Lexicools at SemEval-2023 Task 10: Sexism Lexicon Construction via XAI

Pakawat Nakwijit[1] **Mahmoud Samir**[3] **Matthew Purver**[1,2]

[1] **Cognitive Science Research Group**
School of Electronic Engineering
and Computer Science
Queen Mary University of London, UK
{p.nakwijit, m.purver}@qmul.ac.uk

[2] **Department of Knowledge Technologies**
Jožef Stefan Institute
Ljubljana, Slovenia

[3] **Independent**
mahmoudbadr9199@gmail.com

Abstract

This paper presents our work on the SemEval-2023 Task 10 Explainable Detection of Online Sexism (EDOS) (Kirk et al., 2023) using lexicon-based models. Our approach consists of three main steps: lexicon construction based on Pointwise Mutual Information (PMI) and Shapley value, lexicon augmentation using an unannotated corpus and Large Language Models (LLMs), and, lastly, lexical incorporation for Bag-of-Word (BoW) logistic regression and fine-tuning LLMs. Our results demonstrate that our Shapley approach effectively produces a high-quality lexicon. We also show that simply counting the presence of certain words in our lexicons and comparing the count can outperform a BoW logistic regression in task B/C and fine-tuning BERT in task C. In the end, our classifier achieved F1-scores of 53.34% and 27.31% on the official blind test sets for tasks B and C, respectively. We, additionally, provide an in-depth analysis highlighting model limitations and bias. We also present our attempts to understand the model's behavior based on our constructed lexicons. Our code and the resulting lexicons are open-sourced in our GitHub repository https://github.com/SirBadr/SemEval2023-Task10.

1 Introduction

Social media has become an integral part of people's lives, providing a platform for communication, expression, and sharing of information. However, it has also become a breeding ground for toxic and dangerous behaviour, including the spread of sexist attitudes and discrimination.

European Institute for Gender Equality (2023) defines sexism as the manifestation of gender stereotypes and assumptions that rank one gender as superior to another, which can either be in the form of hostile attitudes or benevolent biases but actually harmful (Barreto and Doyle, 2022). Studies have shown that individuals who engage in online sexist behaviour are likely to exhibit sexist attitudes in their offline lives (Fox et al., 2015), resulting in negative effects on society and potentially causing significant psychological harm to those who are subjected to sexist comments (Swim et al., 2001; Stoop et al., 2019; Jiang et al., 2022).

Despite efforts to regulate content and enforce policies against sexist behaviour, controlling and preventing such behaviour on social media platforms is still a major challenge due to the high volume and velocity of user-generated content, which is impossible for human moderators to keep up. The development of automated methods for detecting sexism on social media has become increasingly important, providing a scalable and efficient solution to this persistent problem.

SemEval-2023 Task 10 Explainable Detection of Online Sexism (EDOS) represents a recent endeavour to develop automated tools to detect and explain sexism text empowering users and moderators to keep healthy and welcoming environments for everybody (Kirk et al., 2023). The shared task consists of three hierarchical subtasks.

- **Subtask A** is a binary sexism classification predicting a post to be either sexist or not.
- **Subtask B** is a multiclass classification for the broad category of a sexist post as a threat, derogation, animosity, or prejudiced discussion.
- **Subtask C** is an 11-class classification for a fine-grained sexist explanation of sexist content.

In this paper, we describe our contribution to the EDOS shared task. We conducted in-depth experiments on lexicon-based models. This type of method was selected because it could provide more transparency and interpretability in how it makes decisions. Unlike black-box models, which can produce accurate results, they are often difficult to understand how they arrived at those results.

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 23–43

This explainability is crucial for improving trustability, transparency, and reducing bias (Das and Rad, 2020), which is a key factor for successfully adopting a model in practice.

To tackle the task, our strategy consists of three parts. The first part is lexicon construction, in which we experimented with two methods; using PMI and Shapley. We then applied augmentation techniques to increase the lexicon vocabulary size using unannotated corpus provided by the task organizer and two generated corpus from BERTweet (Nguyen et al., 2020) and GPT-J (Wang and Komatsuzaki, 2021). Examples of our lexicon are presented in table 1. Lastly, we integrated the resulting lexicons into 3 classifiers; a lexicon-based classifier, a logistic regression, and a fine-tuned large language model. Besides the system description, we did an intensive analysis to observe model limitations, bias, and prediction behaviour based on the resulting lexicons.

We make the following observations based on our experiments:

- Using Shapley value is an effective approach to constructing a lexicon both quantitatively and qualitatively
- Augmentation is important, and using the fill-in-the-blank method outperformed other methods
- Lexicon-based classifier can outperform BoW logistic regression in all metrics.
- In a low-resource setting, a fine-tuned language model might behave similarly to a lexicon-based classifier which is limited to using only surface information.
- Negation is still a challenging issue for a fine-tuned LLM.
- Using lexicons, we observed that different groups of people (e.g. Women of colour, Asian, Muslim, and LGBTQ+) are associated differently in different sexist categories. In addition, our analysis indicates that the model may be affected by the imbalance-representation of each vulnerable group in the training data, such as LGBTQ+, which raises concerns about potential ignorance towards those people.
- Urban slang (mostly sex-related) predominately contributes to the model prediction, so handling the continuous increase of new vocabulary and the shift in meaning will become increasingly important.

2 Background

2.1 Lexicon Construction

Constructing a lexicon is a complex process that requires identifying the words and phrases in a language, their meaning, and their usage in context.

In early work, lexicons were constructed manually by researchers who would collect and analyze words and phrases to create a dictionary. Linguistic Inquiry and Word Count (LIWC) is one of the well-known and widely-used lexicons (Pennebaker et al., 2001). More than thousands of words are manually selected across the 80+ categories, such as functional words (pronouns, articles, conjunctions), psychological words (positive, negative, anxiety, anger, and sadness), and cognitive words (see, hear, and feel). Another example is the Arabic lexicon for sentiment analysis created by Abdulla et al. (2013). Their lexicon was constructed using 300 seed words translated from an existing English lexicon. They then manually added synonyms to increase the vocabulary size. Although these lexicons are detailed and accurate, the manual process of constructing them can be slow, labour-intensive, and expensive, making it difficult to create large-scale lexicons in a timely and efficient manner.

To address these limitations, more automatic approaches have been proposed. For example, Abdulla et al. (2014) selected lexicons based on term frequency from an annotated corpus and translated lexicons from an existing corpus in English. Wang and Xia (2017) constructed a sentiment lexicon by training a sentiment-aware word representation which was then used as the input of the logistic regression classifier to predict word sentiment polarity.

Another line of work in lexicon construction uses statistical approaches. One of the early works is presented by Turney (2002). Turney uses Pointwise Mutual Information (PMI) to measure the similarity between candidate and seed words. The category of the word is then assigned by the difference in the PMI score between positive and negative reference words (as it is used for sentiment analysis). A variant of PMI approaches was proposed in Mohammad et al. (2013), which yielded the best results in SemEval 2013 Task 2 Sentiment Analysis in Twitter.

Sexism Category	N	Lexicons
1. Threats	1,674	whores, skank, whore, pussies, bitches, bitch, pussy, feminism, **ww3**, **uniteamerica**
2. Derogation	5,439	**hillaryclintonisabitch**, **bhahahahahaahahahahaha**, **noodlewhore**, **noodlefoids**, **bitch's**, bitches, bitchy, pussies, **fembots**, bitch
3. Animosity	4,428	**muhahahahahahahahahahahahahahaha**, **pussyfooting**, **bitchtard**, **bitchboi**, bitches, **autothots**, **bitching**, bitchy, **npcunt**, bitch
4. Prejudiced Discussions	2,277	bitch, **whorellywood**, pussypass, feminazis, feminazi, pussy, whore, femoid, pussies, **trumpaccusers**

Table 1: Examples of lexicons for each sexism category, selected from *shapley-aug-berttweet*'s top-10 sexism score. Lexicons that are unique to a particular category are highlighted in bold. More examples are presented in Appendix A.

Recently, Hwang and Lee (2021) presented a new approach to lexicon construction. They based their lexicon on the explainability score from various XAI techniques, such as Local Interpretable Model-Agnostic explanations (LIME), Shapley Additive exPlanations (SHAP), Layer-wise Relevance Propagation (LRP) and Gradient-weighted Class Activation Mapping (Grad-CAM). Their results showed that their lexicons are more robust and of better quality, as determined by human reviewers.

3 System Overview

In this work, our investigation is directed towards task B and C due to the relatively limited number of data available per class in these tasks. Task A, on the other hand, has been approached briefly and serves as a base model for tasks B and C for our lexicon resource.

3.1 Subtask A

To build a classifier for task A, we experimented with a simple fine-tuning approach on 3 Large Language Models (LLMs): BERT (*bert-base-uncased*) (Devlin et al., 2018), BERTweet(*vinai/bertweet-base*) (Nguyen et al., 2020) and TwHIN-BERT(*Twitter/twhin-bert-base*) (Zhang et al., 2022). The fine-tuning parameters for the classifier were set as follows:

- Learning rate: 1e-5
- Optimizer: Adam
- Adam optimizer parameters:
  - Adam epsilon: 1e-8
  - Weight decay rate: 0.2
- Number of epochs: 5
- Batch size: 32
- Learning rate scheduler: Linear warm-up

These parameters were chosen based on prior research and experimentation on a validation set to achieve optimal fine-tuning performance. As a result, we got ***0.796***, *0.740* and *0.776* F1 score from BERT, BERTweet and TwHIN-BERT, respectively. We decided to use fine-tuned BERT as our base model for the rest of the experiments as it provided the best predictions.

3.2 Subtask B and C

In this section, we present our lexicon-based approach to classifying sexism. Our process consists of three main steps: lexicon construction, lexicon augmentation, and incorporating the lexicon information.

In the lexicon construction step, we automatically selected a set of words corresponding to the annotated labels from Task B and C using pointwise mutual information (PMI) and Shapley values. In the augmentation step, we expanded the lexicon by leveraging unannotated data and two large language models, namely, BERTweet and GPT-J (*EleutherAI/gpt-j-6B*). Finally, in the lexical incorporation step, we integrated the lexicons into 3 classifiers: lexicon-based classifier, logistic regression, and fine-tuned LLMs.

3.2.1 Lexicon Construction

We experimented with two approaches to estimate a sexism score of a word to a sexism category: PMI and Shapley value.

Our PMI method is inspired by the work of Turney (2002). A word is assigned a numerical rating by taking the mutual information between the given phrase and a target category.

In contrast to Turney (2002)'s approach, which uses the difference in the PMI score between positive and negative reference words as a categorical score making it more suitable for binary classification tasks but does not be directly applicable to tasks with multiple classes, our study directly uses PMI to circumvent that limitation.

The lexicons of each category are selected from those with a PMI score higher than the 90th and 50th percentile for tasks B and C. These thresholds are cherry-picked based on the validation results to balance lexicon quality, its coverage and the model's performance.

Our Shapley method calculates the importance of words using SHAP (Lundberg and Lee, 2017) from Task A's fine-tuned BERT. The averaging score from all instances in training data is used. Scores from all subtokens are added together to get a score for a word. Lastly, we picked words whose Shapley values exceeded the same percentile threshold as the PMI method.

Please note that a word in our lexicon can be assigned to multiple categories with different scores.

3.2.2 Augmentation

We consider having lexicon augmentation mainly to expand the vocabulary size. The motivation behind this came from the preliminary results, which showed that more than 40% of the sentences in the validation data did not have any words in common with our selected lexicons leading to the lower coverage of the classifier.

To address this issue, three strategies were employed, including using an unannotated corpus, BERTweet, and GPT-J. The primary objective of our strategy was to generate additional data for each sexism category. It was done to address the issue of limited data availability and to achieve a more diverse range of words in the lexicons, thereby obtaining a better representation of the validation data.

Use of the unannotated corpus, we used the million unannotated sentences from Gab given by the task organizer. The data were automatically labelled by our fine-tuned BERT from Task A and filtered out non-sexism sentences. We then calculated Shapley values and PMI from the filtered data. Finally, we selected words with Shapley values with the same threshold, similar to the previous setting. A word was assigned to a category if it had PMI more than the 99th percentile of the existing lexicon in that category.

Use of BERTweet, we masked the labelled sentences in training data on the words that were in our initial lexicons. The sentences were masked with 1-3 empty spaces to handle the possibility of getting subwords as the prediction. We then asked the pre-trained model to fill in the blank. Non-sexism sentences were filtered out based on the prediction from our fine-tuned BERT from Task A. The category was assigned to the new words based on the masked original word. As a result, we got 25,501 and 30,000 new sentences for tasks B and C, respectively.

Use of GPT-J, we prompted the model to generate more sexist sentences with the following prompt:

> These are examples of [sexism category] tweets:
> ⋆ . . . example1. . .
> ⋆ . . . example2. . .
> ⋆ . . . example3. . .

On each prompt, the examples were randomly selected from the training data. The model was run with *temperature=0.9* and *max_length=300*. The generated texts were split into sentences and filtered out sentences with no "*" prefix. The category in the prompt was assigned to the sentence. In total, we got 3,125 and 4,215 new sentences for tasks B and C, respectively. We then applied the similar approach for the unannotated corpus to select a new lexicon.

The examples of lexicons from each approach are presented in Appendix **??**.

3.2.3 Incorporating Lexicons

We implemented 3 classifiers that incorporated lexicon information into its prediction.

Lexicon-based classifier: The input text is first tokenized using NLTK tweet tokenizer and made into lowercase to match our lexicon format. We leave out other preprocessors because we believe that the preprocessor will discard the paralinguistic features that the author used intentionally for a certain semantic meaning, such as using repetitive

characters as intensifiers (for example, suuuuuuck) or intentional misspelling to evade censorship detection (for example, b*tch) (Nakwijit and Purver, 2022).

The text is then matched against the lexicon from each category. The model adds up the scores of all words presented in the lexicon. It then presents the category with the maximum score as the prediction. This is the simplest classifier that presents the possibility of using only surface-level information of a sentence to get the prediction and maintains the highest level of explainability.

BoW logistic regression: It is a a bag-of-words (BoW) logistic regression model. This model represents a word using one-hot vectors and applies logistic regression with an L2 penalty to the input vector. The vocabulary utilized in this model was limited to the lexicons constructed in our study. This approach served as a feature selection mechanism, enabling us to evaluate the extent to which our lexicon captures relevant information.

Fine-tuning BERT: we followed the approach presented by Koufakou et al. (2020). A BERT-based model is fine-tuned with a lexical embedding. The lexical embedding of a word w_i is a vector with 1 on the n^{th} dimension if w_i can be observed on the lexicon of n^{th} category. Otherwise, it is set to 0. The embeddings are given to an LSTM to produce a sentence representation, concatenating with *[CLS]* embedding from BERT and feed to a fully-connected layer to predict a categorical output. We experimented with this approach to understand the possibility of using our lexicon to improve the LLM prediction, which is often presented as state-of-the-art in many tasks.

4 Data and Evaluation Methodology

We only used the labelled and unlabeled corpus provided by the task organizer for all experiments. All labelled data are used to fine-tune BERT for task A. For tasks B and C, the corpus is split into train, validation, and test set. We followed the best practice of data modelling; a lexicon is constructed using only train data; thresholds and other hyperparameters are selected based on validation data; and finally, evaluated on test data. We repeated the split 5 times and reported averaging macro-average F1, precision, and recall.

5 Results and Discussion

5.1 Is Shapley lexicons better than PMI lexicons?

Based on results in table 2, without augmentation, PMI lexicons perform better than Shapley lexicons. It might happen due to the fact that, during our evaluation, we left out the sentences that the model cannot predict due to limited vocabulary. We observed that the lexicon-based classifier from PMI has a missing rate of up to 49.4% while Shapley has only 11.5%. However, once we have augmented the lexicons, we can observe that Shapely lexicons can perform better than the PMI one (increase by 8% in task B and 10.4% in task C).

On qualitative analysis, we observe that Shapley lexicons are more coherent, such as *whores, skank, bitch*. At the same time, PMI includes many words that seem unrelated to the task, such as *iceberg, Australia, senate*. It might be because PMI tends to overestimate the relatedness of word pairs in the low-frequent events (Pesaranghader et al., 2013).

As a result, we can conclude that Shapley lexicons can perform better than PMI quantitatively and qualitatively.

5.2 How effective is the augmentation?

Augmentation can successfully expand the vocabulary size, reducing the missing rate in the lexicon-based classifier from 11.5% to 10.2%, 0%, and 0% (augmented from Gab, GPT-J, and BERTweet, respectively).

It also improves precision and recall, except for GPT-J, which hurts the performance. This drop might be due to the noise of the generated texts. Originally, in the GPT-J setting, we generated 10,856 and 18,465 sentences for tasks B and C; more than 70% of them needed to be discarded because they were non-sexism, according to the prediction from our task-A classifier. It indirectly suggests that the category assigned by the prompt is noisy. On the other hand, BERTweet can greatly improve performance. The increasing performance is even more noticeable in low-resource settings (Task C). A similar finding is also presented in Gao et al. (2023).

5.3 How well do lexicons perform at the task?

Based on results in table 2, 3, and 4, we can observe that by simply counting the presence of certain words in a given text and comparing the count can perform better than a BoW logistic regression in

the overall performance. It might be because the lexicon-based model mostly aids in terms of recall while maintaining the same level of precision.

Compared to the pretraining-fine-tuning baseline, the lexicon-based classifier has a lower F1 (44.9%) in task B compared to the fine-tunned BERT (56.8%). It, however, performs better in task C, where the example per class is much lower (25.4% to 31.6%). It suggests that fine-tuned BERT in a low-resource setting does not behave differently from a lexicon-based classifier which is limited to only surface information.

5.4 Can lexicon be used to improve an existing model?

Yes, results in 3 show that lexicons can be used as feature selection. In our best settings (Logistic Regression with Shapley augmented by BERTweet), we can reduce the number of features from 8,311 to 5,341. In other words, the number of model parameters is reduced by 35.7% while it still maintains 99.9% performance. We can also observe a positive correlation (spearman $\rho = 0.5$) between the performance of the lexicon in the lexicon-based classifier and the logistic regression that use those lexicons as features. Suggests that the quality of the lexicons also plays an important role in the resulting model.

However, *No*, in the fine-tuning setting. The lexicons can only slightly improve fine-tuned BERT up to 1% in task B and worse, the model in task C up to 1%. A more advanced technique is needed.

6 Analysis on Model Predictions

In this section, we analyze the fine-tuned BERT from task A and discuss its limitations and biases in the model's predictions. We also manually investigate the resulting lexicons as an explanation for the model's behaviour. Please refer to appendix C for a more detailed analysis.

6.1 Limitation on Negations

In this subsection, we conducted an experiment to evaluate the impact of negation tokens on the model's performance. Specifically, we manually added and removed negation tokens from over 100 training examples and assessed the resulting effect on the model's prediction. Our analysis revealed that such modifications had minimal impact on the model's classification, with only 5

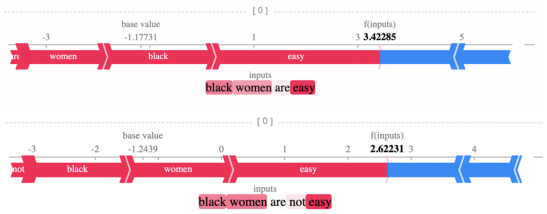


Figure 1: An example when the fine-tuned BERT does not change its prediction when negation was added

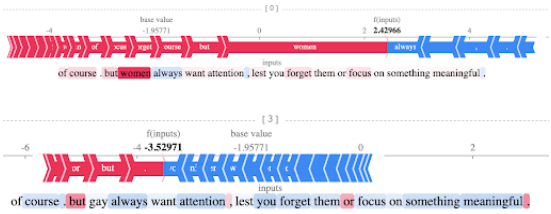


Figure 2: An example of bias towards sexism against women, while ignoring the case of individuals in the LGBTQ+ community.

One example is presented in figure 1, which depicts the Shapely values associated with the addition and removal of negation tokens from an example. The results indicate that the Shapely values were only marginally affected by the presence or absence of negation, with the classification remaining unchanged as either sexist or non-sexist.

More examples are presented in appendix C.1.

6.2 Bias

In this subsection, we address the potential issue of data bias by investigating the inclusivity of our model towards all groups of people. We firmly believe that sexism should not discriminate against any group. To test this, we randomly selected test samples containing feminine words (*"woman"*, and *"mom"*) and replaced them with LGBTQ+ words from the list [*"lesbian"*, *"gay"*, *"lgbtq"*, *"trans"*, *"bi"*]. Our analysis revealed that in 257 out of 407 sexist sentences (63.14

More examples can be found in appendix C.2.

6.3 Analysis on Lexicons

In this subsection, we analyzed the lexicons extracted from the model to identify what kind of lexicons contributes the most to the predictions and how well they are as an explanation of the subtasks.

From Shapley lexicons (without augmentation), we identified a total of 1,335 lexicons. Notably, there is around 4-6% overlap among the sexism

Lexicon-Based Model	Task B (4 classes)			Task C (11 classes)		
	F1	P	R	F1	P	R
With PMI	0.308 ± 0.028	0.305 ± 0.026	0.318 ± 0.031	0.178 ± 0.032	0.191 ± 0.028	0.223 ± 0.028
+ augmentation with Gab	0.313 ± 0.023	0.312 ± 0.023	0.322 ± 0.022	0.169 ± 0.012	0.186 ± 0.017	0.207 ± 0.017
+ augmentation with GPT-J	0.241 ± 0.028	0.265 ± 0.025	0.273 ± 0.029	0.159 ± 0.019	0.173 ± 0.022	0.200 ± 0.016
+ augmentation with BERTweet	**0.369 ± 0.017**	**0.383 ± 0.014**	**0.400 ± 0.018**	**0.212 ± 0.009**	**0.231 ± 0.011**	**0.261 ± 0.013**
With Shapley	0.271 ± 0.019	0.287 ± 0.023	0.286 ± 0.027	0.120 ± 0.010	0.154 ± 0.020	0.140 ± 0.024
+ augmentation with Gab	0.282 ± 0.043	0.302 ± 0.031	0.299 ± 0.053	0.142 ± 0.018	0.178 ± 0.018	0.163 ± 0.037
+ augmentation with GPT-J	0.240 ± 0.016	0.287 ± 0.025	0.280 ± 0.019	0.117 ± 0.008	0.150 ± 0.015	0.141 ± 0.034
+ augmentation with BERTweet	**0.449 ± 0.021**	**0.481 ± 0.024**	**0.498 ± 0.023**	**0.316 ± 0.026**	**0.373 ± 0.017**	**0.383 ± 0.049**

Table 2: Lexicon-based classifier performance: macro-averaging F1, Precision(P) and Recall(R) on PMI and Shapley lexicons and their augmentation by the unannotated corpus (Gab), GPT-J and BERTweet.

BoW Logistic Regression	Task B (4 classes)			Task C (11 classes)		
	F1	P	R	F1	P	R
BoW Baseline	0.397 ± 0.040	0.485 ± 0.041	0.375 ± 0.034	0.250 ± 0.026	0.312 ± 0.076	0.245 ± 0.018
using only PMI lexicon as features	0.209 ± 0.013	0.260 ± 0.094	0.258 ± 0.012	0.167 ± 0.018	0.224 ± 0.047	0.167 ± 0.014
using only Shapley lexicon as features	0.310 ± 0.023	0.443 ± 0.084	0.315 ± 0.014	0.252 ± 0.020	0.305 ± 0.070	0.246 ± 0.012
+ augmentation with Gab	0.312 ± 0.027	0.444 ± 0.086	0.315 ± 0.018	0.252 ± 0.021	0.305 ± 0.070	0.246 ± 0.014
+ augmentation with GPT-J	0.390 ± 0.025	0.455 ± 0.017	0.370 ± 0.025	**0.254 ± 0.029**	0.306 ± 0.079	**0.247 ± 0.019**
+ augmentation with BERTweet	**0.393 ± 0.035**	**0.460 ± 0.044**	**0.371 ± 0.030**	**0.254 ± 0.028**	**0.314 ± 0.075**	0.246 ± 0.019

Table 3: BoW Logistic regression classifier performance: macro-averaging F1, Precision(P), and Recall(R) using PMI and Shapley lexicons and their augmentations as feature selection.

Fine-tuned LLMs	Task B (4 classes)			Task C (11 classes)		
	F1	P	R	F1	P	R
Fine-tuned BERT	0.568 ± 0.029	0.593 ± 0.033	0.553 ± 0.032	0.272 ± 0.007	0.278 ± 0.033	0.290 ± 0.013
+ Shapley lexicons	0.579 ± 0.026	0.593 ± 0.028	0.574 ± 0.031	0.263 ± 0.018	0.279 ± 0.036	0.274 ± 0.019
+ Shapley lexicons augmented with Gab	0.580 ± 0.026	0.593 ± 0.028	0.575 ± 0.030	0.262 ± 0.016	0.277 ± 0.037	0.273 ± 0.016
+ Shapley lexicons augmented with GPT-J	0.577 ± 0.026	0.592 ± 0.029	0.573 ± 0.031	0.262 ± 0.018	0.278 ± 0.038	0.273 ± 0.018
+ Shapley lexicons augmented with BERTweet	0.578 ± 0.022	0.592 ± 0.027	0.573 ± 0.025	0.262 ± 0.019	0.278 ± 0.037	0.273 ± 0.020

Table 4: Fine-tuned BERT classifier performance: macro-averaging F1, Precision(P) and Recall(R) with and without lexical incorporation.

classes except for the *2. Derogation* and *3. Animosity* pair, which had an overlap of 8.49

Furthermore, our analysis revealed that the most common words across all categories are predominantly women-related slang/swear words (e.g, bitch, cunt, pussy, slot, and whore). Lexicons associated with women of color are often found as in *1. Threats* (e.g. nigga, nigger) while Asian-women-related lexicons are predominantly observed in *2. Derogation* (e.g. noodlewhore, noodlefoids). The sexist ideology towards Muslim women is prevalent across all classes (e.g. islamophobe, religonofpeace (use sarcastically), mudslime). Moreover, *1. Threats* and *2. Derogation* are more politically oriented featuring hashtags such as #hillaryclintonisabitch, #uniteamerica, #saveamerica. Lastly, we observe the large contribution from urban slang (e.g. 3/10, cuckolded, slutshamed, willing, hb10, pork, sl00ts, gymmaxxed), highlighting the importance of accommodating the continuous increase of vocabulary and the shift in meaning.

While our lexicon approach provides a useful tool for identifying and analyzing sexist language, it has its downsides. Manually investigating the lexicons suggests that our approach might pick up on dataset artefacts instead of general indicators of sexism. Many words (e.g. hahaha, bubblegum, people's, religonofpeace, #uniteamerica) are truly indicative of sexism in general, but they are over-represented in the training data. This can lead to false positives and misinterpretations of the data.

7 Conclusion

In this paper, we present our experiments on lexicons for sexism classification. The sexism lexicons are constructed using PMI and Shapley values which we, later, show that the Shapley approach can yield higher quantitative and qualitative results compared to PMI. In the augmentation step, we investigated 3 approaches to expand the lexicon's vocabulary size including the use of an unannotated corpus, GPT-J, and BERTweet. Our results indicate that using fill-in-the-blank methods from BERTweet is the most effective method for reducing the missing rate to 0% while also increasing F1 by 17.8% and 19.5% in tasks B and C respectively. Moreover, we also show that by using only lexicons we can outperform BoW logistic regression (in tasks B and C) and fine-tuned BERT (in task C). The lexicons can also be effectively used as feature selection but there is no improvement when it is incorporated into the pre-trained BERT.

Our study highlights two crucial use cases of XAI techniques. The first is to identify potential limitations and biases in the model's predictions. We used adversarial examples to demonstrate that the fine-tuned BERT model has limitations in handling negation. Moreover, our analysis revealed that different groups of people, including Women of color, Asian, Muslim, and LGBTQ+ individuals, are associated differently with various forms of sexism. However, we found that the training data may under-represent some of these vulnerable groups, such as LGBTQ+, which raises concerns about potential ignorance towards them. The second use

case of XAI is to identify and quantify the features that contribute most to the model's predictions and use them as a guideline for better model design. We found that urban slangs are the most significant contributor to the model's predictions, emphasizing the importance of the continuous increase in the new vocabulary and the shifts in meaning.

Lastly, it's important to note that the lexicon-based model we used has some limitations. Our approach heavily relies on the lexicons extracted from the training data, which might pick up on dataset artifacts rather than general indicators. Therefore, the findings from our analysis should be interpreted with caution and validated on other datasets to ensure their generalizability.

Limitations

Our use of lexicons for classification has some limitations that should be considered. Firstly, it is not suitable for capturing context-dependent and indirect meanings, such as idioms, sarcasm, or multi-word expressions. However, we believe it is less of an issue in task B and C where the focus is on explanation rather than classification.

Another major limitation is related to the methodology of our lexicon construction. PMI tends to overvalue the related score in low-frequency events, which can result in a more faulty lexicon in low-resource settings. The Shapley approach could also lead to unrealistic conclusions when features are correlated (Lundberg and Lee, 2017). Furthermore, the explainability of the lexicons is limited by the model they refer to. Therefore, it is important to note that the lexicons constructed by our Shapley approach are limited to the performance of the model trained on task A and may be affected by the same biases present in the model. Finally, it is worth noting that our lexicon-based approach may have picked up on dataset artefacts rather than general indicators, which raises questions about the generality of our constructed lexicons.

Acknowledgements

Matthew Purver is supported by the Slovenian Research Agency via research core funding for the programme Knowledge Technologies (P2-0103) and the project SOVRAG (Hate speech in contemporary conceptualizations of nationalism, racism, gender and migration, J5-3102); and the UK EPSRC via the projects Sodestream (Streamlining Social Decision Making for Improved Internet Standards, EP/S033564/1) and ARCIDUCA (Annotating Reference and Coreference In Dialogue Using Conversational Agents in games, EP/W001632/1).

References

Nawaf Abdulla, Salwa Mohammed, Mahmoud Al-Ayyoub, Mohammed Al-Kabi, et al. 2014. Automatic lexicon construction for arabic sentiment analysis. In *2014 International Conference on Future Internet of Things and Cloud*, pages 547–552. IEEE.

Nawaf A Abdulla, Nizar A Ahmed, Mohammed A Shehab, and Mahmoud Al-Ayyoub. 2013. Arabic sentiment analysis: Lexicon-based and corpus-based. In *2013 IEEE Jordan conference on applied electrical engineering and computing technologies (AEECT)*, pages 1–6. IEEE.

Manuela Barreto and David Matthew Doyle. 2022. Benevolent and hostile sexism in a shifting global context. *Nature reviews psychology*, pages 1–14.

Arun Das and Paul Rad. 2020. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*.

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*.

European Institute for Gender Equality. 2023. European Institute for Gender Equality. https://eige.europa.eu/.

Jesse Fox, Carlos Cruz, and Ji Young Lee. 2015. Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media. *Computers in human behavior*, 52:436–442.

Jun Gao, Changlong Yu, Wei Wang, Huan Zhao, and Ruifeng Xu. 2023. Mask-then-fill: A flexible and effective data augmentation framework for event extraction. *arXiv preprint arXiv:2301.02427*.

Hohyun Hwang and Younghoon Lee. 2021. Semi-supervised learning based on auto-generated lexicon using xai in sentiment analysis. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 593–600.

Aiqi Jiang, Xiaohan Yang, Yang Liu, and Arkaitz Zubiaga. 2022. Swsr: A chinese dataset and lexicon for online sexism detection. *Online Social Networks and Media*, 27:100182.

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.

Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, Viviana Patti, et al. 2020. Hurtbert: Incorporating lexical features with bert for the detection of abusive language. In *Proceedings of the fourth workshop on online abuse and harms*, pages 34–43. Association for Computational Linguistics.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*.

Pakawat Nakwijit and Matthew Purver. 2022. Misspelling semantics in Thai. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 227–236, Marseille, France. European Language Resources Association.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200*.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Ahmad Pesaranghader, Saravanan Muthaiyah, and Ali Pesaranghader. 2013. Improving gloss vector semantic relatedness measure by integrating pointwise mutual information: Optimizing second-order co-occurrence vectors computed from biomedical corpus and umls. In *2013 International Conference on Informatics and Creative Multimedia*, pages 196–201. IEEE.

Wessel Stoop, Florian Kunneman, Antal van den Bosch, and Ben Miller. 2019. Detecting harassment in real-time as conversations develop. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 19–24.

Janet K Swim, Lauri L Hyers, Laurie L Cohen, and Melissa J Ferguson. 2001. Everyday sexism: Evidence for its incidence, nature, and psychological impact from three daily diary studies. *Journal of Social issues*, 57(1):31–53.

Peter D Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. *arXiv preprint cs/0212032*.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Leyi Wang and Rui Xia. 2017. Sentiment lexicon construction with representation learning based on hierarchical sentiment supervision. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 502–510.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2022. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations. *arXiv preprint arXiv:2209.07562*.

A Examples of Task B Lexicons

Randomly selected lexicons in each sexism category from all experimental settings. Lexicons that are unique to a particular category are highlighted in bold.

Sexism Category	N	Lexicons
PMI Lexicons		
1. Threats,	484	**woman.**, **coworkers**, **lovin**, **purposely**, **laughed**, **zog**, **replaceable**, **£20**, **kosher**, **genital**
2. Derogation,	2,418	**threir**, **indians**, **freckles**, **niceguys**, **master**, **contain**, **unshaved**, **events**, **2/3rds**, **shitfuck**
3. Animosity,	1,890	**log**, **misbehaving**, **homie**, **limelight**, **atleast**, **setbullshit**, **appearing**, **nodding**, **bint**, **priorities**
4. Prejudiced Discussions,	608	**sundaymorning**, **boermeester**, **english**, **interviewed**, **histrionics**, **stiletto**, **testify**, **confirmkavanaugh**, **tldw**, **predict**
PMI Lexicons augmented with Gab		
1. Threats,	635	**tate**, **porks**, **headfirst**, **ice**, **explodes**, **africans/latinos**, **drinks**, **stat**, **patriarch**, **hen**
2. Derogation,	3,078	**4.0**, **millennials**, **satisfied**, **sl00ts**, **accidently**, **immorality**, **inability**, **koolaid**, **disconected**, **congresswomen**
3. Animosity,	2,423	**lml**, **quickest**, **ambush**, **pal**, **faggot-ass**, **deus**, **leaning**, **grimace**, **whiz**, **whitepages**
4. Prejudiced Discussions,	775	**coasting**, **prioritizing**, **blue-pilled**, **uncorroborated**, **fraudulent**, **coaching**, **cuckoldry**, **schrödingers**, **genius**, **micheal**
PMI Lexicons augmented with GPT-J		
1. Threats,	2,505	tweet, **continuing**, **turkey**, **pee**, forward, **wrist**, **asphalt**, **flawed**, **massing**, **stomping**
2. Derogation,	4,218	lol, **rejections**, **compensate**, **pontificating**, **ma'am**, **hood**, **eveything**, **lurks**, **parlor**, **bullshit**
3. Animosity,	3,690	**hillbilly**, **1940s**, **terror-spawn**, bullied, **romantically**, **grotto**, **exceedingly**, **ben**, **pulls**, **doxx**
4. Prejudiced Discussions,	2,526	**hypocrite**, **finest**, **reads**, **ads**, **gentleman**, **pic**, **convince**, **warming**, **wars. . .**, **drafted**
PMI Lexicons augmented with BERTweet		
1. Threats,	1,897	rid, **commies**, **bidding**, **jewshit-babbling**, night, dirt, slap, **assaulting/harassing**, **assaulting/harassing**, **srs**
2. Derogation,	6,576	**solace**, **self-admission**, **lacy**, **appointment**, **kathygriffin**, **mouthed**, ka, **explores**, **guff**, **societies**
3. Animosity,	5,360	pound, **exceedingly**, **domina**, **stall**, **12-bore**, **male-female**, **sissy**, **packaged**, **seize**, pies
4. Prejudiced Discussions,	2,538	**transsexualism**, **girl.cigars**, **aspect**, **bets**, **req**, **gaymobiles**, **think/know**, gig, **compatible**, again
Shapley Lexicons		
1. Threats,	198	womens, hahaha, skank, **capital**, **saveamerica**, **islamophobe**, **bleach**, **trump's**, **voterepublican**, **feminism**
2. Derogation,	587	slutty, dykes, enslave, slag, **motivates**, **cunty**, **rawdogging**, woman's, **cuckservative**, **curviest**
3. Animosity,	499	**evedence**, **thotlife**, whore's, **bubblegum**, **dontmakemelaugh**, **thotdom**, **oil**, **slanderer**, cunt, **vagene**
4. Prejudiced Discussions,	243	**enslaving**, **aesthetically**, **people's**, **allen's**, **sprog**, feminazi, **person**, **hellyweird**, **sexuality**, **hire**

Sexism Category	N	Lexicons
Shapley Lexicons augmented with Gab		
1. Threats,	245	**cuntbrained**, **solar**, pussies, **1wga**, sluttier, lmao, **lock**, eats, **howler**, **howiwillchange**
2. Derogation,	1,049	**thermograph**, **1966**, brittanycovington, **obedient**, **losers**, cuhzz, **wino's**, **evolution**, **kidnappers**, **menprovement**
3. Animosity,	915	whore's, chokeslamming, **gabfamsuxsshit**, **hahahahahaha**, **embarrass**, loney, **somebody**, **christiansoldiersunite**, **liberalwhackjob**, **1337dchess**
4. Prejudiced Discussions,	389	**womensequalityday**, venker, feminazis, responsable, **bunnygirl**, kyloren, **gynocentrism**, women's, akafemin, **ugliest**
Shapley Lexicons augmented with GPT-J		
1. Threats,	2,266	whos, hag, **censor**, **poppy**, deserve, happens, calling, rule, isn't, twats
2. Derogation,	2,710	thinking, years, **cuntishness**, fire, **neighborhood**, **female's**, **firms**, sh, nastiest, **op**
3. Animosity,	2,508	**i'ma**, **soycuckfucks**, trying, **reproduce**, **rapefugees**, **prior**, **feministas**, whore's, **jumping**, slut
4. Prejudiced Discussions,	2,191	**man's**, **flattery**, wasn't, those, wear, bitchy, from, tweet, greatest, **louissjw**
Shapley Lexicons augmented with BERTweet		
1. Threats,	1,674	feminism, lr, cut, pussies, cunt, list, sex, **esp**, old, **lon**
2. Derogation,	5,439	stern, thots, **partially**, pussy, **femi**, **remarks**, **fembots**, **facsimile**, foid, **curiously**
3. Animosity,	4,428	try, **fetishes**, **soap**, **rapefugees**, cre, **prostate**, blew, current, **depite**, **whored**
4. Prejudiced Discussions,	2,277	**replies**, **appalling**, cunt, family, womans, **keepamericagreat**, feel, robots, ;, torture

B Examples of Task C Lexicons

Randomly selected lexicons in each sexism category from all experimental settings. Lexicons that are unique to a particular category are highlighted in bold.

Sexism Category	N	Lexicons
PMI Lexicons		
1.1 Threats of Harm,	257	**fill**, **solution**, **sexiest**, **wie**, **somber**, decide, **alice**, **stick**, **meat**, **commies**
1.2 Incitement and Encouragement of Harm,	839	**dump**, **tribe**, feinstein, **apointed**, **calais**, **rapists**, **noose**, **spilt**, **rapists**, islam
2.1 Descriptive Attacks,	1,772	**high-tier**, cunty, **humanly**, tradcucks, competing, **negates**, **librarian**, **kissless**, **raft**, **dyes**
2.2 Aggressive and Emotive Attacks,	1,618	**pest**, **midol**, **cenk**, eastern, **detested**, cultural, anybody, **century**, **she-boon**, **goodlooking**
2.3 Dehumanising Attacks & Overt Sexual Objectification,	701	**handmaids**, **het**, **1/8th**, **swelled**, plague, **hivemind**, **superior**, septic, **bratz**, **mentions**
3.1 Casual Use of Gendered Slurs, Profanities, and Insults,	1,685	**closes**, **she-**, **refusal**, **flowed**, **desperation**, **b4**, **creation**, **disseminate**, **2,500**, **bint**
3.2 Immutable Gender Differences and Dender Stereotypes,	1,246	**friendships**, **score**, confidence, **wxmxn**, wnba, **finds**, manhood, **bodice-ripper**, **starve**, **initiated**
3.3 Backhanded Gendered Compliments,	273	**wakes**, **gym**, pleasure, skinny, aww, tier, **tip**, suit, **virgins**, pleasure
3.4 Condescending Explanations or Unwelcome Advice,	237	**k.**, boss, **monitor**, **cat**, **anecdotes**, card, **indifferent**, **step**, **eyes**, oppress
4.1 Supporting Mistreatment of Individual Women,	396	**ideas**, **hellyweird**, **bombshell**, **brutally**, **schrödingers**, **practice**, **whooped**, guilt, dc
4.2 Supporting Systemic Discrimination Against Women as a Group,	938	**innovator**, insure, **spaces**, **reads**, **challenged**, **activated**, **journalist**, **applying**, **succeeds**, **'equality**
PMI Lexicons augmented with Gab		
1.1 Threats of Harm,	301	**harmed**, **smack**, **topic**, hamburger, **turkey**, gook, **kike**, apple, accident, feet
1.2 Incitement and Encouragement of Harm,	1,005	**arguably**, slavic, **cellmate**, dominance, **matbe**, **negging**, **sufficiently**, encourage, **several**, nigga
2.1 Descriptive Attacks,	2,283	**hobby**, **donkeys**, **movement**, **weaponized**, intimate, evening, **feminazies**, **ifs**, **'spinsters**, **gyow**
2.2 Aggressive and Emotive Attacks,	2,069	**bader**, **acknowledging**, **humor**, **contributed**, **roastie**, misandrists, **wind**, shaped, stillshepersisted, **homely**
2.3 Dehumanising Attacks & Overt Sexual Objectification,	883	**bp**, faces, **nsfw**, shoshanna, **shafted**, **schlong**, correction, **roastbeef**, **kebabs**, settle

Sexism Category	N	Lexicons
3.1 Casual Use of Gendered Slurs, Profanities, and Insults,	2,195	newfound, **biddy**, **neighborhoods**, **transactional**, **shit-stink**, **babylon**, **gabgirls**, **intellectual**, tgirl, **navarro**
3.2 Immutable Gender Differences and Dender Stereotypes,	1,497	**manospere**, played, hired, alcoholic, heaven, **ultimate**, ocean, flaw, stronger, **bossy**
3.3 Backhanded Gendered Compliments,	356	**hotter**, **form**, range, **regardless**, pays, **despised**, **rental**, transfeminine, **freetommy**, couldnt
3.4 Condescending Explanations or Unwelcome Advice,	305	**desexualized**, precious, **stuck**, regular, **pink**, propaganda, 2-3, cats, empowered, **manifestations**
4.1 Supporting Mistreatment of Individual Women,	476	**plays**, **kavanaugh**, **handicapped**, **w**, **loud**, **katy**, **zero**, spit, **wakeupamerica**, dc
4.2 Supporting Systemic Discrimination Against Women as a Group,	1,156	unfair, **objectification**, bothered, /s, **apparently**, **finances**, wage, **purchase**, **smarten**, **mandatory**
PMI Lexicons augmented with GPT-J		
1.1 Threats of Harm,	1,321	worked, **asphalt**, theres, punch, **brat**, rubbing, kidding, **smelt**, gymnastics, drilled
1.2 Incitement and Encouragement of Harm,	1,829	**kicks**, haley, **chain**, **17:16**, shields, **defeat**, **swears**, **arabs**, realised, **nail**
2.1 Descriptive Attacks,	2,719	**pasd**, **olympic**, **lavish**, **hesitating**, **life-like**, 80, army, realizing, **giddiness**, **inconceivable**
2.2 Aggressive and Emotive Attacks,	2,541	girlfriend, but, **ties**, **swam**, **man™**, **christine**, **mouthed**, **acknowledging**, **avoidance**, **jew-jizz**
2.3 Dehumanising Attacks & Overt Sexual Objectification,	1,868	pills, **femoids**, **ponies**, 19, **moossies**, septic, **lighter**, **sterner**, import, weed
3.1 Casual Use of Gendered Slurs, Profanities, and Insults,	2,729	robbed, **codreanu**, **outright**, **martyr**, **dopey**, huh, **shite**, **x**, nails, **spencer**
3.2 Immutable Gender Differences and Dender Stereotypes,	2,318	**guppy**, **perma**, dildo, sisters, **drsleeper**, **activities**, **1983**, **feds**, **low-quality**, simplest
3.3 Backhanded Gendered Compliments,	1,241	anyways, muslims, wondering, pussy, **qualm**, **heavier**, **crow**, delusional, delusional, weak
3.4 Condescending Explanations or Unwelcome Advice,	1,120	jesus, period, mechanism, club, **disappear**, doubt, finding, **comfort**, neighborhood, hookers
4.1 Supporting Mistreatment of Individual Women,	1,566	virginity, **uncorroborated**, **handicapped**, **teases**, choices, **teases**, loud, **evidence**, drunk, **yup**

Sexism Category	N	Lexicons
4.2 Supporting Systemic Discrimination Against Women as a Group,	2,042	**featuring**, remove, retire, damaging, **projects**, **surplus**, **sexualization**, wade, heels, **carbon**
PMI Lexicons augmented with BERTweet		
1.1 Threats of Harm,	800	**decapitated**, **napalm**, chicken, **plant**, **loomer**, **laura**, tf, fried, vodka, and/or
1.2 Incitement and Encouragement of Harm,	2,088	defense, **crib/pad**, curb, **sway**, be, pee, favour, **17:16**, leaders, happening
2.1 Descriptive Attacks,	4,732	womensmarch, **garage**, **conquerors**, **fuels**, **abomination**, **relatives**, **simulation**, realized, **ballot**, **prostitution**
2.2 Aggressive and Emotive Attacks,	4,362	**fapped**, aspects, italy, humor, **unmarried**, bottom, torture, **speeds**, **chong**, **natalie**
2.3 Dehumanising Attacks & Overt Sexual Objectification,	1,996	pills, brought, **scenarios**, drain, **puas**, car, pushing, present, **janitor**, living
3.1 Casual Use of Gendered Slurs, Profanities, and Insults,	4,598	**uncle**, michael, **negroes**, smelling, **exposes**, **wholesome**, manginas, **nahh**, **julius**, worried
3.2 Immutable Gender Differences and Dender Stereotypes,	3,114	miles, **u.**, **weather**, wnba, exceptions, **patient**, lets, **writes**, subreddit, **ex-wives**
3.3 Backhanded Gendered Compliments,	874	russian, body, paul, **shapeless**, **coper**, **developing**, guns, minority, form, shag
3.4 Condescending Explanations or Unwelcome Advice,	757	**monitor**, ironically, x, opposite, wife, guy, say, mechanism, suffering, letting
4.1 Supporting Mistreatment of Individual Women,	1,069	smash, **groped**, **scarily**, fe, **molested**, tie, **kebab**, shop, justice, **testify**
4.2 Supporting Systemic Discrimination Against Women as a Group,	2,528	**pic**, mm, reserved, **todays**, **scares**, sweat, protection, **bestows**, proven, acceptance
Shapley Lexicons		
1.1 Threats of Harm,	301	**explodes**, slapped, speak, cunt, scumbag, yelling, **raff**, ,, force, today
1.2 Incitement and Encouragement of Harm,	864	pussy, **strike**, **hb10**, sub, arse, looked, thats, 1950s, cunt, named
2.1 Descriptive Attacks,	1,834	**congresswomen**, **wahmen**, trannys, **ennemy**, **pussyhats**, **functioning**, **subhuman**, selfish, **romanian**, **depreciating**
2.2 Aggressive and Emotive Attacks,	1,680	**stacies**, envy, **tramp**, **248yrs**, virgins, **trap**, **skanko**, rule, **pussyhat**, vagina
2.3 Dehumanising Attacks & Overt Sexual Objectification,	758	else, **heard**, **jep**, pedo, dream, foids, weak, silly, **sheeit**, incels

Sexism Category	N	Lexicons
3.1 Casual Use of Gendered Slurs, Profanities, and Insults,	1,746	hoe, slut, trannys, whore's, **california**, whoredom, sucking, can't, aids, **slanderer**
3.2 Immutable Gender Differences and Dender Stereotypes,	1,232	ofc, **mars**, virgin, **initiate**, well, haha, flakey, **daygame**, **2incld**, favor
3.3 Backhanded Gendered Compliments,	329	every, wrong, **sensible**, aren't, **faces**, femininity, **dont-makemelaugh**, bitch, care, hotter
3.4 Condescending Explanations or Unwelcome Advice,	281	play, **bullet**, their, :), others, fact, pussy, men, making, 25
4.1 Supporting Mistreatment of Individual Women,	452	**histrionics**, who, girl, whining, some, **presidenttrump**, choices, but, sucker, **night**
4.2 Supporting Systemic Discrimination Against Women as a Group,	1,000	**reads**, women's, **sterilize**, femoid, **giant**, water, **genius**, **stiletto**, **program**, ,
Shapley Lexicons augmented with Gab		
1.1 Threats of Harm,	337	speak, whore, **virginal**, bitch, **anblick**, **purposes**, women, fuckug, **extradite**, too
1.2 Incitement and Encouragement of Harm,	986	femboy, **titwank**, unhinged, quality, **serenawilliams**, says, **implication**, consensual, **neocons**, **server**
2.1 Descriptive Attacks,	2,454	autistic, **dye**, **absolutely**, **fridayfolkday**, earth, **exists**, guns, minetheth, thursdayreads, women's
2.2 Aggressive and Emotive Attacks,	2,202	chauvinism, idenitites, **suffrance**, woman's, bux, skeletonized, trashy, thots, bitchy, **untermensch**
2.3 Dehumanising Attacks & Overt Sexual Objectification,	928	**kazharian**, wear, that's, **humongous**, **arse'ole**, evil, cross-dressers, spread, **pcos**, **jada**
3.1 Casual Use of Gendered Slurs, Profanities, and Insults,	2,343	**ensues**, **pussyfooting**, hahaha, **momma**, [, chokeslamming, vaginocracy, **soccermoms**, **gabfamsuxsshit**, homocommandos
3.2 Immutable Gender Differences and Dender Stereotypes,	1,509	cope, **genuinly**, **kavanuagh**, **celebrity**, **tee**, dropped, **theluckyones**, **feminime**, **pschology**, styxenhexenhammer
3.3 Backhanded Gendered Compliments,	392	them, females, out, womens, she's, russian, needs, slap, thighhighs, bover
3.4 Condescending Explanations or Unwelcome Advice,	346	**lass**, **renders**, rublish, eddieberganza, hookers, feminismishate, play, feminismishate, url, lady
4.1 Supporting Mistreatment of Individual Women,	535	gabfam, gab, chucklefuck, part, **wakepamerica**, **regretted**, pussies, sexuality, **accuser**, skills

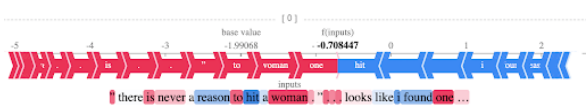
Sexism Category	N	Lexicons
4.2 Supporting Systemic Discrimination Against Women as a Group,	1,227	**hollyweird**, disney, **dealing**, **singing**, hybristophilia, slut, crossdressers, **micheal**, **ripa**, kavs
Shapley Lexicons augmented with GPT-J		
1.1 Threats of Harm,	1,323	moving, bitching, stolen, sluts, ..., wish, wet, relationship, **stroked**, slut
1.2 Incitement and Encouragement of Harm,	1,800	or, wonder, out, females, prostitutes, cuckolded, **uniteamerica**, **final**, damn, it's
2.1 Descriptive Attacks,	2,748	impulses, **undercooked**, cunts, **prone**, **interactions**, submissive, garbage, **operative**, and, after
2.2 Aggressive and Emotive Attacks,	2,612	**condolences**, alimony, **woemen**, **justin**, **poopoo**, perverted, **sasour**, **slender**, disposable, sideways
2.3 Dehumanising Attacks & Overt Sexual Objectification,	1,814	idea, teachers, **fembots**, dehumanized, button, argument, **screams**, argument, assholes, 'bitch
3.1 Casual Use of Gendered Slurs, Profanities, and Insults,	2,701	tranny, **pseudo**, dyke, muslim, **bitchboi**, **cuntfused**, femoids, bitchy, **percentage**, heh
3.2 Immutable Gender Differences and Dender Stereotypes,	2,246	lolol, slapped, **holidays**, **mountains**, **expenctancy**, wear, self, total, rapes, retard
3.3 Backhanded Gendered Compliments,	1,215	**inflames**, **possessed**, **good-guy-with-a-gun**, ain't, blonde, else, wanna, **whack**, wrong, ;
3.4 Condescending Explanations or Unwelcome Advice,	1,109	**symmetry**, go, already, wifey, bitches, girlfriend, so, ..., girls, hoe
4.1 Supporting Mistreatment of Individual Women,	1,517	**presidenttrump**, move, culture, **lnyhbt**, pissed, defend, choke, normie, bitches, damn
4.2 Supporting Systemic Discrimination Against Women as a Group,	2,006	for, wave, finger, lil, **lolgirl1**, **ignominy**, become, legislature, **mentioned**, straight
Shapley Lexicons augmented with BERTweet		
1.1 Threats of Harm,	801	boils, sick, **explodes**, man, meeting, ki, stories, throwing, **kosher**, **alice**
1.2 Incitement and Encouragement of Harm,	2,087	fucker, cunts, dumbass, live, it, re, impregnate, **escalating**, **coworkers**, entry
2.1 Descriptive Attacks,	4,775	female, **towers**, president, threesome, retarded, apples, wish, pussy, school, manipulated
2.2 Aggressive and Emotive Attacks,	4,444	**oniggy**, dykes, pan, ate, **misandrists**, remorse, ici, **infest**, comic, shameful
2.3 Dehumanising Attacks & Overt Sexual Objectification,	2,001	loyal, effect, they'd, g, all, obviously, sue, gunn, **necessity**, because

Sexism Category	N	Lexicons
3.1 Casual Use of Gendered Slurs, Profanities, and Insults,	4,665	**femnazis**, femoids, shekels, self, **slayers**, **muhahahahaha-hahahahahahahahahahahaha**, **geeky**, **muhahahahahahaha-hahahahahahahahaha**, shows, tranny
3.2 Immutable Gender Differences and Dender Stereotypes,	3,101	robber, **psychotics**, kind, what's, you've, **perverted**, n, friendzone, **ppa**, hea
3.3 Backhanded Gendered Compliments,	878	behind, **miniskirts**, getting, trumps, younger, **troops**, single, blonde, fat, star
3.4 Condescending Explanations or Unwelcome Advice,	759	or, her, tone, count, **freedumb**, get, x, bitch, she'd, yes
4.1 Supporting Mistreatment of Individual Women,	1,061	du, communicate, two, yup, bitch, **bail**, better, realize, **assumes**, equal
4.2 Supporting Systemic Discrimination Against Women as a Group,	2,526	politicians, guy, english, for, guess, **femaleprivilege**, ideology, rights, done, co

C Analysis on Model Predictions

C.1 Negations

This subsection presents more examples of negation that the fine-tuned BERT model misclassifies.



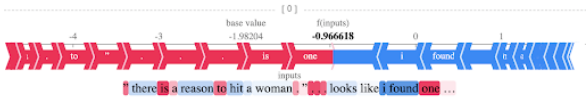


Figure 3: Examples of negation

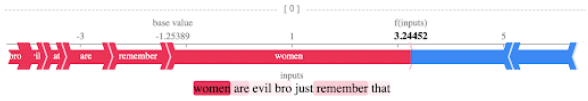
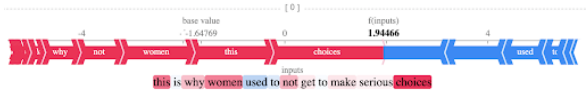


Figure 4: Examples of negation



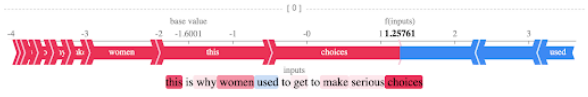
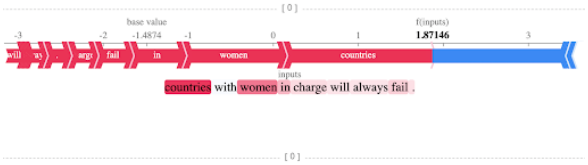


Figure 5: Examples of negation

C.2 Bias

More examples of bias towards sexism agaist women but ignore LGBTQ+ people.



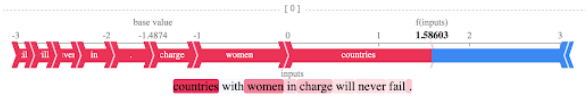
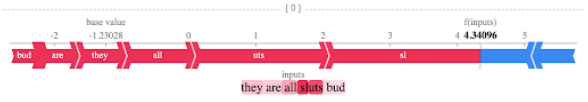


Figure 6: Examples of negation



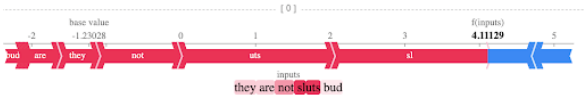


Figure 7: Examples of negation

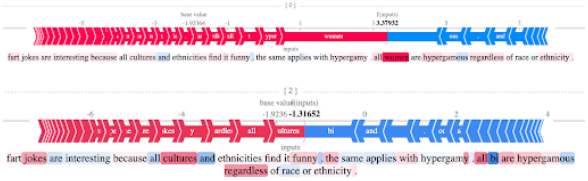


Figure 8: Examples of bias

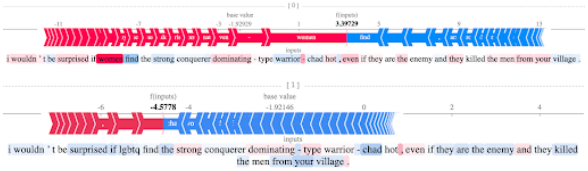


Figure 9: Examples of bias

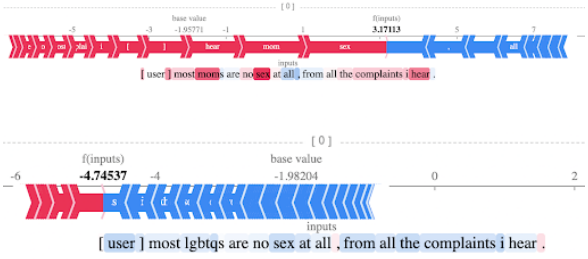


Figure 10: Examples of bias

C.3 Robustness

Robustness in machine learning models refers to the ability to maintain consistent behavior and decisions even in the presence of noise. In online content, automatic sexism detectors may misclassify sexist comments as funny comments when it includes some words like “hahaha”, “lol”, and “funny”. To test the robustness of our model against this kind of attack, we added random funny words from a list [hahaha, hehehe, lol lol, lmao lol, lmfao] to the beginning and end of each sentence from test data and checked if the prediction changed. We found that the model maintained its prediction on 3,815 sentences out of 4,000 sentences, or 4.625%. Overall, this finding suggests that while our model is generally robust against noise in the form of noise.

C.4 Lexicon Overlap

From Shapley lexicons (without augmentation), we identified a total of 1,335 lexicons. Notably, there is around 4-6% overlap among the sexism classes except for the *2. Derogation* and *3. Animosity* pair (8.49%), as illustrated in the plot in figure 11 and 12.

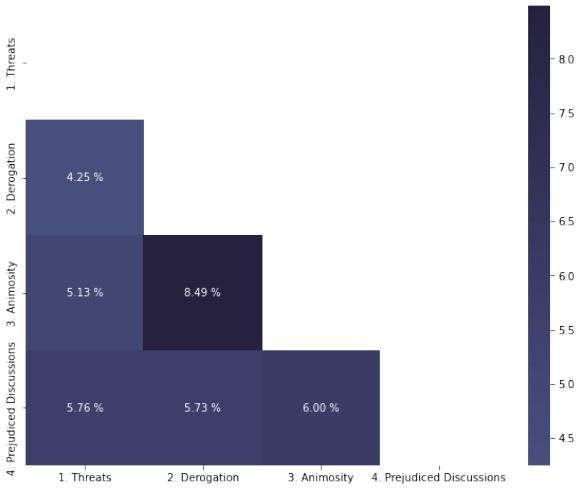


Figure 11: Jaccard Similarity Coefficient between lexicons from all sexism classes in task B from Shapley lexicons (without augmentation)

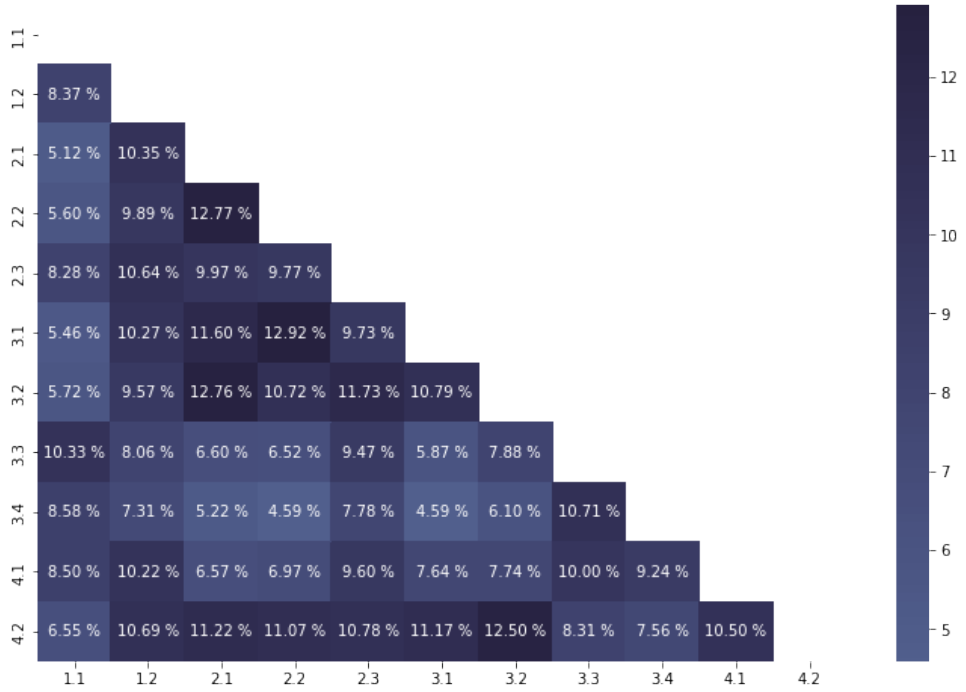


Figure 12: Jaccard Similarity Coefficient between lexicons from all sexism classes in task C from Shapley lexicons (without augmentation)