Automatic Detection of Sexism in Social Media with a Multilingual Approach

Lütfiye Seda Mut Altın and Horacio Saggion

LaSTUS-TALN Research Group, DTIC
Universitat Pompeu Fabra
C/Tànger 122-140, 08018 Barcelona, Spain
lutfiyeseda.mut01@estudiant.upf.edu, horacio.saggion@upf.edu

Abstract. This paper describes the participation of the LaSTUS/TALN team in the shared task EXIST: sEXism Identification in Social neTworks at IberLEF 2021. EXIST promotes the development of classification systems able to spot sexism in social media posts in English and Spanish. More specifically a binary classification task is proposed to classify posts in either sexist or not sexist, while a multi-class task challenges systems to identify the type of sexism present in the post. We address the two tasks with a transformers-based system applying a multi-lingual BERT model. The paper reports and discusses the results obtained by our team.

Keywords: Natural Language Processing · Multilingual-BERT · sexism identification.

1 Introduction

Sexism is defined as actions or attitudes that discriminate against people based solely on their gender [1]. Several studies show that exposure to sexist incidents have a concerning impact on the exposed people (see for example the report from parliamentarians in Europe [2]). Fox et al. [3] showed that higher interactivity with sexist hashtags on social media led to more negative evaluations of female job candidates. More broadly, it has been pointed out that experiencing sexist incidents including gender role stereotypes, prejudice, demeaning comments and sexual objectification has impacts on women's psychological well-being [4]. Moreover, social media provides a space for unregulated gender-based cyber-hate (targeting women in a particular group) in a way that traditional media does not [5]. Due to its consequences and the difficulty of moderation of social media content, automatic detection of sexist language in social media gains a lot of attention.

IberLEF 2021, September 2021, Málaga, Spain.

Detecting sexist language automatically on social media can be difficult, particularly due to its subtle forms and variations.

EXIST: sEXism Identification in Social neTworks shared task aims at identification and classification of sexism in social media posts in English and Spanish [6]. The shared task consists of two sub-tasks [7]:

- **TASK 1 Sexism Identification**: A binary classification task with labels and corresponding examples in English as shown below:
  - **Sexist**: "@fallynfen That is very short skirt"
  - **Non-Sexist**:"@HelenKarenWest This is a super news for the WomensRights"
- **TASK 2 Sexism Categorization**: A much granular categorization of sexism. In particular, five categories as given below along with corresponding examples:
  - **Ideological and Inequality:** This category refers to discrediting the feminist movement, rejecting inequality between men and women or presenting men as victims of gender-based oppression as shown in this example: *"@limitlessleila Ah like the discussions here, so calm. Why? Because there's no feminazi here."*
  - **Stereotyping and Dominance:** Expression of false ideas about women suggesting they are more suitable to fulfill certain roles (mother, wife, family caregiver, faithful, tender, loving, submissive, etc.), or inappropriate for certain tasks (driving, hard-work, etc), or claims that men are somehow superior to women. As the following example illustrates: *"@IWontActMyAge you should learn how to cook for your future husband."*
  - **Objectification:** The text 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 (compliance with beauty standards, hyper-sexualization of female attributes, women's bodies at the disposal of men, etc.). As an example consider: *"@traps4live She is VERY cute...love the short skirt."*
  - **Sexual Violence:** Sexual suggestions, requests for sexual favors or harassment of a sexual nature (rape or sexual assault) are made as illustrated in: *"@dorsenfIfr @nypost You look like a prostitute."*
  - **Misoginy and Non-Sexual Violence:** The text expresses hatred and violence towards women as in: *"@shawnmilrad I mean it's so much shorter just to say I hate women"*

In Section 2 of the paper we present an overview of the related work. Then, in Section 3, we provide information about our model and the data used for training it. In Section 4, we present the results and discuss the performance. Lastly, in Section 5 we give the conclusions.

2 Related Work

Although there is a considerable amount of research in the area of offensive language detection from different aspects such as hate-speech or cyberbullying, the number of studies particularly on sexism is rather low. Previous work on automatic detection of mysogyny include classical machine learning models like Naive Bayes, Support Vector Machines, Logistic Regression as well as ensembles of these, and more recently deep neural networks such as Long Short Term Memory (LSTM), Convolutional Neural Networks (CNN) and so on [8, 9]. Bidirectional Encoder Representations from Transformers (BERT) gains a lot of traction as it outperforms previous models in many cases [10]. More specifically, on the area of aggression detection, BERT implementations has been shown to have better results in comparison with the other methods [11].

The number of studies on sexism identification and especially a fine grained categorization of sexism is limited. Still, there has been several shared tasks on misogyny detection. Automatic Misogyny Identification (AMI) EVALITA 2018, 2020 and AMI IberEval 2018 were shared tasks focused on misogyny detection in Twitter in Spanish, English and Italian [12, 13]. These shared tasks presented a comparison of classical machine learning approaches, ensembles and neural networks. The best performing system of the last one, AMI EVALITA 2020 was composed of fine-tuning of pre-trained models by adapting BERT.

3 Methodology

In our approach, for both Task-1 and Task-2, the Simple Transformers NLP library through the Transformers library by HuggingFace is used [14]. Transformers is an open-source library which provides general-purpose architectures and pre-trained models [15]. As the task data-set is composed of texts both in English and Spanish languages, the multilingual-BERT model by Devlin et al. [10] has been adopted. For this purpose, 'bert-base-multilingual-cased' pre-trained model is used which consists of 12-layer, 768-hidden, 12-heads, totalling 179M parameters which were trained on cased text in the top 104 languages with the largest Wikipedias. BERT's model architecture is a multi-layer bi-directional Transformer encoder, designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. The training data-set provided by the organizers has been split by 10% testing and 90% training. The number of epochs for training was set to 4 (different configurations had no particular effect) and the half precision parameter, *fp16*, was set to false. Moreover, our experiments were run on GPU.

The data-set provided by the shared task organizers consists of 6,977 tweets for training and 3,386 tweets for testing in both Spanish and English languages. In addition, the test set includes 492 "gabs" in English and 490 in Spanish from the uncensored social network "Gab.com". This social network is advertised as *a social network that champions free speech, individual liberty and the free flow of information online.*

4 Results

The results of our submissions are shown in Table 1 with a comparison with the best performing system and the baseline. Task-1 was evaluated using the accuracy and Task-2 was evaluated using the macro-F1 score by the organizers.

For task 1 our submission ranked 30th among 72 submissions and 24th among 63 submissions for task 2. Our results for both tasks has overcome the baseline and given promising results; however were not able to rank in the top performers.

Table 1. Results for Task-1 and Task-2 as accuracy and F1 scores

Task - 1	**Accuracy**	**F1**
LaSTUS / TALN	**0,731**	0,731
Baseline - svm - tfidf	0,685	0,683
Best performing system	0,780	0,780
Task - 2	**Accuracy**	**F1**
LaSTUS / TALN	0,612	**0,523**
Baseline - svm - tfidf	0,522	0,395
Best performing system	0,658	0,579

5 Conclusion and Future Work

This paper describes our approach for the EXIST shared task at IberLef 2021. We used pre-trained multilingual-BERT model for sexism identification and categorization in English and Spanish languages. Our approach obtained promising results, yet improvements are required for future studies. Additional preprocessing steps might also improve the results along with data augmentation steps with other misogyny data-sets.

Acknowledgements

Our work is partly supported from the project Context-aware Multilingual Text Simplification (ConMuTeS) PID2019-109066GB-I00/AEI/10.13039/501100011033 awarded by Ministerio de Ciencia, Innovación y Universidades (MCIU) and by Agencia Estatal de Investigación (AEI) of Spain.

References

1. European Institute of Gender Equality, https://eige.europa.eu/thesaurus/terms/1367
2. Sexism, harassment and violence against women in parliaments in Europe `http://archive.ipu.org/pdf/publications/issuesbrief-e.pdf`

3. Fox, J., Cruz, C., Lee, J. Y. (2015). Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media. Computers in Human Behavior, 52, 436-442.
4. Swim, J. K., Hyers, L. L., Cohen, L. L., Ferguson, M. J. (2001). Everyday sexism: Evidence for its incidence, nature, and psychological impact from three daily diary studies. Journal of Social issues, 57(1), 31-53.
5. Kavanagh, E., Litchfield, C., Osborne, J. (2019). Sporting women and social media: Sexualization, misogyny, and gender-based violence in online spaces. International Journal of Sport Communication, 12(4), 552-572.
6. 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, Salvador Lima, Arturo Montejo-Ráez, Flor Miriam Plaza-de-Arco and Mariona Taulé (eds.): Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), CEUR Workshop Proceedings, 2021
7. 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.
8. Shushkevich, E., Cardiff, J. (2019). Automatic Misogyny Detection in Social Media: A Survey. Computación y Sistemas, 23(4).
9. Fersini, E., Nozza, D., Rosso, P. (2020). Ami@ evalita2020: Automatic misogyny identification. Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020), Online. CEUR. org.
10. Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
11. Kumar, R., Ojha, A. K., Malmasi, S., Zampieri, M. (2020, May). Evaluating aggression identification in social media. In Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (pp. 1-5).
12. Caselli, T., Novielli, N., Patti, V., Rosso, P. (2018, December). Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI). In Proceedings of the Final Workshop (Vol. 12, p. 13).
13. Fersini, E., Rosso, P., Anzovino, M. (2018). Overview of the Task on Automatic Misogyny Identification at IberEval 2018. IberEval@ SEPLN, 2150, 214-228.
14. https://huggingface.co/transformers/
15. Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., ... Rush, A. M. (2020, October). Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 38-45).