MultiAzterTest@Exist-IberLEF 2021: Linguistically Motivated Sexism Identification

Kepa Bengoetxea[1 [0000−0002−0289−6897]] and Itziar Gonzalez-Dios[1 [0000−0003−1048−5403]]

Ixa Group, HiTZ center / University of the Basque Country (UPV/EHU)
{kepa.bengoetxea,itziar.gonzalezd}@ehu.eus

Abstract. Identifying sexism in social networks is the focus of the EXIST-IberLEF 2021 shared task. By participating in this task, the aim of the MultiAzterTest team is to see if linguistically motivated features can help in the detection of sexism. That is why, we present the three approaches: i) an approach based on language models, ii) an approach based on linguistic and stylistic features + machine learning classifiers and iii) an approach combining the previous approaches. The language model approach obtains the best results in the test data. However, the approaches that use linguistic and stylistic features offer more interpretability.

Keywords: Sexism detection · Exist-IberLEF · Language Models · Linguistic features

1 Introduction

Sexism is defined by the Oxford English Dictionary as "prejudice, stereotyping or discrimination, typically against women, on the basis of sex". Sexism, moreover, can be classified as indirect sexism, sexual and physical [26] and categorized as in the Exist-IberLEF shared task [25] as ideological and inequality, stereotyping and dominance, objectification, sexual violence and misogyny and non-sexual violence.

The Natural Language Processing (NLP) community has focused on detecting hate speech [13, 22], and abusive language and offensive language [7] among others but also on the their related outcomes such as misogyny [20] or racism [27]. Sexism has also been addressed and Rodríguez-Sánchez et al. experiment with user, network, and text-based features, machine learning classifiers (logistic regression, support vector machine and random forest), deep recurrent neural networks (BI-LSTM) and transformer-based language models (BERT) [24] to detect it.

In this paper, we test MultiAzterTest-Social (MATS) in the task of detecting sexism in the context of the EXIST 2021 Shared Task [25], a shared task at

IberLEF 2021, September 2021, Málaga, Spain.

IberLef 2021 [1]. MATS is a version of the MultiAzterTest tool [4], which is the trilingual version of AzterTest [5]. MultiAzterTest and AzterTest are open-source NLP based tools and web services for text stylometrics and readability assessment. In addition to the linguistic and stylistic features, MATS includes features to analyse social media texts inspired by Fersini et al. [11] and other improvements. By participating in this shared task, we want to see if a tool that outperforms the state-of-the art results in readability assessment can be applied to other classification tasks, where texts are shorter, include more subjective information and colloquial and informal speech. Linguistic based features have been used to detect fake news [9] and text features have also been taken into account for sexism detection [24], but to other knowledge this is the first time that more than 150 features are taken into account for this task. The aim of using a linguistically motivated tool is to give explanations to the predictions and to be able to analyse the linguistic characteristics of sexism.

This paper is structured as follows: in Section 2 we introduce MultiAzterTest-Social, in Section 3 we describe our approaches and the experimental set-up, in Section 4 we present the results and we conclude and outline the future work in 5.

2 MultiAzterTest-Social

MultiAzterTest-Social is an improvement of MultiAzterTest [4]. MultiAzterTest analyses more than 125 linguistic and stylistic features in Basque (125 features) English (163 features), and Spanish (141 features). Following, we briefly explain how MultiAzterTest works:

- **Preprocessing:** This step carries out all the necessary analysis in raw texts in order to be processed. This includes multilingual parsing (in our case Stanza [23]), syllable splitting, and stopword removing.
- **Linguistic and stylistic profiling:** Based on the previous text analysis, this step calculates the linguistic and stylistic features. These features are grouped in the following types: descriptive and raw features, lexical diversity, classical readability formulae, word frequencies, vocabulary knowledge, morphological information, syntax, semantic information, semantic overlap (semantic similarity), referential cohesion (overlaps) and logical cohesion (connectives). There are five types of indicators: absolute numbers, mean, standard deviation, incidence and ratios.
- **Classification:** Based on the linguistic and stylistic features, a machine learning classifier is applied. This classifier varies depending on the task. In the case of readability assessment, for example, support vector machines seem to be the most adequate.

In order to analyse social media text, we have added more features to Multi-AzterTest and, this way, we have adapted it to MultiAzterTest-Social. Some of the new features for social media are inspired in the features presented by Fersini et al. [11] to profile misogynists. Advanced morpho-syntactic, and named-entities

are based on other readability assessment works e.g. [14], we have created the sentiment analysis and the abusive term features and we have added more descriptive features (descriptive+). Following, we explain the new features:

- **Descriptive+:** We analyse the number of words and sentences per tweet; the number of numerical expressions, its incidence per 1000 words and the ratio of numbers per tweet and per sentence; the number and incidence of each punctuation mark (colon, exclamation mark...); and the number and incidence of special characters.
- **Advanced morpho-syntactic:** We calculate the number and incidence per 1000 words of the types of determinants (definite, indefinite), adjectives (comparative, superlative), pronouns (person and number), causal and intentional verbs and particles, adverbial and prepositional phrases and the ratios of causal/intentional particles to causal/intentional verbs.
- **Named entities:** Stanza's base version detects 4 named entity types (Person, Location, Organisation and Miscellaneous). We calculate the mean of all the entities per sentence and the incidence per 1000 words; the ratio of entities per nouns; and each entity type per all the entities, per sentence and its the incidence per 1000 words.
- **Social media:** These features include the number and ratio of emojis per tweet and sentence; the number and incidence of hashtags/mentions/stretched words, ratio of each of them per sentence and tweet; the number and incidence of mentions, ratio of hashtags per sentence and tweet; and percentage of capital letters per sentence.
- **Sentiment analysis:** We calculate the average positive, negative, neutral or compound score per sentence based on sentiment intensity analyser from VADER [16], the number of positive, negative, and neutral emojis according to the Emoji sentiment lexicon [19] and average sentiment score per sentence.
- **Abusive terms**: We include features for profane words, abusive words and hurt words based on the following resources: i) Luis von Ahn's Research Group's Offensive/Profane Word List [2], ii) the Lexicon of Abusive Words [28], and iii) HurtLex, the multilingual lexicon of words to hurt [3]. Although HurtLex classifies the words in different categories, we take all of them together. We calculate the number, the incidence and the ratio per sentence of the profane, abusive and hurt words.

In Table 1 we show the number of new features MATS analyses. Taking into account the features MultiAzterTest calculates, in this work we have used 280 features for English and 244 for Spanish.

3 Approaches and Experimental Set-up

In this section we present the experiments carried out for the task 1: Sexism Identification. The dataset we have used has been provided by the organisers [25]. The results are calculated using accuracy, as distribution between sexist and non-sexist categories is balanced.

Table 1. The number of linguistic and stylistic features added in English and Spanish to analyse social media text

Feature type	English	Spanish
Descriptive+	20	20
Advanced morpho-syntactic	42	42
Named entities	15	15
Social media	23	23
Sentiment Analysis	8	0
Abusive terms	9	3
Total	117	103

In our experiments, we have tested three approaches: i) a language model (LM), ii) the features of MultiAzterTest-Social together with a machine learning classifier (henceforth, MATS-Sexism) and iii) a combination of the LM and the MATS-Sexism approach.

3.1 Language Model Approach

The LM approach uses the Bidirectional Encoder Representation from Transformer (BERT) [10], exactly the bert-base-uncased model, pre-trained on the BooksCorpus [30] and English Wikipedia for English and BETO (bert-base-spanish-wwm-uncased) [8] for Spanish. Both models are provided by HuggingFace [29]. We have decided to use this approach because BERT achieves state of the art results on many NLP tasks.

This is our experimental setting: we have truncated all texts that had more than 200 tokens and we have added two tokens to mark the beginning and the end of the sequence to each input text, [CLS] and [SEP] respectively. We have padded texts shorter than 200 tokens with zeroes. We have not performed any text augmentation or pre-processing besides standard byte-pair encoding. We have used the PyTorch framework to create our model.

On top of BERT, we have probed with two sequential models: i) a dropout layer to fight overfitting. The dropout probability was set equal to 0.1. On top of the dropout layer, we have added a linear layer and sigmoid activation function. The input dimension of the linear layer was 768 and the output 2 (equal to the number of classes); ii) a linear layer, ReLU activation function and linear layer model. The input dimension of the first linear was 768 and the output 50, and the input dimension of the second linear was 50 and the output 2 (equal to the number of classes).

We have used the cross-entropy loss function for each of the outputs.

We have trained the model in the Google Colaboratory framework. We have split the training data into 80 % for train and 20 % for validation. The training batch size was made equal to 32 and the model was trained for 10 epochs using early stopping technique. We have obtained the best result in the validation data after running 4 epochs, setting the tweet length to 200, and the learning rate to 5e-5 with linear-ReLU-linear sequential model and the Adam optimizer [17].

3.2 Approach Based on Linguistic Features and Machine Learning

The second approach, the MATS-Sexism approach, consists of the outputs of the tool plus a classical machine learning classifier. In order to know which is the most adequate classifier, we have tested the Sequential Minimal Optimization (SMO) [21], Random Forest (RF) [6] and Simple Logistics (SL) [18] classifiers. We have also carried out feature selection with the ten most predictive features according to WEKA [15] based InfoGain attribute evaluator (Table 3), and in the case of SMO, we also have reduced the number of features to 125 and 75. All these preliminary experiments have been done with 10 fold cross-validation.

In Table 2 we present the results of the MATS-Sexism approach with different features and classifiers on the training data. As it happens in readability assessment, SMO is the best classifier [4] and, therefore, SMO will be the classifier of MATS-Sexism. We also see, contrary to what happens in readability assessment, that feature selection and feature reduction are not competitive.

Table 2. MATS-Sexism results in the training data (run2).

Method	**English**	**Spanish**
MATS-Sexism-SMO-All	**68.36**	**65.32**
MATS-Sexism-SMO-125	66.23	63.48
MATS-Sexism-SMO-75	66.00	61.98
MATS-Sexism-SMO-Top10	54.39	58.88
MATS-Sexism-RF-All	62.31	59.95
MATS-RF-Top10	54.45	53.94
MATS-Sexism-SL-All	63.82	61.59
MATS-Sexism-SL-Top10	55.24	59.08

Before continuing with the approaches, let us analyse the most predictive features presented in Table 3 from a linguistic and stylistic point of view. Four of the most predictive features for English are descriptive (word, lemma and syllable length), there are 2 semantic similarity features, and one of the social media features (the percentage of capital letters), sentiment analysis (the VADER compound score), the Flesch readability formula [12] and word frequencies (minimum word frequency).

In the case of Spanish, it is remarkable that 4 features are related to rare words. This can imply that ii) infrequent words have been used or ii) that the spelling of the words was not the correct one and they have not been correctly analysed. The importance of the hashtags is also noticeable (4 features). The use of the first person pronouns and the unclassified miscellaneous named entities play also a role. Finally, 6 out of the 10 features were not in MultiAzterTest and come from the update to social. This shows the validity of the new added features.

Table 3. Top10 features according to InfoGain in English and Spanish.

English	Spanish
word length (std)	number of different rare words (incidence)
lemma length (std)	number of rare words (incidence)
percentage of capital letters	hashtag ratio per sentence
word length without stopwords (std)	hashtag ratio per tweet
VADER compound score per sentence (mean)	hashtag incidence
semantic similarity between adjacent sentences (mean)	rare distint content words (mean)
number of syllables per word (std)	number of hashtags
semantic similarity between all possible pairs of sentences in a paragraph (mean)	rare content words (mean)
Flesch	number of first person pronouns
minimum word frequency per sentence	MISC named-entities per sentence (mean)

3.3 Combination Approach

The third approach is a combination of the results of the LM and MATS-Sexism. To combine the results, we have two options: i) label as sexist if one of the tools tags a tweet as sexist or ii) label as sexist only if both tools consider that a tweet is sexist. We have decided to implement the second option (only if LM and MATS-Sexism agree) in order to give more precision to our predictions (although the official evaluation metric is accuracy). We think that in subjective tasks striving for precision can avoid doing harm.

4 Results in Test Data

In this section we present the results in the test data as provided by the organisers. In total, 72 systems were evaluated in Task1. In Table 4 we present the results of our three approaches together with the baseline results (TF-IDF+SVM), also provided by the organisers.

The LM approach obtains the best results in all the settings: both languages, and English and Spanish on their own. The combination stays in the middle and the MATS-Sexism approach is the worst. Only the LM approach is above the baseline. It is remarkable that all the approaches perform similarly in all the settings: if we rounded numbers, the accuracy of the LMs will be 0.77 in both Spanish and English and also in Spanish and English separately. The MATS-Sexism approach has an accuracy of 0.59-0.60. The combination has more variation, from 0.65 to 0.67. In general, we can say that there is a difference of 17 point between LM and MATS-Sexism, and a difference of 10 between the LM and the combination and 5-8 points between the combination and MATS-Sexism.

Looking at the results, we see that approaches based on distributional information such as the languages models or TF-IDF are very effective when working

Table 4. Results in the test data

Lang.	Method	Accuracy	Precision	Recall	$F1$	Ranking
All	Baseline	0.6845	0.6943	0.6888	0.6832	51
All	LM	**0.7740**	**0.7741**	**0.7727**	**0.7731**	6
All	MATS-Sexism	0.5948	0.5983	0.5974	0.5944	64
All	Comb	0.6582	0.6951	0.6670	0.6482	60
EN	Baseline	0.6889	0.6934	0.691	0.6886	-
EN	LM	**0.7717**	**0.7753**	**0.7683**	**0.7691**	-
EN	MATS-Sexism	0.5996	0.6049	0.6032	0.5989	-
EN	Comb	0.6481	0.6789	0.6571	0.6398	-
ES	Baseline	0.6801	0.6972	0.6853	0.6766	-
ES	LM	**0.7764**	**0.7764**	**0.7769**	**0.7763**	-
ES	MATS-Sexism	0.5898	0.5920	0.5915	0.5897	-
ES	Comb	0.6685	0.7129	0.6770	0.6566	-

with short texts and features that take into account syntactic and discursive information may not be so helpful in these classification tasks. Indeed, they worsen the accuracy.

5 Conclusion and Future Work

In this paper we have presented the results of the MultiAzterTest team at the first task of the Exist-IberLEF 2021 shared task. The aim of these experiments was to see if linguistically motivated features could identify sexism. Looking at our results, we see that distributional approaches are very efficient and linguistic features are not so important when classifying short texts.

However, in the future, the outputs of the linguistically motivated approach can be used to interpret the characteristics of sexism. It would be also an interesting work to test these approaches in longer texts.

Acknowledgments

We acknowledge the following projects: DeepText (KK-2020/00088), DeepReading RTI2018-096846-B-C21 (MCIU/AEI/FEDER, UE), BigKnowledge for Text Mining, BBVA and IXA taldea, A motako ikertalde finkatua (IT1343-19).

References

1. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021. CEUR Workshop Proceedings (2021)
2. von Ahn, L.: Offensive/profane word list. https://www.cs.cmu.edu/ biglou/resources/, accessed: 2021-05-14
3. Bassignana, E., Basile, V., Patti, V.: Hurtlex: A multilingual lexicon of words to hurt. In: 5th Italian Conference on Computational Linguistics, CLiC-it 2018. vol. 2253, pp. 1–6. CEUR-WS (2018)

4. Bengoetxea, K., Gonzalez-Dios, I.: MultiAzterTest: a Multilingual Analyzer on Multiple Levels of Language for Readability Assessment. Manuscript from author (2021)
5. Bengoetxea, K., Gonzalez-Dios, I., Aguirregoitia, A.: AzterTest: Open Source Linguistic and Stylistic Analysis Tool. Procesamiento del Lenguaje Natural **64**, 61–68 (2020)
6. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001)
7. Caselli, T., Basile, V., Mitrović, J., Kartoziya, I., Granitzer, M.: I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In: Proceedings of The 12th Language Resources and Evaluation Conference. pp. 6193–6202 (2020)
8. Cañete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., Pérez, J.: Spanish Pre-Trained BERT Model and Evaluation Data. In: PML4DC at ICLR 2020 (2020)
9. Choudhary, A., Arora, A.: Linguistic feature based learning model for fake news detection and classification. Expert Systems with Applications p. 114171 (2020)
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186 (2019)
11. Fersini, E., Nozza, D., Boifava, G.: Profiling italian misogynist: An empirical study. In: Proceedings of the Workshop on Resources and Techniques for User and Author Profiling in Abusive Language. pp. 9–13 (2020)
12. Flesch, R.: A new readability yardstick. Journal of applied psychology **32**(3), 221 (1948)
13. Fortuna, P., Soler, J., Wanner, L.: Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 6786–6794 (2020)
14. Gonzalez-Dios, I., Aranzabe, M.J., Díaz de Ilarraza, A., Salaberri, H.: Simple or complex? assessing the readability of basque texts. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. pp. 334–344. DCU and ACL, Dublin, Ireland (Aug 2014), https://www.aclweb.org/anthology/C14-1033
15. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter **11**(1), 10–18 (2009)
16. Hutto, C., Gilbert, E.: Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 8 (2014)
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
18. Landwehr, N., Hall, M., Frank, E.: Logistic model trees **95**(1-2), 161–205 (2005)
19. Novak Kralj, P., Smailović, J., Sluban, B., Mozetič, I.: Sentiment of emojis. PloS one **10**(12), e0144296 (2015)
20. Pamungkas, E.W., Basile, V., Patti, V.: Misogyny Detection in Twitter: a Multilingual and Cross-domain study. Information Processing & Management **57**(6), 102360 (2020)
21. Platt, J.: Fast training of support vector machines using sequential minimal optimization. In: Schoelkopf, B., Burges, C., Smola, A. (eds.) Advances in Kernel Methods - Support Vector Learning. MIT Press (1998), http://research.microsoft.com/~jplatt/smo.html

22. Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., Patti, V.: Resources and benchmark corpora for hate speech detection: a systematic review. Language Resources and Evaluation pp. 1–47 (2020)
23. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D.: Stanza: A python natural language processing toolkit for many human languages. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 101–108 (2020)
24. 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)
25. 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)
26. Sharifirad, S., Jacovi, A.: Learning and understanding different categories of sexism using convolutional neural network's filters. In: Proceedings of the 2019 Workshop on Widening NLP. pp. 21–23. Association for Computational Linguistics, Florence, Italy (Aug 2019), https://www.aclweb.org/anthology/W19-3609
27. Waseem, Z.: 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. pp. 138–142 (2016)
28. Wiegand, M., Ruppenhofer, J., Schmidt, A., Greenberg, C.: Inducing a lexicon of abusive words–a feature-based approach. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 1046–1056 (2018)
29. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al.: Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771 (2019)
30. Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S.: Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In: Proceedings of the IEEE international conference on computer vision. pp. 19–27 (2015)