Sexism Identification In Social Networks

Atul Chaudhary[1] and Ritesh Kumar[2]

[1] *Department Computer Science and Engineering*
Indian Institute of Information Technology Surat,
Kholvad Campus. Kamrej Surat, Gujarat, 394190, India
[2] *Department Computer Science and Engineering*
Indian Institute of Information Technology Surat,
Kholvad Campus. Kamrej Surat, Gujarat, 394190, India

Abstract

Sexism is a pervasive issue in society, and it has found a new platform for expression and dissemination in the digital age through social networks. Online social networks provide an environment where individuals can freely express their thoughts and opinions, but unfortunately, this freedom is sometimes misused to propagate sexist content, perpetuating harmful stereotypes and discrimination. In this paper, we present a sexism identification model/system specifically designed for social networks. Specifically, we describe the model submitted for the shared task on "sEXism Identification in Social network (EXIST 2023)" at CLEF2023.The problem concentrates on sexism detection in two languages: English and Spanish. The challenge is a binary classification problem to discover the instances of sexism in the tweets. Overall, the outcome of this project will provide valuable insights into the prevalence and nature of sexism in social networks. It can empower social network administrators, policymakers, and users to take proactive measures to combat sexism, promote gender equality, and create a safer online environment for everyone.

Keywords

Machine learning, Bi-LSTM, Sexism detection, Natural language processing, Sexism identification, social media

1. Introduction

Sexism, which involves discrimination against women, has also become prevalent in digital spaces. Online harassment surveys have shown that women experience harassment on the internet at twice the rate of men due to their gender [1]. Furthermore, a recent study found a correlation between the number of misogynistic tweets and the number of rape cases in the United States [2]. These alarming trends have

[1]CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
EMAIL:atulsci7705@gmail.com (A. Chaudhary);ritesh4rmrvs@gmail.com (R. Kumar)

prompted researchers in Natural Language Processing (NLP) to focus on defining, categorising, and developing solutions for detecting sexism in text.

With the rapid development of internet and mobile communication technologies, social media has become one of the largest sources of data. However, the internet is not an equal space for everyone. Concerns have been raised about the disproportionate levels of abuse experienced by women on social media platforms [3]. Instances of hate speech on Twitter targeting female politicians, journalists, and participants in feminist debates have been documented across different countries. Amnesty International published a report characterising Twitter as a "toxic place" for women, where violence and hate based on gender are promoted. Online gender-based violence can have significant psychological, social, and economic impacts. Women who experience online abuse often alter their online behaviour, self-censor their content, and limit their interactions on platforms out of fear of violence and abuse. By silencing or driving women away from online spaces, online violence can affect their economic outcomes, leading to loss of employment and societal status. Additionally, online gender-based violence may serve as a predictor of violent crimes in the physical world.

The computational understanding of natural language has been instrumental in addressing various issues such as emotion detection, sentiment analysis [4,5], human behaviour detection [6], fake news detection [7,8], question answering [9], and depression and threat detection [10,11] across different forms of media. NLP provides insights into human perspectives and values, enabling us to comprehend sexism and differentiate it from other forms of harassment and hate speech. Researchers have made several attempts to classify sexism [12–16] in order to create more robust datasets and gain a better understanding of sexism from textual data.

It is crucial to address online sexism and gender-based violence, as they impact the well-being, freedom of expression, and economic opportunities of women. Combining insights from NLP research and efforts to create safer digital spaces can contribute to mitigating these harms and fostering more inclusive online environments.

Tasks given to perform in EXIST 2023 challenge consist of:

TASK 1: Sexism Identification

The first task is a binary classification. The systems have to decide whether or not a given tweet contains sexist expressions or behaviours (i.e., it is sexist itself, describes a sexist situation or criticises a sexist behaviour). The following tweets show examples of sexist and not sexist messages.

- SEXIST
- NOT SEXIST

TASK 2: Source Intention

Once a message has been classified as sexist, the second task aims to categorise the message according to the intention of the author, which provides insights into the role played by social networks on the emission and dissemination of sexist messages. In this task, we propose a ternary classification task:

- DIRECT: the intention was to write a message that is sexist by itself or incites to be sexist
- REPORTED: the intention is to report and share a sexist situation suffered by a woman or women in first or third person
- JUDGEMENTAL: the intention was to judge, since the tweet describes sexist situations or behaviours with the aim of condemning them.

TASK 3: Sexism Categorization

Many facets of a woman's life may be the focus of sexist attitudes including domestic and parenting roles, career opportunities, sexual image, and life expectations, to name a few. Automatically detecting which of these facets of women are being more frequently attacked in social networks will facilitate the development of policies to fight against sexism. According to this, each sexist tweet must be categorised in one or more of the following categories

- IDEOLOGICAL AND INEQUALITY: The text discredits the feminist movement, rejects inequality between men and women, or presents men as victims of gender-based oppression.
- STEREOTYPING AND DOMINANCE: The text expresses false ideas about women that suggest 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.
- 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.).
- SEXUAL VIOLENCE: Sexual suggestions, requests for sexual favors or harassment of a sexual nature (rape or sexual assault) are made.
- MISOGYNY AND NON-SEXUAL VIOLENCE: The text expresses hatred and violence towards women.

In this paper we discuss our participation in providing a solution mainly for task 1.

The organisation of the rest of the paper is as follows. Section 2 describes Related Work. In Section 3, we describe the Task and Dataset. Section 4, describe about System Description i.e., Preprocessing, Tokenization, Sequence Padding and Model Architecture. In Section 5, we discuss our Results. Lastly, Conclusion and future direction of our research work is presented in section 6

2. Related Work

In the field of sexism detection and classification, various studies have approached the problem from different perspectives and categorization frameworks. Some studies have included sexism under the broader term of sexual harassment [15] or considered it a form of hate speech [12,19]. Others have employed more direct categorizations, such as "information threat," "indirect harassment," "sexual harassment," or "physical harassment" [20]. Sexism has also been classified as "hostile," "benevolent," or grouped into other categories [13]. Some current studies on sexism identification are closely related to hate speech detection. The introduction of sexism as a classification task was first proposed by Waseem in 2016 [12]. Waseem annotated 16 thousand tweets and categorised them as racist, sexist, or neither. The data was collected from tweets related to the Australian TV show "My Kitchen Rules" using the hashtag #mkr. Waseem employed different methods such as character-level n-grams and word n-grams, using logistic regression with 10-fold cross-validation. Another categorization attempt [16] provided 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. Studies focusing on sexism within hate speech have enhanced results through the use of driven features [21], weakly supervised learning [22], n-grams and linguistic features [12], and typed dependencies

extracted using text parsing [23]. Deep learning approaches have also been employed for sexism classification. These include the use of Convolutional Neural Network (CNN) [14], CNN with Gated Recurrent Unit (GRU) [24], Long short-term memory (LSTM) with various text embeddings [13], and BERT [14,27]. Similarly, sexism classification outside of hate speech or sexual harassment has seen the application of various machine and deep learning classification approaches. Researchers have utilised n-grams and pre-trained embeddings features [20] using SVM, bi-LSTM, and bi-LSTM with attention [13], as well as CNN and RNN [15] algorithms to classify sexism in various categories. Notably, research on the first Spanish dataset (Me Two) [25] for sexism expressions included behavioural analysis on social media. Deep learning techniques using Multilingual BERT (mBERT) [26] outperformed other baselines in this study.

Overall, these studies highlight the diversity of approaches and techniques employed to detect and classify sexism in social networks, ranging from traditional machine learning algorithms to deep learning models. The categorization frameworks have evolved to capture various facets of sexism, including its association with hate speech, sexual harassment, and a wide range of specific categories related to gender discrimination and stereotyping. Future research may further explore these methodologies to enhance the accuracy and effectiveness of sexism identification systems.

3. Task and Dataset

The task organisers of CLEF2023 provided a dataset called EXIST2023 [27,28]. The EXIST2023 dataset contains more than 10,000 tweets or around 10,000 tweets in English and Spanish consisting of 6,920 tweets for training, 1038 tweets for development and 2,076 tweets for testing. The organisers assigned 3 tasks for the participants (Tasks details are already mentioned in Introduction section) but we are participating to provide solution for task 1 only:

Task: Sexism Identification: This subtask is a binary classification problem. The model's objective is to classify whether a given tweet contains sexist expression or not.

Examples of sexist and non-sexist messages:

- SEXIST:
    - “Mujer al volante, tenga cuidado!"
    - “People really try to convince women with little to no ass that they should go out and buy a body. Like bih, I don’t need a fat ass to get a man. Never have."
- NOT SEXIST:
    - “Alguien me explica que zorra hace a la gente en el cajero que se demora tanto."
    - "@messyworldorder it’s honestly so embarrassing to watch and they’ll be like “not all white women are like that”"

Sample distribution of Training, development and Testing dataset

Table 1: Training tweet samples distribution for Task

Language	No. of samples	Percentage (%)
English	3260	47.1
Spanish	3660	52.9

Table 2: development tweet samples distribution for Task

Language	No. of samples	Percentage (%)
English	489	47.1
Spanish	549	52.9

Table 3: Testing tweet samples distribution for Task

Language	No. of samples	Percentage (%)
English	978	47.1
Spanish	1098	52.9

From the above tables we observe that training, development and testing dataset contain equal both English and Spanish dataset in the same ratio.

4. System Description

We begin by segregating training and test dataset into English tweets and Spanish tweets. Since both languages are different and their words might have different meanings for different languages, separating different sources keeps our machine learning models simple and clean. We train machine learning models on newly created datasets for both the languages and then results obtained by these two systems were combined and submitted as a run. We are using Google Colab to perform these tasks.

4.1. Preprocessing

Our first step was to clean the text data in order to get a better vector representation of text data, we applied following text preprocessing techniques to clean text:

1. Removing emoticons from text.
2. Removing unrecognised characters, emojis and stickers from text
3. Removing special characters.
4. Removing Web Addresses from Text e.g., *"@larryelder Happy Veteran's Day to everyone who ever served this great nation, including my Husband and both of our beloved fathers. 🇺🇸🇺🇸🇺🇸🇺🇸 https://t.co/SBsJafPN0G"*
5. Removing Repeating Patterns Like 99, aaaaa, bbbbb,00 etc.
6. Removing Character Length Words like l,9,1, B etc.
7. Fixing Contractions, e.g. converting words like I'll to I will.

We decide to keep stopwords[9] for both English and Spanish tweets, as our experiment yields a slightly better F1 score by keeping them.

4.2. Tokenization and Sequence Padding

To prepare the input data for the model, we apply tokenization using the Tokenizer class from the TensorFlow Keras library. The tokenization step involves converting the text into a sequence of integers representing the words' indices in a predefined vocabulary. We limit the vocabulary size to the most frequent words, as specified by the parameter max_words. The tokenizer is fit on the training data to develop an internal vocabulary.
Sequence Padding To ensure that all input sequences have the same length, we employ sequence padding using the pad_sequences function from the TensorFlow Keras library. The sequences are padded or truncated to a fixed length of max_sequence_length. This step allows us to handle variable-length input sequences efficiently.

4.3. Model Architecture

Our proposed machine learning model consists of several layers that are designed to capture and learn meaningful representations from the input text. The model architecture is as follows:

- Embedding Layer: An embedding layer maps each word in the input sequence to a dense vector representation. We use the Word2Vec embedding technique with embedding_dim dimensions to capture semantic information.
- Bidirectional LSTM Layers: Two bidirectional LSTM layers are stacked to capture both forward and backward contextual information in the text. The first LSTM layer returns sequences, while the second LSTM layer provides higher-level representations.
- Dense Layers: A dense layer with ReLU activation is added to introduce non-linearity and extract complex features from the LSTM outputs.
- Dropout: Dropout regularisation is applied to prevent overfitting. A dropout rate of 0.5 is used to randomly drop a fraction of input units during training.
- Output Layer: The final dense layer with a sigmoid activation function produces a probability between 0 and 1, indicating the likelihood of the input text belonging to the target class (sexist or not).

Model Training: The model is trained using the binary cross-entropy loss function and optimised using the Adam optimizer. The number of training epochs is set to 10, indicating the number of times the entire training dataset is passed through the model during training.
The implementation was done using Python Google Collaboratory (https://colab.research.google.com/), with the following technical specifications: Intel(R) Core (TM) i3-4030U CPU @ 1.90GHz CPU, 4GB of RAM.

5. Result

The results of the Task are presented in terms of F1-Score, as shown in Table 4. The best F1-score we achieved for Task is 0.6355. Table 4 also displays the ranking of our submissions based on the shared task official ranking in (hard-hard) evaluation scenario. The organisers provided a baseline for Task 1.

Table 4. Results for Task - The official Evaluation(hard-hard) measure is F1-Score. The best score obtained by us is mentioned in bold

Run	Rank	ICM-Hard	ICM-Hard Norm	F1
CNLP-NITS-PP_2	60	0.1093	0.4356	0.6409
IIIT SURAT_1	**61**	**0.1042**	**0.4324**	**0.6355**
Awakened_1	62	0.0723	0.412	0.6322

It is interesting to see that our model fails to correctly identify texts that have been labelled as non-sexist due to the presence of words that commonly appear in sexist environments. It also fails to correctly categorise sexist tweets that appear in the same context with non-sexist words (for example, the word friend), short texts or even sometimes subtle or contextual use of sexist's language.

6. Conclusion and Future Work

In this paper, we presented a machine learning model based on a bidirectional LSTM architecture for text classification. The model demonstrated promising results in identifying sexism in tweets or sentences. By leveraging word embeddings, tokenization, and sequence padding techniques, the model effectively captures contextual information and achieves competitive accuracy. This research contributes to the field of natural language processing and provides valuable insights into addressing social issues in online platforms.

Future work can focus on expanding the model to handle multi-class classification problems, incorporating more advanced techniques such as attention mechanisms, and exploring additional preprocessing steps to improve model performance. Additionally, the model can be deployed in real-world applications to mitigate and monitor instances of sexism on social media platforms.

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 WebScience Workshop on Computational Approaches to Social Modeling (ChASM), 2014.
3. M. F. Wright, B. D. Harper, S. Wachs, "The associations between cyberbullying and callous-unemotional traits among adolescents: The moderating effect of online disinhibition," Personality and Individual Differences, vol. 140, pp. 41-45, 2019.
4. 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, no. 2, 2021.
5. 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.
6. 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.
7. N. Ashraf, S. Butt, G. Sidorov, and A. Gelbukh, "CICatCheckThat! 2021: Fake news detection using machine learning and data augmentation," in CLEF 2021 Conference and Labs of the Evaluation Forum, (Bucharest, Romania), 2021.
8. 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.
9. 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.
10. 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.
11. 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.
12. 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.
13. 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.
14. 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.
15. S. Karlekar and M. Bansal, "Safecity: Understanding diverse forms of sexual harassment personal stories," arXiv preprint arXiv:1809.04739, 2018.

16. 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
17. F. Rodríguez-Sánchez, J. C. de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, and T. Donoso, "Overview of exist2021: sexism identification in social networks," Procesamiento del Lenguaje Natural, vol. 67, no. 0, 2021.
18. 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.), "Proceedings of the Iberian Languages Evaluation Forum (IberLEF2021)," CEUR Workshop Proceedings, 2021.
19. 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.
20. 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.
21. 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.
22. 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.
23. P. Burnap and M. L. Williams, "Us and them: Identifying cyberhate on Twitter across multiple protected characteristics," EPJ Data Science, vol. 5, pp. 1-15, 2016.
24. 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.
25. F. Rodríguez-Sánchez, J. Carrillo-deAlbornoz, and L. Plaza, "Automatic classification of sexism in social networks: An empirical study on Twitter data," IEEE Access, vol. 8, pp. 219563-219576, 2020.
26. Magnossao de Paula, A. F., Fray da Silva, R., & Schlicht, I. B. (2021). Sexism Prediction in Spanish and English Tweets Using Monolingual and Multilingual BERT and Ensemble Models. In Proceedings of IberLEF 2021 (Vol. 2943, pp. 356-373).
27. Butt, S., Ashraf, N., Sidorov, G., & Gelbukh, A. (2021). Sexism Identification using BERT and Data Augmentation. In Proceedings of IberLEF 2021 (Vol. 2943, pp. 381-389).
28. Laura Plaza, Jorge Carrillo-de-Albornoz, Roser Morante, Enrique Amigó, Julio Gonzalo, Damiano Spina, Paolo Rosso. Overview of EXIST 2023 – Learning with Disagreement for Sexism Identification and Characterization. Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023). Avi Arampatzis, Evangelos Kanoulas, Theodora Tsikrika, Stefanos Vrochidis, Anastasia Giachanou, Dan Li, Mohammad Aliannejadi, Michalis Vlachos, Guglielmo Faggioli, and Nicola Ferro, Eds. September 2023, Thessaloniki, Greece.
29. Laura Plaza, Jorge Carrillo-de-Albornoz, Roser Morante, Enrique Amigó, Julio Gonzalo, Damiano Spina, Paolo Rosso. Overview of EXIST 2023 – Learning with Disagreement for Sexism Identification and Characterization (Extended Overview). Working Notes of CLEF 2023 -

Conference and Labs of the Evaluation Forum. Mohammad Aliannejadi, Guglielmo Faggioli, Nicola Ferro and Michalis Vlachos, Eds.