Sexism Identification using BERT and Data Augmentation – EXIST2021

Sabur Butt, Noman Ashraf, Grigori Sidorov, and Alexander Gelbukh

CIC, Instituto Politécnico Nacional, Mexico
sabur@nlp.cic.ipn.mx, nomanashraf712@gmail.com, sidorov@cic.ipn.mx, gelbukh@gelbukh.com

Abstract. Sexism is defined as discrimination among females of all ages. We have seen a rise of sexism in social media platforms manifesting itself in many forms. The paper presents best performing machine learning and deep learning algorithms as well as BERT results on "sEXism Identification in Social neTworks (EXIST 2021)" shared task. The task incorporates multilingual dataset containing both Spanish and English tweets. The multilingual nature of the dataset and inconsistencies of the social media text makes it a challenging problem. Considering these challenges the paper focuses on the pre-processing techniques and data augmentation to boost results on various machine learning and deep learning methods. We achieved an F_1 score of 78.02% on the sexism identification task (task 1) and F_1 score of 49.08% on the sexism categorization task (task 2).

Keywords: sexism detection · data augmentation · BERT · machine learning · deep learning.

1 Introduction

Sexism in its basic essence is defined as discrimination among women. However, the manifestation of sexism on social media is far more than just sexism. A survey [1] on online harassment shows that women are harassed on the Internet twice as much as men because of their gender. Similarly, a recent study [2] on rape cases concluded a correlation between the number of misogynistic tweets and the number of rapes in the United States of America. Hence, this social urgency has motivated various Natural Language Processing (NLP) researchers to define, categorize and create novel solutions for sexism detection on text.

Computational understanding of natural language has been used to tackle problems like emotion detection and sentiment analysis [3, 4], human behavior detection [5], fake news detection [6, 7] question answering [8] and depression and

IberLEF 2021, September 2021, Málaga, Spain.

threat detection [9, 10] in all forms of media as it gives us the insight to understand human perspectives and values. On a lexical level, sexism is very difficult to differentiate between multiple types of sexism. Understanding sexism and how it is different from other forms of harassment and hate speech also gives us more potential to restraint the harm caused on digital social platforms. Researchers have made several attempts in classifying sexism [11–15] to achieve more robust datasets or to achieve a better understanding of sexism from the text. Our aim in this study was to create more understanding of the machine and deep learning approaches for sexism in a broad sense, ranging from objectification, explicit misogyny to other types of implicit sexist behaviours.

In this article, we have attempted a shared task on "sEXism Identification in Social neTworks" at Iberian Languages Evaluation Forum (IberLEF 2021) [16, 17]. The first task attempted is sexism identification which is a binary classification problem in a multi-lingual dataset containing both English and Spanish tweets. The second task is titled sexism categorization which aims to categorize the message according to the type of sexism. The second task has five classes and the task is presented in both Spanish and English. The classes of the task are divided into "ideological and inequality", "objectification", "sexual violence", "stereotyping and dominance" and "misogyny and non-sexual violence". Both tasks are attempted and the paper explains the results of various machine learning and deep learning algorithms applied. For better performance, we used an augmented dataset and focused on pre-processing for social media text to enhance machine learning results. Our results show that an augmented dataset enhances both machine learning and deep learning results for sexism detection tasks.

2 Related Work

Some studies included sexism in the umbrella term of sexual harassment [14] or viewed it as a form of hate speech [11, 18]. A more direct categorization has been done in the form of "information threat", "indirect harassment", "sexual harassment" or "physical harassment" [19] and has also been classified as "Hostile", "Benevolent" or "Others" [12]. We have also seen the multi-label classification of sexism [13] where sexism was linked with twenty-three categories including role stereotyping, body shaming, attribute stereotyping, internalized sexism, hyper-sexualization (excluding body shaming), pay gap, hostile work environment (excluding pay gap), denial or trivialization of sexist misconduct, threats, rape, sexual assault (excluding rape), sexual harassment (excluding assault), tone policing, moral policing (excluding tone policing), victim-blaming, slut-shaming, motherhood-related discrimination, menstruation-related discrimination, religion-based sexism, physical violence (excluding sexual violence), gaslighting, mansplaining, and other. Another categorizing attempt [15] gave us sexism detection in the form of benevolent sexism, physical threats, sexual threats, body harassment, masculine harassment, lack of attractiveness harassment, stalking, impersonation and general sexist statements. Among the work that includes sex-

ism in hate speech, researchers have elevated the results by embedding driven features [20], weakly supervised learning [21], n-grams and linguistic features [11] and by extracting typed dependencies using text parsing [22]. We have also seen deep learning approaches for classification using Convolutional Neural Network (CNN) [13], CNN with Gated Recurrent Unit (GRU) [23], Long short-term memory (LSTM) with various text embeddings [12] and BERT [13]. Similarly, sexism classification outside of hate speech or sexual harassment has seen various machine and deep learning classification approaches. Researchers have used n-grams and pre-trained embeddings features [19] using SVM, bi-LSTM and bi-LSTM with attention [12], CNN and RNN [14] algorithms to classify sexism in various categories. Furthermore, a noteworthy work on the first Spanish dataset (MeTwo) [24] on sexism expressions showed us behavioural analyses on social media and deep learning results with Multilingual BERT (mBERT) outperforming other baselines.

3 Dataset

The dataset comprised of Twitter data containing 3,436 tweets in English and 3,541 tweets in Spanish. Table 3 and 4 show us the samples from task 1 and task 2. We used "Back Translation" for the augmentation of the text. Back Translation works by inputting the text in the source language (Spanish and English) and then translating the text to a second language (e.g. English to German). The final step is to translate back the previously translated text into the source language. We augmented the Spanish data by translating it to German and then back to Spanish. Similarly, English data was also translated to German and then back to English. Table 1 and 2 show the complete dataset statistics before and after the augmentation for both task 1 and task 2. For the translation, we used `deep-translator` python library [25].

Table 1: Dataset statistics before and after augmentation for task 1.

Augmentation	**English**	**Spanish**	**Size**
Before	3,436	3,541	6,977
After	6,872	7,082	13,954

4 Methodology

For both task 1 and task 2, we used transformers as well as various machine and deep learning algorithms for analyses. We applied a range of classifiers such as Logistic Regression (LR), Multilayer perceptron (MLP), Random Forest (RF), Support Vector Machine (SVM) [26, 27], 1 Dimensional Convolutional Neural Network (1D-CNN) [28], Long short-term memory (LSTM) [29] and BERT [30]

Table 2: Dataset statistics before and after augmentation for task 2.

Classes	Size
non-sexist	3,580
ideological-inequality	1,720
misogyny-non-sexual-violence	1,350
objectification	958
sexual-violence	1,010
stereotyping-dominance	1,605

Table 3: Samples of task 1 in English and Spanish

Id	Language	Tweet	Classes
18	En	I really just want to be rich but not trophy wife rich, rich with my own	sexist
28	En	@GabSmolders Looks like a cool boss lady	non-sexist
5,442	Es	La Chica Con El Tatuaje De Atrapasueños Entre Las Tetas	sexist
5,757	Es	Odio a los onvres casi el doble cuando están en mi campo laboral.Odio su mansplaining,Odio su machismo.	non-sexist

Table 4: Samples of task 2 in English

Id	Language	Tweet	Classes
192	En	LOLA A BITCH SHE NOT LOYAL	misogyny-non-sexual-violence
193	En	Thank fuck for that. Must be the car's faults then. Silly cunt https://t.co/O4IbNlojfy	non-sexist
511	En	@grandeyikes you do know their in the top 3 while bts aren't..sit your straight dumb blonde ass down.	stereotyping-dominance
567	En	"A prostitute? And risk getting some incurable disease?Do I look like a himbo to you? I at least have standards!" https://t.co/ZGI6AoQxYJ	objectification
632	En	"why is the cafe women-only" so women can chill there without being self-conscious or hit on	ideological-inequality
645	En	@MailOnline Breaking news: women have to suck the D longer due to Covid	sexual-violence

on our augmented dataset. We used the one vs. rest technique for task 2 sexism categorization.

4.1 Pre-processing

We removed the URLs, emails, numbers, digits, currency symbols and punctuation's in the pre-processing phase. All text was processed in lower case and the line breaks were fully stripped. The pre-processing standards were kept the same for both languages and all tasks. We used `Ekphrasis` for pre-processing the dataset [31] and added special tags surrounding important features. The following steps were taken:

1. **hashtags:** We pre-process hashtags by normalizing them to words and wrapping a hashtag tag around them i.e. $< hashtag > i < /hashtag >$ where variable "i" represents the hashtag in the sentence.
2. **all caps:** It is also a very common practice to express emphasis on a certain topic by using all capital words. These all capital words often express anger or excitement. We preserved this information using a special all caps tag which was placed in front and rear of the word before normalizing it to its normal un-capped state i.e. $< allcaps > i < /allcaps >$ where variable "i" represents the all caps word in the sentence.
3. **elongated:** Writing an expanded version of a word is also a very common practice in social media information writing. There are often cases where the user tries to explain the importance of something, goes short of adjectives and use an elongated version of the word i.e. "funyyyy" or "yesss". The elongated tag was added before and after the word i.e. $< elongated > i < /elongated >$ where variable "i" represents the elongated word in the sentence.
4. **repeated:** Writing repeating instances of words or characters is also express strong reactions in a social media text i.e. "????". Repeated tag is used in the same pattern i.e. $< repeated > i < /repeated >$ where variable "i" represents the repeated word in the sentence.
5. **emphasis:** To express emphasis on a certain word, people often try to enclose it in a pair of asterisks. These asterisks show that the user is trying to give more weight to the word. A special emphasis tag is placed i.e. $< emphasis > i < /emphasis >$ where "i" is a variable representing the emphasised word in the sentence.
6. **censored:** People express anger on social media platforms with censored abusive words which is a strong indicator for emotion tasks. In normalization, since it is not a dictionary word, it can be removed. To process this, the word is normalized to the dictionary word and a censored tag is placed with it i.e. $< censored > i < /censored >$ where variable "i" represents the emphasised word in the sentence.
7. **emotional annotations:** Emoticons are extremely important when it comes to capturing emotions in a text. We added a special tag for emotions (happy, annoyed, sad, laughing, tongue-sticking out, wink etc) in the tweets i.e $< annoyed >$, $< laugh >$, $< happy >$ etc.

4.2 Features

We used various feature representations such as word n-gram, char n-gram, and GloVe [32] pre-trained embeddings. Character n-grams and word n-grams have been repeatedly used for tasks like emotion detection, authorship detection, speech analysis and text categorization [33, 34]. GloVe vector representations for word also yield great results for social media text as they have separate embeddings for Twitter.

4.3 Evaluation

The algorithms as suggested by the challenge have been evaluated using accuracy, precision (P), recall (R) and F_1-measure. We used tenfold cross validation for this task which ensures the robustness of our evaluation. The tenfold cross validation takes ten equal size partitions. Out of ten, one subset of the data is retained for testing and the rest for training. This method is repeated ten times with each subset used exactly once as a testing set. The ten results obtained are then averaged to produce estimation.

5 Results

Results of both tasks are shown in Table 5. Our best-performing algorithms are BERT, RF, and MLP for task 1 while BERT, RF, SVM, and MLP performed best on task 2. The results on all algorithms in Table 5 were boosted with the augmented dataset. In both tasks, we observed that with proper pre-processing, machine learning can produce competitive results in comparison to deep learning methods and tends to outperform even deep learning methods such as 1D-CNN as seen in both task 1 and task 2.

Table 5: Results for task 1 and task 2 on the development set

Task	**Model**	**Acc**	**P**	**R**	**F_1**
Task 1	LR	79.92	80.07	81.65	80.67
	MLP	**89.38**	**89.81**	**89.83**	**89.68**
	RF	84.14	83.20	87.12	84.84
	SVM	67.67	66.31	76.44	70.81
	CNN	64.89	–	–	–
	LSTM	62.77	–	–	–
	BERT	81.32	–	–	–
Task 2	LR	16.28	16.42	16.09	15.24
	MLP	67.54	66.83	66.67	66.71
	RF	67.41	71.04	62.12	65.34
	SVM	**70.71**	**71.83**	**68.47**	**69.86**
	BERT	63.31	–	–	–

Table 6 shows the comparison of our results with the top five submitted results in the competition. For task 1, the CIC_1 submission represents BERT, CIC_2 shows RF and CIC_3 shows MLP accuracy and F_1 scores on the test set. For task 2, the CIC_1 submission represents SVM, CIC_2 shows BERT and CIC_3 shows RF accuracy and F_1 scores on the test set. We can see that BERT performed the best on both tasks and RF performed the best among the machine learning models using test set on both tasks.

Table 6: Comparison with top 5 results in the competition for task 1 and task 2

Team name	**Acc**	**F_1**
task1_AI−UPV_1	78.04	78.02
task1_SINAI_TL_1	78.00	77.97
task1_SINAI_TL_3	77.70	77.57
task1_SINAI_TL_2	77.66	77.61
task1_AIT_FHSTP_2	77.54	77.52
task1_CIC_1	**72.78**	**72.70**
task1_CIC_2	63.67	63.66
task1_CIC_3	63.67	63.66

(a) Task 1

Team name	**Acc**	**F_1**
task2_AI−UPV_1	65.77	57.87
task2_LHZ_1	65.09	57.06
task2_SINAI_TL_1	65.27	56.67
task2_SINAI_TL_3	64.97	56.32
task2_QMUL−SDS_1	64.26	55.94
task2_CIC_2	**55.27**	**49.08**
task2_CIC_3	58.38	45.43
task2_CIC_1	56.50	44.89

(b) Task 2

6 Conclusion

In this paper, we discussed possible classification methods for sexism identification and categorization on a multilingual dataset. The paper shows the results of various deep learning and machine learning algorithms for sexism detection. Data augmentation enhanced the results for both the sexism identification task (task 1) and sexism categorization task (task 2). The best performing algorithm on both tasks was BERT with data augmentation achieving an F_1 score of 78.02% on the sexism identification task and F_1 score of 49.08% on the sexism categorization task. Among the machine learning algorithms, RF performed the best and achieved an F_1 score of 63.66% on the sexism identification task and F_1 score of 45.43% on the sexism categorization task. In future, we expect more work on devising approaches for more robust classification methods to mitigate sexism on text. We hope our efforts will help future studies fighting sexism.

Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of the CONACYT, Mexico and grants 20211784, 20211884, and 20211178 of the Secretaría de Investigación y Posgrado of the Instituto

Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico.

References

1. M. Duggan, "Online harassment 2017," 2017.
2. R. Fulper, G. L. Ciampaglia, E. Ferrara, Y. Ahn, A. Flammini, F. Menczer, B. Lewis, and K. Rowe, "Misogynistic language on Twitter and sexual violence," in *Proceedings of the ACM Web Science Workshop on Computational Approaches to Social Modeling (ChASM)*, 2014.
3. I. Ameer, N. Ashraf, G. Sidorov, and H. G. Adorno, "Multi-label emotion classification using content-based features in Twitter," *Computación y Sistemas*, vol. 24, 02 2021.
4. L. Khan, A. Amjad, N. Ashraf, H.-T. Chang, and A. Gelbukh, "Urdu sentiment analysis with deep learning methods," *IEEE Access*, pp. 1–1, 2021.
5. F. Bashir, N. Ashraf, A. Yaqoob, A. Rafiq, and R. U. Mustafa, "Human aggressiveness and reactions towards uncertain decisions," *International Journal of ADVANCED AND APPLIED SCIENCES*, vol. 6, no. 7, pp. 112–116, 2019.
6. N. Ashraf, S. Butt, G. Sidorov, and A. Gelbukh, "CIC at CheckThat! 2021: Fake news detection using machine learning and data augmentation," in *CLEF 2021 – Conference and Labs of the Evaluation Forum*, (Bucharest, Romania), 2021.
7. M. Amjad, G. Sidorov, and A. Zhila, "Data augmentation using machine translation for fake news detection in the Urdu language," in *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 2537–2542.
8. S. Butt, N. Ashraf, M. H. F. Siddiqui, G. Sidorov, and A. Gelbukh, "Transformer-based extractive social media question answering on TweetQA," *Computación y Sistemas*, vol. 25, no. 1, 2021.
9. R. U. Mustafa, N. Ashraf, F. S. Ahmed, J. Ferzund, B. Shahzad, and A. Gelbukh, "A multiclass depression detection in social media based on sentiment analysis," in *17th International Conference on Information Technology–New Generations (ITNG 2020)* (S. Latifi, ed.), (Cham), pp. 659–662, Springer International Publishing, 2020.
10. N. Ashraf, R. Mustafa, G. Sidorov, and A. Gelbukh, "Individual vs. group violent threats classification in online discussions," in *Companion Proceedings of the Web Conference 2020*, WWW '20, (New York, NY, USA), pp. 629–633, Association for Computing Machinery, 2020.
11. Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter," in *Proceedings of the NAACL student research workshop*, pp. 88–93, 2016.
12. A. Jha and R. Mamidi, "When does a compliment become sexist? Analysis and classification of ambivalent sexism using Twitter data," in *Proceedings of the second Workshop on NLP and Computational Social Science*, pp. 7–16, 2017.
13. P. Parikh, H. Abburi, P. Badjatiya, R. Krishnan, N. Chhaya, M. Gupta, and V. Varma, "Multi-label categorization of accounts of sexism using a neural framework," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, (Hong Kong, China), pp. 1642–1652, Association for Computational Linguistics, Nov. 2019.

14. S. Karlekar and M. Bansal, "Safecity: Understanding diverse forms of sexual harassment personal stories," *arXiv preprint arXiv:1809.04739*, 2018.
15. S. Sharifirad and S. Matwin, "When a tweet is actually sexist. a more comprehensive classification of different online harassment categories and the challenges in nlp," *arXiv preprint arXiv:1902.10584*, 2019.
16. F. Rodríguez-Sánchez, J. C. de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, and T. Donoso, "Overview of exist 2021: sexism identification in social networks," *Procesamiento del Lenguaje Natural*, vol. 67, no. 0, 2021.
17. M. Montes, P. Rosso, J. Gonzalo, E. Aragón, R. Agerri, M. Ángel Álvarez Carmona, E. Álvarez Mellado, J. C. de Albornoz, L. Chiruzzo, L. Freitas, H. G. Adorno, Y. Gutiérrez, S. M. J. Zafra, S. Lima, F. M. P. de Arco, and M. T. (eds.) in *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), CEUR Workshop Proceedings*, 2021.
18. P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 759–760, 2017.
19. M. Anzovino, E. Fersini, and P. Rosso, "Automatic identification and classification of misogynistic language on Twitter," in *International Conference on Applications of Natural Language to Information Systems*, pp. 57–64, Springer, 2018.
20. C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in *Proceedings of the 25th International Conference on World Wide Web*, pp. 145–153, 2016.
21. L. Gao, A. Kuppersmith, and R. Huang, "Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach," *arXiv preprint arXiv:1710.07394*, 2017.
22. P. Burnap and M. L. Williams, "Us and them: Identifying cyber hate on Twitter across multiple protected characteristics," *EPJ Data science*, vol. 5, pp. 1–15, 2016.
23. Z. Zhang and L. Luo, "Hate speech detection: A solved problem? The challenging case of long tail on Twitter," *Semantic Web*, vol. 10, no. 5, pp. 925–945, 2019.
24. F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, and L. Plaza, "Automatic classification of sexism in social networks: An empirical study on twitter data," *IEEE Access*, vol. 8, pp. 219563–219576, 2020.
25. N. Baccouri, "Deep-translator." https://pypi.org/project/deep-translator/, 2020.
26. S. Salzberg, "C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993," *Machine Learning*, vol. 16, pp. 235–240, 1994.
27. R. Kohavi, "The power of decision tables," in *Proceedings of the 8th European Conference on Machine Learning*, ECML '95, (Berlin, Heidelberg), pp. 174–189, Springer-Verlag, 1995.
28. Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Doha, Qatar), pp. 1746–1751, Association for Computational Linguistics, Oct. 2014.
29. S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, pp. 1735–1780, Nov. 1997.
30. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018.
31. C. Baziotis, N. Pelekis, and C. Doulkeridis, "Datastories at SemEval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, (Vancouver, Canada), pp. 747–754, Association for Computational Linguistics, August 2017.

32. J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.
33. I. Ameer, G. Sidorov, and R. M. A. Nawab, "Author profiling for age and gender using combinations of features of various types," *Journal of Intelligent & Fuzzy Systems*, vol. 36, no. 5, pp. 4833–4843, 2019.
34. Z. Tüske, R. Schlüter, and H. Ney, "Investigation on lstm recurrent n-gram language models for speech recognition," in *Interspeech*, pp. 3358–3362, 2018.