzou, Algeria*

Abstract

In this paper we present our approach and systems description on Task 1 and 2 on EXIST 2022: sEXism Identification in Social neTworks. Our main contribution is to show the effectiveness of using an ensemble of classifiers based on transformers. In our experiments for Task 1 we tested several models and decided to ensemble the three models that provided the best F1 score for this dataset. We achieved an accuracy of 78.83% for Task 1, ranking our run for Task 1 in the 6th place of the competition. For Task 2 we have used a hierarchical classification approach. First, a binary classification is performed to detect non-sexist tweets and then transformer-based models were used to categorize sexist tweets into the five proposed categories. For this task our system reached an accuracy of 64,65%, ranking in 11th place of the competition.

Keywords

Sexism tweets classification, BERT, BETO, RoBERTa, Transformers, Ensemble

1. Introduction

Sexism is the prejudice, stereotyping or discrimination, typically against women. It refers to a set of expressions and social practices that, based on sexual difference, legitimize and strengthen social inequality between men and women.

With more than 4 billion users spending an average time of 2 hours a day on social media, this computer-based technology became one of the most important ways of global communication. In fact, billions of users spend an average time of 2 hours a day on social media, expressing freely their thoughts and opinions.

We cannot deny the negative and dangerous impacts of social media when it comes to women. Indeed, the online anonymity, invisibility and accessibility allows internet users to freely express sexism and many other types of aggression against women in their comments and the spread of this information is so fast that sexist hashtags become rapidly and dangerously viral.

In this paper we present our contribution to *EXIST 2022: sEXism Identification in Social neTworks* [1]. Given the success and popularity of transformers [2] our idea for Task1 was to combine the predictions of three base models based on transformer to improve the metrics of a single classifier using a majority voting technique. For Task 2, a hierarchical classification approach was carried out. In this case we performed an experimentation with a variety of transformer-based models, but we did not build a meta-classifier ensemble.

In the next section some previous studies are described. In Section 3 we will describe Tasks 1 and 2 and the Corpus provided by the organizers. The experimental methodology and evaluation results are laid out in Section 4 and 5. Finally, in Section 6, we will offer the main results and conclusions of our study and include some perspectives for future works.

IberLEF 2022, September 2022, A Coruña, Spain
EMAIL: vpachon@uhu.es (V. Pachón); mata@uhu.es (J. Mata); wissam.chibane@ummto.dz (W. Chibane); juan.dominguez@dti.uhu.es (JL.Domínguez)

2. Related works

Various studies have theorized that the "online disinhibition effect" [3] encourages overtly sexist behavior. In fact, this effect includes the phenomenon that Internet users, acting under the protection of anonymity, can perform behaviours that they would not normally exhibit in face-to-face situations or in virtual spaces where they are identifiable. Previous studies [4] have shown a close relation between anonymity and higher levels of aggression and rudeness online. Sexism can be hard to identify, sometimes it can be very subtle, or it can be expressed in a misleading way.

Recently, several studies have been carried out in an attempt of automatically detecting misogyny as a form of hate speech. Machine learning models for detection of hate speech against women in English tweets is presented in [5]. [6] Analyzed the hate speech in Spanish tweets against women and immigrants and performed different approaches based on supervised machine learning and deep learning. In [7] a sexism classification dataset in Spanish and in English is proposed. The results showed the outperformance of working with models developed for Spanish texts.

3. Datasets and Tasks

The Corpus provided by the organizers is described at [8]. This Corpus contains two datasets: 6977 tweets for training (3436 in English 3541 in Spanish) and 4368 for test (2208 in English and 2160 in Spanish). For both tasks, the training dataset contains 3377 texts labelled as sexist and 3600 labelled as non-sexist. For the second task, in addition, tweets labelled as sexist were categorized with one of the following labels: “Ideological and inequality”, “Stereotyping and dominance”, “Objectification”, “Sexual violence” or “Misogyny and non-sexual violence”. In the test dataset, the proportion of non-sexist and sexist texts was 2087 and 2281 respectively. The distribution of the tweets labelled as sexist in their subcategories in the training and test dataset is shown in Table 1.

Table 1
Number of labels per class for sexist messages

Label	Number of tweets	
	Training Dataset	Test Dataset
Ideological and inequality	866	621
Stereotyping and dominance	809	464
Objectification	500	324
Sexual violence	517	400
Misogyny and non-sexual violence	685	472

Finally, in Table 2 some examples of the dataset are shown.

Table 2
Examples of tweets of different classes

Lang.	Text	Task1	Task2
EN	Looks like a cool boss lady	Non-sexist	Non-sexist
EN	Gold digger	Sexist	Stereotyping and dominance
EN	Is that the same woman who empowered herself by using men to climb her social ladder	Sexist	Ideological and inequality
SP	Y ni siquiera se que tiene que ver el mansplaining jajaja	Sexist	Ideological and inequality
SP	Es el manual para entender a las mujeres	Sexist	Stereotyping and dominance
SP	Yo no necesito a un hombre, yo soy lo que un hombre necesita	Non-sexist	Non-sexist

4. Methodology

4.1. Task 1

The methodology used in our work is based on creating an ensemble as a set of pre-trained transformers model. Ensemble approach is one of the most advanced solutions to many supervised learning tasks. These methods improve the prediction performance of a single model by combining the predictions of several models [9]. Each ensemble method requires a proper fusion of several learners to generate the final prediction model. The voting classifier estimator, created by merging different classification models predictions, is a stronger meta-classifier that balances the weaknesses of the individual classifiers [10]. In our case, a majority voting based on equal weights has been implemented. The predictions of the different models are combined to obtain a single prediction.

Since some models only work in a certain language, an English dataset was created by converting all tweets to lowercase and translating them from Spanish to English and vice versa. In the training process, we fine-tune ten pre-trained models on the training dataset using an 80%-20% training-validation split for each model. The test dataset provided by the organizers were used to test the performance of the models and obtain their metrics. The models, available in the Huggingface (https://huggingface.co/) transformers library, were the following:

- *dehatebert-mono-spanish* [11]. This model is used to detect hate speech in Spanish language
- *roberta-large-bne* [12]. A transformer-based masked language model for the Spanish language based on the RoBERTa large model and pre-trained using the National Library of Spain
- *bert-base-uncased* [13]. BERT is a transformer model pretrained on a large corpus of English data
- *sentiment-roberta-large-english* [14]. This model is a fine-tuned of RoBERTa-large and allows binary sentiment analysis for English texts
- *roberta-base-bne-finetuned-cyberbullying-spanish* (https://huggingface.co/JonatanGk/roberta-base-bne-finetuned-cyberbullying-spanish). This model is a fine-tuned version for detection of ciberbullying on Spanish texts
- *distilbert-base-uncased* [15]. This model is a distilled version of the BERT base model
- *bert-base-multilingual-uncased* [13]. This model is BERT multilingual base model
- *bert-base-spanish-wwm-uncased* [16]. This model (BETO) is a BERT Spanish version
- *roberta-base* [17]. This is the original RoBERTa base model
- *distilroberta-base* [18]. This model is a distilled version of the RoBERTa-base model

Before feeding the texts to the classifiers, we performed a simple preprocessing that consisted of lowercasing the texts, removing links, URLs and user mentions. Hashtags were kept because, in our opinion, they contain important information for this task.

The models that provided the best accuracy score in test dataset were selected to create the ensemble based on majority voting.

All the models were trained with 3 epochs, 32 batch size, 128 token length and a learning rate of 2e-5. Early stopping was used to avoid overfitting while training.

4.2. Task 2

For the second task, we performed a hierarchical classification where a model from Task 1 classifies between sexist and non-sexist and another model is trained to detect the specific categories of sexism. To obtain the prediction in the test phase, we first use the binary classifier to detect whether a tweet is sexist or not applied. If the text is labelled as non-sexist, this label is used for the final classification. If the binary classifier labels the text as sexist, the tweet is fed to the multiclass classifier to obtain a prediction of the sexism category.

For this task we only fine-tuned three pre-trained models (two using an English training dataset and one using a Spanish training dataset). In this case, we didn't create an ensemble model and only the predictions of the model that obtained the best values in the accuracy and F1 measures were submitted.

The models were *roberta-large-bne*, *roberta-large* y *distilroberta-base* and the same pre-processing was performed as for Task 1. In addition, the same hyperparameters were used to fine-tuning the training of the systems.

5. Results

To run our experiments, we trained each model with the dataset in the appropriate language. The results over the test dataset for Task 1 are presented in Tables 3, 4 and 5. Table 6 shows the results for Task 2.

Table 3
Results with pre-processing

Model	Language Corpus	Acc.	F1	ROC AUC	PR AUC
roberta-large-bne	SP	0.75	0.78	0.750	0.697
bert-base-multilingual-uncased	EN/SP	0.75	0.76	0.749	0.702
bert-base-spanish-wwm-uncased	SP	0.76	0.78	0.757	0.704
dehatebert-mono-spanish	SP	0.74	0.76	0.731	0.681
roberta-base-bne-finetuned-cyberbullying-spanish	SP	0.72	0.73	0.716	0.671
sentiment-roberta-large-english	EN	**0.77**	**0.79**	**0.768**	**0.719**
distilroberta-base	EN	0.75	0.77	0.751	0.704
bert-base-uncased	EN	0.75	0.77	0.748	0.696
roberta-base	EN	**0.77**	**0.78**	**0.764**	**0.715**
distilbert-base-uncased	EN	0.76	0.77	0.755	0.708

Table 4
Results without pre-processing

Model	Language Corpus	Acc.	F1	ROC AUC	PR AUC
roberta-large-bne	SP	0.75	0.75	0.747	0.706
bert-base-multilingual-uncased	EN/SP	0.75	0.76	0.749	0.702
bert-base-spanish-wwm-uncased	SP	**0.76**	**0.78**	**0.771**	**0.718**
dehatebert-mono-spanish	SP	0.73	0.76	0.734	0.673
roberta-base-bne-finetuned-cyberbullying-spanish	SP	0.71	0.74	0.711	0.665
sentiment-roberta-large-english	EN	**0.77**	**0.78**	**0.770**	**0.725**
distilroberta-base	EN	0.70	0.73	0.701	0.657
bert-base-uncased	EN	0.75	0.78	0.751	0.696
roberta-base	EN	0.77	0.78	0.755	0.702
distilbert-base-uncased	EN	0.75	0.77	0.749	0.699

As it can be seen in Tables 3 and 4, best ROC AUC scores were performed by *sentiment-roberta-large-english* (with and without preprocessing), *roberta-base* (with preprocessing) and *bert-base-spanish-wwm-uncased* (without preprocessing). To select the best combination of these models, an ensemble with some possible combinations of three models was created and the one that produced the best ROC AUC score was selected (see Table 5). *Sentiment-roberta-large-english* (without preprocessing), *roberta-base* (with preprocessing) and *bert-base-spanish-wwm-uncased* (without preprocessing) scored the best metrics values. Finally, each of these three models was trained again with all the tweets as follows: *sentiment-roberta-large-english* was trained without preprocessing with the dataset in English, *roberta-base* was trained with the dataset in English with preprocessing and *bert-base-spanish-wwm-uncased* was trained with the dataset in Spanish without preprocessing.

Table 5
Some ensemble combinations

Set of models	Accuracy	F1	ROC AUC	PR AUC
sentiment-roberta-large-english (without preprocessing) + roberta-base + bert-base-spanish-wwm-uncased	**0.80**	**0.81**	**0.794**	**0.744**
sentiment-roberta-large-english (with preprocessing) + roberta-base + bert-base-spanish-wwm-uncased	0.78	0.80	0.782	0.730
sentiment-roberta-large-english (with preprocessing) + sentiment-roberta-large-english (without preprocessing) + roberta-base	0.79	0.80	0.784	0.736
sentiment-roberta-large-english (with preprocessing) + sentiment-roberta-large-english (without preprocessing) + bert-base-spanish-wwm-uncased	0.79	0.80	0.789	0.740

Table 6
Results for Task 2 over test dataset

Model	Lang. Corpus	Accuracy	F1
roberta-large-bne	SP	0.6185	0.6044
distilroberta-base	EN	0.6536	0.6467
roberta-large	EN	**0.6654**	**0.6510**

Our results in the competition for both subtasks among the participants are shown in Table 7 and Table 8. For the first subtask, the best model run was obtained by the described ensemble, achieving the sixth place. We also submitted other two runs pretraining *bert-base-spanish-wwm-uncased* (10th place) and *sentiment-roberta-large-english* with preprocessing (15th place). For the second task, predictions submitted were obtained by training *roberta-large* model. In this task, we ranked 11th among the participants.

Table 7
Official results for Task 1

Ranking	System	Accuracy	F1 score
1	task1_avacaondata_1	0.7996	0.7978
5	task1_CIMATCOLMEX_2	0.7883	0.7877
6	**task1_I2C_1**	**0.7883**	**0.7880**
10	**task1_I2C_3**	**0.7807**	**0.7788**
15	**task1_I2C_2**	**0.7656**	**0.7656**
25	task1_shm2022_1	0.7533	0.7530
44	Majority Class	0.4905	0.4872

Table 8
Official results for Task 2

Ranking	System	Accuracy	F1 score
1	task2_avacaondata_1	0.7013	0.5106
10	task2_ThangCIC_8	0.6626	0.4706
11	**task2_I2C_1**	**0.6465**	**0.4700**
17	task2_AI-UPV_3	0.6267	0.4516
30	Majority Class	0.5539	0.1018

6. Conclusions

In this paper we presented our proposal for sexism detection in English and Spanish language tweets and the results obtained in the shared tasks for EXIST 2022. For the binary classification task of distinguishing between sexist and non-sexist, our proposal was an ensemble of classifiers in Spanish and English based on transformers. This model got very promising results with an 0.7883 of accuracy and a 0.7880 F1-score and with 6th best run in the ranking. For the sexism categorization task our proposal was a hierarchical approach obtaining 0.6465 accuracy score and 0.4700 F1-score.

With this approach we prove the accuracy and usability of ensembles in training deep learning models mixing different languages corpus. In future works we plan to explore other techniques of creating ensemble as well as study the best hyperparameters to train the models that are part of the ensemble and apply this technique to other problems of classification such as hate-speech or racism identification in social networks.

7. References

[1] F. Rodríguez-Sánchez, J. Carrillo-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, vol. 69, septiembre 2022

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I.Polosukhin. 2017. Attention is all You Need.

[3] F. Jesse, C. Cruz, and J.Y. Lee. 2015. "Perpetuating Online Sexism Offline: Anonymity, Interactivity, and the Effects of Sexist Hashtags on Social Media." Computers in Human Behavior 52:436-442.doi:10.1016/j.chb.2015.06.024.

[4] C.L. Schoffstall and R. Cohen. 2011. "Cyber Aggression: The Relation between Online Offenders and Offline Social Competence." Social Development 20 (3): 587-604. doi:10.1111/j.1467-9507.2011.00609.x. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9507.2011.00609.x.

[5] R. Ahluwalia, H. Soni, E. Callow, A. Nascimento, and M. De Cock. 2018. "Detecting Hate Speech Against Women in English Tweets." In 194: Accademia University Press. doi:10.4000/books.aaccademia.4698. http://books.openedition.org/aaccademia/4698.

[6] F.M. Plaza-Del-Arco, M.D. Molina-González, L. Ureña-López, and M. Martín-Valdivia. 2020. "Detecting Misogyny and Xenophobia in Spanish Tweets using Language Technologies." ACM Transactions on Internet Technology 20 (2): 1-19. doi:10.1145/3369869. http://dl.acm.org/citation.cfm?id&#61;3369869.

[7] M. Anzovino, E. Fersini, and P. Rosso. 2018. "Automatic Identification and Classification of Misogynistic Language on Twitter." In Natural Language Processing and Information Systems, 57-64. Cham: Springer International Publishing. doi:10.1007/978-3-319-91947-8_6. http://link.springer.com/10.1007/978-3-319-91947-8_6.

[8] F. Rodríguez-Sánchez, J. Carrillo-de-Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, and T.Donoso. 2021. "Overview of EXIST 2021: sEXism Identification in Social neTworks." Procesamiento Del Lenguaje Natural 67 (0): 195-207. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6389.

[9] S. Omer and L. Rokach. 2018. "Ensemble Learning: A Survey." WIREs Data Mining and Knowledge Discovery 8 (4): e1249. doi:10.1002/widm.1249. https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1249.

[10] I. Lazaros and D.G. Wat. On Ensemble Techniques of Weight- Constrained Neural Networks Ensemble Deep Learning Models for Forecast Ing Crypt Ocurrency T Ime-Series Panagiot is E Pint Elas Ensemble of Deep Recurrent Neural Net Works for Ident Ifying Enhancers Via Dinucleot Ide Physicoche… Ivy Yeh T He Rhet Oric and Realit Y of Ant Hropomorphism in Art Ificial Int Elligence.

[11] S. SAluru,, B. Mathew, P. Saha, and A. Mukherjee. 2020. "Deep Learning Models for Multilingual Hate Speech Detection.". https://arxiv.org/abs/2004.06465.

[12] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C.P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A.Gonzalez-Agirre, and M. Villegas. MarIA: Spanish Language Models MarIA: Modelos Del Lenguaje en español.

[13] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova. 2018. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." CoRR abs/1810.04805.

[14] J.Hartmann, M. Heitmann, and C. Schamp. More than a Feeling: Accuracy and application of Sentiment Analysis Christian Siebert 2.

[15] V.Sanh, L. Debut, J. Chaumond, and T. Wolf. 2019. "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter.".

[16] J. Cañete, G. Chaperon, R. Fuentes, J.H. Ho, H. Kang, and J. Pérez. 2020. "Spanish Pre-Trained BERT Model and Evaluation Data.".

[17] Y.Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O.Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach.".

[18] V. Sanh, L. Debut, J. Chaumond, T. Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv 2019.