Detection of Sexism on Social Media with Multiple Simple Transformers

Chirayu Jhakal[1,*,†], Khushi Singal[1,†], Manan Suri[1,†], Divya Chaudhary[2,†], Bijendra Kumar[1,†] and Ian Gorton[2,†]

[1]*Netaji Subhas University of Technology (NSUT), Dwarka Sector-3, Dwarka, Delhi, 110078*
[2]*Khoury College of Computer Sciences, Northeastern University, Seattle, USA*

Abstract
Social media platforms have become virtual communication channels, allowing users to voice their thoughts and opinions. However, this openness and features of anonymity have also given rise to the proliferation of harmful and offensive content, including sexism. This research aims at proposing a methodology and explores the use of different simple transformers. Monolingual Simple Transformers such as BERT, RoBERTa[1], BERTweet, DistilBERT, XLNet were evaluated on the EXIST2023 shared task challenge at the IberLEF2023 dataset. It was observed that RoBERTa has given the best results among all other transformers. The proposed approach has great scope for the efficient detection of sexist content on social media, aiding in the development of effective content moderation systems.

Keywords
Sexism Detection, Simple Transformer Models, Natural Language Processing,

1. Introduction

Social media platforms have been revolutionary in the way people communicate and express themselves in this digital age. These platforms have become a ubiquitous part of our daily lives, enabling users to share thoughts and opinions and engage in social interactions. However, along with the numerous benefits, the openness and features like anonymity and accessibility of social media have also given rise to a concerning issue - the rapid increase of offensive and harmful content, including sexism.

Sexism, defined as discrimination, stereotyping, or prejudice based on gender, continues to be a pervasive problem in society. The presence of such content on social media not only perpetuates harmful gender biases on possibly young and impressionable minds but also undermines the inclusivity and safety of these online spaces. Consequently, there is a pressing need to develop effective methods for detecting and mitigating sexist content on social media.

In this work, we mainly focus on using Natural Language Processing techniques and state-of-the-art models for sexism detection, which aims to identify if a specific sentence contains sexist

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
*Chirayu Jhakal and Khushi Singal are Corresponding authors.
†These authors contributed equally.
jhakalchirayu@gmail.com (C. Jhakal); khushisingal7@gmail.com (K. Singal); manansuri27@gmail.com (M. Suri); d.chaudhary@northeastern.edu (D. Chaudhary); bizender@nsut.ac.in (B. Kumar); i.gorton@northeastern.edu (I. Gorton)

content. The salient features of the proposed methodology are:

- The use of transformer-based language models, such as BERT (Bidirectional Encoder Representations from Transformers)[2], RoBERTa (Robustly Optimized BERT)[2][1], and BERTweet, etc. These models, pre-trained on vast amounts of textual data, have demonstrated exceptional capabilities in understanding and processing the complexities of human language[2].
- The methodology involves training and fine-tuning the transformer models on the EXIST2023 dataset.

2. Related Work

There has been an increase in the work and interest in sexism detection as all social media platforms want to limit offensive content and make their platforms more inclusive. Examples of such work are from the EXIST2022 edition: In [3], the use of Multilingual Models [3]and Data Augmentation has been employed, and researchers have leveraged multilingual models to detect sexist content across different languages[3] [4]. To overcome the lack of data in specific languages, data augmentation techniques have been employed [3] [4] [5] [6]. In [7], the English dataset and the Spanish dataset have been trained separately using a HuggingFace transformer [7] [8] [6]. [4] uses the ensemble of 5 classification monolingual models and back-translation has been used for augmenting the data; while [5] used datasets of different languages such as French, German, and Italian and translated them into the pivot languages and BERTweet and BETO have been used for English and Spanish respectively in a gradual unfreezing-discriminative fine-tuning fashion [9] [10]. In [11], pre-processing methods like TFIDF (Term Frequency-Inverse Document Frequency), Bag of words, and word2vec have been employed [11] and have seen the use of basic algorithms like Naive Bayes, Support Vector Machines (SVM), and Linear Regression to achieve remarkable results. In [12], a framework for sexism detection on social media via ByT5 and TabNet has been implemented.

3. Method

3.1. Dataset

The EXIST2023 dataset comprises posts from social media platforms like Twitter and Gab, accompanied by annotations categorizing them into different types of sexism. The dataset is divided into separate train and test partitions. The training set contains 6920 instances in both English (3,260) and Spanish (3,660) languages. Meanwhile, the test set consists of 2,076 instances, with 978 posts in English and 1098 posts in Spanish. Each data instance is labeled with binary labels (for task 1) to determine whether it is classified as sexist or non-sexist.

3.2. Exploratory Analysis

Before proceeding with the preprocessing steps, we performed an exploratory analysis of the dataset[13]. This analysis involved gaining insights into the distribution of sexist content,

identifying common patterns, and understanding the characteristics of the data. We examined the frequency of sexist posts, the prevalent forms of sexist language and expressions, and any potential biases in the dataset.

3.3. Preprocessing

The preprocessing began with the segregation of English data and Spanish data because both languages are structurally and semantically quite different, so applying various language models to the data would be easy and efficient [7] [4] [10] [9]. After cleaning the data various preprocessing techniques were applied [14]:

- Removing web addresses from text.
- Removing emoticons from the text.
- Removing unrecognized characters, emojis, and stickers from text.
- Removing special characters.
- Removing repeating patterns like aaaaa, bbbbb, 00 etc.
- Stemming words using Porter stemmer

3.4. Pre-trained Transformer Models

- RoBERTa: Robustly Optimized BERT (RoBERTa) [1] is an optimized variant of BERT that incorporates additional pre-training techniques. We fine-tuned the pre-trained RoBERTa model on our dataset to specifically detect and classify instances of sexist content. In English models, we applied RoBERTa after pre-processing and got decent results but in the Spanish model, we got even better results [15] [7] [4].
- BERT: Bidirectional Encoder Representations from Transformers (BERT) [16] is a widely used transformer model that captures bidirectional contextual information. We fine-tuned the pre-trained BERT model on our dataset to identify sexist content with improved accuracy. BERT didn't give satisfactory results on the English dataset but performed better on the Spanish dataset [17] [18] [9].
- Distill-BERT[19]: DistilBERT is a distilled version of the BERT model that offers a lighter and faster alternative while maintaining comparable performance. We fine-tuned the pre-trained DistilBERT model on our dataset to detect sexist content efficiently. DistilBERT was applied only on the Spanish dataset and gave poor results [6].
- BERTweet [20]: We fine-tuned the pre-trained BERTweet model on our dataset, specifically designed to handle social media text. BERTweet employs a specialized vocabulary and tokenization scheme tailored to social media language, enabling it to effectively detect and classify instances of sexism in social media posts [7] [17] [5].
- CamemBERT: CamemBERT is a transformer model specifically designed for the French language. It is trained on large-scale French corpora and exhibits strong language understanding capabilities. We fine-tuned the pre-trained CamemBERT model on our dataset to accurately detect and categorize instances of sexism in French social media posts. CamemBERT was only applied to the Spanish dataset and it gave the best results [7] [8] [6].

- XLNet: XLNet is a transformer model that employs permutation-based training, allowing it to capture bidirectional and context-aware representations effectively. We fine-tuned the pre-trained XLNet model on our dataset to enhance the detection and classification of sexist content. XlNet on the English dataset gave the best results among other models [7] [8] [6].

4. Results

In this study, we aimed to detect sexism in tweets written in both English and Spanish using pre-trained transformer models. Specifically, we employed four different models for English tweets, namely RoBERTa, BERT, BERTweet, and XLNet, while for Spanish tweets, we utilized BERT, DistilBERT, RoBERTa, and CamemBERT. The performance of each model was evaluated based on its accuracy in identifying instances of sexism in the tweets.

For the English language, the RoBERTa model achieved an accuracy of 59.02% in detecting sexism, while the BERT model achieved an accuracy of 56.16%. The BERTweet model, designed specifically for tweets, achieved an accuracy of 69.60%, outperforming both RoBERTa and BERT. The XLNet model, which incorporates a permutation-based approach, demonstrated the highest accuracy among the English models, achieving 71.34%.

In the case of Spanish tweets, the BERT model achieved an accuracy of 62.24% in detecting sexism, followed by DistilBERT with an accuracy of 58.38%. The RoBERTa model exhibited an accuracy of 62.77%, while the CamemBERT model demonstrated the highest accuracy among the Spanish models, achieving 69.05%.

These results indicate that the choice of pre-trained model has a significant impact on the performance of sexism detection in tweets. While all models achieved relatively moderate accuracy, the BERTweet model for English and the CamemBERT model for Spanish exhibited the highest accuracies, suggesting their effectiveness in identifying instances of sexism in tweets.

It is important to note that accuracy alone does not capture the full picture of model performance, and other evaluation metrics, such as precision, recall, and F1 score, should be considered for a comprehensive analysis. Furthermore, the generalizability of these models to different datasets and domains should be further investigated to assess their robustness and applicability in real-world scenarios.

Table 1
English Classification Models and Accuracy

English Models	Accuracy (%)
RoBERTa	59.02
BERT	56.16
BERTweet	69.60
XLNet	71.34

Table 2
Spanish Classification Models and Accuracy

Spanish Models	Accuracy (%)
BERT	62.24
DistilBERT	58.38
RoBERTa	62.77
CamemBERT	69.05

5. Discussion

The results of our study demonstrate the performance of various pre-trained transformer models in detecting sexism in tweets written in both English and Spanish. Overall, the models exhibited varying levels of accuracy, indicating their effectiveness in identifying instances of sexism in social media content.

For the English language, the BERTweet and XLNet models performed relatively better than the RoBERTa[1] and BERT models[16]. This observation suggests that models specifically designed for processing Twitter data, such as BERTweet, may be more suitable for capturing the nuances and informal language commonly used in tweets. The XLNet model, which utilizes a permutation-based approach, outperformed the other models, possibly due to its ability to capture long-range dependencies in the text.

In the case of Spanish tweets, the CamemBERT model displayed the highest accuracy among the evaluated models. This indicates that CamemBERT, which is specifically trained on Spanish text, is effective in capturing the linguistic characteristics and context-specific aspects of Spanish tweets. However, it is worth noting that the accuracies achieved by the Spanish models were relatively lower compared to the English models, suggesting the need for further research and improvement in detecting sexism in Spanish-language tweets.

It is important to consider the limitations of our study. First, the evaluation was performed on a specific dataset, and the results may not be directly generalizable to other datasets or real-world scenarios. Additionally, the performance of the models may vary depending on the nature of the tweets, the distribution of sexism-related content, and cultural or contextual factors. Further research is needed to assess the robustness and generalizability of these models across diverse datasets and contexts.

Moreover, accuracy alone may not be sufficient to fully evaluate the performance of sexism detection models. Additional metrics such as precision, recall, and F1 score should be considered to assess the models' ability to correctly identify instances of sexism while minimizing false positives and false negatives.

In conclusion, our study highlights the effectiveness of pre-trained transformer models in detecting sexism in tweets, both in English and Spanish. The results demonstrate the importance of using language-specific models and models designed for social media data to achieve higher accuracy. This research contributes to the development of automated systems for identifying and addressing sexism in online communication, ultimately fostering a more inclusive and respectful digital environment.

6. Conclusion

In this research, we investigated the detection of sexism in tweets using pre-trained transformer models for both English and Spanish languages. Our results demonstrate that these models can be effective in identifying instances of sexism in social media content. The BERTweet model performed well in capturing the nuances of English tweets, while the CamemBERT model showed promise for Spanish tweets. Additionally, the XLNet model exhibited superior performance among the English models, highlighting the effectiveness of permutation-based approaches. However, it is important to note that the accuracies achieved, especially for Spanish models, can still be improved.

The findings of this study have implications for developing automated systems that can detect and mitigate sexism in online communication. By leveraging pre-trained transformer models, we can gain insights into the prevalence of sexism and take steps toward fostering a more inclusive and respectful digital environment.

7. Future Work

While this research provides valuable insights into the detection of sexism in tweets, there are several avenues for future work that can enhance the accuracy and robustness of the models.

Firstly, data augmentation techniques can be employed to improve model performance [3] [4] [5] [6]. By increasing the diversity and quantity of training data through techniques such as back-translation, word replacement, or text synthesis, we can potentially reduce the model's bias and enhance its ability to detect subtle forms of sexism.

Secondly, ensemble modeling can be explored to leverage the strengths of multiple models and improve overall performance. By combining predictions from different models, either by majority voting or weighted averaging, we can potentially achieve higher accuracy and mitigate the limitations of individual models.

Furthermore, it is important to expand the evaluation of sexism detection models to different languages and cultural contexts. The linguistic characteristics and contextual nuances can significantly vary across languages, necessitating the development of language-specific models and datasets.

Additionally, further research should focus on addressing the issue of bias in the models. It is crucial to identify and mitigate any biases encoded in the pre-trained models to ensure fair and equitable detection of sexism.

Finally, it would be beneficial to conduct user studies and assess the real-world impact of automated systems in addressing sexism in online spaces. Understanding user perceptions, reactions, and potential ethical concerns will guide the development of more effective and responsible solutions.

By pursuing these avenues, we can advance the field of sexism detection in social media and contribute to the development of robust and inclusive technologies.

References

[1] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[2] I. Turc, M.-W. Chang, K. Lee, K. Toutanova, Well-read students learn better: On the importance of pre-training compact models, arXiv preprint arXiv:1908.08962 (2019).

[3] D. Liakhovets, M. Schütz, J. Böck, M. Andresel, A. Kirchknopf, A. Babic, D. Slijpcevic, J. Lampert, A. Schindler, M. Zeppelzauer, Transfer learning for automatic sexism detection with multilingual transformer models, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022). CEUR-WS. org, A Coruna, Spain, 2022.

[4] V. Ahuir, J. Á. González, L.-F. Hurtado, Enhancing sexism identification and categorization in low-data situations (2022).

[5] R. L. Tamayo, R. O. Bueno, Are examples worth more than language? (2022).

[6] D. García-Baena, M. Á. García-Cumbreras, S. M. Jiménez-Zafra, M. García-Vega, Sinai at exist 2022: Exploring data augmentation and machine translation for sexism identification (2022).

[7] A. Vaca-Serrano, Detecting and classifying sexism by ensembling transformers models, language 2 (2022) 1.

[8] V. P. Álvarez, J. M. Vázquez, W. Chibane, J. L. Domínguez, Automatic sexism identification using an ensemble of pretrained transformers (2020).

[9] F. M. Plaza-del Arco, M.-D. Molina-González, L. A. Ureña-López, M.-T. Martín-Valdivia, Exploring the use of different linguistic phenomena for sexism identification in social networks (2022).

[10] K. Bengoetxea, A. Aguirregoitia, Multiaztertest@ exist-iberlef2022: Sexism identification in social networks (2022).

[11] A. Rizvi, A. Jamatia, Nit-agartala-nlp-team at exist 2022: Sexism identification in social networks (2022).

[12] A. Younus, M. A. Qureshi, A framework for sexism detection on social media via byt5 and tabnet (2022).

[13] J. W. Tukey, et al., Exploratory data analysis, volume 2, Reading, MA, 1977.

[14] H. T. Ta, A. B. S. Rahman, L. Najjar, A. Gelbukh, Transfer learning from multilingual deberta for sexism identification, in: CEUR Workshop Proceedings, volume 3202, CEUR-WS, 2022.

[15] A. F. M. de Paula, R. F. da Silva, Detection and classification of sexism on social media using multiple languages, transformers, and ensemble models, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022) co-located with the XXXVIII International Conference of the Spanish Society for Natural Language Processing (SEPLN 2022), CEUR Workshop proceedings, volume 3202, 2022, pp. 1–11.

[16] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[17] E. Villa-Cueva, F. Sanchez-Vega, A. P. López-Monroy, Bi-ensembles of transformer for online bilingual sexism detection, in: CEUR Workshop Proceedings, volume 3202, 2022.

[18] G. Shimi, J. Mahibha, D. Thenmozhi, Sexism identification in social media using deep learning models, in: CEUR Workshop Proceedings, volume 3202, 2022.

[19] A. Danday, T. S. Murthy, Twitter data analysis using distill bert and graph based convolution neural network during disaster (2022).
[20] D. Q. Nguyen, T. Vu, A. T. Nguyen, Bertweet: A pre-trained language model for english tweets, arXiv preprint arXiv:2005.10200 (2020).