Ethics Therapy Project: Detection and Text Style Transfer of Sexist Language in Social Media Comments

Ahana Chattopadhyay, Araya Kiros Hailemariam,
Aubin Medjaed, Elisabeth Olamisan, Hanady Yasmine,
Julien Schmitz, Kira Grudinina, Tian Fang, Zhengjian Li
Institut des sciences du Digital, Management & Cognition, Université de Lorraine
{ahana.chattopadhyay1,araya-kiros.hailemariam5,aubin.medjaed8,
elisabeth.olamisan3,hanady.yasmine4,julien.schmitz7,
kira.grudinina7,tian.fang1,zhengjian.li3}@etu.univ-lorraine.fr

Project Website Github Repository Hugging Face

Abstract

Sexist language is pervasive across contemporary social media platforms. This project focused on identifying and altering sexist language in user-generated comments on these platforms, thereby contributing to the advancement of tolerant and respectful discourse. The research, distinctively based on data in the English language, involved an in-depth analysis of a corpus used for training deep learning models. Specifically, the dataset employed is based on the EDOS, and a self-built corpus named ComMit2k, containing 2,000 mitigated comments, was developed for the mitigation task. Furthermore, the report details the multifaceted approaches that were implemented to address the issue of sexist language. This project was executed using Hugging Face, with both the code and datasets being made available on GitHub. Additionally, careful consideration of the ethical aspects of this project has been undertaken.

Warning: This project involves research on offensive and sexist comments, as well as AI-generated content. Liability for any discomfort or offense caused by the content is not assumed. Please proceed with caution.

1 Introduction

According to the Pew Research Center's the State of Online Harassment report (Vogels, 2021), 41% of the respondents in the survey have personally encountered online harassment, while a majority of the participants (66%) have witnessed such behavior being directed at others. Furthermore, 8% of those surveyed reported being singled out for harassment based on their gender.

This project focuses on addressing sexist speech in social media. Sexist language, or sexist discourse, is a linguistic pattern that designates the female gender as its primary target group, executing a narrative that debases and belittles them, with an underlying objective of inflicting indelible harm upon women (Lillian, 2007). In the latest advancements in linguistic research, the notion referred to as "gender-biased language " has undergone a significant expansion that the demeaning, ignorance and stereotype towards both men and women is included (MacArthur et al., 2020). From a broader perspective, numerous academic investigations position sexist language as a subcategory of hate speech (Lillian, 2007), especially in research on hate speech detection (Waseem and Hovy, 2016; Kennedy et al., 2020; Sachdeva et al., 2022). It is noteworthy that isolating sexist language as an independent research task is relatively infrequent.

This research aims to mitigate the repercussions of aggressive sexist speech. The approach involves converting offensive remarks into forms that encompass constructive criticism or neutral expressions, all the while maintaining fidelity to the original semantics and contextual nuances. Consider the following example, classified as "casual use of gendered slurs, profanities, and insults":

> Women should be in the kitchen !!

By executing tasks A, B, and C with the assistance of several models (cf. Table 1, Table 2) the transformed output is presented as follows:

> Women have traditionally played a significant role in the kitchen.

In pursuit of this objective, the endeavor aims to contribute scholarly insights to discourse moderation and digital ethics. Within the purview of this study, the principal objectives encompass three key tasks:

Task A: Binary classification The first step is oriented towards the identification of sexist dis-

course. The models are trained using binary classification on a dataset named "EDOS dataset" that is already labeled.

Task B: Multi-class classification The second step is to train models on multi-class classification of each sexist comment tagged with one of the four labels based on the level of sexism it possesses.

Task C: Text style transfer The third step incorporates feeding another dataset, named "CommMit2k", with mitigated version of the previously classified sexist comments, into a text style transfer model.

As shown in Figure 1, an interface on Hugging Face has been developed to address the aforementioned tasks, utilizing two RoBERTa classifiers and one BART model, all of which exhibit superior performance compared to other researched models.

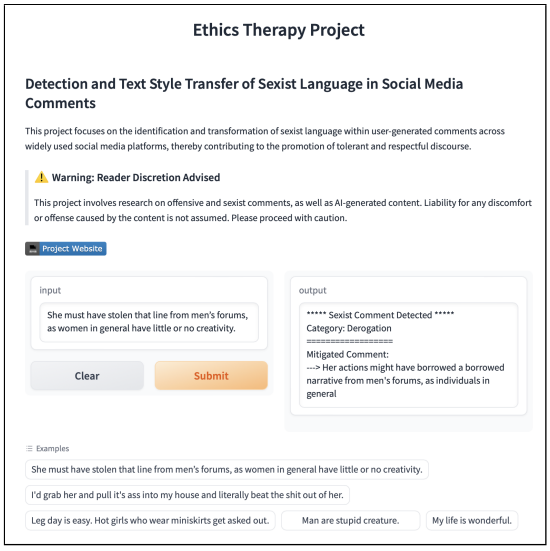

Figure 1: Interface of the models

This project has navigated the complexities of transforming sexist language into constructive or neutral expressions while preserving the underlying meaning and contextual nuances, contributing to innovative approaches in mitigating online harassment and fostering a culture of respectful discourse in digital spaces.

2 Related Works

In this research, tasks encompassing sexist language detection and text style transfer necessitate an exploration of recent advancements in natural language processing (NLP).

This section primarily addresses the classification task, specifically the detection of sexist comments. Altering sexist comments to render them less offensive presents a significant challenge, and relevant studies on this subject were scarce. However, this task aligns with text style transfer,thus studies related to this broader topic are referenced.

For the identification of sexism in tasks A and B, attention was directed towards research utilizing traditional machine learning approaches. An SVM, employing TF (term frequency) and TF-IDF (term frequency-inverse document frequency) features of n-grams, facilitates pattern recognition and decision boundary establishment based on term frequency and significance (Kondragunta et al., 2023).

Model	F_1 **score**
SBERT (Das et al., 2023)	0.6726
GPT-3-FT-Curie (Knospe, 2023)	0.8540
XLNet (Zhang and Wang, 2023)	0.8224
T5-Large (Tavan and Najafi, 2023)	0.7566
SVM + TF-IDF (Kondragunta et al., 2023)	0.7460
CNN-BiLSTM (Vetagiri et al., 2023)	0.7300
Bi-LSTM (Knospe, 2023)	0.6700
TextCNN (Zhang and Wang, 2023)	0.4916

Table 1: Results from different models in Task A

Model	F_1 **score**
GPT-3-FT-Curie (Knospe, 2023)	0.6470
DeBERTa (Kondragunta et al., 2023)	0.6120
RoBERTa (Al-Azzawi et al., 2023)	0.5933
SVM + TF (Kondragunta et al., 2023)	0.4757

Table 2: Results from different models in Task B

The efficacy of TextCNN (Zhang and Wang, 2023) is attributed to its ability to capture complex text representations using convolutional kernels of various window sizes, enabling the acquisition of comprehensive features at different n-grams levels.

In the paper (Das et al., 2023), embeddings were created for an entire sentence or paragraph, rather than for each word. These embeddings are intended to reflect the semantic meaning, context, and relationships between sentences. Sentences with similar meanings generate similar embeddings. To achieve this objective, a pre-trained SBERT model was utilized, incorporating a modified Masked Language Modeling (MLM) approach. In MLM, a specific proportion of words within the input text is randomly obscured or masked. Subsequently,

the SBERT model undergoes training to reconstruct or predict the concealed or corrupted segments, leveraging contextual information derived from the unmasked sentences that remain. In contrast to the SBERT model, which utilizes masked language modelling (MLM) involving token masking during training, XLNet adopts an autoregressive mechanism (Zhang and Wang, 2023). XLNet generates various contexts by permuting the input sequence, thereby creating multiple instances for predicting individual tokens. By sampling diverse permutations of the token sequence, XLNet treats each permutation as a distinct context for predicting subsequent tokens. This approach is presumed to offer a more comprehensive understanding of word and sentence interrelationships within a sequence. DeBERTa showcased promising performance in contextual analysis and comprehension (Kondragunta et al., 2023). DeBERTa's core principle revolves around the independent regulation of word positions and their content. This disentanglement strategy facilitates a nuanced exploration of the intricate relationships between words and their positional significance within the text. Such disentanglement augments the model's capacity to grasp contextual intricacies. Moreover, (Al-Azzawi et al., 2023) employed an alternate model within the RoBERTa lineage, integrating data augmentation methodologies. The next paper investigates the conjecture that incorporating dependency information via Graph Convolutional Networks (GCNs) could yield a more profound understanding of stylistic intricacies within sexist content (Tavan and Najafi, 2023). Their investigation employs T5-large as a benchmark to assess the effectiveness of their proposed GCN-centric approach. The benchmark's performance also demonstrated notable results.

BiLSTM with Gloves embeddings (Knospe, 2023). In this case, embeddings are also used as in (Das et al., 2023), but at the word level - word embeddings. GloVe (Global Vectors for Word Representation) provides pre-trained word embeddings based on co-occurrence statistics in a corpus. GloVe embeddings often capture semantic and syntactic relationships between words, providing a strong starting point for the model to understand language nuances. Another work also uses pre-trained embeddings (Vetagiri et al., 2023). Pre-trained word embeddings are fed into the CNN-BiLSTM system. The first is a CNN layer, which is intended to identify specific correlations and patterns in the input text. A BiLSTM layer, the second, is created to identify long-term dependencies in the input text. In the research, the choice was made to utilize BiLSTM without relying on pre-trained embeddings, allowing the model to acquire word representations through independent learning. Throughout the training process, the BiLSTM network generates embeddings by discerning patterns and relationships within the input sequences. This approach enables the model to adapt to the specific nuances inherent in the task.

According to the paper (Knospe, 2023), the GPT-3 model underwent fine-tuning by being trained on examples that taught it how to determine if a text completion contained sexist language or context. As a result of this training, the model became more skilled at identifying sexism in text and more sensitive towards it. Given the promising outcomes of the experiment with GPT-3 for detecting sexism without hyperparameter optimization (Knospe, 2023), the decision has been made to utilize this model for the development of a dataset aimed at mitigating sexist comments.. In addition, the associated paper underscores the significance of analytical reasoning in understanding the value of interpreting the model's inner workings, identifying biases or dependencies on specific words, and potentially improving the model's robustness or fairness. This is particularly critical for gaining a more nuanced understanding of why specific instances are attributed to particular classes, as even a well-defined taxonomy may not necessarily assume a leading role in classification. The authors have undertaken an in-depth analysis of the impact of replacing salient words with semantically related ones on the model's predictions, thereby elucidating the role of these words (e.g., women, woman, female, girls, females, ladies, and girl) in the classification of the given post. This has prompted us to consider a comparable strategy for softening the tone of a comment by substituting salient words. Such a replacement may potentially alter the overall tenor of the comment in some cases.

Moreover, hybrid approaches that often amalgamate multiple machine learning models are frequently employed to enhance performance in detecting hate speech (e.g. Pitsilis et al., 2018; Badjatiya et al., 2017).

Text style transfer involves the intricate task of rephrasing sentences to align with different stylis-

tic preferences while maintaining their original meaning. This process encompasses altering linguistic properties like formality, politeness, gender, toxicity, and offensiveness. While many models rely on encoder-decoder architecture, alternative approaches, such as employing Generative Adversarial Networks (GANs) or using prompts with models including GPT, offer viable solutions, particularly when datasets are limited in text style transformation research (Jin et al., 2022; Toshevska and Gievska, 2022; Li and Liang, 2021; Qin and Eisner, 2021, e.g.).

3 Corpus

This research undertakes the analysis of sexist language present in English-language comments on social media platforms, aiming to identify and transform such language. Two datasets were utilized in order to achieve these objectives. The first, called EDOS (Explainable Detection of Online Sexism) dataset[1], encompasses 20,000 comments from sources such as Gab and Reddit, with a binary classification of 4,854 sexist and 15,146 non-sexist remarks (Kirk et al., 2023). This dataset is integral to the execution of tasks A and B within the project. The second dataset, named ComMit2k,was specifically developed for task C, offering a tailored corpus for the study.

3.1 EDOS

For tasks A and B, existing annotations from the EDOS dataset are utilized. In order to mitigate implicit biases, a team of annotators, all of whom self-identify as women, were recruited. To address the issue of class imbalance in random sampling from online platforms, a mixed and diversified sampling approach was employed by the research team. This approach included community-based sampling and six different techniques designed to enhance the proportion of sexist content in the dataset. It also ensured coverage of various types of sexism and the attainment of lexical and topical diversity.

To illustrate the application of these tasks, consider the following examples:

Task A: Examples

> **Sexist:** "I did. You have to have the bravery to escalate, touch her boobs etc, and work on escalating it further."
>
> **Non-sexist:** "He should be arrested for Assault and based on his claim that he's raped women, he should be investigated. His bragging could be a confession."

For Task B, the focus shifts to categorizing comments into specific types of sexist behavior.

Task B: Categories

The categories include "Threats, plans to harm and incitement," "Derogation," "Animosity," and "Prejudiced discussions." Table 5 provides an overview of the taxonomy for classifying such comments, along with descriptions and examples for each category.

3.2 ComMit2k

For the text style transfer task, a dataset named ComMit2k is created (Comments Mitigation 2000). It cointains less offensive alternatives for three classified categories: derogation, animosity, and prejudiced discussions. The 2000 utterances were randomly selected from the EDOS dataset, focusing on these specific categories.

However, it is important to note the boundaries of this mitigation approach. The sexist comments categorized as "threats of harm" and "incitement and encouragement of harm" are excluded from the ComMit2k dataset. The reason for this exclusion is that such comments involve direct harm to individuals. Due to their nature, these comments cannot be rewritten in a way that removes the harmful intent without significantly altering their original meaning. After outlining the scope and limitations of the ComMit2k dataset, it is illustrative to present specific examples of how offensive comments have been transformed into more constructive and respectful discourse. This process exemplifies the core goal of the text style transfer task. Consider the following example:

> "How about I talk shit to an idiot like you? Are you a nazi like that bitch?"

Through the mitigation approach, this comment is transformed into:

> "I have some concerns and disagreements with your perspective and would like to discuss them constructively."

Following the initial phase, the methodology involved a manual review to ensure the accuracy and

[1] Available at https://github.com/rewire-online/edos

relevance of the content. This included the careful annotation of newly created mitigated comments by three annotators. Once agreement among these annotators was reached, the agreed-upon data was utilized for the training of the model. The process of manual annotation was conducted in adherence to established guidelines (cf. Appendix C).

Given the absence of gold data for this task, the validity of the manual annotation was of significant interest, so it was vital to ensure the reliability of it. In order to proceed with reliability measurement , coefficients of agreement were used such as observed agreement A_0, the π coefficient and Cohen's kappa κ.

The annotation task was undertaken by a team of three annotators: A, B and C. The outcomes of their annotation process are presented in Table 3.

Metric	Pair	Value
A_0	A_annot, B_annot	**0.82**
	A_annot, C_annot	**0.80**
	B_annot, C_annot	**0.77**
π Coefficient	A_annot, B_annot	**0.77**
	A_annot, B_annot	**0.23**
	A_annot, C_annot	**0.69**
	A_annot, C_annot	**0.37**
	B_annot, C_annot	**0.74**
	B_annot, C_annot	**0.13**
	Overall	**0.24**
Cohen's Kappa (κ)	A_annot, B_annot	**0.77**
	A_annot, B_annot	**0.24**
	A_annot, C_annot	**0.69**
	A_annot, C_annot	**0.37**
	B_annot, C_annot	**0.73**
	B_annot, C_annot	**0.15**
	Overall	**0.26**

Table 3: Inter-annotator agreement metrics

Observed Agreement A_0 showed relatively high consistency among annotators, with A_annot and B_annot having the highest agreement at 0.82.

π Coefficient values were generally lower, indicating some agreements might have been coincidental. The overall π value is 0.24, suggesting room for improvement in consistency beyond chance agreement.

Cohen's Kappa κ, a more stringent measure. The overallκ value is 0.26, indicating moderate agreement when accounting for chance.

To conclude, the data indicated a decent level of agreement among annotators, especially between A_annot and B_annot. However, the lower π and κ values suggest that some of the agreements might have been due to chance. This implies a need for further standardization or training in the annotation process to enhance consistency and reduce coincidental agreements.

The methodology thus introduces a comprehensive framework for mitigating offensive language in online comments, combining diverse perspectives with rigorous standards and reliability measures to foster a more respectful and inclusive online dialogue environment.

4 Ethical concerns

The chosen dataset, EDOS, contains sexist data from two major social media platforms: Reddit and Gab. For its construction, tools were utilized by the original team to automatically collect data from other websites, a method known as "scrapping," followed by labeling. While this practice is not strictly illegal in France, many websites have policies regarding data usage. Platforms like Reddit and Gab permit data collection for research purposes, provided steps are taken to protect user anonymity. The dataset adheres to the rights of users and the usage policies of both websites by excluding any means of identifying the users behind the comments. Contained within the dataset are graphic descriptions of sexist comments, some with elements of extremism (e.g., "threats of harm"), presenting several ethical risks that necessitate responsible addressing. The primary objective of this research is to facilitate a text style transfer while preserving the core semantic content of the original comments. However, a subset of comments in the dataset is characterized by elevated toxicity levels. These comments, marked by explicit derogatory content, have the potential to inflict harm and perpetuate harmfully prejudiced narratives, making their inclusion ethically problematic. By excluding high-toxicity comments from the translation phase, the aim was to uphold ethical standards, minimize potential harm, and prioritize the responsible application of research outcomes. This method demonstrated a commitment to conducting research that is not only technically proficient but also ethically sound, contributing to safer and more responsible digital discourse.

Another concern is the potential threat to the "expression of free speech". Despite clear definitions of what sexism is and the nuances of each label, such actions might be perceived as infringing on fundamental freedoms. The text style transfer model was applied exclusively to the remaining sample as a moderation method, respecting the semantic features of the original data. Users will retain the ultimate choice to implement the system based on personal opinion and judgment. In French law, sexism and cyberbullying are criminal offenses with potential prison sentences, as outlined in articles 222-33 of the French Penal Code and L 1153-1 of the Labour Code. The model is intended not to conceal the perpetrator's crimes but to protect online populations that may fall victim to such offenses. From a broader perspective, there exists a high possibility that the state-of-the-art model could be utilized for specific political, social, or commercial gains. If the system of suppression or moderation is handled maliciously, it could negatively impact society. Responsible handling of the model is strongly advocated, with the sole purpose of benefiting humanity.

5 Methodology

Task Description In this project, a systematic approach was outlined for the detection and mitigation of sexist comments within textual data. This multi-step process, as depicted in Figure 2, encompasses several tasks, namely :

1. Data Preprocessing
2. Task A : Binary Sexism Detection
3. Task B : Multi-Class Severity Classification
4. Task C : Text Style Transfer
5. Comprehensive Model Evaluation

Step 1: Data Preprocessing In the data preprocessing phase, several key steps were undertaken to refine the text data for subsequent natural language processing tasks. The initial phase involved the removal of emojis and URLs from the text corpus. Emojis, while expressive and conveying sentiment in text, were not the primary focus of the project's analysis. Hence, the decision to remove them was made to streamline the text corpus and focus on the textual content relevant to the project's objectives. URLs were also removed due to their potential lack

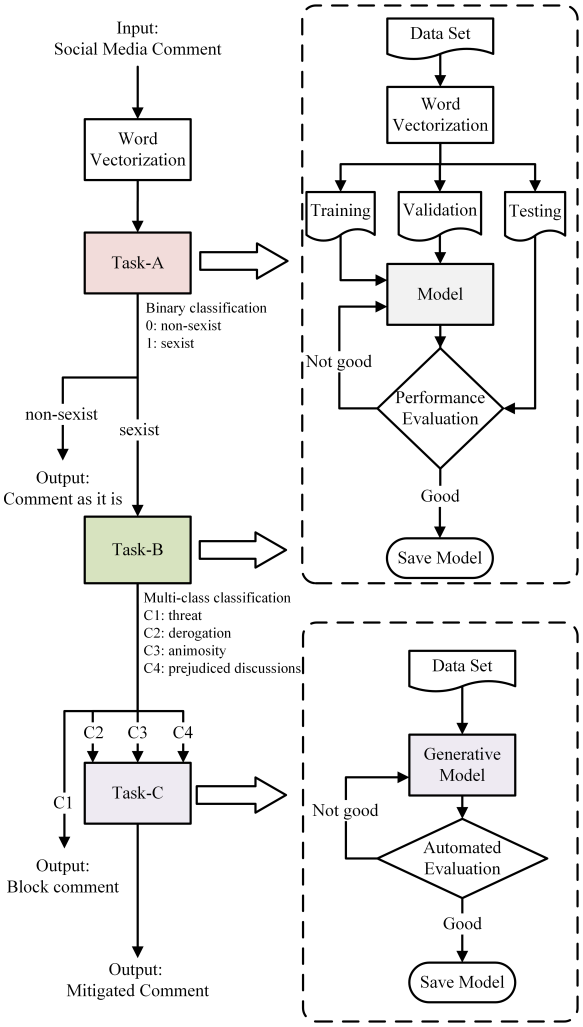

Figure 2: An overview of the current models

of relevance or contribution to the linguistic context of the dataset.

Contrary to traditional preprocessing practices, the exclusion of stop words and the conversion of text to lowercase were intentionally omitted in this specific workflow. This decision stemmed from the utilization of a BERT tokenization approach in the implementation part. BERT, as a state-of-the-art transformer-based model, operates on subword tokenization and contextual word embeddings. Hence, the removal of stop words, which are commonly occurring words but might carry contextual significance, was bypassed. Additionally, the casing of the text was preserved to allow BERT's contextual understanding of text nuances, considering its ability to comprehend letter case variations and maintain context in its embeddings.

This tailored preprocessing pipeline prioritized the preservation of contextual information and linguistic nuances within the text data, aligning with

the requirements and functioning principles of the BERT tokenization methodology.

Step 2: Task A: Binary Sexism Detection In this phase, comments are classified into sexist or non-sexist text. The approach involves training models on meticulously labeled datasets, which contain examples of both sexist and non-sexist comments. These models leverage advanced techniques to analyze various textual features, including vocabulary choices, specific phrases, and contextual cues. The models that were used are :

- SVM TF-IDF
- RoBERTa
- DeBERTa
- BiLSTM

To guide this phase, the latest advancements in the field of sexism detection are incorporated, building on insights gained from recent research. Notably, the dataset exhibits an inherent class imbalance, with 24% of comments being sexist and 76% non-sexist. To assess model performance, the Macro F_1 score is used as the primary evaluation metric. The aim was to develop a robust and precise sexism detection model that can effectively handle the challenges posed by imbalanced datasets.

Step 3: Task B: Multi-Class Severity Classification Similar to Task A, this task also employs the same four models that have been previously delineated. This task aims to design, implement, and enhance the best-performing models, as identified in the recent review of related works (cf. Section 2). These models served as the foundation for multi-class classification. The goal was to categorize each comment labeled as "sexist" into one of the following classes: "threats, plans to harm and incitement," "derogation," "prejudiced discussion," or "animosity." The core focus was on enhancing and adapting these models to the specific context of the project to maximize performance. For this task, the Macro F_1 is also used as an evaluation metric.

Step 4: Task C: Text Style Transfer For this task, two models, BART and Switch, were employed, utilizing the ComMit2k dataset. This dataset comprises 'less offensive' iterations of sexist comments that lack the inherent toxicity levels associated with 'threats of harm' and 'incitement and encouragement of harm,' since these specific elements are immutable and necessitate proactive blocking measures. The BLEU score was utilized to evaluate the quality of the generated mitigated text for task C.

Step 5: Implementation The objective in this step was to establish an interface or API integration, which facilitates the transmission of identified sexist comments to the BART and Switch models, thereby enabling content modification in real-time.

6 Results

Within this section, the performance evaluation of diverse models across multiple tasks, assessed through F_1 Macro and BLEU scores shown in Table 4, illuminates their efficacy and relative strengths in addressing distinct task complexities.

Model	F_1 Macro		BLEU
	Task A	Task B	Task C
SVM TF-IDF	0.83	0.50	-
RoBERTa	0.83	0.66	-
DeBERTa	0.81	0.60	-
BiLSTM	0.65	0.65	-
BART	-	-	0.17
Switch	-	-	0.09

Table 4: Evaluation Metrics for All Tasks

Across Task A, the comparison reveals that SVM TF-IDF, RoBERTa and DeBERTa exhibit relatively comparable F_1 Macro scores (0.83, 0.83, and 0.81, respectively), while their underlying architectures differ significantly. The SVM model relies on traditional TF-IDF representation coupled with linear classification, whereas RoBERTa and DeBERTa leverage transformer architectures, enabling attention-based mechanisms and contextual embeddings. While the SVM TF-IDF model achieves competitive F_1 scores by incorporating a custom lexicon during the training phrase, its static vocabulary and shallow linguistic comprehension may limit adaptability in dynamic real-world scenarios. In contrast, RoBERTa and DeBERTa, with their transformer architectures, offer deeper contextual understanding, potentially providing greater robustness across diverse and evolving language patterns.

In Task B, BiLSTM attained an F1-Macro score of 0.65, while RoBERTa achieved a slightly higher

score of 0.66. RoBERTa's transformer-based architecture likely gave it an edge in capturing nuanced language patterns.

Employing the BLEU (Bilingual Evaluation Understudy) metric for assessing the quality of generated mitigated text is crucial, given its established role in measuring linguistic similarity, widely recognized in evaluating machine translation systems. Its adoption for evaluating mitigated text ensures a standardized and reliable assessment of the generated content's quality and coherence compared to the original text. The granularity of Task C's assessment accentuates the impact of model architectures on performance. BART, leveraging a sequence-to-sequence architecture with denoising autoencoding, exhibits a superior BLEU score of 0.17 compared to the Switch model's score of 0.09, which employs a mixture-of-experts framework.

The comparison across various tasks highlights the competitive performance of models such as SVM TF-IDF, RoBERTa, and DeBERTa, showcasing the significance of architecture in language understanding. RoBERTa and DeBERTa's transformer architectures offer deeper contextual comprehension, while traditional models such as SVM TF-IDF might have limitations in adapting to dynamic language patterns. In specific tasks, RoBERTa's transformer architecture outperformed others, emphasizing the impact of model selection on task performance, as also seen in BLEU scores for mitigated text evaluation in Task C.

7 Implementation

The implementation phase of this research project involved utilizing two RoBERTa classifiers and a BART model to effectively mitigate the identified issues. These models were implemented using the Hugging Face library and integrated into the Gradio Interface framework. Specifically, these models are implemented within Hugging Face's Spaces, leveraging a server configuration equipped with 2 vCPUs and 16GB of RAM. This setup provided the computational power necessary for efficient model execution.

Integrating these advanced models into the Hugging Face environment significantly contributed to achieving the research objectives. As a result, the link to access the interface showcasing these implementations is available at the beginning of this report.

8 Future works

To enhance the scope and effectiveness of the project, several avenues for future work are proposed.

A thorough examination of the impact of emojis on sentiment and meaning is essential. Recognizing their nuanced role could significantly influence the accurate identification of sexist language in comments.

Conducting statistical analyses on both the raw data and the mitigated comments is recommended to provide valuable insights into the effectiveness of the approach and guide further improvements.

An investigation into the concept of positive sexism is also suggested, to ensure a comprehensive understanding of linguistic nuances. Positive sexism, which includes attitudes, behaviors, or language that may seem favorable toward a particular gender but can still reinforce stereotypes and contribute to inequality, requires careful consideration. For instance, a statement like 'Women are naturally better caregivers than men' exemplifies positive sexism.

Addressing the challenge of hallucinations in Large Language Models (LLMs) is identified as a critical area. Exploring fine-tuning strategies and potential ensemble approaches for LLMs is proposed. Additionally, the integration of multilingual datasets is recommended for creating a more globally applicable model.

The complexities of sarcasm and irony in language are suggested as areas of focus, considering their unique challenges in accurate interpretation. Fine-tuning models, such as LLaMa, and exploring their implementation could contribute to improved model performance.

Finally, expanding the project to address other forms of discrimination, such as racism, is aligned with the broader goal of fostering an inclusive and respectful online environment. These proposed future directions collectively aim to enhance the project's impact, ensuring its adaptability to diverse linguistic nuances and the evolving dynamics of online discourse.

References

Sana Sabah Al-Azzawi, György Kovács, Filip Nilsson, Tosin Adewumi, and Marcus Liwicki. 2023. NLP-LTU at SemEval-2023 Task 10: The Impact of Data Augmentation and Semi-Supervised Learning Tech-

niques on Text Classification Performance on an Imbalanced Dataset. ArXiv:2304.12847 [cs].

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 - WWW '17 Companion*, pages 759–760, Perth, Australia. ACM Press.

Amit Das, Nilanjana Raychawdhary, Tathagata Bhattacharya, Gerry Dozier, and Cheryl D. Seals. 2023. AU_nlp at SemEval-2023 Task 10: Explainable Detection of Online Sexism Using Fine-tuned RoBERTa. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 707–717.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep Learning for Text Style Transfer: A Survey. *Computational Linguistics*, 48(1):155–205.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing Hate Speech Classifiers with Post-hoc Explanation. ArXiv:2005.02439 [cs].

Hannah Rose Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 Task 10: Explainable Detection of Online Sexism. ArXiv:2303.04222 [cs].

Anders Knospe. 2023. Counterfactual Replacement Analysis for Interpretation of Blackbox Sexism Classification Models.

Murali Manohar Kondragunta, Amber Chen, Karlo Slot, Sanne Weering, and Tommaso Caselli. 2023. SKAM at SemEval-2023 Task 10: Linguistic Feature Integration and Continuous Pretraining for Online Sexism Detection and Classification. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1805–1817.

Loïc Lannelongue, Jason Grealey, and Michael Inouye. 2021. Green algorithms: Quantifying the carbon footprint of computation. *Advanced Science*, 8(12):2100707.

Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. 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 4582–4597, Online. Association for Computational Linguistics.

Donna L. Lillian. 2007. A thorn by any other name: sexist discourse as hate speech. *Discourse & Society*, 18(6):719–740.

Heather J. MacArthur, Jessica L. Cundiff, and Matthias R. Mehl. 2020. Estimating the Prevalence of Gender-Biased Language in Undergraduates' Everyday Speech. *Sex Roles*, 82(1):81–93. Rate: 5.

Georgios K. Pitsilis, Heri Ramampiaro, and Helge Langseth. 2018. Effective hate-speech detection in Twitter data using recurrent neural networks. *Applied Intelligence*, 48(12):4730–4742.

Guanghui Qin and Jason Eisner. 2021. Learning How to Ask: Querying LMs with Mixtures of Soft Prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212, Online. Association for Computational Linguistics.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. The Measuring Hate Speech Corpus: Leveraging Rasch Measurement Theory for Data Perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.

Ehsan Tavan and Maryam Najafi. 2023. MarSan at SemEval-2023 Task 10: Can Adversarial Training with help of a Graph Convolutional Network Detect Explainable Sexism? In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1011–1020.

Martina Toshevska and Sonja Gievska. 2022. A Review of Text Style Transfer Using Deep Learning. *IEEE Transactions on Artificial Intelligence*, 3(5):669–684. Conference Name: IEEE Transactions on Artificial Intelligence rate: 5.

Advaitha Vetagiri, Prottay Adhikary, Partha Pakray, and Amitava Das. 2023. CNLP-NITS at SemEval-2023 Task 10: Online sexism prediction, PREDHATE! In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 815–822.

Emily A. Vogels. 2021. The state of online harassment.

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, San Diego, California. Association for Computational Linguistics.

Yao Zhang and Liqing Wang. 2023. HHS at SemEval-2023 Task 10: A Comparative Analysis of Sexism Detection Based on the RoBERTa Model. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 963–968.

Category	Description	Example
1. Threats, plans to harm and incitement (443 comments)	Language expressing intent or encouraging actions causing significant harm, including physical, sexual, or privacy harm to women.	*"tell that cunt to come to my house and take it herself. I wouldn't need a gun to plant her bitch ass 6 feet down. (2024)"*
2. Derogation (2271 comments)	Language that explicitly disrespects, dehumanizes, or insults women, including negative descriptions, stereotypes, objectification, emotional negativity, and dehumanizing comparisons, directed at both specific women and women in general.	*"Tell the friend to dump the bitch. Let him watch how easily she gets another man to simp for her. (77)"*
3. Animosity (1665 comments)	Language conveying implicit or subtle sexism, stereotypes, or descriptive statements, which may include benevolent sexism presented as compliments.	*"How about I talk shit to an idiot like you? Are you a nazi like that bitch? (62)"*
4. Prejudiced discussions (475 comments)	Language that denies the presence of discrimination, justifies sexist treatment, and incorporates elements of gender inequality denial, excusing mistreatment of women, and the ideology of male victimhood.	*"It was rape because she regretted it afterwards, duh. There's no way she could have sex with a nerd. Sheesh. (13058)"*

Table 5: Taxonomy of Sexism Categories

A Carbon footprint

Using a rough estimate of the number of hours spent training different models, the carbon footprint and energy consumption were averaged to 10.61 kg of CO_2 emitted and 206.88 kWh used. This calculation was performed using "Green Algorithms" (Lannelongue et al., 2021). An average of 165 hours of development and training on personal computers was recorded, divided into 73 hours for the binary classification task, 56 hours for the multi-classification task, and 36 hours for the text style transfer task.

These results are equivalent to the average usage of a passenger car traveling up to 60.62 km, or 21 percent of a Paris-London flight.

B Corpus: EDOS

Table 5 presents a taxonomy of sexism categories, each defined with a description and illustrated with an example. It includes four categories, each category is supported by the number of comments it includes and an example comment to provide context.

C Corpus: ComMit2k

While engaged in the annotation of the dataset, a comprehensive guideline was established to assist in the process of manual annotation, ensuring consistency and accuracy across the entire dataset. This guideline provided clear instructions and criteria for annotators to follow, streamlining the annotation task and enhancing the quality of the dataset for further analysis and research purposes.

- Elimination of offensive language (e.g., explicit derogatory terms).
- Prohibition of individual objectification.
- Avoidance of generalizations; personal opinions, if necessary, are prefaced at the comment's start.
- In cases of descriptive attacks, these comments are not shown to the user. Instead, they are replaced with constructive messages emphasizing individual respect and avoidance of generalizations.