DS at SemEval-2023 Task 10: Explaining Online Sexism using Transformer based approach

Madisetty Padmavathi
4WE Tech Solutions
Hyderabad, Telangana
India
padmadatascience0@gmail.com

Abstract

In this paper, I describe the approach used in the SemEval 2023 - Task 10 Explainable Detection of Online Sexism (EDOS) competition (Kirk et al., 2023). I use different transformer models, including BERT and RoBERTa which were fine-tuned on the EDOS dataset to classify text into different categories of sexism. I participated in three subtasks: subtask A is to classify given text as either sexist or not, while subtask B is to identify the specific category of sexism, such as (1) threats, (2) derogation, (3) animosity, (4) prejudiced discussions. Finally, subtask C involves predicting a fine-grained vector representation of sexism, which included information about the severity, target and type of sexism present in the text. The use of transformer models allows the system to learn from the input data and make predictions on unseen text. By fine-tuning the models on the EDOS dataset, the system can improve its performance on the specific task of detecting online sexism. I got the following macro F1 scores: subtask A:77.16, subtask B: 46.11, and subtask C: 30.2.

1 Introduction

Sexism refers to any form of discrimination or prejudice against an individual based on their gender and can manifest in various ways, including harassment, stereotyping, and unequal treatment. It can be directed towards women, as well as individuals who identify as non-binary or gender nonconforming. The intersection of sexism with other forms of discrimination, such as racism or homophobia, can result in even greater harm to individuals who experience multiple forms of marginalization. These actions not only harm individual women but also create a toxic online culture that perpetuates social injustices and asymmetries. Online sexism also contributes to the underrepresentation of women and other marginalized groups in certain online spaces and industries, limiting their opportunities to participate fully in the digital world.

Although automated tools are now extensively employed to detect and evaluate sexist content at a large scale, most of these tools merely provide generic and high-level classifications without any additional clarification. This is a common limitation of many automated tools that are used to detect and evaluate sexist content online. While these tools can be useful in identifying patterns of sexist language or behavior, they often lack the contextual understanding necessary to provide more nuanced classifications or explanations. This can be a problem because online sexism can take many different forms, and what may be considered sexist in one context may not be in another. Therefore, it is essential to have tools that can accurately identify and contextualize instances of online sexism, rather than relying solely on high-level classifications.

To enhance interpretability, trust, and comprehension of the decisions made by automated tools, it is important to flag the sexist content along with an explanation of why it is considered sexist. This approach empowers both the moderators and users of the online space. providing explanations for why certain content is flagged as sexist is important for enhancing the interpretability, trust, and comprehension of the decisions made by automated tools. This approach can help moderators and users better understand why certain content is being flagged, and how they can work to improve the online space to prevent future instances of online sexism.

In addition to providing explanations for flagged content, it is also important to ensure that the explanations are clear and accessible to a wide range of users. This may involve providing additional context or examples to help users better understand why certain language or behavior is considered sexist, as well as using plain language and avoiding technical jargon to ensure that the explanations are easy to understand. By empowering both moder-

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1102–1106
July 13-14, 2023 ©2023 Association for Computational Linguistics

ators and users to better understand and address instances of online sexism, we can work towards creating safer and more inclusive online spaces that are welcoming to everyone, regardless of their gender or other identity attributes.

In this paper, I describe the approach used in three subtasks of the SemEval 2023 Task 10: Explainable Detection of Online Sexism (EDOS) (Kirk et al., 2023). The following are the details of three hierarchical subtasks.

- subtask A - Binary Sexism Detection: a two-class classification where the proposed model has to predict whether a post is sexist or not sexist.

- subtask B - Category of Sexism: for sexist posts, a four-class classification where the proposed model has to predict one of four categories: (1) threats, (2) derogation, (3) animosity, (4) prejudiced discussions.

- subtask C - Fine-grained Vector of Sexism: for posts which are sexist, an 11-class classification where the proposed model has to predict one of 11 fine-grained vectors.

2 Related Work

I divide the related work into two components: *hate speech detection* and *online sexism detection*.

2.1 Hate Speech Detection

Hate speech in online user comments is addressed in Djuric et al. (2015). The authors use neural language models to learn the distributional low-dimensional representations of the comments. A method to detect hate speech on Twitter is proposed in Waseem and Hovy (2016). They analyzed the impact of various other-linguistic features in conjunction with character n-grams for hate speech detection. The most indicative words are presented in a dictionary. Deep Learning based methods for hate speech detection in tweets is discussed in Badjatiya et al. (2017). The authors experimented on a benchmark dataset of 16K annotated tweets to show that such deep learning methods outperform state-of-the-art char/word n-gram methods by 18 F1 points. A method for offensive language detection is presented in Davidson et al. (2017). Their results show that fine-grained labels can help in the task of hate speech detection and highlight some of the key challenges to accurate classification.

A survey on hate speech detection using Natural Language Processing is explained in Schmidt and Wiegand (2017). The authors investigate about the key areas that have been automatically recognized using Natural Language Processing. Hate speech detection on Facebook is explained in Del Vigna12 et al. (2017). The authors implemented two classifiers for the Italian language. The first one is based on Support Vector Machines (SVM) and the second one is on Long Short Term Memory Recurrent Neural Networks (LSTM - RNN) . For detecting the hate expressions on Twitter the authors (Watanabe et al., 2018) proposed an approach, which is based on unigrams and patterns that are automatically collected from the training set. The authors (Gröndahl et al., 2018) describe that adversarial training does not completely mitigate the attacks, and using character-level features makes the models systematically more attack-resistant than using word-level features.

Challenges and solutions for hate speech are addressed in MacAvaney et al. (2019). The authors proposed a multi-view SVM approach that achieves near state-of-the-art performance, it produces very easily and explicable solutions than other methods. They also explained both technical and practical challenges. Deep learning based methods (Madisetty and Desarkar, 2018) are used to detect aggression in social media. The authors use CNN, LSTM, Bi-LSTM in their approach. The authors (Sap et al., 2019) show dialect and race priming as ways to reduce the racial bias in annotation, showing that when annotators are made explicitly aware of the tweet's dialect they are significantly less likely to label the tweet as offensive. A method to detect hate speech against women is demonstrated in Saha et al. (2018). They generate three types of features: Sentence Embeddings, TF-IDF Vectors, and BOW Vectors to represent each tweet. These features are then concatenated and fed into the machine learning models.

2.2 Online Sexism Detection

Now I describe the literature related to online sexism detection. A method to detect sexism using Twitter data is explained in Jha and Mamidi (2017). The authors used Support Vector Machines (SVM), sequence-to-sequence models, and FastText classifier. They achieved the best F1 score using FastText classifier. Automatic Identification of Misogyny on Twitter is explained in Frenda et al. (2019). The authors aim to analyze the differences and analogies between two aspects of online hate speech against women: misogyny and sexist behavior. They used a machine learning approach to detect automatically misogynistic and sexist tweets against women in the English Language. Automatic Classification of Sexism is described in Rodríguez-Sánchez et al. (2020). The authors have made it possible to detect using deep learning approaches. They discussed the performance of automatic classification methods to deal with different types of sexism and the generalizability of their task to other subdomains, such as misogyny. An Expert Annotated Dataset for the Detection of Online Misogyny is created by Guest et al. (2021). The researchers achieved the accuracy of 0.93 and an F1 of 0.43 using the binary classification task. They made datasets freely available for future researchers. The Misogyny dataset is created in Zeinert et al. (2021). The authors done three contributions, first one explained about the complete design of their iterative annotation process and codebook. The second one is that developed a comprehensive taxonomy of labels for annotating different types of misogynistic language in online posts, which can help to identify and categorize instances of online misogyny. Finally, the third one is the introduction of high-quality dataset of annotated posts sampled from social media posts. However, all the above methods do not have explanations in their methods.

3 System Description

3.1 Data Pre-processing

To prepare the data, each sample in the dataset is first concatenated with a given token in the format [CLS] + text + [SEP] sexism sentence [SEP]. This concatenation format included adding the [CLS] token for classification and two [SEP] tokens to identify the nature of the post. The resulting sentence is then tokenized, and the [CLS] and [SEP] tokens are added to the beginning and end of the sentence. The tokenized sentence is then truncated or padded to a maximum length of 512 to ensure uniformity across all samples.

3.2 BERT

BERT (Devlin et al., 2018)(Bidirectional Encoder Representations from Transformers) is a pre-trained language model developed by Google that can be fine-tuned on various NLP tasks such as question answering, sentiment analysis, text classification, and more. Pre-training refers to the initial stage of training BERT on a large corpus of text, such as Wikipedia or a web crawl, in an unsupervised manner. This pre-training allows BERT to learn contextual representations of words and sentences, which can then be fine-tuned on specific downstream tasks with relatively little data. The fine-tuning process involves training BERT on a smaller labeled dataset specific to the task at hand, which allows it to learn to perform the specific task accurately. The BERT base model has 12 transformer layers, 12 self-attention heads and 110 million parameters whereas BERT large model has 24 transformer layers, 24 self-attention heads and 340 million parameters.

3.3 RoBERTa

RoBERTa (Liu et al., 2019)(Robustly Optimized BERT pre-training approach) is a pre-trained language model developed by Facebook AI Research (FAIR) that is based on the same architecture as BERT. RoBERTa is pre-trained on a large corpus of English text using a self-supervised learning approach, which involves predicting masked words and predicting the next sentence in a given text. The pre-training allows RoBERTa to learn contextual representations of words and sentences, which can be fine-tuned on a variety of downstream NLP tasks such as text classification, sentiment analysis, and question answering. RoBERTa's learned inner representation of English can be used to extract features that are useful for downstream tasks. For example, if you have a dataset of labeled sentences for sentiment analysis, you can use RoBERTa to extract features from each sentence and then train a standard classifier (such as logistic regression or a neural network) using these features as inputs. This allows the model to leverage RoBERTa's pre-trained knowledge of English to improve its performance on the downstream task.

4 Experiments

Our approach (RoBERTa) utilizes the pre-trained weights of the RoBERTa model and fine-tunes it on the specific task of detecting sexism in text. Now I describe the fine-tuning process and loss functions used in subtask A for the proposed model. To fine-tune the model, a fully connected layer was added on top of the pooled output, and a sigmoid activation function was used to output the probability of each class as either 1 or 0. The proposed

Table 1: Performance Analysis on Test data

Task Name	Model	Macro F1 score	Competition Rank
subtask A	RoBERTa	77.16	109
subtask B	RoBERTa	46.11	76
subtask C	RoBERTa	30.2	66

model used an ensemble loss function that combined two loss functions, cross-entropy loss, and negative log-likelihood loss, in different weights proportion. The proportional weight functions used were 0.8/0.2 and 0.2/0.8 for RoBERTa ensemble loss A and RoBERTa ensemble loss B, respectively. The combination of these two loss functions was aimed at embedding the benefits derived from each loss function into the model. In contrast, the BERT model (baseline) used a single cross-entropy loss function. The proposed model outperformed the baseline model, suggesting that the ensemble loss function used in the proposed model was more effective in fine-tuning the model for the task of detecting sexism in text. The same model is applied to subtask B and subtask C (fine-grained classification).

The experiments were conducted on the training and dev sets provided by the task organizers for each respective subtask. The training set for subtask A consists of 14000 samples, and subtasks B and C consist of 3398 and 3398 samples, respectively. The test set for subtask A consists of 4000 samples, and subtasks B and C consist of 970 samples, respectively. Additionally, a dev set for subtask A consists of 2800 samples, and subtasks B and C consist of 970 samples, respectively. For all the subtasks I used the train set for training, the dev set for validation, and the test set for testing. The models were fine-tuned using the Adam optimizer for 25 epochs with a batch size of 32 and a learning rate of 5e-6, with a default epsilon value of 1e-6. I used RoBERTa large in our experiments. The evaluation of the models was performed using the macro F1 score for all the subtasks, including A, B, and C.

Table 1 shows the results of three tasks for test data. I achieve a 77.16 macro F1 score for subtask A, 46.11 score for subtask B and 30.2 score for subtask C. Table 2 shows the results of three tasks for validation data. From the tables, we can observe that RoBERTa model performs better compared to BERT model.

Table 2: Performance Analysis on Validation data

Task Name	Model	Macro F1 score	Competition Rank
subtask A	BERT	75.40	103
subtask A	RoBERTa	78.26	
subtask B	BERT	44.36	63
subtask B	RoBERTa	47.82	
subtask C	BERT	26.46	46
subtask C	RoBERTa	28.66	

5 Conclusion

In this paper, I experiment with different transformer models to make predictions over the EDOS shared task data. Explaining sexism is an important research topic to explore. I got a highest F1 score on the dev and test sets using a RoBERTa model which outperformed a BERT model. I participated in all the 3 subtasks. The performance of the system is evaluated using the macro F1 score.

References

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Fabio Del Vigna12, Andrea Cimino23, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the first Italian conference on cybersecurity (ITASEC17)*, pages 86–95.

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*.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30.

Simona Frenda, Bilal Ghanem, Manuel Montes-y Gómez, and Paolo Rosso. 2019. Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *Journal of Intelligent & Fuzzy Systems*, 36(5):4743–4752.

Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N Asokan. 2018. All you need is" love" evading hate speech detection. In *Proceedings of the 11th ACM workshop on artificial intelligence and security*, pages 2–12.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350.

Akshita Jha and Radhika Mamidi. 2017. 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*, pages 7–16.

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 (SemEval-2023)*. Association for Computational Linguistics.

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*.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

Sreekanth Madisetty and Maunendra Sankar Desarkar. 2018. Aggression detection in social media using deep neural networks. In *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, pages 120–127.

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.

Punyajoy Saha, Binny Mathew, Pawan Goyal, and Animesh Mukherjee. 2018. Hateminers: Detecting hate speech against women. *arXiv preprint arXiv:1812.06700*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Hajime Watanabe, Mondher Bouazizi, and Tomoaki Ohtsuki. 2018. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE access*, 6:13825–13835.

Philine Zeinert, Nanna Inie, and Leon Derczynski. 2021. Annotating online misogyny. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3181–3197.