Brainstormers_msec at SemEval-2023 Task 10: Detection of sexism related comments in social media using deep learning

C.Jerin Mahibha and **C.M.Swaathi** and **R.Jeevitha** and **R.Princy Martina**
Meenakshi Sundararajan Engineering College, Chennai
[jerinmahibha, cmswaathi0306, rahijee365,princymartina2001]@gmail.com

Durairaj Thenmozhi
Sri Sivasubramaniya Nadar College of Engineering , Chennai
theni_d@ssn.edu.in

Abstract

Social media is the media through which people share their thoughts and opinions. This has both its pros and cons which depends on the type of information being conveyed. If any information conveyed over social media hurts or affects a person, such information can be removed as it may disturb their mental health and may decrease their self confidence. During the last decade, hateful and sexist content towards women is being increasingly spread on social networks. The exposure to sexist speech has serious consequences to women's life and limits their freedom of speech. Sexism is expressed in very different forms: it includes subtle stereotypes and attitudes that, although frequently unnoticed, are extremely harmful for both women and society. Sexist comments have a major impact on women being subjected to it. We as a team participated in the shared task Explainable Detection of Online Sexism (EDOS) at SemEval 2023 and have proposed a model which identifies the sexist comments and its type from English social media posts using the data set shared for the task. Different tranformer models like BERT, DistilBERT, and RoBERTa are used by the proposed model for implementing all the three tasks shared by EDOS. On using the BERT model, macro F1 score of 0.8073, 0.5876 and 0.3729 are achieved for Task A, Task B and Task C respectively.

1 Introduction

There is a lot of information being spread in social media. These can be categorised as necessary and unnecessary information. Any information that helps in the growth of a persona and society is considered as necessary information. Any information that is in any way not useful to the society or harms the livelihood and mental health of an individual is considered as unnecessary information which can be removed for the benefit of the society.

Hateful and sexist content towards women that are increasingly spread in the social media networks has a serious impact on their life and limits their freedom of speech. An exhaustive analysis is required to understand how sexist behaviour and attitudes prevail in social media platforms (Rodríguez-Sánchez et al., 2020). Online violence against women can have a serious impact on them as it may affect their mental health or bring hesitation in voicing out their points in social media. Hence it is important to find whether the content that is present in social media is sexist and such content should be removed (Shimi et al., 2022). It is necessary to explain the reason behind the classification of text as sexist.Thus an explainable classification of sexist comments become necessary.

By silencing or pushing women out of online spaces, online violence can affect the economic outcomes of those who depend on these platforms for their livelihoods[1]. It can also lead to loss of employment and societal status, in cases where online violence impacts their reputation.

At this time, although research is beginning to delve into the incidence of sexist content in social media, experimental research examining the effects of these messages is scarce.

Hashtags on Twitter are used to establish topical links. By using hashtags before words or phrases in their posts, users categorise tweets. By searching hashtags, users can find tweets relevant to a specific topic posted by other users on Twitter. Thus, Twitter enables users to disseminate their tweets publicly, and other users can easily track and view those tweets. On Twitter, several sexist hashtags have gone viral. Some sexist hashtags, such as #GetBackInTheKitchen, #LiesToldByFemales and #IHateFemalesWho, promote stereotypes and hostility toward women. Others, such as #RulesForGirls and #MyGirlfriendNotAllowedTo, suggest that men are responsible for regulating women's

[1]https://gehweb.ucsd.edu/social-media-sexist-online-gender-based-violence/

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1114–1120

behaviour and that women should submit to male authority. Some of the most appalling, such as #ThatsWhatSlutsDo and #ItsNotRapeIf, promote rape myths and dehumanize women as objects whose only function is sex. Not just twitter, this type of behaviour can be seen in many other social media platforms as well. In order to avoid the negative impacts such kind of behaviour has, it is necessary to find such comments and remove them (Sivamanikandan et al., 2022). It is equally necessary to explain why this action must be taken. This task aims in identifying the content that are sexist and explains why it is marked as sexist.

The shared task Explainable Detection of Online Sexism (EDOS) (Kirk et al., 2023) was a part of SemEval 2023 which is based on English comments. The task contains three hierarchical subtasks namely Task A, Task B and Task C.

Task A: Sexist vs. Non-Sexist

The first task was a straightforward Binary Classification problem, in which the model had to predict whether a given phrase is sexist or not. For example, *"Women belong to kitchen"* is sexist, whereas *"A doctor cares for his patient"* is non-sexist.

Task B: Categories of sexism

The second task was a multi-label classification problem, where the aim was to identify the categories of sexism. There are four types of classification where systems have to predict one of four categories: threats, derogation, animosity and prejudiced discussions.For example, *"Choose to follow me or oppose me. If you oppose, then I will rape you and kill too."* is considered as Threat, whereas *"Girls are always more docile than boys and want to please the opposite gender."* is considered to be of type Animosity.

Task C: Fine grained vector of sexism

The final task is to predict the sexist post in one of the 11 categories: Threats of harm, Incitement and encouragement of harm, Descriptive attacks, Aggressive and emotive attacks, Dehumanising attacks and overt sexual objectification, Causal use of gendered slurs, profanities and insults, Immutable gender differences and gender stereotypes, Backhanded gendered compliments, Condescending explanations or unwelcome advice, Supporting mistreatment of individual women, and Supporting systemic discrimination against women as a group. For example, *"Looking down your neck, I recognized that God is actually generous towards the creation."* is considered to be of type Dehumanisation and overt sexual objectification, whereas *"Women are not hired for the job in this poster, since they won't fit traditionally into male workplace."* fall in the class, Supporting systemic discrimination against women.

Various transformer based models are tried for implementing all the three classification tasks and the model that gave the highest accuracy was selected for submission. Considering all the three tasks the training of the proposed model was done using the training dataset provided for the corresponding task. This model was then tested with a testing dataset provided by the shared task, based on which the task was evaluated.

2 Related Works

Rodríguez-Sánchez et al. (2020) had carried out an empirical study on automatic classification of sexism in social networks using Bi-LSTM and mBERT. Sexism had been identified and auto classified from social media text by (Rodríguez-Sánchez et al., 2020) and (Shimi et al., 2022). Deep Learning Models had been used for sexism identification considering languages English and Spanish by Shimi et al. (2022). BERT multilingual and LaBSE models had been used to implement the desired targets by fine tuning. Fersini et al. (2019) had detected sexist memes on the web which had included different types of sexism against women like shaming, stereotyping and objectification to violence. Traditional machine learning algorithms including SVM, Naive Bayes, Multi-Layer Perceptron and Random Forest had been used for the classification process. Different deep learning techniques had been used for implementing intent classification (Purohit et al., 2015) as a multi-class classification problem for identifying different types of sexual assaults. An analysis of Women Safety in Indian Cities (Kumar and Aggarwal, 2019) had been done using machine learning models. Deep learning techniques like convolutional neural network (CNN), long short-term memory (LSTM) and bidirectional LSTM (Bi-LSTM) had been used to classify sexual violences from twitter and it had been observed that sexual assaults by a family member at own home is a more serious concern than harassment by a stranger at public places (Khatua et al., 2018). The intent in Twitter conversation on sexual assaults had been detected using an approach which had used deep neural network model

Task	Category	Tranining dataset	Evaluation dataset
Task A	Sexism	3398	486
	Non Sexism	10602	1514
Task B	Threats	310	44
	Derogation	1590	227
	Animosity	1165	167
	Prejudiced discussions	333	48
Task C	Threats of harm	56	8
	Incitement and encouragement of harm	254	36
	Descriptive attacks	717	102
	Aggressive and emotive attacks	673	96
	Dehumanising attacks & overt sexual objectification	200	29
	Casual use of gendered slurs, profanities, and insults	637	91
	Immutable gender differences and gender stereotypes	417	60
	Backhanded gendered compliments	64	9
	Condescending explanations or unwelcome advice	47	7
	Supporting mistreatment of individual women	75	11
	Supporting systemic discrimination against women as a group	258	37

Table 1: Data Distribution

as the feature extractor with a traditional logistic regression classifier (Pandey et al., 2018).

Offensive text from social media posts had been identified using linear regression algorithm and transformer based deep learning techniques (Mahibha et al., 2021). Kumar Sharma et al. (2018) had used machine learning techniques to detect insulting comments on social networking platforms. It had used binary level classification of text from online users which were in the form of comments, and status/post into two categories namely "Bully" and "Non-Bully". The pipeline used had been involved in the extraction of suitable dataset from various online sources, preprocessing, ground truth building, feature engineering and selection, classification. Rini et al. (2020) had done a systematic literature review of hate speech detection with text mining and had shown the process of hate speech detection using a variety of algorithmic methods and features. Classification of online toxic comments had been implemented using different machine learning algorithms like Logistic Regression, Random Forest, SVM, Naive Bayes, Decision tree, and KNN. The comments had been grouped into six categories namely threat, insult, toxic, severe toxic, obscene, or identity hate (Islam et al., 2020). Online aggression detection on Twitter data had been implemented using streamDM classification algorithms that are designed for incremental training, namely Hoeffding Tree, Adaptive Random Forest, and Streaming Logistic Regression (Kumar Sharma et al., 2018). The framework could also be extended to detect variety of behaviors like bullying, aggression, abuse, offense, racism, sexism.

The related works shows few works that are carried out on sexism identification. It is also found that continuous research are being carried out in related fields like identifying insulting comments, hate speech, toxic comments and intent classification which can be used as a base for identifying comments representing sexism from social media text. It could also be observed that, as the tweet and its contents have inconsistent structure, data preprocessing will help to improve the accuracy of the training model.

3 Data set

The data set that are used to implement sexism detection was the training, evaluation and the test dataset that were provided by the organisers of the shared task. Each instance of the training dataset had the following informations attached to it:

1. Label specifying whether the text is sexist or not
2. Category of sexism related to the sexist text
3. Sub category of sexism related text or rephrase

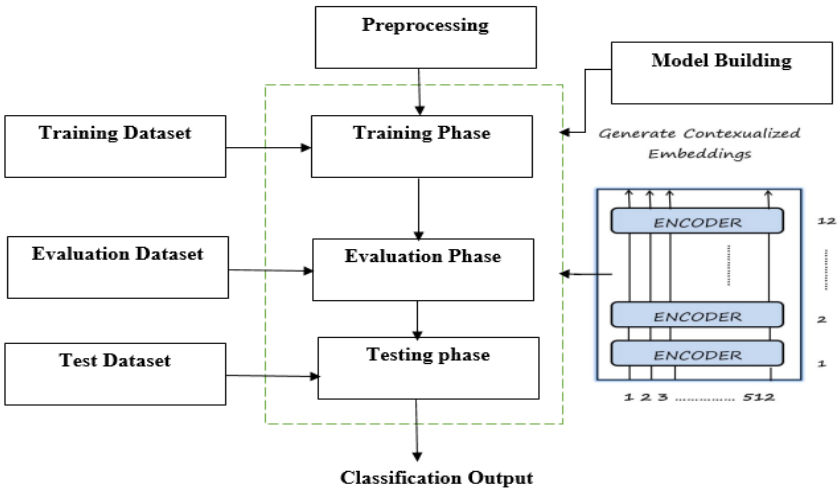

Figure 1: Proposed Architecture

The data distribution of the training and development dataset for the three tasks are shown in Table 1. The training dataset of Task A had 14000 instances of which 3398 instances were under the sexism category and 10602 instances were under the non sexism category. The development dataset of the same task had 486 and 1514 instances under the sexism and non sexism category respectively. This shows the unbalanced nature of the data set. The test data had 4060 instances for which the predictions had to be done using the proposed model. Task B was supported with training dataset which had 310, 1590, 1165 and 333 instances under the categories (i) threats, plans to harm and incitement, (ii) derogation, (iii) animosity and (iv) prejudiced discussions respectively. The training dataset of Task C had classifed the categories identified in Task B into 11 different subcategories.

4 System Description

Initially the three dataset provided by the task organizers, namely training dataset, development dataset and testing dataset were collected. Training dataset is preprocessed where unnecessary digits, characters and white spaces are removed using tokenization and it is followed by an encoding process. Then the model is created. In this system, three models were used namely BERT, DistilBERT and RoBERTa. The preprocessed dataset along with the model created is used for the training phase. Each model is then evaluated using a development dataset. The BERT model that provided the highest accuracy is taken as final run for submission and was used to find the predictions for the testing dataset.

The proposed architecture is represented in Figure 1. The removal of unnecessary information is taken care by the preprocessing phase. All the three sets of data namely training, evaluation and test dataset are preprocessed. This is followed by the process of model building, where pretarined models namely BERT, DistilBERT and RoBERTa were used. In the training phase the pretained models are trained using the preprocessed training dataset. The evaluation of the trained model is carried out in the evaluation phase using the evaluation dataset which makes use of the accuracy as the parameter of evaluation. Fine tuning of hyper parameters are performed to improve the accuracy of the proposed system. The labels for the text in the dataset are predicted during the testing phase. Contextual embeddings are generated and are used during the training of the model.

4.1 BERT

BERT (Devlin et al., 2018) is an open source machine learning framework for natural language processing (NLP). BERT, which stands for Bidirectional Encoder Representations from Transformers, is based on Transformers, a deep learning model in which every output element is connected to every input element, and the weightings between them are dynamically calculated based upon their connection. BERT is designed to read the input text in both directions at once. Using this bidirectional capability, BERT is pre-trained on two different, but related, NLP tasks: Masked Language Modeling

and Next Sentence Prediction.

BERT is made possible by Google's research on Transformers. By looking at all surrounding words, the Transformer allows the BERT model to understand the full context of the word, and therefore better understand searcher intent. BERT is currently being used at Google to optimise the interpretation of user search queries. BERT understands the context which helps it to interpret patterns that different languages share without having to understand the language completely.

4.2 DistilBERT

DistilBERT (Sanh et al., 2019) is a general-purpose pre-trained version of BERT which had been pre-trained on the same corpus as BERT in a self supervised fashion. Distil-BERT has 97% of BERT's performance while being trained on half of the parameters of BERT. BERT-base has 110 parameters and BERT-large has 340 parameters, which are hard to deal with. For this problem's solution, distillation techniques are used to reduce the size of these large models.

We have used "distilbert–base-cased" model to implement the classification task of identifying sexism from social media text which comprises of 6-layer, 768-hidden layers and also 12-heads, 65M parameters. It is a smaller version than BERT which is incredibly less expensive and quicker to train than BERT.

4.3 RoBERTa

RoBERTa (Liu et al., 2019) is a transformer model pre-trained on a large corpus of English data and is based on BERT model and modifies key hyper parameters and training is implemented with larger mini-batches and learning rates4. RoBERTa is a Robust BERT method which has been trained on a far extra large data set and for a whole lot of large quantities of iterations with a bigger batch length of 8k.

The "RoBERTa–base" model was also used for the task which is a pretrained model on English language using a masked language modelling (MLM) objective. This model is case-sensitive and it comprises 12-layers, 768-hidden layers, 12-heads and 125M parameters.

5 Results

The metrics that was considered for the evaluation of all the three tasks was macro-F1 score. F1 score is an overall measure of a model's accuracy that combines precision and recall. A high F1 score means that the classification has resulted with low number of false positives and low false negatives.

Tasks	Model	F1-Score	Accuracy
Task A	DistilBERT	0.79	0.85
	RoBERTa	0.80	0.86
	BERT	0.81	0.87
Task B	DistilBERT	0.57	0.59
	RoBERTa	0.57	0.60
	BERT	0.59	0.61
Task C	DistilBERT	0.35	0.50
	RoBERTa	0.37	0.52
	BERT	0.37	0.53

Table 2: Performance score

The values of the performance metrics namely F1 score and accuracy obtained for different models are shown in Table 2. It could be found that the BERT model outperformed the other models considering all the three tasks namely Task A, B and C.

All the three tasks were evaluated based on the macro F1 score obtained by the proposed model. The proposed model resulted in an macro F1 score of 0.81 based on which Task A was evaluated and we were ranked 61 on the leader board. The accuracy obtained for Task A was 0.87. Task B had resulted in a macro-F1 score of 0.59 and accuracy obtained was 0.61. We ranked 49 on the leader board under this category. Macro-F1 score and accuracy achieved by the proposed model for Task C were 0.37 and 0.53 respectively and we obtained 39th rank on the leader board.

6 Error Analysis

The F1 score obtained for Task A, Task B and Task C using the proposed BERT model shows that more false positive and false negative classification has occurred. One reason for this could be considered as the data imbalance nature of the dataset. Considering Task A, the number of instances for the class labelled non-sexiest is higher, and the F1 score, precision, and recall associated with this class are high when compared to the class sexiest. The same could also be found with Tasks B and C. This represents that the number of misclassifications increases when the number of instances for training is lower, which is associated with data imbalance. Data augmentation could be considered

SNo.	Text	Task	Predicted Label	Actual Label
1.	Also at the end he says women are more insecure and we need to capitalize and take advantage of that	Task A	Non Sexist	Sexist
2.	Well i guess according to feminists women aren t human then	Task A	Sexist	Non Sexist
3.	Bitch waited for her Christmas gifts first you gon die	Task B	Derogation	Threats, Plans to Harm and Incitement
4.	Every woman wants to be a model. Itś codeword for I get everything for free and people want me	Task B	Animosity	Derogation
5.	Not our fault she chose to date an abusive scumbag	Task B	Derogation	Prejudiced Discussions
6.	Women don't care about intelligence, good looks and big dick is all that matters	Task C	immutable gender differences and gender stereotypes	Descriptive attacks
7.	Fake it till you make it. For all she knows you're already drowning in slut	Task C	Aggressive and emotive attacks	casual use of gendered slurs, profanities and insults
8.	The thirst is strong. You can't find a woman ugly enough to not be getting dick.	Task C	Descriptive attacks	Dehumanising attacks and overt sexual objectification

Table 3: Error Analysis

to improve the model's performance. The presence of indirect sexism markers in the text and sarcastic posts could also be considered reasons for the poor performance of the model. Examples of texts that are misclassified due to the above reasons are shown in table 3. Considering the first text of the table, it does not have any specific sexism marker and is classified as non-sexist instead of the correct label of sexist. The second text in the table has the word 'feminists, which acts as a sexism marker. This sentence has been classified as sexiest, but based on the context, it falls under the category of non-sexist. The table also shows texts that are sarcastic which are misclassified. All the example texts show that sexism markers and sarcasm play a major role in the process of classification.

7 Conclusions

Sexism detection has become an important area of research as it is interlinked with different areas of application that includes sentiment analysis, opinion mining, offensive and hate speech detection. Having this in mind SemEval 2023 had come up with the task of sexism detection which was represented by three tasks namely sexism detection, identifying the categories of sexism related text and identifying whether the text or its rephrase is sexism. We have applied binary classification for Task A, B and C. And we have used BERT transformer model for training and testing the dataset. All the three tasks were implemented considering the language English.

Dataset for sexism detection could be created with contextual information which can help in effectively detecting sexism. Usage of hybrid approaches where different deep learning models are combined can also facilitate efficient detection of sexism from text. Often it could be observed that sexism is not in the text, but could be detected from the intonation or facial expression, which has made mulitmodal sexism detection also as a promising research area.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elisabetta Fersini, Francesca Gasparini, and Silvia Corchs. 2019. Detecting sexist meme on the web: A study on textual and visual cues. In *2019 8th International Conference on Affective Computing and In-*

telligent Interaction Workshops and Demos (ACIIW), pages 226–231.

Md Manowarul Islam, Md Ashraf Uddin, Linta Islam, Arnisha Akter, Selina Sharmin, and Uzzal Kumar Acharjee. 2020. Cyberbullying detection on social networks using machine learning approaches. In *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, pages 1–6.

Aparup Khatua, Erik Cambria, and Apalak Khatua. 2018. Sounds of silence breakers: Exploring sexual violence on twitter. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 397–400.

Hannah Rose Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 Task 10: Explainable Detection of Online Sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Deepak Kumar and Shivani Aggarwal. 2019. Analysis of women safety in indian cities using machine learning on tweets. In *2019 Amity International Conference on Artificial Intelligence (AICAI)*, pages 159–162.

Hitesh Kumar Sharma, K Kshitiz, and Shailendra. 2018. Nlp and machine learning techniques for detecting insulting comments on social networking platforms. In *2018 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, pages 265–272.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jerin Mahibha, Sampath Kayalvizhi, and Durairaj Thenmozhi. 2021. Offensive language identification using machine learning and deep learning techniques.

Rahul Pandey, Hemant Purohit, Bonnie Stabile, and Aubrey Grant. 2018. Distributional semantics approach to detect intent in twitter conversations on sexual assaults. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 270–277.

Hemant Purohit, Guozhu Dong, Valerie Shalin, Krishnaprasad Thirunarayan, and Amit Sheth. 2015. Intent classification of short-text on social media. In *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, pages 222–228.

Rini Rini, Ema Utami, and Anggit Dwi Hartanto. 2020. Systematic literature review of hate speech detection with text mining. In *2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS)*, pages 1–6.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, and Laura Plaza. 2020. Automatic classification of sexism in social networks: An empirical study on twitter data. *IEEE Access*, 8:219563–219576.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Gersome Shimi, Jerin Mahibha, and Durairaj Thenmozhi. 2022. Sexism identification in social media using deep learning models.

S Sivamanikandan, V Santhosh, N Sanjaykumar, Thenmozhi Durairaj, et al. 2022. scubemsec@ lt-ediacl2022: detection of depression using transformer models. In *Proceedings of the second workshop on language technology for equality, diversity and inclusion*, pages 212–217.