# UNEDBiasTeam at IberLEF 2021's EXIST Task: Detecting Sexism Using Bias Techniques

Francisco-Javier Rodrigo-Ginés, Jorge Carrillo-de-Albornoz, and Laura Plaza

NLP & IR Group, UNED, 28040 Madrid, Spain
frodrigo@invi.uned.es; {jcalbornoz, lplaza}@lsi.uned.es

**Abstract.** Detecting and tackling sexist messages in social media is important for encouraging better behaviours in our society as well as to contribute to effective equality between men and women. In this paper we present our participation in the sEXism Identification in Social neTworks (EXIST) task at IberLEF'2021 [1]. Our approach to solve the task is based on considering the sexism as a subset of bias. Our work consisted in transferring lexical features commonly associated with bias, and analyzing how well they serve to detect sexism in social networks. The results show that these types of features do not have much statistical correlation with these types of short sexist messages.

**Keywords:** Natural Language Processing · Sexism detection · Bias detection.

CEUR-WS.org/Vol-2943/exist_paper20.pdf

## 1 Introduction

The emergence of social networks in recent years has allowed people from different countries and cultural backgrounds to communicate with each other freely using the same communication channels. This fact, which is a positive consequence of web technologies, also carries a series of disadvantages. The disinhibition generated by anonymity makes users feel free to say things that they would never say in person, including hateful and sexist messages [2].

For this reason, creating systems capable of automatically detecting sexist messages on social networks is very important for encouraging better behaviors in society and fight against discrimination and inequality. It is also important to analyze and detect sexism as it is presented in this task, identifying not only hateful messages, but also messages that discredit the feminist movement, that deny equality between men and women, or that present women as objects.

On the other hand, the study and interpretation of bias is a wide research field and it has been studied from many different perspectives. The bias in a text is not only found in words choice, but also in what information is omitted

*IberLEF 2021, September 2021, Málaga, Spain.*
Copyright © 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

and what is commissioned, in what labels are placed on the entities that appear in the text, and it may even be present in who read the text according to its socio-cultural context. In this work, we focus on studying bias from a lexical point of view.

According to the Oxford English Dictionary, sexism may be understood defined as any type of prejudice, stereotype or discrimination, generally against women, that is carried out on the basis of sex. If we compare this definition with the definition of bias (inclination or prejudice for or against a person or group, especially in a way considered unfair) we will realize that sexism is a subset of it. Our approach is based on transferring features commonly used in the task of bias and propaganda detection, and studying the applicability of these features with the detection of sexist messages published in social media.

In our approach, we study how well these extracted characteristics model sexism. To do this, we add them to several classic machine learning models based on TF-IDF, and a bidirectional Long Short-Term Memory (Bi-LSTM) model with a word-embeddings layer. The results obtained show that the correlation between the lexical characteristics extracted, and the class of the text (binary or multiclass) is minimal, resulting in very little or inexistant improvements in the learning step performance. The limitations encountered are discussed in Section 4.

The rest of this paper is structured as follows. In Section 2, we explain the data used in our system and the pre-processing done. Section 3 presents the details of the proposed systems. In Section 4 we present and analyze the results obtained in our experiments and in the EXIST competition. Finally, the paper concludes in Section 5 with conclusions and future work.

## 2 Data and Task description

The sEXism Identification in Social neTworks (EXIST) task consits on automatically identifing sexism content on social networks such as Twitter and Gab.com. The aim of this task is to detect sexism content in a broad sense, from explicit and offensive misogyny to other subtle expressions that involve implicit sexist expressions and behaviours [3].

The task is divided into two sub-tasks. The first sub-task is a binary classification in which the system have to classify the given text as "sexist" (it is sexist itself, describes a sexist situation or criticizes a sexist behaviour) or "non-sexist". The second sub-task aims to classify the text, once a message is classified as sexist, according to any of the following type of sexism:

- **Ideological and inequality**: the message discredits the feminist movement, rejects inequality between men and women, or presents men as victims of gender-based oppression.
- **Stereotyping and dominance**: the message expresses false ideas about women that suggest they are more suitable to fulfill certain roles, or inappropriate for certain tasks, or claims that men are somehow superior to women.

– **Objectification**: the message presents women as objects apart from their dignity and personal aspects, or assumes or describes certain physical qualities that women must have in order to fulfill traditional gender roles.
– **Sexual violence**: the message includes or describes sexual suggestions, requests for sexual favors or harassment of a sexual nature.
– **Misogyny non-sexual violence**: the message expresses hatred and violence towards women.

The dataset provided contains 11,345 instances of text in both English and Spanish. The texts were extracted from the social networks Twitter (tweets) and Gab (gabs). The training set contains 6,977 tweets while the test set contains 3,386 tweets and 982 gabs. The distribution between both languages has been balanced.

## 3 Using bias techniques to detect sexism

In this section we present the system pipeline, which consist in a text preprocessing step, feature engineering, and describe a learning step using both traditional ML techniques and novel deep learning approaches.

### 3.1 Data pre-processing

To assist the feature extraction step, and the TF-IDF computation, the following pre-processing has been applied to the text:

1. Converting the text to lowercase.
2. Removing punctuation and digits.
3. Removing hashtags and mentions.
4. Tokenizing the text.
5. Removing stop-words. The NLTK library has been used to obtain the stopwords in both English and Spanish.
6. Lemmatization. The Stanza library has been used, both for English and Spanish texts.

For the Bi-LSTM with a word embedding layer model developed, this preprocessing is not applied, only the text has been converted to lowercase

### 3.2 Feature engineering

For both languages, a custom lexicon of biased words is used. We have used the lexicon of sexist words used in [6] for texts in Spanish. For English texts we have replicated the method described in [6]: (1) we have selected five seed words for hateful speech toward women (*slut, whore, bitch, floozy, tramp*) from the Hatebase.org website, (2) we have used the GloVe word embeddings trained with tweets to alleviate data sparseness, and generate more terms, and (3) we have removed repeating terms, resulting in a lexicon of 48 terms.

Also, we have included Hurtlex [7], a multilingual lexicon of offensive, aggressive, and hateful words. The words are divided into the following 17 subsets: PS (negative stereotypes ethnic slurs), RCI (locations and demonyms), PA (professions and occupations), DDF (physical disabilities and diversity), DDP (cognitive disabilities and diversity), DMC (moral and behavioural defects), IS (words related to social and economic disadvantage), OR (plants), AN (animals), ASM (male genitalia), ASF (female genitalia), PR (words related to prostitution), OM (words related to homosexuality), QAS (with potential negative connotations), CDS (derogatory words), RE (felonies and words related to crime and immoral behaviour), SVP (words related to the seven deadly sins of the Christian tradition). We aim to know how these categories interact with the sexism categories given for the second sub-task. We provide a more detailed analysis in Section 4.

Finally, other features related to the text context have been extracted (i.e. the count of mentions and hashtags for each tweet/gab), along with the sentiment and the PoS tagging. In Table 1 we describe every feature extracted:

Table 1: Description of features extracted.

| Feature | Value | Description |
|---|---|---|
| Sentiment | [0, 1] | Sentiment value as labelled by NLTK toolkit for English messages, and the Spanish Sentiment Analysis library for the Spanish texts. |
| PoS Tagging | percentage | The ratio of a PoS tag or a bigram of PoS tags in a statement. |
| Number of mentions | [0, n] | The number of mentions included in the tweet/gab. |
| Number of hashtags | [0, n] | The number of hashtags included in the tweet/gab. |
| Biased words | [0, n] | The number of words in the statement that occur in the bias word lexicon |
| Biased words distance | [0, n] | The average distance amongst bias words in a the given text. |
| Hurtlex occurrences | [0, n] | The number of occurrences for each Hurtlex subset. |
| Report verbs | boolean | True if the given text contains at least one word from the report verb list [5]. |
| Implicative verbs | boolean | True if the given text contains at least one word from the implicative verb list [8]. |
| Assertive verbs | boolean | True if the given text contains at least one word from the assertive verb list [9]. |
| Factive verbs | boolean | True if the given text contains at least one word from the factive verb list [9]. |
| Positive words | boolean | True if the given text contains at least one word from the positive words list [10]. |

| | | |
|---|---|---|
| Negative words | boolean | True if the given text contains at least one word from the negative words list [10]. |
| Weak subjective words | boolean | True if the given text contains at least one word from the weak subjective words list [11]. |
| Strong subjective words | boolean | True if the given text contains at least one word from the strong subjective words list [11]. |
| Hedge words | boolean | True if the given text contains at least one word from the hedge words list [12]. |

### 3.3 Learning step

We have developed both traditional methods using TF-IDF features and deep learning based methods using word embeddings. In the following subsections we describe the classification systems in detail:

**Traditional Machine Learning methods.** We have opted for some classical ML classifiers such as Logistic Regression (LR), and Support Vector Machine (SVM), using the skLearn library, since they are widely used for this type of task [13], [14]. We have developed models based on TF-IDF attributes, based only on the bias features extracted, and models based on TF-IDF attributes (both with unigrams and unigrams + bigrams) along with the features extracted. No hyper-parameter tuning has been carried out, skLearn default values has been used.

**Bidirectional Long Short-Term Memory (Bi-LSTM).** We have experimented with some Deep Neural Network approaches such as Bi-LSTM that has been succcesfully used for NLP classification before [15] [16]. With Bi-LSTMs we aim to capture long range dependencies in texts [16].

We have developed a Bi-LSTM model whose first layer performs word embeddings. After this first layer, we add a dropout layer (0.3 dropout rate) and a fully-connected output layer with one neuron per predicted class. The Adam optimizer is used along with binary cross entropy as loss function for task 1, or categorical loss function for task 2. Besides, 10 epochs were executed for training the models for task 1, and 50 epochs for task 2.

Also, another model with a parallel Dense layer for the bias features extracted has been implemented. This model has been designed as showed in Fig. 1.

The word embeddings used in the embedding layers are the following:

- GloVe_Twitter_27B: 2B tweets, 27B tokens, 1.2M vocab, uncased, 200 dimensions vectors.
- GloVe_SBWC (Spanish Billion Word Corpus): 0.85M vocab, uncased, 300 dimensions vectors.
- FastText_SUC (Spanish Unannotated Corpora): 3B tokens, uncased, 300 dimensions vectors.

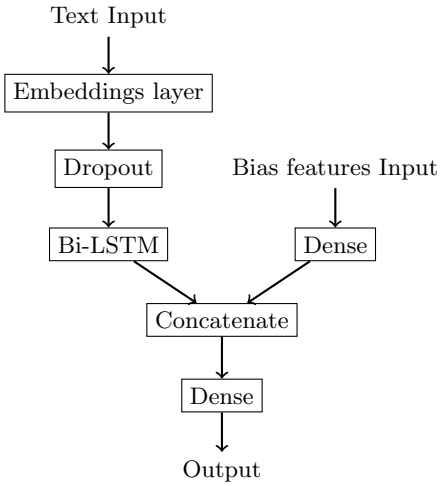


Fig. 1: Bi-LSTM model architecture

The models that have generated the runs sent to the organizers and that are analyzed in the following section have been trained with the GloVe_Twitter_27B embedding for English, and the FastText_SUC embedding for Spanish.

## 4 Results and Discussion

In this section we show the results obtained for both tasks and we do an analysis of them.

### 4.1 Official results

A total of 72 runs were submitted for task 1, our approaches were ranked $44^{\text{th}}$ for the Bi-LSTM model with a word embedding layer, $51^{\text{st}}$ for the Bi-LSTM model with a word embedding layer with bias features, and $65^{\text{th}}$ for the Logistic Regression model with only bias features.

Our results are summarized and compared with the best systems and the baseline system in the following table:

Table 3: Official results for task 1. Our results are highlighted in bold.

| Ranking | System | Accuracy | F1 score |
|---|---|---|---|
| 1 | task1_AI-UPV_1 | 0.7804 | 0.7802 |
| 2 | task1_SINAI_TL_1 | 0.78 | 0.7797 |
| 3 | task1_SINAI_TL_3 | 0.777 | 0.7757 |
| **44** | **Bi-LSTM - embeddings layer** | **0.7056** | **0.7056** |
| **51** | **Bi-LSTM - embeddings layer + bias** | **0.6905** | **0.6898** |
| 52 | Baseline SVM TF-IDF | 0.6845 | 0.6832 |
| **65** | **LR with bias features** | **0.543** | **0.5359** |
| 66 | Majority Class | 0.5222 | 0.3431 |

A total of 63 runs were submitted for task 2, our approaches were ranked $37^{\text{th}}$ for the Bi-LSTM model with a word embeddings layer and bias features, $39^{\text{st}}$ for the Bi-LSTM model with a word embeddings layer, and $57^{\text{th}}$ for the Logistic Regression model with only bias features.

Our results are summarized and compared with the best systems and the baseline system in the following table:

Table 4: Official results for task 2. Our results are highlighted in bold.

| Ranking | System | Accuracy | F1 score |
|---|---|---|---|
| 1 | task2_AI-UPV_1 | 0.6577 | 0.5787 |
| 2 | task2_LHZ_1 | 0.6509 | 0.5706 |
| 3 | task2_SINAI_TL_1 | 0.6527 | 0.5667 |
| **37** | **Bi-LSTM - embeddings layer + bias** | **0.5797** | **0.4704** |
| **39** | **Bi-LSTM - embeddings layer** | **0.5689** | **0.4621** |
| 51 | Baseline SVM TF-IDF | 0.5222 | 0.395 |
| **57** | **LR with bias features** | **0.4444** | **0.165** |
| 62 | Majority Class | 0.4778 | 0.1078 |

As we can see in both cases, the incorporation of bias features is not statistically significant for the improvement or deterioration of the models. Furthermore, the Logistic Regression method with only bias features performs worse than the baseline method, and just above selecting the majority class for each prediction in both tasks.

### 4.2 Extended results

In this subsection we show the results obtained for systems that were not sent to the competition. We also analyze the difference in classification performance according to the language of the texts. All these results have been obtained in the test set.

Table 5: Extended results for Task 1. Best results highlighted in bold.

| System | Accuracy English | F1 score English | Accuracy Spanish | F1 score Spanish |
|---|---|---|---|---|
| LR – TF-IDF | 0.7020 | 0.7019 | 0.7019 | 0.7007 |
| LR – Bias features | 0.5693 | 0.5686 | 0.5162 | 0.4939 |
| LR – Bias + TF-IDF (unigrams) | 0.6997 | 0.6997 | 0.6796 | 0.6786 |
| LR – Bias + TF-IDF (uni + bigrams) | **0.7024** | **0.7023** | 0.6866 | 0.6862 |
| SVM – TF-IDF | 0.6889 | 0.6886 | 0.6801 | 0.6766 |
| SVM – Bias features | 0.5525 | 0.5254 | 0.5593 | 0.5571 |
| SVM – Bias + TF-IDF (unigrams) | 0.6881 | 0.6873 | 0.6974 | 0.6969 |
| SVM – Bias + TF-IDF (uni + bigrams) | 0.6805 | 0.6804 | 0.7010 | 0.7008 |
| Bi-LSTM – Word_Embeddings | 0.6825 | 0.6840 | **0.7292** | **0.7292** |
| Bi-LSTM – Bias + Word_Embeddings | 0.6857 | 0.6849 | 0.6954 | 0.7278 |

Table 6: Extended results for Task 2. Best results highlighted in bold.

| System | Accuracy English | F1 score English | Accuracy Spanish | F1 score Spanish |
|---|---|---|---|---|
| LR – TF-IDF | 0.5569 | 0.3601 | 0.5792 | 0.3858 |
| LR – Bias features | 0.4212 | 0.1785 | 0.4681 | 0.1384 |
| LR – Bias + TF-IDF (uni) | 0.5516 | 0.3822 | 0.5769 | 0.3893 |
| LR – Bias + TF-IDF (uni + bigrams) | 0.5489 | 0.3813 | 0.5792 | 0.3844 |
| SVM – TF-IDF | 0.4841 | 0.3788 | 0.5611 | 0.4088 |
| SVM – Bias features | 0.4755 | 0.1091 | 0.4769 | 0.1486 |
| SVM – Bias + TF-IDF (uni) | 0.5412 | 0.3615 | 0.5688 | 0.3621 |
| SVM – Bias + TF-IDF (uni + bigrams) | 0.5446 | 0.3662 | 0.5691 | 0.3606 |
| Bi-LSTM – Word_Embeddings | 0.5208 | 0.4240 | **0.6181** | **0.5038** |
| Bi-LSTM – Bias + Word_Embeddings | **0.5571** | **0.4501** | 0.6028 | 0.4846 |

If we look at the results, in both cases the models that incorporate bias features perform equally or slightly better than their analogues without these characteristics, while for the results in Spanish, the models that incorporate the bias features worsen significantly. We believe this may be due to the fact that in the process of translating the lexicons from English to Spanish, we have lost some relevant information.

### 4.3 Discussion

The results obtained show that the correlation between the bias characteristics extracted, and the class of the text (binary or multiclass) do not behave significantly better than other models not using such characteristics. We believe that the use of these features is limited by the difference between the language

for which these lexicons were created and the language being analyzed (formal vs. informal / slang), having to literally translate the lexicons from English to Spanish, and the appearance of unintended bias.

Furthermore, as expected, the models based on deep learning techniques perform better than both of the classic ML models implemented (LR and SVM). This improvement is most clearly seen in subtask 2 (multiclass classification).

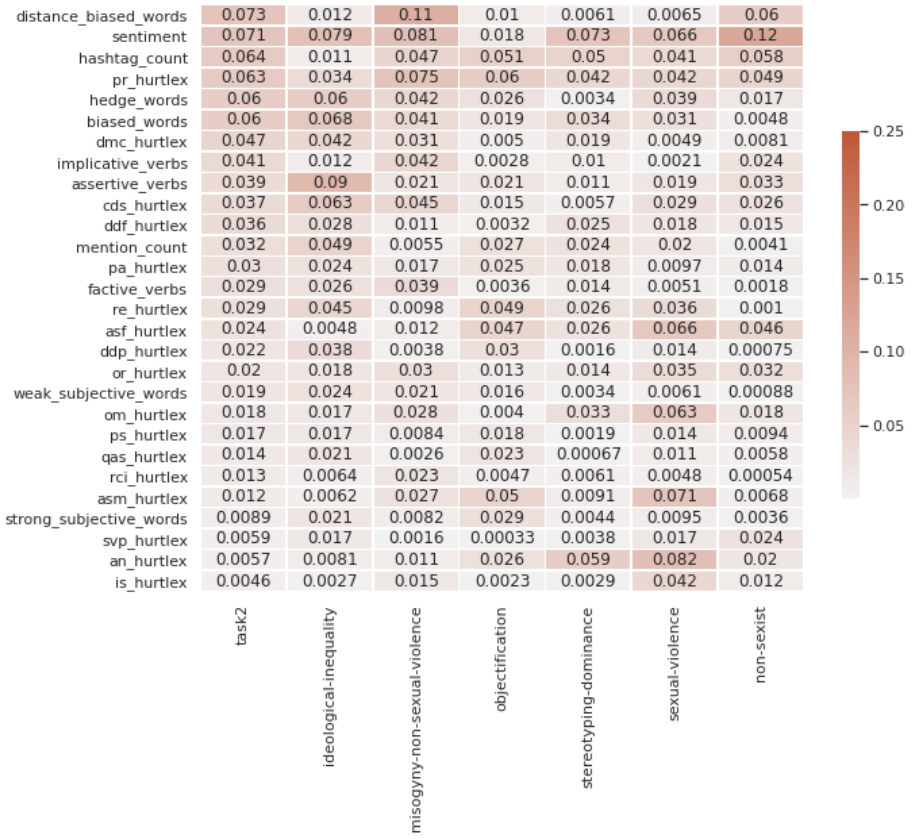


Fig. 2: Pearson correlation matrix between features extracted and task 2 labels. Please note that the coefficients are shown in absolute value in order to sort by correlation with the task 2 column.

Sexism does not only have to be latent in hateful messages, this is seen in the categories described by the organizers for task 2. In fact, if we look at the following Pearson correlation matrix, we will realize that all the features extracted from Hurtlex have a higher correlation with the label sexual-violence to a greater extent. It is logical since Hurtlex is a lexicon of offensive, aggressive

and/or hateful words; and this label is the one that has the most relationship with this type of words. If we look at the subsets of Hurtlex that most correlate with this category are: AN (animals), ASM and ASF (male and female genitalia), OM (words related to homosexuality), and PR (words related to prostitution).

Furthermore, we believe that our models are heavily affected by unintended bias. This problem leads the models to associate unreasonably high sexist scores to a non-sexist text only because it contains certain terms, called identity terms [17], [18]. By categorizing texts that include this type of identity terms, which in our case are aggressive and hateful terms, as a sexist message we generate a large number of false positives.

## 5 Conclusions and Future work

In this paper, we present the systems we have developed as part of our participation in the EXIST competition. Specifically, we have participated in both sub-tasks proposed. In order to solve these tasks, we have implemented classical Machine Learning models and novel Deep Learning methods such as a Bi-LSTM, incorporating lexical features from bias detection and offensive lexicons.

We have found that adding bias features to any model do not make it behave significantly better than the same models not using those features. These small performance improvements are canceled in the classification of Spanish texts due to the loss of information produced by translating the bias lexicons with an automatic translation system.

Our next steps will focus on exploring more features from others lexicons related to bias understanding (misinformation, hoaxes, propaganda), and study how to translate better the current English lexicons to Spanish. Also, we aim to investigate different bias mitigation strategies.

**Acknowledgements.** 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. IberLEF 2021 proceedings can be tentatively referred to as: 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.
2. Wright, Michelle F.: The relationship between young adults' beliefs about anonymity and subsequent cyber aggression. Cyberpsychology, Behavior, and Social Networking, 858–862 (2013)

3. Francisco Rodríguez-Sánchez, Jorge Carrillo-de-Albornoz, Laura Plaza, Julio Gonzalo, Paolo Rosso, Miriam Comet, Trinidad Donoso. Overview of EXIST 2021: sEXism Identification in Social neTworks. Procesamiento del Lenguaje Natural, vol 67, septiembre 2021.
4. Hube, C., and Fetahu, B.: Detecting biased statements in wikipedia. In Companion Proceedings of the Web Conference, 1779–1786 (2018).
5. Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky: Linguistic Models for Analyzing and Detecting Biased Language. ACL, 1650–1659 (2013).
6. Plaza-Del-Arco, F. M., Molina-González, M. D., Ureña-López, L. A., and Martín-Valdivia, M. T.: Detecting misogyny and xenophobia in Spanish tweets using language technologies. ACM Transactions on Internet Technology (TOIT), 20(2), 1–19 (2020).
7. Bassignana, E., Basile, V., and Patti, V.: Hurtlex: A multilingual lexicon of words to hurt. In 5th Italian Conference on Computational Linguistics (CLiC-it 2018), Vol. 2253, 1–6 (2018).
8. Lauri Karttunen: Implicative verbs. Language (1971).
9. Joan B Hooper: On assertive predicates. Indiana University Linguistics Club (1974).
10. Bing Liu, Minqing Hu, and Junsheng Cheng: Opinion observer: analyzing and comparing opinions on the web. In Proceedings of the 14th international conference on World Wide Web, 342–351 (2005).
11. Ellen Riloff and Janyce Wiebe: Learning extraction patterns for subjective expressions. In Proceedings of the 2003 conference on Empirical methods in natural language processing. Association for Computational Linguistics, 105–112 (2003).
12. Ken Hyland: Metadiscourse. Wiley Online Library (2005).
13. Fares, M. Fares, S. Oepen, and E. Velldal: Word vectors, reuse, and replicability: Towards a community repository of large-text resources, In Proc. NoDaLiDa/WS, 271–276 (2017).
14. H. Liu, F. Chiroma, and E. Haig: Identification and classification of misogynous tweets using multi-classifier fusion, I In Proc. IberEval, 268–273 (2018).
15. E. Shushkevich and J. Cardiff: Classifying misogynistic tweets using a blended model: The AMI shared task in IBEREVAL 2018, In Proc. IberEval, 255–259 (2018).
16. Rodríguez-Sánchez, F., Carrillo-de-Albornoz, J., and Plaza, L.: Automatic Classification of Sexism in Social Networks: An Empirical Study on Twitter Data. IEEE Access, 8, 219563–219576 (2020).
17. Nozza, D., Volpetti, C., and Fersini, E.: Unintended bias in misogyny detection. In IEEE/WIC/ACM International Conference on Web Intelligence, 149–155 (2019).
18. Garrido-Muñoz, I., Montejo-Ráez, A., Martínez-Santiago, F., and Ureña-López, L. A.: A Survey on Bias in Deep NLP. Applied Sciences, 3184 (2021).