Explainable Detection of Online Sexism*

Jannatul Ferdous
Computer Science (CS)
BRAC University
Dhaka, Bangladesh
jannatul.ferdos@g.bracu.ac.bd

Dipannita Baidya
Computer Science (CS)
BRAC University
Dhaka, Bangladesh
dipannita.baidya@g.bracu.ac.bd

Shantanu Das
Computer Science (CS)
BRAC University
Dhaka, Bangladesh
shantanu.das@g.bracu.ac.bd

Navid Hasan Rafi
Computer Science and Engineering (CSE)
BRAC University
Dhaka, Bangladesh
navid.hasan.rafi@g.bracu.ac.bd

***Abstract*—Online sexism is a pervasive and negative phenomena. The widespread detection of sexism can be helped by automated technologies. Binary detection, on the other hand, fails to offer concise justifications for why something is sexist and ignores the diversity of sexist material. SemEval Task 10 on the Explainable Detection of Online Sexism (EDOS) was created to address this problem. In this study, we focus on the interpretability, trustworthiness, and comprehension of the classification-related judgements made by models. Two smaller assignments make up the main task. Identifying Binary Sexism Detection is the first task. The second job outlines the Sexism Category.**

***Index Terms*—sexist, not sexist, Naive Bayes, Logistic Regression and Neural Network Models (Keras)**

I. Introduction

In this day and age, online sexism is a rising issue. Considering the volume of texts available online. Finding online Sexism is a significant NLP task with several societal repercussions. Sexism is characterized as any form of maltreatment or unfavorable opinion aimed against women because of their gender or because of their gender in combination with one or more additional identification characteristics (such as black women or trans women). Automated sexism detection technologies are widely accessible, however they frequently fail to address the problem of why the text is sexist. These automated technologies must be capable of both identifying sexist content and stating why it is sexist. We made an effort to concentrate on the interpretability, trustworthiness, and comprehension of model decisions. One of the key initiatives towards ensuring the security of the web will be the creation of a computerised framework that is capable of reading, evaluating, and grouping language that is sexist. In an effort to address this issue, Explainable Detection of Online Sexism at SemEval 2023 (Kirk et al., 2023) investigates several classification methods that can classify the dataset compiled from writings from Reddit and Gab. Two smaller tasks make up the overall assignment. The first assignment is titled Binary Sexism Detection, and its major goal is to determine whether or not a text contains sexism. The Category of Sexism is discussed in the second assignment. So, what category does the text fall under if it is sexist? The categories include threats, insults, enmity (a profound sense of animosity or hostility), and biassed discourse (an unjustified sentiment of hatred for a person or group due to ethnicity, sexual orientation, faith, etc.). Thus, it becomes a four-class assignment.

II. Data Collected

The entirety of Task 10: Explainable Detection of Online Sexism (EDOS) data was gathered from Reddit and Gab. The 14,000 data points offered for testing are unbalanced for texts that are sexist and non-sexist are listed in the "rewire-id, text, label-sexist, label-category, label-vector" column. For Task A, however, we are just interested in the columns "text" and "label-sexist." This dataset's goal is to create models that can identify between sexist and non-sexist material. Rewire-id is solely used to identify each text in a particular way. After analyzing the dataset we get 10,602 non-sexist posts and 3,398 sexist posts. Gab- 34M publicly accessible Gab posts were gathered between August 2016 and October 2018.Three numerous academic research have utilized this data (e.g., Kennedy et al., 2018; Cinelli et al., 2021). To establish the pool, we select 1M entries at random. Reddit - Here, we have compiled a record of around 81 subreddits which are expected to have sexist content located from the previous works by the authors (Guest et al., 2021; Farrell et al., Zuckerberg, 2018; Jones et al., 2020; Qian et al., 2019; Ging, 2017; Ribeiro et al., 2021) and online platforms. Then, we collected all such comments from August 2016 to October 2018 in these subreddits by using the Reddit API.5. Also, we manually viewed each of the subreddits and designated it to one of the four categories given below (Incels, Men Going Their Own Way, Men's Rights Activists, Pick Up Artists) which are based on the prior work done by Lily (2016). Now, to make sure

that our dataset is not excessively prejudiced regarding niche semantic impressions and subjects, we have kept under control our samples to those 24 subreddits with no less than 100k comments following in a dataset of about 42M comments.

III. Methodology

In this research project, our main goal is to build automated tools for the Explainable Detection of Online Sexism (EDOS), which will help us battle the widespread and harmful problem of online sexism. Now to achieve our goal, the first thing we will do is to compile a large dataset. The dataset was taken from well known Reddit and Gab. We balanced the dataset carefully to achieve a fair representation of both sexist and non-sexist content. We developed a novel two-level classification that combines Binary Sexism Detection and Category of Sexism. We have designed our classification system very carefully to identify and separate sexist content accurately. To do this, we have used valuable information from existing research and real examples in our dataset.We use various smart computer methods, like Logistic Regression, Naive Bayes, and Neural Networks (specifically LSTM-based models in the Keras framework), to classify the content accurately. These models have the capacity to recognize complex correlations and patterns in the data, enabling them to offer illuminating explanations for their predictions. We continue to train and test the models using the selected dataset after having the classification and models in place. Performance metrics are calculated to evaluate how well each model categorizes sexist content, including accuracy, precision, recall, and F1-score. Furthermore, by using cutting-edge explainability approaches like LIME, SHAP, and the investigation of attention processes, the models' interpretability is improved. By revealing which passages in the text have the most influence on the classifications made by the models, these strategies let us get important insights into how the models make decisions. We carefully analyze biases in our models' predictions to find and correct any potential biases, ensuring the validity and dependability of our models. In summary, our research aims to significantly contribute to reducing online misogyny and promoting a secure digital environment. We seek to provide a full understanding of online sexism and its numerous manifestations by creating explainable models, a thorough classification, and a diversified and annotated dataset.

IV. Literature Review

By examining diverse strategies, techniques, and challenges encountered in this area, this review of the literature aims to look at the studies that have previously been conducted on detecting online sexism. We reviewed a wide range of sexism and misogyny-related academic literature. For example, A number of recent research, including Farrell et al. (2019), Parikh et al. (2021), Guest et al. (2021),Zeinert et al. (2021), Jha and Mamidi (2017), Samory et al. (2021), and Rodrguez-Sánchez et al. (2021), suggest taxonomies of sexist or misogynist material. Although every taxonomy is unique, there are some common differences among them, including (i) construction (either they are founded on hypothesis or empirical evidence); (ii) scope (the way most fundamental classification for sexism is characterized and which groups are included); and (iii) structure (whether it's groups have been progressively ordered).

Taxonomy - From this examination of the literature, we developed a preliminary taxonomy. Using empirical items from our dataset (see 3), the taxonomy was further improved using a grounded theory technique (Glaser and Strauss, 2017) to merge or modify the schema. It has two related tasks:

Task 1: Binary Sexism- Our taxonomy divides content into two categories, sexist and non-sexist, at the first level. Sexist concept can be described as any maltreatment, whether implicit or overt, directed against women because of their gender or because of how their gender is linked with one or more other identity factors (for instance, Black women or Muslim women). Instead of focusing on misogyny, our taxonomy emphasizes sexism. According to Ussher (2016), misogyny is defined as "expressions of hate towards women," but sexism also encompasses less overtly articulated subtle forms of abuse and discrimination that can nevertheless do great harm to women.

Task 2: Sexism Category- At the second level of our taxonomy, sexist content is divided into four theoretically and statistically distinct categories. We purposely chose not to split categories by the putative effect on the recipient or the supposed motivation of the speaker because the harm caused by content is distinctive and speaker intent is difficult to discern, particularly in the absence of further background.

(1) Threats, malicious intent, and incitement: Language that either expresses an individual's intention to harm or incites serious harm and violence against women, or that encourages others to do the same. Threats of injury to the body or to the sex or privacy are included. (2) Derogatory language: Phrases that specifically disparage, dehumanise, belittle, or insult women. It contains disparaging remarks and gender stereotypes, objectification of women's body, vehemently negative emotional comments, and dehumanising comparisons. It covers disparaging remarks made about particular women as well as women in general. (3) Animosity is the use of language to subtly or overtly express sexism, stereotypes, or descriptive claims. It contains benign sexism, or sexism that is presented as a complement. (4) Prejudiced Discussion: Wording that rationalizes sexism and refutes the presence of prejudice. It comprises the denial of gender disparity and its explanation, the justification of women's abuse, and the idea of male victimization.

V. Conclusion

We made three key contributions in this research. In order to classify sexism in a way that is easier to understand, we first established a new taxonomy with two hierarchical levels: binary sexism detection and category of sexism. A survey of earlier taxonomies and social science literature serves as the foundation for this taxonomy, which is then experimentally verified using the data collection of values from the two

social networking websites. Thus, it offers a sociotechnical assessment of the complex and varied online sexism scene. Second, we produced a finest dataset that was sampled using a variety of ways to improve the variation of the content which are explained by the women specialists who self-identified as such in order to guarantee uniform labeling. To lessen the impact that a labeling budget (both in terms of economic and conceptual cost to analysts) has on the effectiveness of trained systems, this labeled dataset is combined with a bigger unlabelled dataset. We provided the basic models and systems that were summarized for our SemEval task before concluding. This research shows how the models work well when combined with multi-task learning and ongoing pre-training. However, there are still many more chances for the sexist content detection to be improved. We are hoping that by using our research and resources, future solutions that make everyone's online experience safer may be designed.

REFERENCES

[1] Papers with Code - Attention at SemEval-2023 Task 10: Explainable Detection of Online Sexism (EDOS). (n.d.). Paperswithcode.com. Retrieved June 24, 2023.

[2] Aliyu, Saminu Mohammad, Abdulmumin, I., Muhammad, S. H., Ahmad, I. S., Salahudeen, Saheed Abdullahi, Yusuf, A., and Lawan, Falalu Ibrahim. (2023). HausaNLP at SemEval-2023 Task 10: Transfer Learning, Synthetic Data and Side-Information for Multi-Level Sexism Classification.

[3] Chernyshev, K., Garanina, E., Bayram, D., Zheng, Q., and Edman, L. (2023, June 8). LCT-1 at SemEval-2023 Task 10: Pre-training and Multi-task Learning for Sexism Detection and Classification. ArXiv.org.

[4] Kirk, H. R., Yin, W., Vidgen, B., and Röttger, P. (2023). SemEval-2023 Task 10: Explainable Detection of Online Sexism.