UMUTeam at EXIST 2021: Sexist Language Identification based on Linguistic Features and Transformers in Spanish and English

José Antonio García-Díaz[1] [0000−0002−3651−2660],
Ricardo Colomo-Palacios[2] [0000−0002−1555−9726], and
Rafael Valencia-García[1] [0000−0003−2457−1791]

[1] Facultad de Informática, Universidad de Murcia,
Campus de Espinardo, 30100, Spain
{joseantonio.garcia8,valencia}@um.es
[2] Faculty of Computer Sciences, Østfold University College, Halden, Norway
ricardo.colomo-palacios@hiof.no

Abstract. Sexism is harmful behaviour that can make women feel worthless promoting self-censorship and gender inequality. In the digital era, misogynists have found in social networks a place in which they can spread their oppressive discourse towards women. Although this particular form of oppressive speech is banned and punished on most social networks, its identification is quite challenging due to the large number of messages posted everyday. Moreover, sexist comments can be unnoticed as condescends or friendly statements which hinders its identification even for humans. With the aim of improving automatic sexist identification on social networks, we participate in EXIST-2021. This shared task involves the identification and categorisation of sexism language on Spanish and English documents compiled from micro-blogging platforms. Specifically, two tasks were proposed, one concerning a binary classification of sexism utterances and another regarding multi-class identification of sexist traits. Our proposal for solving both tasks is grounded on the combination of linguistic features and state-of-the-art transformers by means of ensembles and multi-input neural networks. To address the multi-language problem, we tackle the problem independently by language to put the results together at the end. Our best result was achieved in task 1 with an accuracy of 75.14% and 61.70% for task 2.

Keywords: Sexism Identification · Document Classification · Feature Engineering · Natural Language Processing.

IberLEF 2021, September 2021, Málaga, Spain.

1 Introduction

This manuscript describes the participation of the UMUTeam in the shared task EXIST 2021 [12] proposed at IberLeF 2021 [7], focused on the identification and categorisation of sexist language. Despite all the benefits of the raising of the Web 2.0. when it comes to reducing barriers to communication, misogynists have found in social networks a place in which they can spread their oppressive discourse towards women, making these networks an intimidating place. Sexism and misogyny are particular forms of discriminatory speech and conduct in which women are the victims of the harassment. It is worth mentioning the identification of sexist speech is more challenging than other forms of hate-speech, as sexist and misogynistic messages can go unnoticed within funny or complacent comments. Therefore, some sexist messages are subtle to distinguish even for humans if we are not aware of them, just how messages with a condescending treatment could be. Sexist behaviour includes stereotyping, ideological issues, or sexual violence [11].

The objective of EXIST 2021 shared task is the identification and categorisation of sexism behaviours from a broad sense. Specifically, the organisers of the shared task proposed two challenges. On the one hand, a binary sexism classification task and, on the other, a multiclass sexism categorisation task that consisted in a fine-grained classification of those messages classified as sexist in the following traits: (1) ideological and inequality, (2) stereotyping and dominance, (3) objectification, (4) sexual violence, and (5) misogyny and non-sexual violence. To evaluate both tasks, participants were encouraged to build an automatic classification system and evaluated them on a corpus composed by messages from social networks written in Spanish and English.

Our participation is grounded on the usage of linguistic features combined with state-of-the-art transformers. As part of the doctoral thesis of one of the members of the UMUTeam, we are evaluating a tool for extracting linguistic features. This tool is inspired in LIWC [13] but designed to the Spanish from scratch. Our hypothesis is that the usage of linguistic features could be combined with statistical features, such as n-grams or any form of embeddings to improve the accuracy of the results and, at the same time, providing interpretable features. Therefore, one of the runs consisted into the usage of linguistic features in isolation as baseline and then we combined them with BERT in another run and finally we stack an ensemble for the last run.

The remainder of this manuscript is organised as follows. First, in Section 2 some experiments and corpus regarding misogyny and sexist behaviour are discussed. Next, in Section 3 we give some insights regarding the corpus that was made available to the participants. Our pipeline is described in Section 4. In Section 5 we show the results achieved by our team and compare them with the best results achieved by the rest of the participants and the baselines proposed. Finally, the conclusions and future research directions are shown in Section 6.

2 Background information

The identification of aggressive, hateful, and oppressive speech have been recurrent tasks in NLP workshops. For our point of view, this trend is caused by two main factors: (1) They are challenging, as they involve the identification of subjective information and figurative language, and of course (2) the benefits that its automatic identification has for society. We can find, therefore, previously shared tasks focused on hate-speech based on racial, gender, religious, disability, sexual-orientation, and gender traits. Due to the scope of this shared task we will focus on sexism and misogyny identification. As far as our knowledge goes, the most relevant work regarding misogyny identification is the Automatic Misogyny Identification (AMI) that has been proposed at [3, 4] with datasets documents written in English, Spanish and Italian and focused on misogyny identification and categorisation of misogynist traits. In both tasks, the organisers note that misogynistic categorisation *still remains a challenging problem.* From a wider perspective, the SemEval shared task HateEval 2019 [1], focused on hate speech against immigrants and women. This task was also focused on a multilingual perspective, with tweets written in Spanish and English. It consisted in two subtasks: a binary classification problem towards Hate Speech Detection against immigrants and women, and a task focused on detecting if the hate-speech was directed to specific individuals or to wider groups, and if the text contains aggressive behaviour but that may not be linked to hate speech. Organisers also highlighted the challenging of identifying hate speech in micro-blogging texts.

Our research group has also focused on misogyny identification with the release of the Spanish MisoCorpus-2020 [6], focused on Spanish. The MisoCorpus-2020 contains three subsets: VARW, SELA, and DDSS, focused, respectively, on the identification of misogynistic messages towards female politicians and journalists, cultural and background differences between misogyny among European Spanish countries and Latin America countries, and tweets that contains specific misogynistic traits. In this work we evaluate the combination of linguistic features and sentence word embeddings from fastText. The results shown that linguistic features regarding offensive language, grammatical gender, spelling mistakes, punctuation symbols, and jargon from social networks are effective for misogyny identification.

3 Corpus

The dataset is composed by documents written in Spanish and English compiled between December 2020 and February 2021 that contains expressions used to underestimate the role of women in our society. The main data-source is Twitter but Gab was used to extend the dataset. Training and testing have a temporal separation as tweets were selected based on time in order to determine which ones belong to each split. According to the description of the task, a subset of the data were analysed in depth by two experts in gender issues.

The resulting dataset contains 6977 tweets for training and 3386 tweets for testing, where both datasets were randomly selected from the 9000 and 4000 labelled sets, to ensure class balancing for Task 1. This dataset was enlarged with 492 gabs in English and 490 in Spanish from the uncensored social network Gab.com following a similar procedure as described before. This set will be included in the EXIST test set in order to measure the difference between social networks.

Each document was labelled by five annotators. The final label was selected using a majority vote. However, tweets with 3-2 votes were manually reviewed by two experts of different gender. The reader can find more details regarding the dataset compilation in the overview of the task [12].

To evaluate the reliability of our proposal we extracted a validation split consisting in the 20% of the training dataset. Table 1 and Table 2 depict the label distribution we used for task 1 and 2 respectively. We can observe that the distribution of the sexism identification task is slightly imbalanced with more documents labelled as non-sexist. For subtask 2, we can observe that three of the sexist traits namely ideological inequality, stereotyping and dominance, and misogyny non-sexual violence have similar number of documents but the labels objectification and sexual violence have less examples.

Table 1. Dataset distribution for subtask 1. Sexism identification

Label	Total	Train	Val
Spanish			
non-sexist	1800	1446	354
sexist	1636	1387	354
English			
non-sexist	1800	1443	357
sexist	1636	1306	330

4 Methodology

Our proposal is grounded on the combination of linguistic features with state-of-the art transformers. During our experimentation, we also evaluated word and sentence embeddings.

For the linguistic features we use UMUTextStats [5, 6]. This tool is inspired in LIWC [13] but designed for the Spanish language. Although LIWC has available a Spanish version that has been evaluated in different domains, such as satire identification [10], it does not take into account specific Spanish linguistic phenomena that UMUTextStats does. Specifically, UMUTextStats handles a total of 365 linguistic variables classified in the following groups: (1) Phonetics, (2) Morphosyntax, (3) Correction and style, (4) Semantics, (5) Pragmatics and figurative language [9], (6) Stylometry, (7) Lexical, (8) Psycho linguistic processes, (9), and (10) Social media.

For the transformers we use BERT. Specifically, the large cased version for those documents written in English and BETO [2] for those tweets written in

Table 2. Dataset distribution for subtask 2. Sexism categorisation

Label	Total	Train	Val
Spanish			
non-sexist	1800	1446	354
ideological-inequality	480	388	92
stereotyping-dominance	443	357	86
misogyny-non-sexual-violence	401	320	81
objectification	244	193	54
sexual-violence	173	129	44
English			
non-sexist	1800	1443	357
ideological-inequality	386	315	76
stereotyping-dominance	366	290	71
misogyny-non-sexual-violence	344	279	70
objectification	284	214	65
sexual-violence	256	208	48

Spanish. In addition, we train and evaluate models applying word and sentences embeddings from fastText, word2vec, and gloVe for the Spanish documents and from fastText from the English documents. These features were trained with recurrent, convolutional, and vanilla neural networks. Multiple feature sets in the same neural network were also evaluated using the functional API of Keras. It is worth noting that the evaluation of other neural networks architectures such as convolutional neural networks was performed because provided good results in the past for conducting sentiment analysis tasks [8].

Our strategy in this shared task consisted in dealing with documents written in Spanish and English separately, by splitting them in two datasets and evaluated with different models to merge the results just before the final submission.

For the hyperparameter optimization we proceeded as follows. We evaluate a total of 100 neural networks per feature set in isolation and in combination, both for Spanish and English. As there was a slightly imbalance among the classes, the best models were selected based on weighted F1 score. The features evaluated were linguistic features (LF), sentence embeddings from fastText (SE), sentence embeddings from BERT (BE), and word embeddings (WE) from fastText for the English dataset and fastText, gloVe, and word2vec for the Spanish dataset. The majority of neural networks evaluated consisted in multilayer perceptrons (MLP) with different number of hidden layers (between 1 and 8), and different number of neurons (8, 16, 48, 64, 128, 256) organised in different shapes, including *funnel*, *rhombus*, *long_funnel*, *brick*, *diamond*, and *triangle*. In case of WE, convolutional and recurrent networks were also evaluated including Bidirectional Long Short Term Memory (LSTM) and Bidirectional Gated Recurrent Unit (BiGRU). We use the functional API of Keras to feed multiple inputs for each neural network and thus, the combination of several features in the same network were also evaluated. For all test, we tried different dropout rate to avoid overfitting (0, 0.1, 0.2, and 0.3), and several activation functions including *relu*, *sigmoid*, *tanh*, *selu*, and *elu*. We also included an early stopping mechanism. The results of each parameter set can be viewed at https://github.com/Smolky/exist-2021.

5 Results

We participate with three runs. The first one consisted in the usage of the linguistic features in isolation. As commented earlier, this run was used to set a baseline to evaluate the thesis objectives of the doctoral student of the team. The second one consisted in the combination of linguistic features and a BERT-based model, and the third run consisted in an ensemble of neural networks composed from LF, SE, BE, WE, and BERT.

First, the results based on the first task concerning sexism identification are depicted in Table 3.

Table 3. Comparison of the UMUTeam with the best three runs and the baselines for misogyny identification task

Rank	Team	Accuracy	f1-score
1	AI-UPV_1	78.04	78.02
2	SINAI_TL_1	78.00	77.97
3	SINAI_TL_2	77.77	77.57
22	UMUTEAM_3	75.14	75.14
27	UMUTEAM_2	74.40	74.40
52	SVM TFIDF (baseline)	68.45	68.32
63	UMUTEAM_1	59.66	59.64
66	Majority Class (baseline)	52.22	34.31

In regards of the official results for task 1 (misogyny identification), we achieved our best result with the ensemble model (run 3) with an accuracy and F1-score of 75.14%, reaching position 22 in the official rank. The overall best result was achieved by AI-UPV, with an accuracy of 78.04% and a F1-score of 78.02%. Our second run, based on BERT and LF, achieved a 74.4% of accuracy and F1-score, reaching position 27. Finally, our first run, based on the usage of linguistic features in isolation, achieved position 63, with an accuracy of 59.66% and a F1-score of 59.64%. Note that this result does not outperform the baseline result consisted in a bag of words based on TF-IDF score. The poor reliability of linguistic features in isolation is not sparingly due to the fact that UMUTextStats is focused on the Spanish language. In fact, the results only with the Spanish partition achieved an accuracy of 61.94% whereas with the English partition only a 57.43% of accuracy was obtained. The feature selection stage for the English dataset mainly selected those linguistic features based on Corpus statistics such as the length of the text or the Type-Token Ratio (TTR).

Figure 1 depicts the Mutual Information of the top ranked LF for the task 1 for the Spanish split compared by label. We can observe that there are no important differences between linguistic features among classes except for lexical related to sex and female groups that appears mostly in sexist posts.

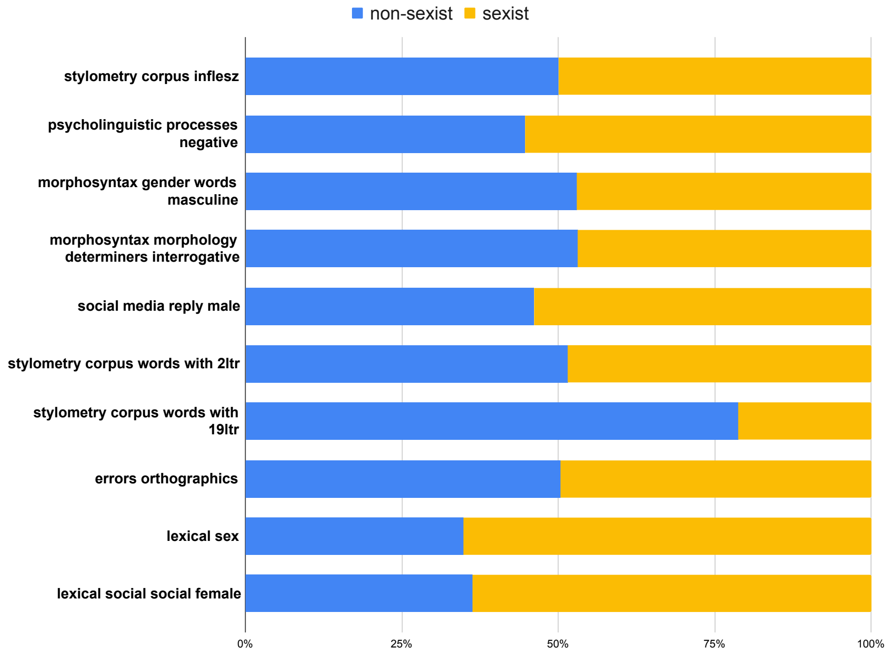

Fig. 1. Mutual Information of the top-ten linguistic features for the Spanish split for task 1: Sexism identification.

Second, the result of the second task (sexism categorisation) are depicted in Table 4.

For task 2 (see Table 4), our best result was achieved with the combination of BERT and LF with an F1-score of 53.62%, reaching position 18 in the official rank. Similarly to task 1, our best run is not far for the best result achieved by UPV with an F1-Score of 57.87%. However, contrary to the first task, our third run based on the ensemble model, achieved lower result than our second run. The first run, consisted on the linguistic features, achieved lower results falling below the baseline. In this task, however, it draw our attention the reliability of the LF were similar for Spanish and English, achieving an F1-score of 28.0% for Spanish and 27.21% for English.

Figure 2 depicts the Mutual Information of the linguistic features for the task 2 for the Spanish split compared by label. However, these results must be viewed with caution, due to the limited results of the linguistic features in isolation for the task sexism categorisation. We can observe that lexical words regarding sex

Table 4. Comparison of the UMUTeam with the best three runs and the baselines for misogyny categorisation task

Rank	Team	Accuracy	f1-score
1	AI-UPV_1	65.77	57.87
2	LHZ_1	65.09	57.06
3	SINAI_TL_1	65.27	56.67
18	UMUTEAM_2	61.70	53.62
23	UMUTEAM_3	59.11	52.40
51	SVM TFIDF (baseline)	52.22	39.50
56	UMUTEAM_1	29.05	28.12
62	Majority Class (baseline)	47.78	10.78

appears less frequently in sexism messages categorised as stereotyping and dominance. Out of the different sexist traits, we can notice that negative statements (*psycholinguistic processes negative*) appear most frequently in documents labelled as misogynist but without sexual violence. The rest of the features does not show significant differences among the sexist traits.

6 Conclusions

In this manuscript we have detailed the participation of the UMUTeam in the EXIST 2021 shared task with three runs that combined linguistic features with state-of-the-art transformers. We are very happy with the opportunity that we have been given to participate in these tasks and we hope to repeat it in the future. We are aware that the results achieved by the LF in isolation are below reliable baselines such as n-grams based on TF-IDF. We are currently evaluating the labels of the test set in order to detect weakness and to improve our pipeline. Moreover, we are currently implementing more advanced ensembles by training deep-learning models that learn from the predictions of each individual model. We also have learned another way to create sentence-fixed embeddings from fine-tuned BERT models that are more easy to combine with other kind of features.

For future research directions we observed that misogyny categorisation was performed as multiclass, in which all labels are considered mutually exclusive. However, we consider that it will be interesting to evaluate this sexist speech as a multi-label task. However, we are aware that this proposal would imply to relabel the dataset. Another research direction is to incorporate contextual features to the classification in order to provide a context to the documents.

7 Acknowledgments

This work was supported by the Spanish National Research Agency (AEI) through project LaTe4PSP (PID2019-107652RB-I00/AEI/10.13039/501100011033). In addition, José Antonio García-Díaz has been supported by Banco Santander and University of Murcia through the industrial doctorate programme.

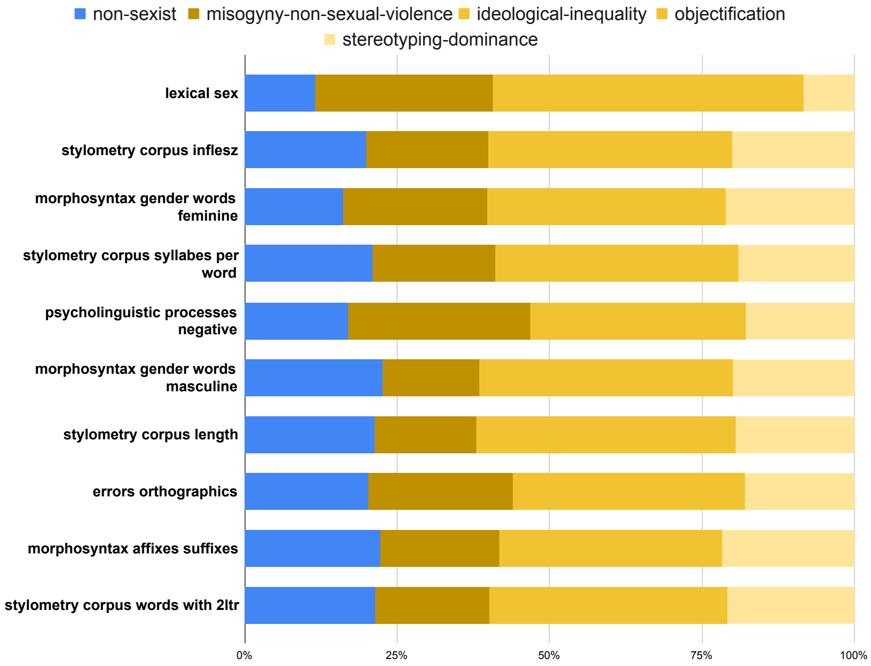

Fig. 2. Mutual Information of the top-ten linguistic features for the Spanish split for task 2: Sexism categorisation.

References

1. Basile, V., Bosco, C., Fersini, E., Debora, N., Patti, V., Pardo, F.M.R., Rosso, P., Sanguinetti, M., et al.: Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In: 13th International Workshop on Semantic Evaluation. pp. 54–63. Association for Computational Linguistics (2019)
2. Cañete, J., Chaperon, G., Fuentes, R., Pérez, J.: Spanish pre-trained bert model and evaluation data. PML4DC at ICLR **2020** (2020)
3. 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)
4. Fersini, E., Rosso, P., Anzovino, M.: Overview of the task on automatic misogyny identification at ibereval 2018. IberEval@ SEPLN **2150**, 214–228 (2018)
5. García-Díaz, J.A., Cánovas-García, M., Valencia-García, R.: Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in latin america. Future Generation Computer Systems **112**, 614–657 (2020). https://doi.org/10.1016/j.future.2020.06.019
6. García-Díaz, J.A., Cánovas-García, M., Colomo-Palacios, R., Valencia-García, R.: Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings. Future Generation Computer Systems **114**, 506 – 518 (2021). https://doi.org/10.1016/j.future.2020.08.032, http://www.sciencedirect.com/science/article/pii/S0167739X20301928
7. Montes, M., Rosso, P., Gonzalo, J., Aragón, E., Agerri, R., Álvarez-Carmona, M.Á., Álvarez Mellado, E., Carrillo-de Albornoz, J., Chiruzzo, L., Freitas, L., Gómez Adorno, H., Gutiérrez, Y., Jiménez Zafra, S.M., Lima, S., Plaza-de Arco, F.M., Taulé, M.: Proceedings of the iberian languages evaluation forum (iberlef 2021). In: CEUR workshop (2021)
8. Paredes-Valverde, M.A., Colomo-Palacios, R., Salas-Zárate, M.d.P., Valencia-García, R.: Sentiment analysis in spanish for improvement of products and services: a deep learning approach. Scientific Programming **2017** (2017)
9. del Pilar Salas-Zárate, M., Alor-Hernández, G., Sánchez-Cervantes, J.L., Paredes-Valverde, M.A., García-Alcaraz, J.L., Valencia-García, R.: Review of english literature on figurative language applied to social networks. Knowledge and Information Systems **62**(6), 2105–2137 (2020)
10. del Pilar Salas-Zárate, M., Paredes-Valverde, M.A., Rodriguez-García, M.Á., Valencia-García, R., Alor-Hernández, G.: Automatic detection of satire in twitter: A psycholinguistic-based approach. Knowledge-Based Systems **128**, 20–33 (2017)
11. Rodríguez-Sánchez, F., Carrillo-de Albornoz, J., Plaza, L.: Automatic classification of sexism in social networks: An empirical study on twitter data. IEEE Access **8**, 219563–219576 (2020)
12. 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)
13. Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: Liwc and computerized text analysis methods. Journal of language and social psychology **29**(1), 24–54 (2010)