Sexism Identification In Social Media Using Deep Learning Models

Gersome Shimi[1,*], Jerin Mahibha[2,*] and Durairaj Thenmozhi[3,*]

[1]*Madras Christian College, Chennai, India*
[2]*Meenakshi Sundararajan Engineering College, Chennai, India*
[3]*Sri Sivasubramaniya Nadar College of Engineering, Chennai, India*

Abstract

Every minute, millions of information are posted in social media especially in Twitter, which includes sensitive information. Sexism identification in social media is a tedious task. This paper is a challenge given by sEXism Identification in Social neTworks(EXIST) at IberLEF 2022 to identify sexism in the specified dataset obtained from tweets. The dataset contains text in English and Spanish languages with sexism expressions. The specified text is identified as sexist and non-sexist by LaBSE model and multilingual BERT classification model. The predicted text and the performance are measured in different scaling and it is found that LaBSE model performed better. The F1 score of 75.30% is attained in Task1 related to identification of sexism.

Keywords

Sexism identification, Classification, Multilingual, Spanish, Twitter, Deep Learning, Natural Language Processing(NLP), BERT, LaBSE

1. Introduction

Hard copied information which was traditionally used, have started to decline in the recent decades. The information is shared to their inmates, well-wishers, colleagues and opponents by social media, prominently through Twitter, Facebook, Instagram and as such. The information may contain sensitive data also, which may hurt or mislead people who read those text. Even though some policy restrictions are given by social media, people don't follow it up. It made up the need, to scrutinize the information before it is shared, in social media. The information may create impact on the people who read and follow social media. The identification of text is an important element to perform this task.

Even though many social media sites exist to exhibit people's opinions and thoughts, Twitter is the trendy one especially to share textual data. Every day, text messages are shared by many people including the eminent personalities and famous politicians, who may be their role models, well-wishers, sports stars, film stars etc. People can also text them directly.

According to the recent research, for every second, millions of data are shared in social media, the data can be structured, unstructured or the combination of both. These data should be organized, channelized and analyzed, which is a tedious task. Since these data are enormous

IberLEF 2022, September 2022, A Coruña, Spain

*These authors contributed equally.

gshimi2022@gmail.com (G. Shimi); jerinmahibha@gmail.com (J. Mahibha); d_theni@ssn.edu.in (D. Thenmozhi)

CEUR Workshop Proceedings (CEUR-WS.org)

and not are in a common format. The text in twitter also contains discriminated contents even though it has enforced policies to ban hatred or discriminated content. These contents should be trimmed before it get posted in the social media, since it may harm the people emotionally, physically and sentimentally. It may also affect people's beliefs, relationships, harmony in the society and the environment.

Britannica defines Sexism as a prejudice or discrimination based on sex or gender, particularly against women and girls. The women discrimination is widely available in social media [1]. Among all discrimination sexism meant to be one of the highlighted problem. The detection of sexism is a arduous task, which it is very difficult to identify. Since the content is hidden and popped up in non-identical formats [2].

Keeping an eye on these problems, the shared Task1 by EXIST at IberLEF2002 Task1[3], concentrates on sexism identification to classify tweets. The classification should be performed by binary classification. The content is available in English and Spanish. This paper overviews the overall approaches made to solve this task.

2. Related Work

Sexism identification on a multilingual dataset with text representing tweets in Spanish and English languages had been implemented using different machine learning, deep learning and transformer models. The challenges associated with the multilingual nature of the dataset and the inconsistencies in the structure of the tweets had been handled by the process of pre-processing [2]. The multilingual BERT transformer model had been used for the process of classification by Altin et al. [4].

The use of a Multi-Task Learning approach to identify sexism related tweets had been implemented by [5] and [6]. A shared representation [6] had been used for multiple tasks like sentiment analysis and offensive language detection and were learned in parallel to support the process of sexism identification.

Graph Convolutional Neural Networks (GCN) exploring different edge creation strategies and graph embeddings based ensemble models had been used for the process of sexism identification [7]. It had been shown that the syntactic relationship between words encoded in the graph had an impact on the result of the model.

The performance of different pre-trained transformers, such as BERT and roBERTa had been compared with traditional Machine Learning approaches, such as SVM, Logistic Regression, SGD-based classifier and XGBoost for identifying sexism from text data [8]. It had been shown that the transformer models outperformed the traditional models. The impact of linguistically motivated features on the detection of sexism had been represented by machine language classifiers with an approach based on linguistic and stylistic features. [9].

A recurrent neural model for sexism detection and classification had been implemented by combining RoBERTa representations with linguistic features like Empath, Hurtlex, and Perspective API [10]. The emoticon and hashtag representations had also been used to infuse external knowledge-specific features into the learning process. Rodríguez-Sánchez et al. [11] had developed a Spanish data set for sexist expression and attitude by collecting posts from Twitter and used both machine learning and deep learning models to identify sexism related text and has shown that deep learning algorithms had provided better results.

Language-agnostic BERT Sentence Embedding(LaBSE) pretrained, multilingual model.When implemented for evaluating two different languages, LaBSE achieved more efficiency than other deep learning model. [10]. Different deep neural network architectures like Long-Short-Term Memory (LSTMs) and Convolutional Neural Networks (CNNs) and transformer models like BERT and DistilBert had been used to identify sexism from the dataset of tweets and gabs provided by the sEXism Identification in Social neTworks (EXIST) task in IberLEF 2021. BERT based model and multi filter CNN model had shown better performance and the use of data augmentation to improve the result also had been depicted [12]. The use of ensembled model developed by combining different pretained transformer models for identifying sexism related text had been implemented by Angel Felipe Magnossão de Paula and etal. [8]. The same task had been implemented with three different BERT based models and the classification had been done by combining the individual results using a voting based mechanism [13]. Two transformer models namely multilingual BERT and XLM-R had been used for the process of sexism identification and had used two approaches for the implementation. The first approach had used unsupervised pre-training with additional data and the other had used supervised fine-tuning with additional and augmented data. The XLM-R model when used with supervised fine tuning had reported better performance for the task [14].

From the study we observed the tweets have inconsistent structure, data preprocessing will help to improve accuracy of the training model. Deep learning algorithm can identify the sexism from tweets. Categorical data are be handled by logical regression.

3. Task Description

IberLEF is a shared evaluation campaign for Natural Language Processing (NLP) in Spanish and other Iberian languages. The second edition of sEXism Identification in Social neTwork task at IberLEF 2022 contains, the Task1 aims on classification of text from twitter dataset. The tweets may or may not contain sexist expression or behaviours. The first sub task Task1, is to perform the binary classification in tweets. The text should be analyzed and result should be either sexist or IberLEF is a shared evaluation camp non-sexist.

- Sexist-describes a sexist situation or criticizes a sexist behaviour.
- Non-Sexist- describes the text does not criticizes a sexist situation or behavior.

3.1. Dataset Description

The dataset is offered by EXIST at IberLEF 2022 that are collected from tweets in twitter containing implicit sexism behaviours. The training data set contains 6977 entries with features test_case, id, source, language, test and Task1. The test data contains 1058 entries with labels namely test_case, id, source, language and text . Identifying the text from social media is the need of an hour. In accordance with this IberLEF 2022 aims on sexism identification of text in English and Spanish languages from twitter. The given text is identified and labeled as sexist or non-sexist.

Table 1
Sample Data

Sample Tweet
@BirdoOnDaLow @bonsorlol YOU'RE RUNNING AWAY LIKE A BITCH!???!!!!!
@MrGee54 Meet me in Temecula was a wild ass time
I Donâ€™t Hate Women I Am Simply On 70 Mg Of Adderall And Poasting My Stream Of Consciousness

Table 2
Dataset

	Source	Size	English	Spanish	sexist	Non-sexist
Train data	Twitter	6977	3437	3540	3377	3600
Test data	Twitter	1058	527	531	-	-

4. Methodology

Multilingual Language Model(MLM) are pretrained models that are trained on large corpora in various domain. The model need not be trained from start, but it can be fine tuned to achieve the desired target. BERT multilingual and LaBSE are MLM models which can be implemented desired target by fine tuning.

The BERT multilingual model is a pretrained model for 104 languages. This transformer model is case sensitive which performs on self supervised technique. We imparted the parameters num_labels, use_cuda, num_train_epochs to train our model. The model is trained aiming on masked language modeling and next sentence prediction [15].

Language-agnostic BERT Sentence Embedding called LaBSE, is a multilingual, embedding model works for cross-lingual sentence embedding for 109 languages. The model is trained by 17 billion monolingual sentences and 6 billion bilingual sentences. The model appears to be effective for low resource language.

The Task1 EXIST is offered by IberLEF2022 is to classify and label the text as sexist and non-sexist. The text incorporates both English and Spanish. The classification is performed using deep learning model. The dataset is trained using BERT multilingual model and LaBSE model. The results are compared and found LaBSE model performed better BERT multilingual model.

4.1. Preprocessing

The training data is preprocessed before applying to the model to avoid overfitting. The preprocessing is performed by eliminating the stop words, punctuations, email, numbers, special characters, tags which are not appropriate to identify data. All the characters are converted to lower case. The preprocessing step will help the model to perform better.

Table 3
Submission score

Run	Model	Precision	Recall	F1 Score
1	LaBSE	0.735	0.7338	0.7329
2	BERT Multilingual	0.7006	0.6984	0.6968

4.2. Model Building

Among the available information, the text and Task1 labels are selected from the training dataset. The Task1 label comprises either sexist or non-sexist data. Using LabelBinarizer, a SciKit learn class used to binarize labels, converts multiclass labels to binary labels. Using this binarizer the Task1 label is converted to categorical data.

The shared task from EXIST at IberLEF2022, Task1 focus on identification of sexism data given in bilingual languages English and Spanish. As per the instruction specified in the task We carried out the experiment by binary classification model in deep learning. The two deep learning models BERT multilingual and LaBSE were used to accomplish the task.

In Run1 we tried the experiment with LaBSE classification model. The needed labels are selected from training dataset. Before the instances are applied to the model, they are preprocessed by eliminating the unrelated characters which will affect the accuracy of the model. The BERT multilingual classification model is fine tuned by feeding the training dataset. The model is trained for one epoch. The eval dataset is supplied as the input to the model and the text is predict as sexist or non-sexist.

In Run2 we selected BERT multilingual classification model to perform the classification. The needed labels are selected from training dataset. Before the instances are applied to the model, they are preprocessed by eliminating the unrelated characters which will affect the accuracy of the model. The BERT multilingual classification model is fine tuned by feeding the training dataset. The model is trained for 5 epochs. The eval dataset is supplied as the input to the model and the text is predict as sexist or non-sexist.

5. Predictions

Comparing the results of two deep learning model, BERT multilingual and LaBSE and we arrived at the conclusion that the LaBSE performed well on sexism identification. The performance of the model is measured with the metrics accuracy, recall, F1 Score. For Run1-BERT multiingual model, the accuracy and F1 score are obtained as 75.33%, and 75.30% respectively. These measures placed in the leader board with the rank 25 for LaBSE model. The accuracy and F1 score are obtained as 71.83% for BERT multilingual model, placed in the lead board with the rank 37.

6. Conclusion

The Task1, EXIST is a subtask provided by IberLEF2022 to perform binary classification on tweets from twitter. We experimented the task by predefined deep learning models BERT multilingual model and LaBSE model. We evaluated the performance metrics for both the models and found LaBSE model produced competitive good result when compared to other deep learning models, to identify sexism for the dataset offered by EXIST at IberLEF2022 with languages English and Spanish. The model is trained for one epoch only. The best performance can be achieved by training the model for more epochs.

We have focused only on bilingual language English and Spanish, as suggested by EXIST at IberLEF 2022. Further research can be made to find the performance of the model in more languages.

References

[1] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[2] S. Butt, N. Ashraf, G. Sidorov, A. Gelbukh, Sexism identification using bert and data augmentation-exist2021, in: International Conference of the Spanish Society for Natural Language Processing SEPLN 2021, IberLEF 2021, 2021.

[3] F. Rodríguez-Sánchez, J. Carrillo-de-Albornoz, L. Plaza, A. Mendieta-Aragón, G. Marco-Remón, M. Makeienko, M. Plaza, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2022: sexism identification in social networks, Procesamiento del Lenguaje Natural 69 (2022).

[4] L. S. M. Altin, H. Saggion, Automatic detection of sexism in social media with a multilingual approach, in: IberLEF@ SEPLN, 2021.

[5] F. Rodrıguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, A multi-task and multilingual model for sexism identification in social networks (2021).

[6] F. M. Plaza-del Arco, M. D. Molina-González, L. Alfonso, Sexism identification in social networks using a multi-task learning system (2021).

[7] R. Wilkens, D. Ognibene, Mb-courage@ exist: Gcn classification for sexism identification in social networks, IberLEF@ EXIST (2021).

[8] A. F. Magnossão de Paula, R. Fray da Silva, I. Baris Schlicht, Sexism prediction in spanish and english tweets using monolingual and multilingual bert and ensemble models, arXiv e-prints (2021) arXiv–2111.

[9] K. Bengoetxea, I. Gonzalez-Dios, Multiaztertest@ exist-iberlef 2021: Linguistically motivated sexism identification (2021).

[10] H. Abburi, S. Sehgal, H. Maheshwari, V. Varma, Knowledge-based neural framework for sexism detection and classification (2021).

[11] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, Automatic classification of sexism in social networks: An empirical study on twitter data, IEEE Access 8 (2020) 219563–219576. doi:10.1109/ACCESS.2020.3042604.

[12] A. Kalra, A. Zubiaga, Sexism identification in tweets and gabs using deep neural networks, arXiv preprint arXiv:2111.03612 (2021).

[13] C. Feng, A simple voting mechanism for online sexist content identification, arXiv preprint arXiv:2105.14309 (2021).

[14] S. Mina, B. Jaqueline, L. Daria, S. Djordje, K. Armin, H. Manuel, B. Johannes, S. Sven, S. Alexander, Z. Matthias, Automatic sexism detection with multilingual transformer models, arXiv preprint arXiv:2106.04908 (2021).

[15] M. Schütz, J. Boeck, D. Liakhovets, D. Slijepcevic, A. Kirchknopf, M. Hecht, J. Bogensperger, S. Schlarb, A. Schindler, M. Zeppelzauer, Automatic sexism detection with multilingual transformer models, CoRR abs/2106.04908 (2021). URL: https://arxiv.org/abs/2106.04908. arXiv:2106.04908.