Sexism Identification In Tweets Using Machine Learning Approaches

Murari Sreekumar, Shreyas Karthik, Durairaj Thenmozhi, Shriram Gopalakrishnan and Krithika Swaminathan

Sri Sivasubramaniya Nadar College Of Engineering, Rajiv Gandhi Salai (OMR), Kalavakkam 603 110, Tamil Nadu, India

Abstract

Sexism poses significant challenges in sentiment analysis, as it can manifest in subtle and nuanced ways, often embedded within seemingly benign language. On social media, where communications are frequently code-mixed, particularly in Dravidian languages, there is an increasing demand for identifying sexist content to ensure healthy online interactions. The EXIST 2024 shared task aims to detect sexism in Spanish and English tweets collected from social media platforms. Various traditional machine learning approaches are employed to identify whether the comments contain sexist content in Spanish and English languages. Utilizing Support Vector Machines (SVM), Random Forest and Logistic Regression as a classifier, we achieve F1 scores of 0.6299, 0.6074 and 0.5518 respectively for English dataset.

Keywords

Sexism Identification, Traditional Machine Learning Algorithms, Natural Language Processing, Sentiment Analysis, Text Analytics

1. Introduction

Sexism is prejudice or discrimination based on one's sex or gender. Sexism can affect anyone, but primarily affects women and girls. It has been linked to gender roles and stereotypes, and may include the belief that one sex or gender is intrinsically superior to another. With the advent of social media people have begun misusing the freedom speech and expression and instead have engaged in lot of hate speech on women politicians, journalists, personalities etc. This has especially risen in social media platforms such as twitter during the pandemic time [1].

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 [2][3]. it is crucial to address these aspects of sexism in social networks and hence Natural Language Processing research is crucial in providing insights into identifying the tweets and classifying them as Sexist and Non-Sexist. Computational understanding of natural language has been used in addressing issues such as sentiment analysis[4], human behaviour detection, fake news detection[5], question answering and depression and threat detection across different forms of media.

Our research paper presents various innovative solutions contributing to the field of sexism identification in significant ways:

- Annotated Datasets: We leverage a vast dataset annotated by multiple annotators so that it ensures model's robustness and improves the accuracy of sexism identification.
- Optimized Approach: The models used in this research like Support Vector Machines (SVM), Logistic Regression, and Random Forest have their hyperparameters tuned to their finest level so that it effectively identifies sexiset tweets.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

murari2310237@ssn.edu.in (M. Sreekumar); shreyas2310140@ssn.edu.in (S. Karthik); theni_d@ssn.edu.in (D. Thenmozhi); shriram2310156@ssn.edu.in (S. Gopalakrishnan); krithika2010039@ssn.edu.in (K. Swaminathan)

(0000-0003-0681-6628 (D. Thenmozhi)

- This project can be used for real time applications in social media platforms like Twitter, Instagram, Facebook, LinkedIn etc in order to maintain a healthy and safe online environment.

The task that we have performed in EXIST 2024 is Sexism Identification in Tweets. In this task, the systems have to decide whether the tweets are Sexist or Not Sexist.

In this research paper, we have discussed the research works that we have done for Task 1. The rest of the paper is organised as follows: Section 2 presents a literature survey explaining the key theories and concepts, research methodologies and the trends and patterns common in the field of sexism identification. Section 3 describes the different datasets used and the task performed. Section 4 talks about the methodology like preprocessing, lemmatization, vectorization and the various models used for our task. Section 5 talks about our results and performance analysis with other teams participating in the task. Finally, in Section 6 we talk about the conclusions and the future prospects of the research work.

2. Related Work

Various works in the field of Sexism Identification were studied and diverse methodologies and approach for sexism identification and classification were employed to solve this issue. Significant efforts have been made by researchers around the world to develop annotated datasets and apply deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). In addition to these, various transformer based models like BERT have been used as they have consistently provided excellent accuracy in identifying sexist tweets.

Rodríguez-Sánchez et al. (2020) [6] undertook a research on automatic classification of sexism in social networks. They specialized mainly on Twitter data in Spanish. They developed the MeTwo dataset that labels the tweets into sexist, non-sexist and doubtful. This is the first dataset in Spanish used to identify sexism in a broad sense, ranging from hostile to subtle sexism.To classify the tweets into three categories, they have used various traditional Machine Learning models like Support Vector Machine (SVM), Logistic Regression, Random Forest, and Naive Bayes. Various advanced deep learning models like Bidirectional Encoder Representations from Transformers (BERT), Bidirectional Long Short Term Memory (Bi-LSTM), Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have also been used. This research done by them can be used in fields such as misogyny detection in tweets and various other texts.

Davidson et al. (2017) [7] in their research worked to distinguish hate speech from offensive language on social media. They collected the tweets and labeled them into three categories namely hate speech, offensive language and neither. First, they converted all the text into lowercase, stemmed the text to obtain the root words using PorterStemmer, create bigram, unigram and trigram features using TF-IDF. They used Penn Part-Of-Speech (POS) tagging and included count indicators for r hashtags, mentions, retweets, and URLs, as well as features for the number of characters, words, and syllables in each tweet. Then various models like Logistic Regression, naive Bayes, decision trees, random forests, and linear SVMs. These models successfully classified racist and homophobic slurs as hate speech, while sexist language was more frequently categorized as offensive.

Harika Abburi et al. (2021) [8] worked on Fine-Grained Multi-label Sexism Classification Using a Semi-Supervised Multi-level Neural Approach. They initially employed the technique of Self-training, which is a semi supervised learning approach that helps augment the set of labeled instances by selectively adding unlabeled samples. Then it applies the models to the unlabeled instances and identifies a subset of them to be added to the training set, along with the predicted labels. To address categories with scarce labeled data, they propose a multi-level training approach. The model trains initially on a reduced set of broader categories (coarse), then refines its understanding on the full set of fine-grained categories. To begin with, the data was tested on various Traditional Machine Learning models like logistic regression (LR), Support Vector Machine (SVM) and Random Forests (RF) Classifiers. These were applied on two feature sets namely TF-IDF on word unigrams and bigrams (Word ngrams) and the average of the ELMo vectors. Then various Deep Learning techniques like BiLSTM, BERT and

Table 1
Distribution of tweet samples across training, development and testing for each language

Task	Language	No. of samples	Percentage (%)
Training	English	3260	47.1
Training	Spanish	3660	52.9
Development	English	489	47.1
Development	Spanish	549	52.9
Testing	English	978	47.1
Testing	Spanish	1089	52.9

other CNN based architectures were used. Thus, this approach can be used to analyze online sexism by using unlabeled data and various Deep Learning and Neural Network models.

S Sharifirad et al. (2019) worked on a comprehensive classification of different online harassment categories and explain its challenges using NLP. The tweets have been classified into Indirect Harassment, Information Threat, Sexual Threat and Non Sexist. They have used various classification methods like bigrams, threegrams, Two Character Grams, Word2Vec, Doc2Vec, Long Short Term Memory (LSTM) among others. These techniques help identify boundaries between words or phrases in text, especially in languages without explicit word separators. By analyzing sequences of words, n-grams can be used to predict the next word in a sequence, which is useful for tasks like text generation. They have used neural networks and the traditional machine learning technique Naive Bayes. The tweets were classified correctly in their categories with accuracy ranging from 0.66 to 0.91 for LSTM.

Thus, it is found that while significant progress has been made in identifying and mitigating various forms of sexism on social networks, many existing studies primarily focus on explicit instances of sexist language. However, the detection and analysis of more subtle, implicit forms of sexism remain under-explored. Additionally, the intersection of sexism with other forms of discrimination, such as racism or homophobia, has not been thoroughly investigated. This research aims to address these gaps by developing more sophisticated algorithms that can identify both explicit and implicit sexist content, considering the broader context of intersectional discrimination in social network environments. In addition to these, we also aim to integrate these techniques in various social media platforms to ensure safe and healthy online environments.

3. Task and Dataset

The task organizers of CLEF2024 provided a dataset called EXIST2024 [9][10]. The EXIST2024 dataset contains exactly 6920 tweets for training, 1038 tweets for development and 2076 tweets for testing which adds upto to an overall of more than 10000 tweets.

From the above table, it can be observed that the training, development and testing dataset contain English and Spanish tweets in the same ratio.

TASK 1: Sexism Identification in Tweets 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 criticizes a sexist behaviour). The following tweets show examples of sexist and not sexist messages. The opinions of Six annotators were also given. These annotators classified the tweets into Sexist and Non Sexist using "YES" and "NO". The opinion given by the majority of the annotators was taken into account for every tweet and then used for identifying whether a tweet is sexist.

Table 2
Examples of Sexist and Non-Sexist Statements

Sexist	Non-Sexist
"Mujer al volante, tenga cuidado!"	"Alguien me explica que zorra hace la gente en el cajero que se demora tanto."
"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."	"@messyworldorder it's honestly so embarrassing to watch and they'll be like 'not all white women are like that'"

4. Methodology

We trained the traditional machine learning models such as Support Vector Machine (SVM) [11, 12], Random Forest and Logistic Regression on the training dataset, evaluated the models on the dev dataset and submitted our runs by applying the ML models on the test dataset.

4.1. Preprocessing

Our first step was to clean the data given in order to improve the performance of the machine learning models:

1. Converting the text to lowercase: This ensures consistency in text data. By doing this the vocabulary size is reduced and it reduces the computational requirements.
2. Removing punctuation marks:They often point to external resources that are not relevant to the context of the text being analyzed.
3. Removing http links and emoticons:These do not contribute to the semantic meaning of the text.
4. Removing twitter mentions like `@username`
5. Removing all the numbers from the tweet column: These do not contribute towards sexist words.
6. Removing stop words like "a", "an", "the", "is" and so on to improve the accuracy of the models.

4.2. Lemmatization

Lemmatization is a crucial step in preprocessing data where the words in the text are converted to the base form. We have preferred to use Lemmatization as it followed grammatical rules better than Stemming. This process involves:

- Identifying the part of speech: Understanding whether a word is a noun, verb, adjective, etc., which helps in determining the correct lemma.
- Morphological analysis: Analyzing the structure and form of the word to convert it to its base form.

4.3. Vectorization

In order to ensure that the data is understood well by the model we need to convert the data into a format that machines can understand,typically vectors or array of numbers. Among vectorization techniques we found TF-IDF vectorization to give a better accuracy. Basically it adjusts the frequency of words by how commonly they appear across all documents, giving more weight to less common but significant words.

4.4. Model Evaluation

We have used three models using hard-hard labels such as Sexist and Non-Sexist. They are:

1. Support Vector Machines: A supervised machine learning algorithm that we used for classification and regression tasks. It operates by creating a decision boundary that separates n-dimensional spaces into classes so that a new data point can be assigned to its relevant category.
2. Logistic Regression: It is a regression model mainly used for classification problems. Logistic regression models the probability that a given input belongs to a particular class. It uses the logistic function, also known as the sigmoid function, to map any real-valued number into the range [0, 1].
3. Random Forest: It is an ensemble learning method in which multiple decision trees are built during training and merges their results to improve accuracy and over-fitting .
4. Decision Trees: A tree-like model of decisions and their possible consequences, including outcomes, resource costs, and utility.
5. Hyper-parameter tuning: This is an essential step that helps in optimizing the performance of the models used for classifying the tweets. Hyper-parameters are configurations external to the model that cannot be learned from the data, such as learning rate, batch size, and the number of layers in a neural network. Since the data used in NLP is highly complex and multi-dimensional hyper-parameter tuning is used to identify optimal hyper-parameter configurations in order to make the models more efficient and accurate. There are various methods of hyper-parameter tuning like GridSearchCV, RandomSearchCV, Bayesian Optimization and Gradient-based Optimization. We have used GridSearchCV for our research.
   For SVM, we have tuned the hyper-parameters like regularization parameter (C) and the kernel parameters, such as the gamma parameter for the radial basis function (RBF) kernel. For Logistic Regression, we have tuned hyper-parameters like the regularization strength (often denoted as C). Regularization techniques such as L1 (lasso) and L2 (ridge) are also tuned to improve model generalization.

5. Results and Performance Analysis

5.1. Performance Analysis

Scikit-learn, also known as sklearn, is an open-source, machine learning and data modeling library for Python. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python libraries, NumPy and SciPy. The sklearn metrics library also provides the classification report for evaluation of the performance of the model. The performance is measured using the following metrics:

1) Precision: Precision is defined as the ratio of true positives to sum of true and false positives.

2) Recall: Recall is defined as the ratio of true positives to sum of true positives and false negatives.

3) F1-Score: The F1 is the weighted harmonic mean of precision and recall. The closer the value of F1 is to 1, better is the performance of the model.

The result of the task is represented in the form of the table below. Among the 3 being used SVM had the best F1 score of 0.6299. The next best F1 score came from Random Forest which is of 0.6074. Logistic Regression had an F1 score 0.5518. Table 4 also displays the ranking of our submissions based on the shared task official ranking in (hard-hard) evaluation scenario.

6. Reflections

Through this paper we learnt about important methods in the filed of natural language processing and the steps involved in it .We learnt through this task that SVM in general is a very good model for text classification as they are particularly effective in cases where the number of dimensions (features) is

Table 3
Ranks based on the f1 score of our 3 models in comparison with others

Run	Rank	ICM-Hard	ICM-Hard Norm	F1
FraunhoferSIT_1	55	0.2334	0.6191	0.6447
The-Three_Musketeers_2	56	0.2171	0.6108	0.6299
The-Three_Musketeers_3	57	0.2130	0.6087	0.6074
maven_2	58	0.1926	0.5983	0.6512
The-Three_Musketeers_1	60	0.1184	0.5604	0.5518

greater than the number of samples. This makes them suitable for applications like text classification, where each word can be considered a feature. While SVMs work with linear hyperplanes by default, the 'kernel trick' allows them to handle non-linear relationships between features. This is crucial for text, where complex semantic relationships exist between words.

7. Conclusions

Through the scope of the paper we have explored traditional models to perform classification of Sexist and Not-Sexist speech on the given data by EXIST in English Language. It was noted that the SVM had the best F1 score of 0.6299. This research contributes to the field of natural language processing and provides valuable insights into addressing social issues in online platforms. Future work can be done in incorporating more advanced techniques and also introduce more pre-processing techniques in order to improve the performance of the model. Additionally the model can be deployed in real world applications in order to monitor sexist tweets on social 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. We hope these efforts will contribute towards fight against sexism.

References

[1] N. Dehingia, J. McAuley, L. McDougal, E. Reed, J. G. Silverman, L. Urada, A. Raj, Violence against women on twitter in india: Testing a taxonomy for online misogyny and measuring its prevalence during covid-19, PLoS one 18 (2023) e0292121.

[2] A. Chaudhary, R. Kumar, Sexism identification in social networks, Working Notes of CLEF (2023).

[3] R. Ouedraogo, D. Stenzel, How domestic violence is a threat to economic development, IMF Blog Insights & Analysis on Economics and Finance (2021).

[4] L. Khan, A. Amjad, N. Ashraf, H.-T. Chang, A. Gelbukh, Urdu sentiment analysis with deep learning methods, IEEE access 9 (2021) 97803–97812.

[5] Z. Khanam, B. Alwasel, H. Sirafi, M. Rashid, Fake news detection using machine learning approaches, in: IOP conference series: materials science and engineering, volume 1099, IOP Publishing, 2021, p. 012040.

[6] 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.

[7] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, in: Proceedings of the international AAAI conference on web and social media, volume 11, 2017, pp. 512–515.

[8] P. Parikh, H. Abburi, N. Chhaya, M. Gupta, V. Varma, Categorizing sexism and misogyny through neural approaches, ACM Transactions on the Web (TWEB) 15 (2021) 1–31.

[9] L. Plaza, J. Carrillo-de Albornoz, E. Amigó, J. Gonzalo, R. Morante, P. Rosso, D. Spina, B. Chulvi,

A. Maeso, V. Ruiz, Exist 2024: sexism identification in social networks and memes, in: European Conference on Information Retrieval, Springer, 2024, pp. 498–504.
[10] L. Plaza, J. Carrillo-de-Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Social Networks and Memes (Extended Overview), in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.
[11] S. L. Salzberg, C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993, 1994.
[12] T. Pranckevičius, V. Marcinkevičius, Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification, Baltic Journal of Modern Computing 5 (2017) 221.