

Article

# A Transparent Pipeline for Identifying Sexism in Social Media: Combining Explainability with Model Prediction

**Hadi Mohammadi** , **Anastasia Giachanou** \*  and **Ayoub Bagheri** 

Department of Methodology and Statistics, Utrecht University, 3584 CS Utrecht, The Netherlands;

h.mohammadi@uu.nl (H.M.); a.bagheri@uu.nl (A.B.)

\* Correspondence: a.giachanou@uu.nl

**Featured Application:** We show illustrative examples of sexist language to describe the taxonomy and explainability analysis.

**Abstract:** In this study, we present a new approach that combines multiple Bidirectional Encoder Representations from Transformers (BERT) architectures with a Convolutional Neural Network (CNN) framework designed for sexism detection in text at a granular level. Our method relies on the analysis and identification of the most important terms contributing to sexist content using Shapley Additive Explanations (SHAP) values. This approach involves defining a range of Sexism Scores based on both model predictions and explainability, moving beyond binary classification to provide a deeper understanding of the sexism-detection process. Additionally, it enables us to identify specific parts of a sentence and their respective contributions to this range, which can be valuable for decision makers and future research. In conclusion, this study introduces an innovative method for enhancing the clarity of large language models (LLMs), which is particularly relevant in sensitive domains such as sexism detection. The incorporation of explainability into the model represents a significant advancement in this field. The objective of our study is to bridge the gap between advanced technology and human comprehension by providing a framework for creating AI models that are both efficient and transparent. This approach could serve as a pipeline for future studies to incorporate explainability into language models.



**Citation:** Mohammadi, H.; Giachanou, A.; Bagheri, A. A Transparent Pipeline for Identifying Sexism in Social Media: Combining Explainability with Model Prediction. *Appl. Sci.* **2024**, *14*, 0.

<https://doi.org/>

Academic Editor: Firstname Lastname

Received: 20 August 2024

Revised: 11 September 2024

Accepted: 19 September 2024

Published:



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The digital age has significantly transformed communication, especially with the emergence of online platforms facilitating real-time interactions. Although social media platforms can be a useful tool for instant communication, they also facilitate the sharing of harmful content, such as hate speech and sexist comments. According to Kurasawa et al. [1], gender-based online violence (GBOV) involves digital forms of harassment and abuse targeted at women, providing a significant challenge to maintaining safe and respectful digital environments. Content regulations on these platforms [2], within the framework of the European Union's Digital Services Act, highlight the challenge of effectively managing harmful content. The increasing rate of online sexism poses a serious threat to women's mental health [3]. This challenge is worsened by social media platforms, where sexist ideologies can spread rapidly, leading to an increase in sexist comments [4].

Recognizing and addressing sexism in online spaces is crucial for advancing gender equality, as emphasized by Lee et al. [5] in their study on the role of online forums in movement dynamics and communication. Creating a safe digital space is about more than just protecting individuals from gender-based harm; it is also about creating an environment where gender equality is possible. Identifying sexist content, on the other

hand, is a challenging task. The sensitive and context-dependent nature of language presents challenges for accurately distinguishing between sexist and non-sexist content, as shown by Feng [6] in their development of a voting mechanism for identifying online sexist content. While studies such as Schütz et al. [7] and Ortiz [3] have extensively studied the detection of online sexism in various contexts, including in non-English ones, these models often rely on a binary sexist/non-sexist classification [8–10].

To the best of our knowledge, no study has addressed the problem of sexism detection as a regression task. It is essential to address this gap for a thorough analysis that can effectively moderate content. Treating sexism detection as a regression task within a continuous range allows users and platform moderators to assess the intensity of sexism, recognizing that not all posts cause the same level of harm. This approach provides a more nuanced and detailed understanding than a binary classification, emphasizing the varying degrees of impact. Also integrating explainable AI (XAI) techniques in sexism detection as shown by Mehta and Passi [11] and Gil Bermejo et al. [12] can provide deeper insights into the rationale behind these classifications, thereby enhancing the transparency and trustworthiness of the models. This knowledge is essential for ensuring responsible and transparent content management on social media, particularly because these models often function as black boxes, making the integration of explainability crucial. The challenge exists in overcoming people's reluctance to trust machine-based decisions, which often arises from their lack of understanding of how these decisions are made. Even though these models may achieve high performance, their decisions need to be reviewed to ensure that users are not unfairly blocked and to minimize false negatives, especially given the sensitivity of gender discrimination issues. The European Union's General Data Protection Regulation (GDPR), emphasizing a "right to explanation" as highlighted by Hoofnagle et al. [13], underscores the need for clarity and transparency in these models, as argued by Mathew et al. [14]. A more detailed classification of sexist content allows policymakers to address online sexism with greater precision, reducing gender-based harm. This granular detection should be distinguished from explainability, as each serves a different but complementary purpose in managing online content.

While automated detection tools are crucial in identifying complex social issues such as sexism, recent models such as transformers function as black boxes, offering little to no insight into the reasoning behind their classifications. As highlighted in the review by Velankar et al. [15], and the analysis of online content challenges by Jiang [16], while the lack of transparency in these models does not necessarily affect their accuracy in detecting sexist content, the 'black box' nature of these models makes it difficult to understand their rationale and identify potential errors. This lack of transparency and explainability can lead to mistrust from users, which in turn might reduce their willingness to engage with or rely on the system, thereby undermining its overall effectiveness in detecting online sexism.

The need to understand the decision process of the models has led to the development of various techniques to make these models more interpretable, as highlighted in recent surveys on the state of explainable natural language processing (XNLP) tasks [17,18]. XNLP offers an exciting response to the challenges of interpretability and trust in the context of online harm detection, including sexism detection. XNLP enables users to understand and critically evaluate the outcomes of automated detection systems by providing insights into the decision-making process of AI models. This is particularly important in sensitive fields like the detection of sexism, where the importance of false positives or negatives exceeds simple errors. In this scenario, a false negative (failure to identify a sexist comment) poses the risk of overlooking harmful behavior, whereas a false positive (incorrectly labeling a statement as sexist) may wrongly implicate individuals or content. These errors carry significant ethical and social consequences, in addition to being crucial from a data accuracy perspective.

Our research proposes the application of the XNLP technique to detect and understand sexism in social media posts. By prioritizing explainability, we aim to clarify the decision-making processes of algorithms, particularly when dealing with large language models

(LLMs). Some researchers have tackled the problem of detecting sexist text using ensemble models, as shown by Mohammadi et al. [19] in their study on ensembling transformer models for identifying online sexism. Their approach combines multiple transformer models to improve detection accuracy. We also propose an ensemble method for this task; however, our method incorporates explainability techniques such as SHAP values to provide insights into the decision-making process of each model, offering a more transparent and interpretable solution compared to the previous work. The ensemble approach has advantages because it combines predictions from multiple pre-trained models, such as Bidirectional Encoder Representations from Transformers (BERT) and DistilBERT, enhancing the overall performance of the system. In summary, our study aims to contribute to the growing body of research on sexism detection by providing a novel ensemble approach that integrates technical accuracy with explainability.

The remainder of this paper is organized as follows: we start with a review of the most relevant works in Section 2. Then, we describe the data and our methodology in Section 3, which includes the data analysis, data preparation, and augmentation, and our ensemble model design; we then introduce our approach about explainability analysis in Section 3.5. Subsequently, we discuss experimentation and results in Section 4, covering the hyperparameter tuning of the model and training progress, along with several examples of how explainability impacts the performance metrics. The paper concludes with the discussion, conclusion, and future work in Section 5.

## 2. Related Work

The detection of online sexism has received much attention in recent years. Some approaches have relied on machine learning algorithms and hand-crafted features to determine whether text is sexist or not [20,21]. These methods, however, often do not capture the context that is crucial for such a complex task. Lopez-Lopez et al. [22] proposed integrating transformer-based models with traditional machine learning approaches for detecting sexism in social networks, highlighting the dynamic nature of these methodologies. The quality and diversity of datasets, whether exclusively in English or multilingual, can create challenges that may affect the applicability of the results. This issue is addressed further by Samory et al. [23], who study sexism detection using psychological factors and adversarial samples to improve dataset annotation for more reliable sexism-detection methods.

Significant improvements in detecting online sexism have been made in recent years [24,25]. This investigation has extended beyond English, as evidenced by the work of researchers Mohammadi et al. [19], Jiang et al. [26], and de Paula et al. [9], who applied LLMs to datasets in multiple languages. Despite these advancements, a common issue persists: all of these studies focus on a binary classification—sexist or not sexist—without delving into the underlying rationale behind why certain texts are perceived to contain sexist content. Das et al. [27] presented a method for addressing this limitation by combining user gender information with textual features, which improves classification performance over the typical binary categorization. Furthermore, this need led to competitions such as Explainable Detection of Online Sexism ([\(https://codalab.lisn.upsaclay.fr/competitions/7124\)](https://codalab.lisn.upsaclay.fr/competitions/7124), (accessed on)), which attempted to address this issue through supervisors in more detailed categories [28]. Tasneem et al. [29], Kiritchenko et al. [30], and Lamsiyah et al. [31] have investigated the effectiveness of transfer learning models for explainable online sexism detection, the ethical and human rights perspective in confronting online abuse, and semi-supervised multi-task learning for explainable online sexism detection, respectively.

Recent advances in natural language processing (NLP) and machine learning have provided opportunities for more effective detection of online sexism. Deep learning, context-aware algorithms, and lexicon-based sentiment analysis have been employed to enhance the discernment of the nuances in sexist language. For instance, the utilization of BERT for sentiment analysis within textual data [32], alongside the refinement of sentiment-analysis methodologies to incorporate finer details, exemplifies notable advancements within this domain [33].

Furthermore, the introduction of sentimental and context-aware recurrent convolutional neural network (CNN) highlights advancements in handling complex language structures [34]. However, these methods face challenges, particularly in dealing with ambiguous or indirect expressions of sexism. Furthermore, it is crucial to address the issue of bias in the training data utilized to train such models, as it can significantly influence their performance and reliability.

The usage of techniques such as Shapley Additive Explanations (SHAP) [35] and Local Interpretable Model-agnostic Explanations (LIME) [36] has increased the popularity of explainability in AI, especially in NLP. These methods enable more transparency and understanding of model predictions by offering insights into the machine learning models' decision-making process. To provide more detailed knowledge of how specific features affect model output, SHAP has been used, for example, to interpret complicated NLP models. Also, LIME has played a crucial role in clarifying the logic underlying specific forecasts, facilitating comprehension and confidence in AI judgments.

Studies that emphasize the importance of human–AI collaboration show that the role of human involvement in AI-driven content filtering has grown in popularity. For example, Lai et al. [37] and Molina and Sundar [38] show how human oversight can drastically minimize errors in AI moderation systems. Furthermore, the Rallabandi et al. [39] study emphasizes the significance of balancing human moderation with algorithmic action in content moderation, especially in sensitive areas such as online harm detection.

Our study combines explainability with the power of transformer-based models to advance the field of sexism detection. In contrast to the previous studies that focused on performance, our approach emphasizes comprehending the decision process in sexism detection. This enables us to not only detect sexism in various forms but also to provide clear explanations for these classifications.

### 3. Materials And Methods

#### 3.1. Data

For our study, we utilized the data by SemEval Task 10 [40], which includes labeled datasets from Gab and Reddit (<https://github.com/rewire-online/edos> (accessed on)). Gab is recognized as a social networking platform supporting free speech and harboring a diverse user base, thereby hosting content spanning a wide range. Conversely, Reddit serves as a network of communities wherein individuals engage with topics aligning with their interests, hobbies, and passions. The labeled dataset consists of 14,000 posts. The tasks include a binary classifier for categorizing posts as sexist or non-sexist (Subtask A), a four-class classification system for sexist posts (Subtask B), and an 11-class system for more specific labels of sexism (Subtask C). These subtasks ensure that texts labeled as sexist are given specific reasons for the classification. The overview of tasks and datasets is shown in Table 1.

**Table 1.** Data summary.

Category		Records
	<b>Task A</b>	
Not sexist		10,602
Sexist		3398
<b>Total</b>		<b>14,000</b>
	<b>Task B</b>	
1. Threats, plans to harm, and incitement		310
2. Derogation		1590
3. Animosity		1165
4. Prejudiced discussion		333
<b>Total</b>		<b>3398</b>

**Table 1.** Cont.

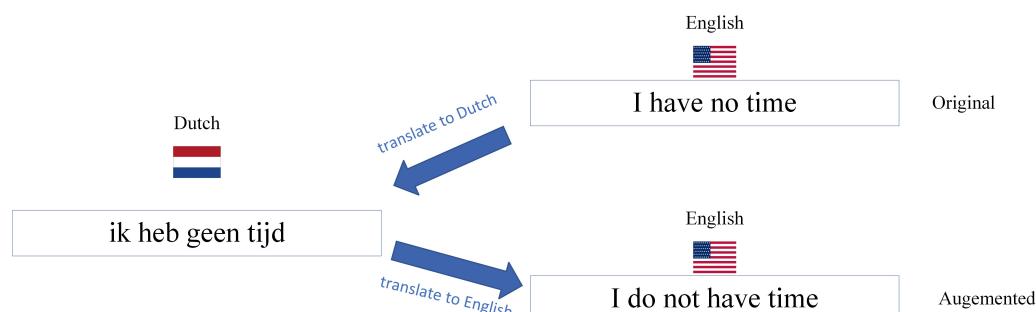
Category	Records
Task C	
1.1 Threats of harm	56
1.2 Incitement and encouragement of harm	254
2.1 Descriptive attacks	717
2.2 Aggressive and emotive attacks	673
2.3 Dehumanising attacks and overt sexual objectification	200
3.1 Casual use of gendered slurs, profanities, and insults	637
3.2 Immutable gender differences and gender stereotypes	417
3.3 Backhanded gendered compliments	64
3.4 Condescending explanations or unwelcome advice	47
4.1 Supporting mistreatment of individual women	75
4.2 Supporting systemic discrimination against women as a group	258
<b>Total</b>	<b>3398</b>

### 3.2. Data Preparation and Augmentation

The first step in the data-preparation process was to convert all of the text to lowercase so that the analysis could recognize words uniformly. Additionally we removed URLs, special characters and punctuation. Then, we applied tokenization and lemmatization to divide the text into units.

We applied various augmentation strategies due to limited data availability in certain classes and the necessity to develop more robust models. Using techniques like ‘random deletion’ to remove random words trained the model to understand incomplete data, and techniques such as ‘synonym replacement’ for word and phrase replacement expanded the model’s understanding of context. Different spellings were also used to simulate errors found in the real world. The dataset was significantly imbalanced, particularly in Task A, where 76% of the training data was ‘Not sexist’ (10,602 instances) and only 24% was ‘Sexist’ (3398 instances). This imbalance was similarly reflected in the test data and persisted across Task B and Task C categories. For example, some classes in Task C, like ‘Threats of harm’, had only 56 training instances, representing a mere 2% of the data. The training dataset contained 14,000 records. After augmentation, this number increased to 42,000 records, significantly enhancing the data volume available for training.

To further address the class imbalance, we used back translation to increase the number of ‘sexist’ texts. This means we first translated the ‘sexist’ texts into Dutch and then back into English. Dutch was selected for this purpose based on previous research indicating its efficacy for such tasks and due to its linguistic similarity to English, as both languages belong to the West Germanic language family [41]. Figure 1 illustrates an example of our back translation augmentation approach.

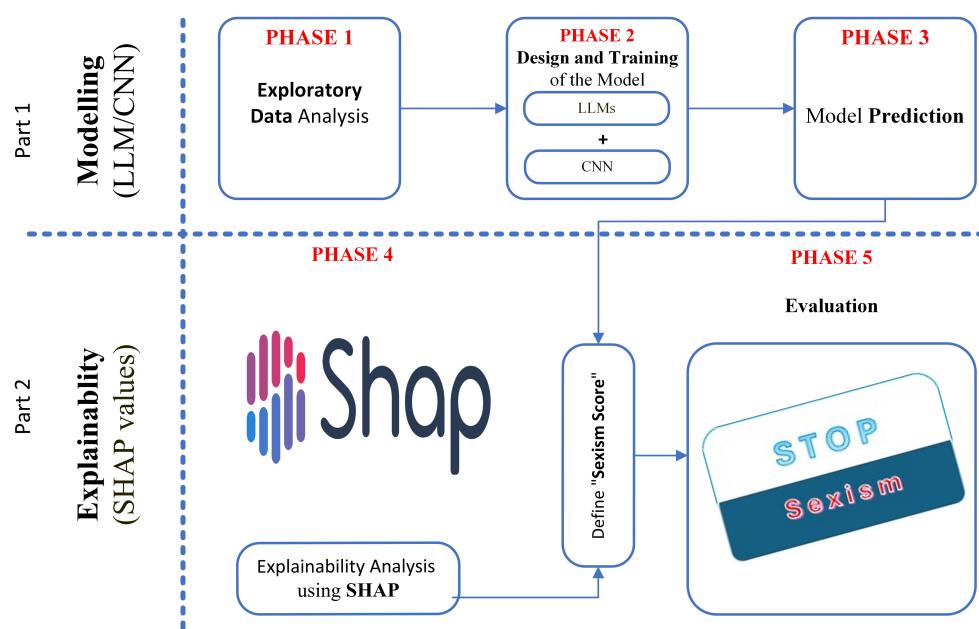


**Figure 1.** Back translation method for data augmentation (English ↔ Dutch).

To further address the class imbalance, we used some other techniques like stratified K-fold cross-validation ([https://scikit-learn.org/stable/modules/generated/sklearn.model\\_selection.StratifiedKFold.html](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html), (accessed on)) method to ensure an accurate class distribution throughout the data segments. Strategies including RandomOverSampler, SMOTE ([https://imbalanced-learn.org/stable/references/generated/imblearn.over\\_sampling.SMOTE.html](https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html), (accessed on)), and differential class weighting were used during training to address the class imbalance [42].

### 3.3. Methodology

In this study, we propose a novel methodology that uses explainability to define a Sexism Score, which can lead to more specific results and identification. This approach is structured into two parts, where the various phases are shown in Figure 2.



**Figure 2.** Research methodology.

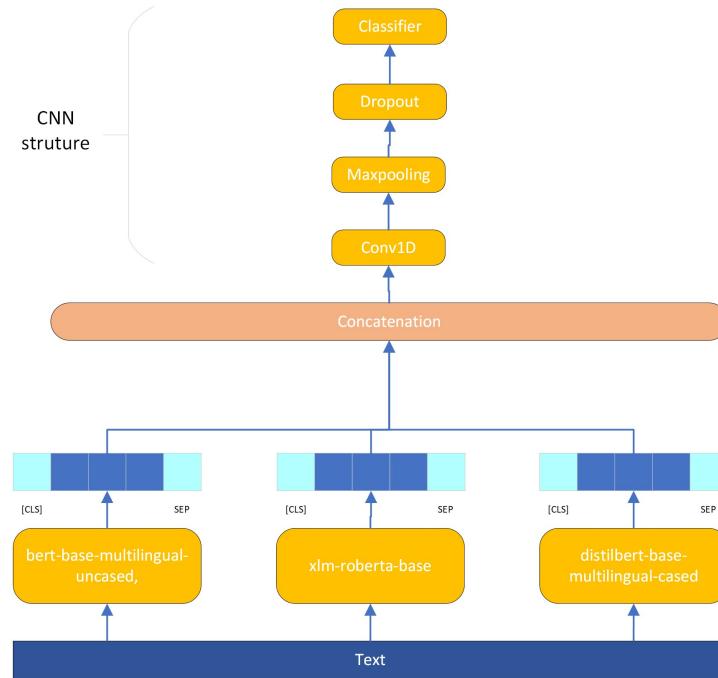
As illustrated in Figure 2, the first part focuses on the sexism detection and starts with dataset analysis. This phase is followed by preprocessing and data augmentation. Then we develop an ensemble model by combining various versions of BERT with a CNN architecture (detailed in Section 3.4).

In the second part of our methodology, we generate SHAP values to identify the most influential tokens in the model's decision-making process [35] and we integrate the explainability scores into model prediction to define the Sexism Score, which we elaborate more on in Section 3.5.

### 3.4. **CustomBERT**—Ensemble Model Design

We developed the CustomBERT model, which combines different BERT versions and a CNN, inspired by the way CNNs recognize similarities in images, as investigated by Xu and Vaziri-Pashkam [43]. This approach is motivated by the need to capture diverse linguistic patterns from different transformer models. Each BERT variant offers distinct strengths: BERT multilingual excels in handling various aspect of language, XLM-RoBERTa is particularly adept at cross tasks, and DistilBERT provides efficiency with minimal loss in performance. By using these models, our ensemble aims to capture a wider range of semantic features. This method identifies similarities between various pre-trained transformer models using text input, similar to how CNNs identify similarities in images. The combination of transformer models and CNN introduces additional layers of representation

learning, allowing the system to extract more nuanced textual features that a single model might miss. Figure 3 shows the detailed architecture of our approach.



**Figure 3.** Architecture of our CustomBERT model.

By using the advantages of various transformer models, this structure aims to provide an improved and flexible method to evaluate new text. We used BERT's bidirectional nature to fully understand the context of words in sentences [44]. After the pre-trained model processes the input, its output undergoes two subsequent processes. First, we perform concatenation by merging the outputs from the pre-trained model. Then, we pass this concatenated input through a CNN layer, which is designed to identify important features from both the text data and the token scores. CNN is chosen here to effectively capture local dependencies in the concatenated transformer outputs, enabling the model to recognize intricate relationships within the text that are critical for tasks like sexism detection. The pseudocode of Algorithm 1 is as follows:

---

**Algorithm 1** CustomBERT model for sexism detection.

---

**Require:** Input sentence  
**Ensure:** Output classes

```

1: function PRE-TRAIN LANGUAGE MODELS(sentence)
2:   bert_output  $\leftarrow$  BERT Model(sentence)
3:   xlmroberta_output  $\leftarrow$  XLMRoBERTa Model(sentence)
4:   distilbert_output  $\leftarrow$  DistilBERT Model(sentence)
5:   return bert_output, xlmroberta_output, distilbert_output
6: end function
7: function CUSTOMBERT(tokens)
8:   bert_output, xlmroberta_output, distilbert_output  $\leftarrow$ 
9:     PRE-TRAIN LANGUAGE MODELS OUTPUTS(sentence)
10:   concatenated_outputs  $\leftarrow$  concatenate(bert_output, xlmroberta_output, distilbert_output)
11:   cnn_output  $\leftarrow$  apply Conv1D layer, filters=64, kernel=3, activation='relu'(concatenated_outputs)
12:   flattened_cnn_output  $\leftarrow$  flatten(cnn_output)
13:   output  $\leftarrow$  apply Dense layer with 1 unit, activation='sigmoid'(flattened_cnn_output)
14:   return class
15: end function
16: final output  $\leftarrow$  CUSTOMBERT(input sentence)
  
```

---

As shown in Algorithm 1, the CustomBERT model combines the outputs of various pre-trained transformer models with token scores before classification. It consists of *BERT multilingual* [44], *XLM-RoBERTa* [45], and *DistilBERT*, a smaller but efficient version of BERT [46]. The main task for these transformer models is to encode the input sentence into a dense vector representation, capturing contextualized meaning and semantic nuances. We combine the outputs from each transformer model by concatenating their vector representations, which are then processed further through a convolutional neural network (CNN) and process them through a Conv1D layer ([https://keras.io/api/layers/convolution\\_layers/convolution1d/](https://keras.io/api/layers/convolution_layers/convolution1d/), (accessed on)), followed by MaxPooling1D ([https://www.tensorflow.org/api\\_docs/python/tf/keras/layers/MaxPooling1D](https://www.tensorflow.org/api_docs/python/tf/keras/layers/MaxPooling1D), (accessed on)) and a flattening step. This preprocessing stage primes the data for subsequent dense layers, using binary and multiclass classification across diverse tasks.

### 3.5. Explainability Analysis

In this section, we introduce the symbols and parameters used throughout our explainability methodology, followed by a detailed description of the techniques applied.

**Table 2.** Definitions of symbols and parameters.

Symbol	Description
$t$	Token in the sentences
$S_t$	SHAP value for token $t$
$T$	Set of all tokens
$f(t)$	Model prediction without the token $t$
$SI_t$	SHAP Importance for token $t$
$N_t$	Number of sentences in which token $t$ appears
$S_t(i)$	SHAP value for token $t$ in the $i$ -th sentence
$\mu$	Mean of the SHAP scores
$\sigma$	Standard deviation of the SHAP scores
$IR_t$	Importance Ratio for token $t$
$CI_k$	Cumulative Importance up to the $k$ -th token
$T_c$	Threshold for cumulative importance, set to 0.95
$Label(i)$	Modified predicted hard label indicator for sentence $i$
$SS(i)$	Sexsim Score for sentence $i$

We utilized SHAP to integrate the explainability component into our methodology. These values are instrumental in understanding the contribution of each token to the model's predictions. SHAP generates scores indicative of the importance of a token in the prediction [35]. SHAP values are based on cooperative game theory and provide a method to attribute the output of the model to its input features. The application of SHAP values allows us to identify the key factors influencing the classification process. By analyzing these values, we can assign scores to the most influential tokens based on their perceived impact, leading to a better understanding of the textual content.

The SHAP value for each token is calculated as follows: let  $S_t$  denote the SHAP value for token  $t$ ,  $T$  be the set of all tokens,  $T \setminus \{t\}$  be a subset of tokens excluding  $t$ , and  $f(T \setminus \{t\} \cup \{t\})$  and  $f(T \setminus \{t\})$  be the model predictions with and without the token  $t$ , respectively. The SHAP value  $S_t$  is computed as:

$$S_t = \sum_{T \setminus \{t\} \subseteq T} \frac{|T \setminus \{t\}|! (|T| - |T \setminus \{t\}| - 1)!}{|T|!} [f(T \setminus \{t\} \cup \{t\}) - f(T \setminus \{t\})] \quad (1)$$

Here, the term  $\frac{|T \setminus \{t\}|! (|T| - |T \setminus \{t\}| - 1)!}{|T|!}$  represents the weight assigned to the difference in model outputs, ensuring that the contribution of token  $t$  is fairly distributed among

all possible combinations of tokens. This weight is derived from the concept of Shapley values in cooperative game theory, which ensures a fair distribution of the total gain (or loss) among all contributors.

To determine the most influential tokens based on SHAP values, we first calculate the SHAP importance for each token  $t$  across all samples where the token appears in the set of all tokens  $T$  in the sentences from  $i = 1$  to  $N_t$ , which represent all sentences in the dataset.  $S_i(t)$  represents the SHAP value for token  $t$  in the  $i$ -th sentence. The SHAP importance for token  $t$ , denoted as  $SI_t$ , is computed as:

$$SI_t = \frac{1}{N_t} \sum_{i=1}^{N_t} |S_i(t)| \cdot \mathbb{I}(y_i = \hat{y}_i) \quad (2)$$

In this formula,  $N_t$  is the number of sentences in which token  $t$  appears, and  $S_i(t)$  is the SHAP value for token  $t$  in the  $i$ -th sentence. The term  $\mathbb{I}(y_i = \hat{y}_i)$  is an indicator function that equals 1 if the predicted class  $\hat{y}_i$  matches the true class  $y_i$ , and 0 otherwise. This ensures that we only consider SHAP values from sentences where the model's prediction is correct, thereby focusing on the tokens that truly influence accurate predictions.

We remove data points that fall outside 99.7% of the SHAP value distribution to exclude outliers. This step ensures that our analysis focuses on the most representative data. Specifically, we filter the SHAP scores  $s$  to satisfy the following condition: where  $-\mu$  is the mean of the SHAP scores,  $-\sigma$  is the standard deviation of the SHAP scores.

$$\mu - 3\sigma \leq s \leq \mu + 3\sigma \quad (3)$$

After filtering outliers, we normalize the remaining SHAP values to derive the importance ratio for each token. The importance ratio for token  $t$  is defined as:

$$IR_t = \frac{SI_t}{\sum_{k \in T} SI_k} \quad (4)$$

This step converts the SHAP values into a proportional format, where each ratio represents the token's share of the total impact on the model. Following the normalization, we calculate the cumulative importance of the tokens. Let  $K$  be the total number of tokens sorted by descending importance, and  $IR_i$  be the importance ratio of the  $i$ -th token. The cumulative importance is given by:

$$CI_k = \sum_{i=1}^k IR_i \quad \text{such that} \quad CI_k \leq T_c \quad (5)$$

We establish a threshold,  $T_c = 0.95$ , to select tokens based on their importance. Tokens are selected such that their cumulative importance is less than or equal to the threshold  $T_c$ , focusing on those that contribute to the first 95% of total SHAP importance.

The selected tokens are then employed to calculate a 'Sexism Score' for each text entry in our dataset. This score combines the SHAP scores with a modified predicted hard label. The hard Label(i) is changed to -1 for 'No' labels and 1 for 'Yes' labels.

$$\text{Label}(i) = \begin{cases} 1 & \text{if prediction = YES} \\ -1 & \text{if prediction = NO} \end{cases} \quad (6)$$

Let  $SS(i)$  be the Sexism Score for sentence  $i$  derived from the summed absolute SHAP values of the selected tokens, and  $\text{Label}(i)$  be the modified label indicator. The Sexism Score for sentence  $i$  is calculated as follows:

$$SS(i) = \sum_{t \in N_i} SI_t(i) \times \text{Label}(i) \quad (7)$$

After calculating the Sexism Scores for the training dataset, we proceed to evaluate the test dataset. For each sentence in the test dataset, we utilize the trained model to generate predictions. Using the SHAP scores computed during the training phase, we calculate the Sexism Score for each sentence in the test dataset. This is achieved by summing the SHAP values of the selected effective tokens based on Table A1.

Subsequently, we categorize the test dataset into different bins based on the calculated Sexism Scores. This binning allows us to compare the model's performance across varying levels of detected sexism. By analyzing the model's predictive accuracy and other performance metrics within these bins, we can know how well the model generalizes to new data, particularly in identifying and handling sexist content. The overall steps of test dataset evaluation and performance comparison are shown below:

1. **Model Prediction on Test Dataset:** For each sentence  $i$  in the test dataset, use the trained model to generate predictions  $\hat{y}_i^{\text{test}}$ .
2. **Sexism Score Calculation for Test Dataset:** Calculate the Sexism Score  $SS^{\text{test}}(i)$  for each sentence  $i$  in the test dataset using the SHAP scores from the training dataset. This is achieved as follows:

$$SS^{\text{test}}(i) = \sum_{t \in T_i} SI_t \times \text{Label}(\hat{y}_i^{\text{test}}) \quad (8)$$

where  $SI_t$  are the SHAP importance scores from the training dataset, and  $\text{Label}(\hat{y}_i^{\text{test}})$  is the predicted label for the test sentence  $i$ .

3. **Binning Based on Sexism Scores:** Divide the test dataset into bins based on the calculated Sexism Scores  $SS^{\text{test}}(i)$ . Define bins  $B_k$  such that:

$$B_k = \{i \mid \alpha_{k-1} \leq SS^{\text{test}}(i) < \alpha_k\} \quad (9)$$

where  $\alpha_k$  are the thresholds for the bins.

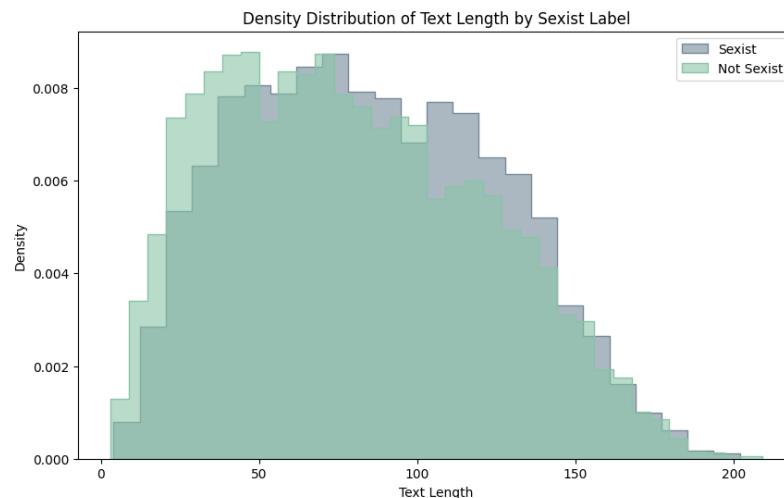
4. **Performance Comparison:** For each bin  $B_k$ , evaluate the model's performance by calculating metrics such as accuracy, precision, recall, and F1 score. Compare these metrics across different bins to assess the model's performance in handling varying levels of detected sexism.

By following these steps, we can understand the model's effectiveness and robustness in identifying and handling sexist content in the test dataset, highlighting any potential biases or areas for improvement.

## 4. Results

### 4.1. Exploratory Data Analysis (EDA)

In this part, we analyze the data by examining the text length distribution, frequency of unique words and n-grams across different categories of sexist texts, and sentiment analysis of common word pairs within each category. We aim to uncover linguistic patterns and characteristics that distinguish sexist texts from non-sexist ones. First, we looked at the relationship between text length and the presence of sexist content. We calculated the text length based on the number of characters in each text, after removing stop words, punctuation, etc. We divided the text length into five ranges, [2–50, 51–100, 101–150, 151–200, 201+], to simplify the analysis. A density distribution of text lengths within each label category (sexist or non-sexist) is shown below.



**Figure 4.** Density distribution of text lengths by sexist label.

We used several statistical methods, including logistic regression, chi-square tests, and *t*-tests, to analyze how text length influences sexist labeling. Table 3 show the results from the logistic regression, and Table 4 show the results from the 1 chi-square tests, and *t*-tests show that longer texts are more likely to be labelled as sexist; however, the logistic regression results indicate a low R-squared value, near 0.004, suggesting that text length alone is not a strong predictor. More details are shown in Tables 3 and 4:

**Table 3.** Logistic regression results.

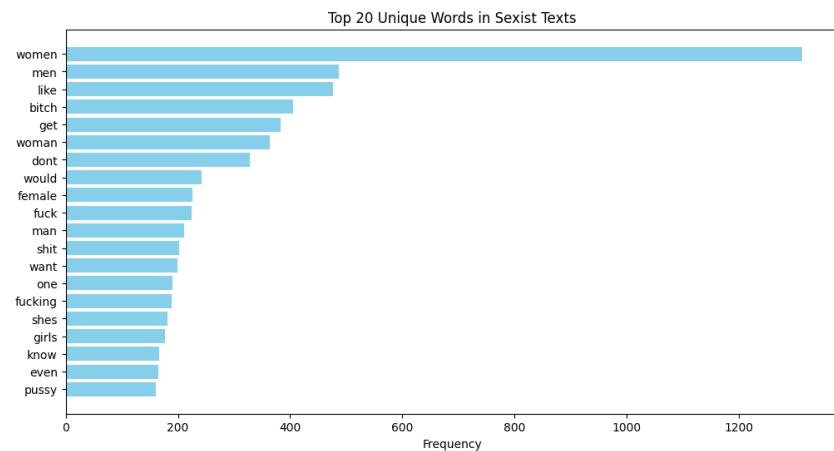
Parameter	Coefficient	Standard Error	z-Value	p-Value	95% Confidence Interval
Intercept	-1.431497	0.044	-32.262	$2.37 \times 10^{-228}$	[−1.518, −1.345]
Text Length	0.003579	0.000	7.532	$4.98 \times 10^{-14}$	[0.003, 0.005]
Log-Likelihood:	−7730.043				
R-Squared:	0.003650				

**Table 4.** Chi-Square and T-test results.

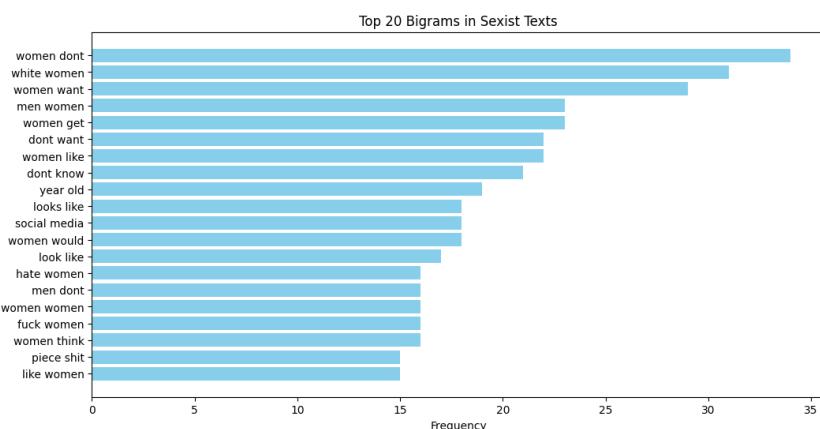
Statistic	Value
Chi2 Statistic	65.703
p-Value	$1.83 \times 10^{-13}$
Degrees of Freedom	4
T-Statistic	7.562
p-Value	$4.21 \times 10^{-14}$
Mean Text Length (Sexist)	64.83
Mean Text Length (Non-Sexist)	58.29
Standard Deviation (Sexist)	51.41
Standard Deviation (Non-Sexist)	50.77

Based on the results above, with a *p*-value of  $4.98 \times 10^{-14}$ ,  $1.83 \times 10^{-13}$ , and  $4.21 \times 10^{-14}$  for logistic regression, chi-square test, and *t*-test, respectively, it is evident that text length is a significant factor in determining whether a text is labeled as sexist. All three tests consistently indicate a strong and statistically significant relationship between text length and sexist labeling.

To further analyze the linguistic features of sexist content, we intend to understand the typical language used in it. We implemented n-gram analysis to identify common bigrams and unique words in sexist texts. Figures 5 and 6 show the top 20 unique words and bigrams in sexist texts.



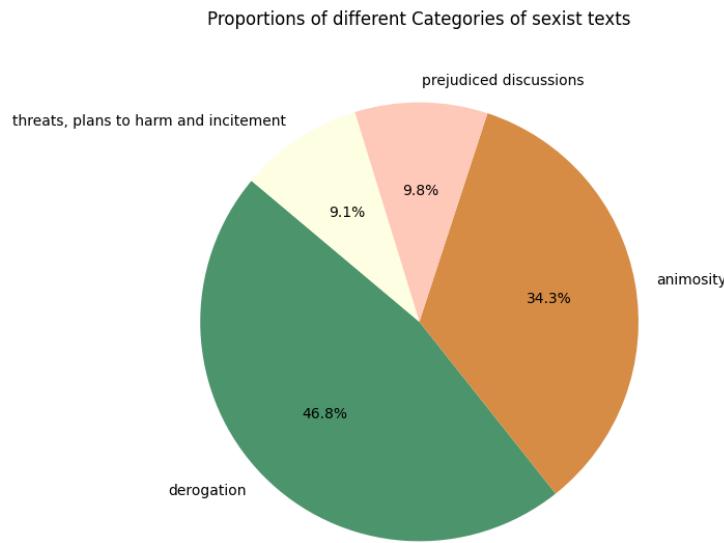
**Figure 5.** Top 20 unique words in sexist texts.



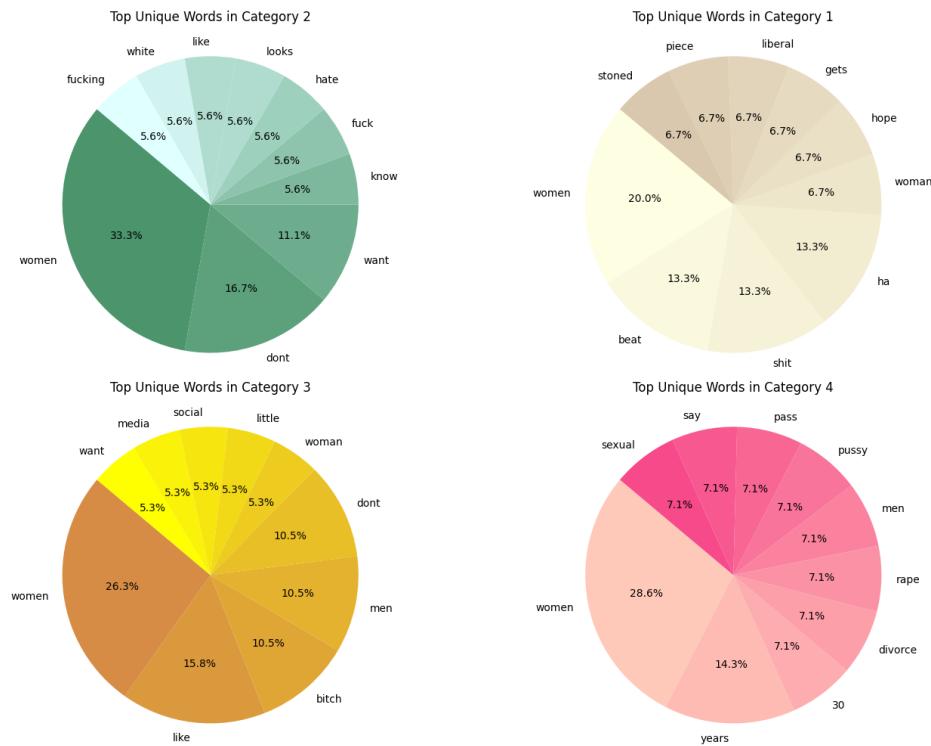
**Figure 6.** Top 20 bigrams in sexist texts.

As we can see in Figures 5 and 6, the word 'women' was the most frequently occurring word in sexist texts, highlighting the targeted nature of these texts. We also notice that some words indicate aggressive and derogatory language. The sexist text is categorized into 4 main categories including 'threats, plans to harm, and incitement', 'derogation', 'animosity', and 'prejudiced discussions' by Task B. Figure 7 demonstrates the proportions of different types of sexist content.

As we can see, 'derogation' was the most prevalent, comprising 46.8% of the sexist texts. In Figure 8, we identify the top unique words for each category of sexist content to understand common words associated with each category. In category 1, common words included 'women', 'beat', and 'shit', reflecting violent and harmful intentions. Frequent words were 'women', 'don't', and 'want', indicating derogatory and dismissive language associated with category 2. In category 3, words like 'women', 'like', and 'bitch' were prevalent, showing animosity. Top words included 'women', 'years', and 'rape/sexual/divorce', indicating prejudiced discussions often revolving around stereotypes and harmful myths for category 4.

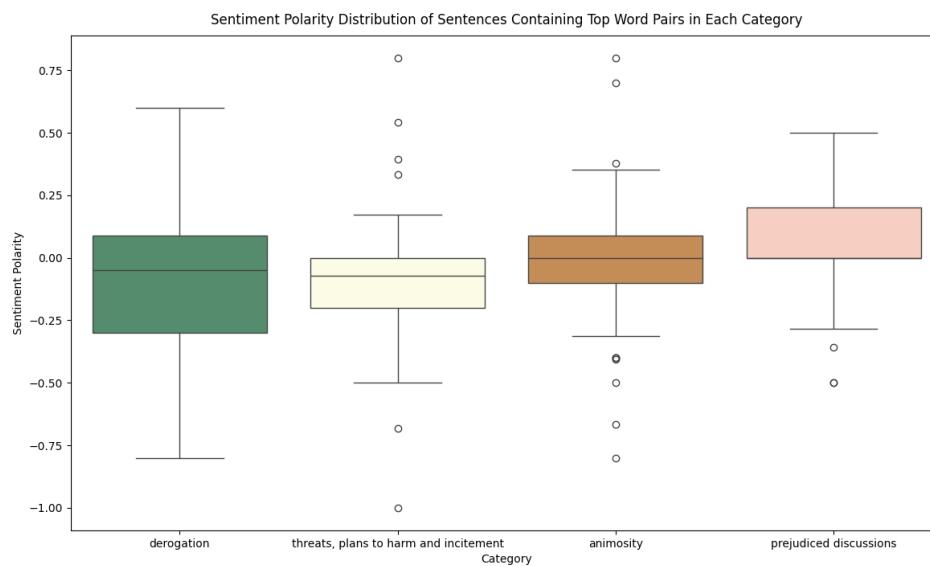


**Figure 7.** Proportions of different categories of sexist texts.



**Figure 8.** Top unique words for each category of sexist content.

Here, we analyze the sentiment polarity of sentences containing top word pairs in each category of sexist content to understand the emotional impact of sexist content and how it varies across different types of sexism. Sentiment polarity refers to the classification of a sentence as positive, negative, or neutral, which allows us to assess the emotional tone associated with the content [47]. Figure 9 shows the distribution of sentiment scores within each category.



**Figure 9.** Sentiment polarity scores.

Based on the sentiment polarity distribution shown in the box plot, the analysis of the minimum and maximum scores within the interquartile range (IQR) reveals significant insights. For the first category, the IQR ranges from  $-0.2$  to  $0.0$ , indicating that most comments are moderately negative to neutral, reflecting the mixed nature of this category. In the derogation category, the IQR spans from  $-0.3$  to  $0.09$ , showing that most comments are generally negative to slightly positive, suggesting derogatory remarks often contain a blend of sentiments. Animosity has an IQR from  $-0.1$  to  $0.089$ , indicating that while extreme sentiments exist, the majority of comments are slightly negative to neutral. Prejudiced discussions have the narrowest IQR from  $0.0$  to  $0.202$ , suggesting that sentiments within this category are more consistently neutral to positive. These IQRs reveal that while extreme sentiments exist, the bulk of comments tend to be less extreme and range from moderately negative to neutral, with derogation and threats categories showing the most variability in negative sentiments.

#### 4.2. Model Training and Optimization

The experimental steps of this study followed a progressive approach, initially based on the methodology introduced by Mohammadi et al. [48]. We further tested this approach using two datasets: EXIST 2023 (<http://nlp.uned.es/exist2023/>, (accessed on)) and EDOS (<https://github.com/rewire-online/edos>, (accessed on)). While EDOS mainly focuses on content from Gab and Reddit, EXIST is based on Twitter, a more mainstream and diverse platform. This combination of datasets allows us to test the model's ability to generalize across both niche and widely used social media environments. The results of the model on the EXIST dataset can be found in [48]. However, although the first task in both datasets is similar (sexism detection), we did not compare the results due to the completely different dataset compositions and annotation processes. Initially, we started with more traditional models as our baseline. Over time, we improved and tested these models with different variations and techniques until we developed the final structure of our model. Furthermore, upon publication of this article, the comprehensive final model, accompanied by all requisite materials, will be accessible on GitHub (<https://github.com/hadimh93/Explainable-Sexim-Detection>, (accessed on)).

During the training phase, we used the Adam optimizer [49]. We selected the model's hyperparameters, such as the learning rate and batch size, using a random search approach with Keras Tuner (<https://blog.tensorflow.org/2020/01/hyperparameter-tuning-with-keras-tuner.html>, (accessed on)). The learning rate was set at  $3 \times 10^{-5}$ , enabled by a TensorFlow-based learning rate scheduler ([https://www.tensorflow.org/api\\_docs/](https://www.tensorflow.org/api_docs/)

[python/tf/keras/callbacks/LearningRateScheduler](https://tf.keras.callbacks.LearningRateScheduler), (accessed on)), including a 200-step warm-up phase. To enhance efficiency and learning, we chose an early stopping mechanism ([https://www.tensorflow.org/api\\_docs/python/tf/keras/callbacks/EarlyStopping](https://www.tensorflow.org/api_docs/python/tf/keras/callbacks/EarlyStopping), (accessed on)) and mixed precision training ([https://www.tensorflow.org/guide/mixed\\_precision](https://www.tensorflow.org/guide/mixed_precision), (accessed on)). Our preprocessing step considered a tokenization limit of 512. Subsequently, we selected important hyperparameters, such as the learning rate and batch size, using a random search approach with Keras Tuner. The learning rate followed a cosine decay schedule ([https://keras.io/api/optimizers/learning\\_rate\\_schedules/cosine\\_decay/](https://keras.io/api/optimizers/learning_rate_schedules/cosine_decay/), (accessed on)), completed by a warm-up period, which was calibrated to the dataset and the number of epochs. We used early stopping based on validation losses to prevent overfitting [50].

The binary cross-entropy loss function ([https://www.tensorflow.org/api\\_docs/python/tf/keras/losses/BinaryCrossentropy](https://www.tensorflow.org/api_docs/python/tf/keras/losses/BinaryCrossentropy), (accessed on)) and the Adam optimizer (<https://keras.io/api/optimizers/adam/>, (accessed on)) were used for training, employing the 'mixed float16' precision training policy ([https://keras.io/api/mixed\\_precision/](https://keras.io/api/mixed_precision/), (accessed on)). A custom function was developed for model structure, employing various transformers like *bert-base-multilingual-uncased*, *xlm-roberta-base*, and *distilbert-base-multilingual-cased*. These transformers' outputs were amalgamated for binary classification with L2 regularization ([https://www.tensorflow.org/api\\_docs/python/tf/keras/regularizers/L2](https://www.tensorflow.org/api_docs/python/tf/keras/regularizers/L2), (accessed on)). A comprehensive summary of the hyperparameters we examined is presented in Table 5.

**Table 5.** Summary of model parameters and hyperparameters.

Parameter	Description
Tokenization Max Length	512 tokens
Learning Rate Range	$1 \times 10^{-5}$ to $1 \times 10^{-4}$ (Default: $3 \times 10^{-5}$ )
Batch Sizes	32, 64, 128
Learning Rate Scheduler	Cosine decay schedule
Warm up Steps	200 steps
Early Stopping Patience	5 epochs
Loss Function	Binary cross-entropy
Optimizer	Adam
Precision Training Policy	Mixed float16

In evaluating the performance of various models, we considered several key metrics, including accuracy, precision, recall, and F1 score across multiple tasks. Table 6 summarizes the results of our experiments, comparing the performance of different models on Tasks A (binary classification of posts as sexist or non-sexist), B (four-class classification of sexist posts), and C (11-class system for more specific labels of sexism).

According to Table 6, the CustomBERT model consistently outperforms the individual models across all tasks. For Task A, the CustomBERT model achieves the highest accuracy of 0.79, precision of 0.77, recall of 0.79, and F1 score of 0.76. Similarly, for Task B, the CustomBERT model leads with an accuracy of 0.71, precision of 0.69, recall of 0.71, and F1 score of 0.68. For Task C, the CustomBERT model again surpasses the others, showing superior performance with an accuracy of 0.67, precision of 0.65, recall of 0.67, and F1 score of 0.64.

The consistent performance improvements observed with the CustomBERT model highlight the benefits of an ensemble approach, effectively combining the strengths of various models to achieve better overall results. This ensemble strategy, therefore, was selected as the final model for our application due to its robust performance across multiple evaluation metrics and tasks.

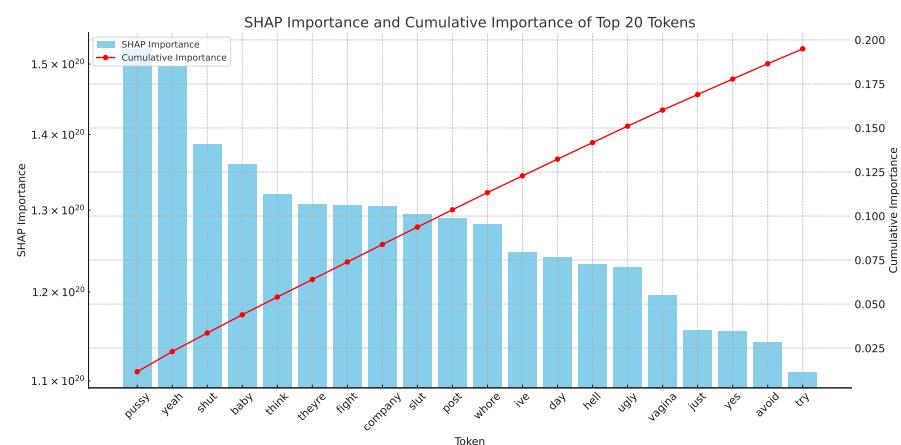
**Table 6.** Performance results of CustomBERT and baselines on Tasks A, B, and C.

Model	Accuracy	Precision	Recall	F1 Score
<b>TaskA</b>				
Logistic Regression	0.70	0.42	0.70	0.46
XGBOOST	0.72	0.45	0.72	0.49
BERT	0.76	0.57	0.76	0.65
XLM-RoBERTa	0.76	0.57	0.76	0.65
DistilBERT	0.77	0.74	0.77	0.72
<b>CustomBERT</b>	<b>0.79</b>	<b>0.77</b>	<b>0.79</b>	<b>0.76</b>
<b>TaskB</b>				
Logistic Regression	0.54	0.25	0.54	0.23
XGBOOST	0.56	0.27	0.56	0.21
BERT	0.68	0.52	0.68	0.59
XLM-RoBERTa	0.67	0.51	0.67	0.58
DistilBERT	0.69	0.66	0.69	0.65
<b>CustomBERT</b>	<b>0.71</b>	<b>0.69</b>	<b>0.71</b>	<b>0.68</b>
<b>TaskC</b>				
Logistic Regression	0.40	0.10	0.40	0.07
XGBOOST	0.42	0.12	0.42	0.08
BERT	0.63	0.44	0.63	0.52
XLM-RoBERTa	0.64	0.46	0.64	0.53
DistilBERT	0.65	0.62	0.65	0.60
<b>CustomBERT</b>	<b>0.67</b>	<b>0.65</b>	<b>0.67</b>	<b>0.64</b>

#### 4.3. Explainability Results

In this section, we explore the explainability of our model by analyzing the SHAP values to understand the contribution of individual tokens to the model's predictions. Explainability is crucial for ensuring that our model's decisions are transparent and interpretable, particularly in sensitive applications such as detecting sexist content.

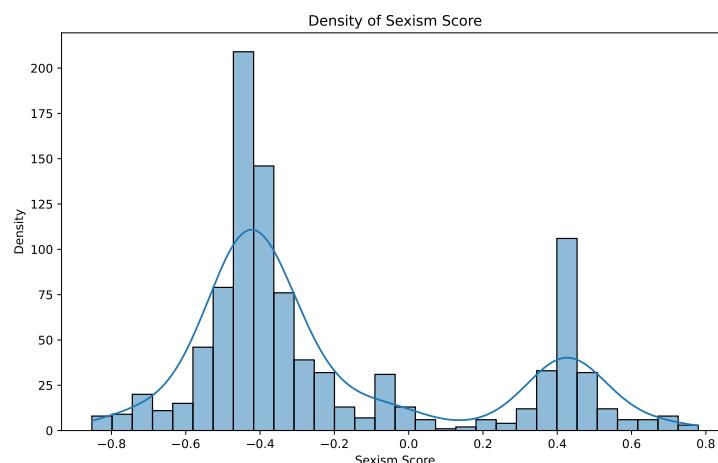
Figure 10 illustrates the SHAP importance and cumulative importance of the top 20 tokens. These tokens significantly contribute to the model's prediction outcomes, as described by Equations (1)–(5). Notably, the most effective tokens predominantly consist of offensive language directed towards women. Additionally, we computed the Sexism Score for texts labeled as 'YES' (indicative of sexism) using Equation (7). To provide a more comprehensive understanding, we present the statistical summary of the Sexism Scores for these texts in Table 7.

**Figure 10.** Cumulative importance of top 20 tokens.

**Table 7.** Statistical summary of SHAP scores.

Statistic	Value
Count	14,000
Mean	-0.271917
Standard Deviation	0.458922
Minimum	-0.963626
25th Percentile	-0.526098
50th Percentile (Median)	-0.483276
75th Percentile	-0.297328
Maximum	0.962766

This statistical summary provides insights into the distribution of Sexism Scores among texts labeled as sexist. The mean score indicates a moderate level of sexism on average, as evidenced by the standard deviation. The median and percentile values offer additional details about the central tendency and variability of the scores. To take a closer look at the distribution of Sexism Scores, we created a histogram, as shown in Figure 11.

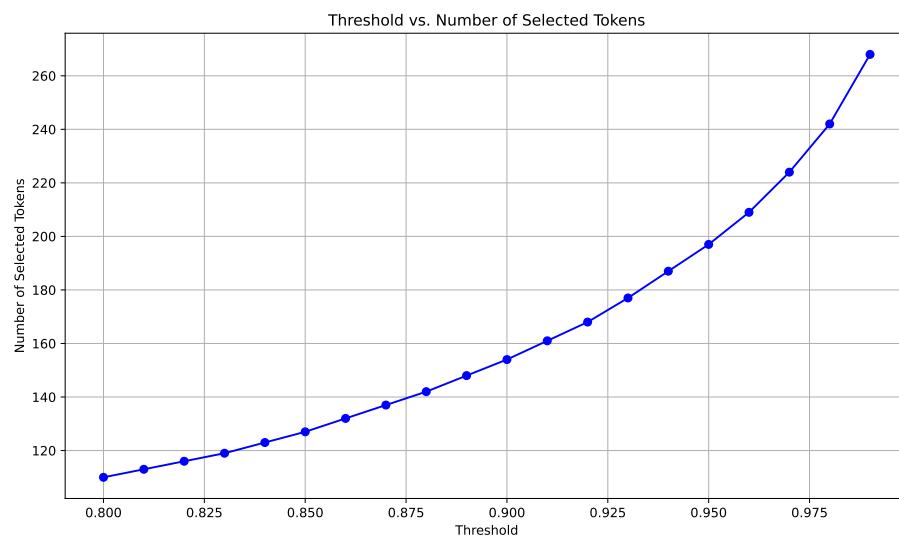
**Figure 11.** Distribution of Sexism Scores for texts labeled as sexist.

The histogram in Figure 11 depicts the density of Sexism Scores across the dataset. The x-axis represents the Sexism Scores, while the y-axis indicates the density of texts with those scores. Notably, the distribution features two prominent peaks, suggesting a bimodal distribution of Sexism Scores.

The first peak, centered around  $-0.4$ , corresponds to texts that have been labeled as 'NO' but contain tokens with relatively high SHAP importance in negative contexts, resulting in negative Sexism Scores. The second peak, centered around  $0.4$ , represents texts labeled as 'YES' with high SHAP importance in positive contexts. Additionally, the histogram shows a notable gap around the zero mark, indicating a clear separation between texts identified as sexist and non-sexist by the model.

In this study, we opted to set the threshold at 0.95, indicating that we considered tokens contributing to 95% of the cumulative importance, based on Formula (4). Consequently, the number of selected tokens amounted to 197. With this configuration, we identified the most influential tokens and their contributions to the model's decision-making process regarding sexist content. A comprehensive list of all effective tokens can be found in Appendix B.

Figure 12 shows the relationship between the threshold and the number of selected tokens. As the threshold increases, the number of tokens contributing to the cumulative importance also increases, with a marked rise observed near the higher thresholds. At a threshold of 0.95, we capture the most significant tokens influencing the model's output.



**Figure 12.** Threshold vs. number of selected tokens.

Next, we will validate the effectiveness of the selected tokens and the threshold choice in accurately identifying sexist content. By ensuring that the selected tokens are both influential and relevant to the model's predictions, we enhance the model's interpretability and reliability.

To evaluate the impact of Sexism Scores on model performance, we segmented the data into bins based on Sexism Score ranges and calculated the performance metrics for each bin. The results for the test dataset are summarized in Table 8.

As depicted in Table 8, there is a positive correlation between the Sexism Score and the performance metrics. Lower and higher Sexism Scores generally correspond to better model performance, indicating increased confidence and accuracy in identifying strongly sexist content. This trend suggests that the model is more reliable in detecting content with higher Sexism Scores, which aligns with its design to emphasize the most impactful tokens.

**Table 8.** Performance metrics for the test dataset in each bin across Tasks A, B, and C.

Bin	Accuracy	Precision	Recall	F1 Score
(−0.332, −0.0644] $\cup$ (0.734, 0.984]	0.79	0.78	0.79	0.78
(−0.0644, 0.202] $\cup$ (0.202, 0.468]	0.78	0.76	0.78	0.77
(0.468, 0.734]	0.77	0.75	0.77	0.76
<b>All data (Task A)</b>	<b>0.79</b>	<b>0.77</b>	<b>0.79</b>	<b>0.76</b>
(−0.332, −0.0644] $\cup$ (0.734, 0.984]	0.72	0.70	0.72	0.71
(−0.0644, 0.202] $\cup$ (0.202, 0.468]	0.71	0.69	0.71	0.70
(0.468, 0.734]	0.70	0.68	0.70	0.69
<b>All data (Task B)</b>	<b>0.71</b>	<b>0.69</b>	<b>0.71</b>	<b>0.68</b>
(−0.332, −0.0644] $\cup$ (0.734, 0.984]	0.68	0.66	0.68	0.67
(−0.0644, 0.202] $\cup$ (0.202, 0.468]	0.67	0.65	0.67	0.66
(0.468, 0.734]	0.66	0.64	0.66	0.65
<b>All data (Task C)</b>	<b>0.67</b>	<b>0.65</b>	<b>0.67</b>	<b>0.64</b>

#### 4.4. Increasing Model Efficiency

To demonstrate the efficiency improvements achieved by our model, we compared its performance and runtime when processing the entire dataset versus only sentences with high Sexism Scores. Table 9 provides a detailed comparison of these scenarios.

**Table 9.** Efficiency comparison.

Dataset Portion	Accuracy	Precision	Recall	F1 Score	Runtime (s)
All Sentences	0.79	0.77	0.79	0.76	1578 (26.3 min)
High Sexism Score (Top 80%)	0.79	0.75	0.76	0.73	1342 (22.4 min)

As shown, by focusing on sentences with higher Sexism Scores, the runtime is reduced by 13% while maintaining comparable performance metrics. This significant reduction in processing time demonstrates that prioritizing sentences with higher Sexism Scores can enhance model efficiency without substantially compromising accuracy, precision, recall, or F1 score.

This approach leverages the insights gained from the explainability analysis, where tokens with higher SHAP values were identified as more impactful in the model's decision-making process. By concentrating computational resources on these high-impact tokens, we achieve a more efficient processing pipeline. Details of the computing environment used for all experiments can be found in the Appendix A.

#### 4.5. Usability For Decision Makers

The pipeline consists of two key elements: (1) the model's true prediction based on the annotated data, and (2) the SHAP value analysis highlighting the most influential tokens in the sentence. By introducing the sexism range which combines both predictions and influential tokens, decision makers are provided with a transparent view of why a certain sentence is classified as sexist. This can improve their ability to make informed decisions quickly, focusing on the most relevant content. As a result, organizations can have better confidence in their moderation efforts and lessen their dependency on costly extensive manual annotation.

Moreover, this approach addresses one of the key limitations of traditional machine learning models—black-box decision making—by providing a human-interpretable explanation of how the model arrived at a decision. This transparency is essential for maintaining ethical standards and avoiding potential biases in automated content moderation, as well as mitigating the risk of false positives or negatives, which are particularly critical when dealing with sensitive content like sexism.

Also, human annotation, especially for sensitive topics like sexism, is a resource-intensive process, both in terms of time and cost. This approach offers a more transparent and interpretable pipeline where decision makers do not need to manually evaluate every sentence. Instead, they can focus on sentences flagged by the model with the highest Sexism Scores, and use SHAP values to validate the most significant parts of the text. This not only reduces the time required for manual review but also provides a clear rationale for each decision, enhancing trust in the system.

While this paper mainly evaluates the model on data from Gab and Reddit, future work will focus on validating the model across more platforms, to ensure its generalizability. This will help address concerns about cross-platform validation, expanding the model's usability in diverse real-world settings. Additionally, by simplifying the complexity of the system and making the SHAP explanations more actionable, organizations without significant computational resources can implement the model more effectively, ensuring practical utility without sacrificing interpretability or performance.

## 5. Discussion and Conclusions

The present study introduces a new methodology for detecting sexism in textual content by using explainability to define a Sexism Score. This approach integrates ensemble modelling and SHAP values for understanding and identifying sexist language. Our methodology shows advancements in the field of sexism detection. The ensemble model design, CustomBERT, which combined various BERT versions with a CNN architecture,

capitalized on the strengths of multiple transformer models, enhancing the overall accuracy and robustness of the system.

The explainability analysis using SHAP values provided a deeper understanding of the model's decision-making process. By identifying the most influential tokens, we could assign a meaningful Sexism Score to each text entry, thereby offering a transparent and interpretable metric for sexism detection. This aspect is crucial, as it not only improves the trustworthiness of the model but also aids in highlighting specific elements within the text that contribute to its classification as sexist.

Experimental results indicated that our CustomBERT model outperforms individual transformer models across all tasks. Also, the implementation of Sexism Scores based on SHAP values showed a clear correlation between these scores and model performance. Texts with higher Sexism Scores were more reliably identified as sexist, highlighting the efficacy of our explainability-driven approach. Moreover, the efficiency improvements observed by prioritizing high-scoring sentences underscore the practical benefits of this methodology in real-world applications, where processing time and computational resources are critical considerations.

In conclusion, this study presents an interpretable framework for sexism detection in textual content. By integrating a sophisticated ensemble modeling approach, and a thorough explainability analysis, we have developed a model that provides valuable insights into its decision-making process. The introduction of Sexism Scores enhances the model's transparency and interpretability, making it a valuable tool for both academic research and practical applications in combating online harassment and promoting respectful discourse.

Future work can explore the application of this methodology to other forms of hate speech and biased language, further refining the explainability components to address diverse linguistic and cultural contexts. Additionally, integrating user feedback into the explainability analysis could enhance the model's adaptability and accuracy, ensuring its continued relevance and effectiveness in dynamic online environments.

**Author Contributions:** Conceptualization, H.M., A.G. and A.B.; Methodology, H.M., A.G. and A.B.; Resources, H.M.; Writing – original draft, H.M.; Writing – review & editing, A.G. and A.B.; Supervision, A.G. and A.B. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** [REDACTED]

**Informed Consent Statement:** [REDACTED]

**Data Availability Statement:** [REDACTED]

**Conflicts of Interest:** [REDACTED]

## Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolutional Neural Network
LLM	Large Language Model
NLP	Natural Language Processing
SHAP	Shapley Additive Explanations
XAI	Explainable Artificial Intelligence
XNLP	Explainable Natural Language Processing

## Appendix A. System Configuration

The experiments were conducted on a SURF (<https://www.surf.nl/en>, (accessed on)) Azure cloud instance with the following configuration:

- Operating System: Ubuntu 20.04 server

- Instance Type: GPU 24 Core 220 GB RAM 1x A100 (Standard\_NC24ads\_A100\_v4)

## Appendix B. Effective Tokens

The following table lists all the effective tokens identified by our model:

**Table A1.** List of effective tokens.

Token	SHAP Importance	Importance Ratio	Cumulative Importance
pussy	$1.52 \cdot 10^{20}$	$1.16 \cdot 10^{-2}$	$1.16 \cdot 10^{-2}$
yeah	$1.5 \cdot 10^{20}$	$1.14 \cdot 10^{-2}$	$2.3 \cdot 10^{-2}$
shut	$1.39 \cdot 10^{20}$	$1.06 \cdot 10^{-2}$	$3.36 \cdot 10^{-2}$
baby	$1.36 \cdot 10^{20}$	$1.04 \cdot 10^{-2}$	$4.39 \cdot 10^{-2}$
think	$1.32 \cdot 10^{20}$	$1.01 \cdot 10^{-2}$	$5.4 \cdot 10^{-2}$
theyre	$1.31 \cdot 10^{20}$	$9.97 \cdot 10^{-3}$	$6.4 \cdot 10^{-2}$
fight	$1.31 \cdot 10^{20}$	$9.95 \cdot 10^{-3}$	$7.39 \cdot 10^{-2}$
company	$1.31 \cdot 10^{20}$	$9.94 \cdot 10^{-3}$	$8.39 \cdot 10^{-2}$
slut	$1.3 \cdot 10^{20}$	$9.87 \cdot 10^{-3}$	$9.37 \cdot 10^{-2}$
post	$1.29 \cdot 10^{20}$	$9.82 \cdot 10^{-3}$	$0.1$
whore	$1.28 \cdot 10^{20}$	$9.77 \cdot 10^{-3}$	$0.11$
ive	$1.25 \cdot 10^{20}$	$9.51 \cdot 10^{-3}$	$0.12$
day	$1.24 \cdot 10^{20}$	$9.46 \cdot 10^{-3}$	$0.13$
hell	$1.23 \cdot 10^{20}$	$9.39 \cdot 10^{-3}$	$0.14$
ugly	$1.23 \cdot 10^{20}$	$9.37 \cdot 10^{-3}$	$0.15$
vagina	$1.2 \cdot 10^{20}$	$9.12 \cdot 10^{-3}$	$0.16$
just	$1.16 \cdot 10^{20}$	$8.81 \cdot 10^{-3}$	$0.17$
yes	$1.16 \cdot 10^{20}$	$8.8 \cdot 10^{-3}$	$0.18$
avoid	$1.14 \cdot 10^{20}$	$8.71 \cdot 10^{-3}$	$0.19$
try	$1.11 \cdot 10^{20}$	$8.45 \cdot 10^{-3}$	$0.19$
credit	$1.1 \cdot 10^{20}$	$8.39 \cdot 10^{-3}$	$0.2$
talk	$1.08 \cdot 10^{20}$	$8.2 \cdot 10^{-3}$	$0.21$
lady	$1.07 \cdot 10^{20}$	$8.17 \cdot 10^{-3}$	$0.22$
turn	$1.06 \cdot 10^{20}$	$8.05 \cdot 10^{-3}$	$0.23$
sidebar	$1.05 \cdot 10^{20}$	$7.96 \cdot 10^{-3}$	$0.24$
wine	$1.04 \cdot 10^{20}$	$7.95 \cdot 10^{-3}$	$0.24$
fucking	$1.04 \cdot 10^{20}$	$7.93 \cdot 10^{-3}$	$0.25$
work	$1.03 \cdot 10^{20}$	$7.83 \cdot 10^{-3}$	$0.26$
crap	$1.02 \cdot 10^{20}$	$7.77 \cdot 10^{-3}$	$0.27$
rape	$1.02 \cdot 10^{20}$	$7.75 \cdot 10^{-3}$	$0.27$
equality	$1.01 \cdot 10^{20}$	$7.71 \cdot 10^{-3}$	$0.28$
feminism	$9.91 \cdot 10^{19}$	$7.55 \cdot 10^{-3}$	$0.29$
raped	$9.9 \cdot 10^{19}$	$7.54 \cdot 10^{-3}$	$0.3$
youre	$9.71 \cdot 10^{19}$	$7.4 \cdot 10^{-3}$	$0.31$
far	$9.7 \cdot 10^{19}$	$7.38 \cdot 10^{-3}$	$0.31$
making	$9.63 \cdot 10^{19}$	$7.34 \cdot 10^{-3}$	$0.32$
little	$9.35 \cdot 10^{19}$	$7.13 \cdot 10^{-3}$	$0.33$
sex	$9.28 \cdot 10^{19}$	$7.07 \cdot 10^{-3}$	$0.33$
personality	$9.26 \cdot 10^{19}$	$7.05 \cdot 10^{-3}$	$0.34$
commie	$9.23 \cdot 10^{19}$	$7.03 \cdot 10^{-3}$	$0.35$
muslim	$9.15 \cdot 10^{19}$	$6.97 \cdot 10^{-3}$	$0.36$
face	$9.08 \cdot 10^{19}$	$6.91 \cdot 10^{-3}$	$0.36$
cunt	$9.01 \cdot 10^{19}$	$6.86 \cdot 10^{-3}$	$0.37$
maybe	$8.93 \cdot 10^{19}$	$6.8 \cdot 10^{-3}$	$0.38$
girl	$8.88 \cdot 10^{19}$	$6.77 \cdot 10^{-3}$	$0.38$
bitch	$8.84 \cdot 10^{19}$	$6.73 \cdot 10^{-3}$	$0.39$
liberty	$8.83 \cdot 10^{19}$	$6.72 \cdot 10^{-3}$	$0.4$
kino	$8.61 \cdot 10^{19}$	$6.56 \cdot 10^{-3}$	$0.4$
life	$8.58 \cdot 10^{19}$	$6.53 \cdot 10^{-3}$	$0.41$
wtf	$8.56 \cdot 10^{19}$	$6.52 \cdot 10^{-3}$	$0.42$
stop	$8.53 \cdot 10^{19}$	$6.5 \cdot 10^{-3}$	$0.42$
dad	$8.51 \cdot 10^{19}$	$6.48 \cdot 10^{-3}$	$0.43$
burning	$8.35 \cdot 10^{19}$	$6.36 \cdot 10^{-3}$	$0.43$
sending	$8.28 \cdot 10^{19}$	$6.31 \cdot 10^{-3}$	$0.44$
guess	$8.28 \cdot 10^{19}$	$6.3 \cdot 10^{-3}$	$0.45$
guy	$8.25 \cdot 10^{19}$	$6.28 \cdot 10^{-3}$	$0.45$
rapist	$8.11 \cdot 10^{19}$	$6.17 \cdot 10^{-3}$	$0.46$
tbh	$7.85 \cdot 10^{19}$	$5.98 \cdot 10^{-3}$	$0.47$
rule	$7.77 \cdot 10^{19}$	$5.92 \cdot 10^{-3}$	$0.47$
care	$7.72 \cdot 10^{19}$	$5.88 \cdot 10^{-3}$	$0.48$

Token	SHAP Importance	Importance Ratio	Cumulative Importance
run	$7.65 \cdot 10^{19}$	$5.82 \cdot 10^{-3}$	0.48
dirty	$7.44 \cdot 10^{19}$	$5.67 \cdot 10^{-3}$	0.49
islam	$7.09 \cdot 10^{19}$	$5.4 \cdot 10^{-3}$	0.49
having	$6.82 \cdot 10^{19}$	$5.19 \cdot 10^{-3}$	0.5
college	$6.79 \cdot 10^{19}$	$5.17 \cdot 10^{-3}$	0.5
feminist	$6.69 \cdot 10^{19}$	$5.1 \cdot 10^{-3}$	0.51
tell	$6.54 \cdot 10^{19}$	$4.98 \cdot 10^{-3}$	0.52
make	$6.49 \cdot 10^{19}$	$4.94 \cdot 10^{-3}$	0.52
come	$6.42 \cdot 10^{19}$	$4.89 \cdot 10^{-3}$	0.52
role	$6.4 \cdot 10^{19}$	$4.88 \cdot 10^{-3}$	0.53
way	$6.33 \cdot 10^{19}$	$4.82 \cdot 10^{-3}$	0.53
whale	$6.11 \cdot 10^{19}$	$4.65 \cdot 10^{-3}$	0.54
red	$6.11 \cdot 10^{19}$	$4.65 \cdot 10^{-3}$	0.54
attractive	$5.99 \cdot 10^{19}$	$4.56 \cdot 10^{-3}$	0.55
theyd	$5.9 \cdot 10^{19}$	$4.5 \cdot 10^{-3}$	0.55
logic	$5.76 \cdot 10^{19}$	$4.39 \cdot 10^{-3}$	0.56
white	$5.6 \cdot 10^{19}$	$4.27 \cdot 10^{-3}$	0.56
liberal	$5.58 \cdot 10^{19}$	$4.25 \cdot 10^{-3}$	0.57
forget	$5.54 \cdot 10^{19}$	$4.22 \cdot 10^{-3}$	0.57
treat	$5.43 \cdot 10^{19}$	$4.14 \cdot 10^{-3}$	0.57
trump	$5.42 \cdot 10^{19}$	$4.13 \cdot 10^{-3}$	0.58
oppression	$5.29 \cdot 10^{19}$	$4.03 \cdot 10^{-3}$	0.58
left	$5.29 \cdot 10^{19}$	$4.03 \cdot 10^{-3}$	0.59
expect	$5.24 \cdot 10^{19}$	$3.99 \cdot 10^{-3}$	0.59
funny	$5.24 \cdot 10^{19}$	$3.99 \cdot 10^{-3}$	0.59
different	$5.15 \cdot 10^{19}$	$3.92 \cdot 10^{-3}$	0.6
marriage	$5.1 \cdot 10^{19}$	$3.89 \cdot 10^{-3}$	0.6
natural	$5.1 \cdot 10^{19}$	$3.88 \cdot 10^{-3}$	0.61
hope	$5.09 \cdot 10^{19}$	$3.88 \cdot 10^{-3}$	0.61
cuck	$4.98 \cdot 10^{19}$	$3.79 \cdot 10^{-3}$	0.61
surprised	$4.98 \cdot 10^{19}$	$3.79 \cdot 10^{-3}$	0.62
selfish	$4.93 \cdot 10^{19}$	$3.76 \cdot 10^{-3}$	0.62
picture	$4.88 \cdot 10^{19}$	$3.72 \cdot 10^{-3}$	0.62
wonder	$4.88 \cdot 10^{19}$	$3.71 \cdot 10^{-3}$	0.63
getting	$4.78 \cdot 10^{19}$	$3.64 \cdot 10^{-3}$	0.63
did	$4.73 \cdot 10^{19}$	$3.6 \cdot 10^{-3}$	0.64
rt	$4.71 \cdot 10^{19}$	$3.59 \cdot 10^{-3}$	0.64
dead	$4.67 \cdot 10^{19}$	$3.55 \cdot 10^{-3}$	0.64
rest	$4.59 \cdot 10^{19}$	$3.49 \cdot 10^{-3}$	0.65
suck	$4.58 \cdot 10^{19}$	$3.49 \cdot 10^{-3}$	0.65
vote	$4.41 \cdot 10^{19}$	$3.36 \cdot 10^{-3}$	0.65
course	$4.38 \cdot 10^{19}$	$3.34 \cdot 10^{-3}$	0.66
number	$4.32 \cdot 10^{19}$	$3.29 \cdot 10^{-3}$	0.66
idiot	$4.29 \cdot 10^{19}$	$3.27 \cdot 10^{-3}$	0.66
hard	$4.15 \cdot 10^{19}$	$3.16 \cdot 10^{-3}$	0.67
soros	$4.11 \cdot 10^{19}$	$3.13 \cdot 10^{-3}$	0.67
report	$4.09 \cdot 10^{19}$	$3.11 \cdot 10^{-3}$	0.67
begin	$4.08 \cdot 10^{19}$	$3.11 \cdot 10^{-3}$	0.68
space	$4.08 \cdot 10^{19}$	$3.11 \cdot 10^{-3}$	0.68
away	$4.07 \cdot 10^{19}$	$3.1 \cdot 10^{-3}$	0.68
spend	$4.06 \cdot 10^{19}$	$3.09 \cdot 10^{-3}$	0.69
ball	$4.04 \cdot 10^{19}$	$3.08 \cdot 10^{-3}$	0.69
fucked	$3.98 \cdot 10^{19}$	$3.03 \cdot 10^{-3}$	0.69
monkey	$3.96 \cdot 10^{19}$	$3.02 \cdot 10^{-3}$	0.69
enemy	$3.96 \cdot 10^{19}$	$3.01 \cdot 10^{-3}$	0.7
wait	$3.95 \cdot 10^{19}$	$3.01 \cdot 10^{-3}$	0.7
wont	$3.89 \cdot 10^{19}$	$2.96 \cdot 10^{-3}$	0.7
waiting	$3.86 \cdot 10^{19}$	$2.94 \cdot 10^{-3}$	0.71
oh	$3.84 \cdot 10^{19}$	$2.93 \cdot 10^{-3}$	0.71
send	$3.83 \cdot 10^{19}$	$2.91 \cdot 10^{-3}$	0.71

Token	SHAP Importance	Importance Ratio	Cumulative Importance
id	$3.78 \cdot 10^{19}$	$2.88 \cdot 10^{-3}$	0.71
going	$3.77 \cdot 10^{19}$	$2.87 \cdot 10^{-3}$	0.72
wife	$3.74 \cdot 10^{19}$	$2.85 \cdot 10^{-3}$	0.72
foid	$3.66 \cdot 10^{19}$	$2.79 \cdot 10^{-3}$	0.72
thanks	$3.64 \cdot 10^{19}$	$2.78 \cdot 10^{-3}$	0.73
thing	$3.6 \cdot 10^{19}$	$2.74 \cdot 10^{-3}$	0.73
hate	$3.58 \cdot 10^{19}$	$2.73 \cdot 10^{-3}$	0.73
place	$3.49 \cdot 10^{19}$	$2.66 \cdot 10^{-3}$	0.73
current	$3.47 \cdot 10^{19}$	$2.64 \cdot 10^{-3}$	0.74
easily	$3.45 \cdot 10^{19}$	$2.63 \cdot 10^{-3}$	0.74
need	$3.41 \cdot 10^{19}$	$2.6 \cdot 10^{-3}$	0.74
really	$3.28 \cdot 10^{19}$	$2.5 \cdot 10^{-3}$	0.74
word	$3.26 \cdot 10^{19}$	$2.48 \cdot 10^{-3}$	0.75
thank	$3.25 \cdot 10^{19}$	$2.48 \cdot 10^{-3}$	0.75
say	$3.25 \cdot 10^{19}$	$2.48 \cdot 10^{-3}$	0.75
lying	$3.25 \cdot 10^{19}$	$2.47 \cdot 10^{-3}$	0.75
mean	$3.24 \cdot 10^{19}$	$2.47 \cdot 10^{-3}$	0.76
female	$3.18 \cdot 10^{19}$	$2.42 \cdot 10^{-3}$	0.76
state	$3.15 \cdot 10^{19}$	$2.4 \cdot 10^{-3}$	0.76
men	$3.14 \cdot 10^{19}$	$2.39 \cdot 10^{-3}$	0.76
actually	$3.13 \cdot 10^{19}$	$2.38 \cdot 10^{-3}$	0.77
ground	$3.13 \cdot 10^{19}$	$2.38 \cdot 10^{-3}$	0.77
12	$3.12 \cdot 10^{19}$	$2.38 \cdot 10^{-3}$	0.77
deserve	$3.07 \cdot 10^{19}$	$2.34 \cdot 10^{-3}$	0.77
exist	$3.06 \cdot 10^{19}$	$2.33 \cdot 10^{-3}$	0.78
wouldnt	$3.02 \cdot 10^{19}$	$2.3 \cdot 10^{-3}$	0.78
hang	$2.98 \cdot 10^{19}$	$2.27 \cdot 10^{-3}$	0.78
reason	$2.97 \cdot 10^{19}$	$2.26 \cdot 10^{-3}$	0.78
lmao	$2.94 \cdot 10^{19}$	$2.24 \cdot 10^{-3}$	0.79
daily	$2.91 \cdot 10^{19}$	$2.22 \cdot 10^{-3}$	0.79
stand	$2.87 \cdot 10^{19}$	$2.19 \cdot 10^{-3}$	0.79
wall	$2.85 \cdot 10^{19}$	$2.17 \cdot 10^{-3}$	0.79
youll	$2.82 \cdot 10^{19}$	$2.15 \cdot 10^{-3}$	0.79
ha	$2.8 \cdot 10^{19}$	$2.13 \cdot 10^{-3}$	0.8
fact	$2.8 \cdot 10^{19}$	$2.13 \cdot 10^{-3}$	0.8
potential	$2.77 \cdot 10^{19}$	$2.11 \cdot 10^{-3}$	0.8
damage	$2.76 \cdot 10^{19}$	$2.1 \cdot 10^{-3}$	0.8
gender	$2.75 \cdot 10^{19}$	$2.09 \cdot 10^{-3}$	0.8
agree	$2.74 \cdot 10^{19}$	$2.08 \cdot 10^{-3}$	0.81
giving	$2.71 \cdot 10^{19}$	$2.06 \cdot 10^{-3}$	0.81
trap	$2.71 \cdot 10^{19}$	$2.06 \cdot 10^{-3}$	0.81
use	$2.7 \cdot 10^{19}$	$2.06 \cdot 10^{-3}$	0.81
imagine	$2.69 \cdot 10^{19}$	$2.05 \cdot 10^{-3}$	0.81
thats	$2.68 \cdot 10^{19}$	$2.04 \cdot 10^{-3}$	0.82
art	$2.68 \cdot 10^{19}$	$2.04 \cdot 10^{-3}$	0.82
ring	$2.66 \cdot 10^{19}$	$2.02 \cdot 10^{-3}$	0.82
lot	$2.62 \cdot 10^{19}$	$2 \cdot 10^{-3}$	0.82
matter	$2.62 \cdot 10^{19}$	$1.99 \cdot 10^{-3}$	0.82
smart	$2.61 \cdot 10^{19}$	$1.99 \cdot 10^{-3}$	0.83
chance	$2.61 \cdot 10^{19}$	$1.99 \cdot 10^{-3}$	0.83
inside	$2.6 \cdot 10^{19}$	$1.98 \cdot 10^{-3}$	0.83
difference	$2.59 \cdot 10^{19}$	$1.97 \cdot 10^{-3}$	0.83
shes	$2.58 \cdot 10^{19}$	$1.96 \cdot 10^{-3}$	0.83
trying	$2.56 \cdot 10^{19}$	$1.95 \cdot 10^{-3}$	0.84
user	$2.54 \cdot 10^{19}$	$1.93 \cdot 10^{-3}$	0.84
want	$2.53 \cdot 10^{19}$	$1.93 \cdot 10^{-3}$	0.84
cope	$2.51 \cdot 10^{19}$	$1.91 \cdot 10^{-3}$	0.84
future	$2.51 \cdot 10^{19}$	$1.91 \cdot 10^{-3}$	0.84
got	$2.5 \cdot 10^{19}$	$1.9 \cdot 10^{-3}$	0.85
mom	$2.5 \cdot 10^{19}$	$1.9 \cdot 10^{-3}$	0.85

Token	SHAP Importance	Importance Ratio	Cumulative Importance
rich	$2.44 \cdot 10^{19}$	$1.85 \cdot 10^{-3}$	0.85
femininity	$2.42 \cdot 10^{19}$	$1.84 \cdot 10^{-3}$	0.85
friend	$2.41 \cdot 10^{19}$	$1.84 \cdot 10^{-3}$	0.85
instead	$2.4 \cdot 10^{19}$	$1.83 \cdot 10^{-3}$	0.86
clinton	$2.39 \cdot 10^{19}$	$1.82 \cdot 10^{-3}$	0.86
boring	$2.35 \cdot 10^{19}$	$1.79 \cdot 10^{-3}$	0.86
immediately	$2.32 \cdot 10^{19}$	$1.77 \cdot 10^{-3}$	0.86
plan	$2.31 \cdot 10^{19}$	$1.76 \cdot 10^{-3}$	0.86
working	$2.31 \cdot 10^{19}$	$1.76 \cdot 10^{-3}$	0.86
sister	$2.3 \cdot 10^{19}$	$1.75 \cdot 10^{-3}$	0.87
toe	$2.28 \cdot 10^{19}$	$1.73 \cdot 10^{-3}$	0.87
behavior	$2.27 \cdot 10^{19}$	$1.73 \cdot 10^{-3}$	0.87
blame	$2.27 \cdot 10^{19}$	$1.73 \cdot 10^{-3}$	0.87
bos	$2.24 \cdot 10^{19}$	$1.71 \cdot 10^{-3}$	0.87
fought	$2.23 \cdot 10^{19}$	$1.69 \cdot 10^{-3}$	0.87
dick	$2.23 \cdot 10^{19}$	$1.69 \cdot 10^{-3}$	0.88
feminine	$2.2 \cdot 10^{19}$	$1.68 \cdot 10^{-3}$	0.88
mgtow	$2.18 \cdot 10^{19}$	$1.66 \cdot 10^{-3}$	0.88
mother	$2.14 \cdot 10^{19}$	$1.63 \cdot 10^{-3}$	0.88
thinking	$2.14 \cdot 10^{19}$	$1.63 \cdot 10^{-3}$	0.88
american	$2.13 \cdot 10^{19}$	$1.62 \cdot 10^{-3}$	0.88
bang	$2.09 \cdot 10^{19}$	$1.59 \cdot 10^{-3}$	0.89
self	$2.08 \cdot 10^{19}$	$1.59 \cdot 10^{-3}$	0.89
problem	$2.08 \cdot 10^{19}$	$1.59 \cdot 10^{-3}$	0.89
male	$2.06 \cdot 10^{19}$	$1.57 \cdot 10^{-3}$	0.89
young	$2.04 \cdot 10^{19}$	$1.55 \cdot 10^{-3}$	0.89
jew	$2.04 \cdot 10^{19}$	$1.55 \cdot 10^{-3}$	0.89
worst	$2.03 \cdot 10^{19}$	$1.55 \cdot 10^{-3}$	0.9
attention	$2.03 \cdot 10^{19}$	$1.55 \cdot 10^{-3}$	0.9
said	$2.01 \cdot 10^{19}$	$1.53 \cdot 10^{-3}$	0.9
start	$2 \cdot 10^{19}$	$1.52 \cdot 10^{-3}$	0.9
shit	$2 \cdot 10^{19}$	$1.52 \cdot 10^{-3}$	0.9
black	$1.98 \cdot 10^{19}$	$1.51 \cdot 10^{-3}$	0.9
waste	$1.98 \cdot 10^{19}$	$1.5 \cdot 10^{-3}$	0.9
fuck	$1.96 \cdot 10^{19}$	$1.49 \cdot 10^{-3}$	0.91
kavanaugh	$1.94 \cdot 10^{19}$	$1.48 \cdot 10^{-3}$	0.91
lol	$1.93 \cdot 10^{19}$	$1.47 \cdot 10^{-3}$	0.91
le	$1.93 \cdot 10^{19}$	$1.47 \cdot 10^{-3}$	0.91
wild	$1.93 \cdot 10^{19}$	$1.47 \cdot 10^{-3}$	0.91
lead	$1.88 \cdot 10^{19}$	$1.43 \cdot 10^{-3}$	0.91
virgin	$1.82 \cdot 10^{19}$	$1.39 \cdot 10^{-3}$	0.92
realize	$1.8 \cdot 10^{19}$	$1.37 \cdot 10^{-3}$	0.92
dont	$1.79 \cdot 10^{19}$	$1.36 \cdot 10^{-3}$	0.92
hanging	$1.77 \cdot 10^{19}$	$1.35 \cdot 10^{-3}$	0.92
typical	$1.74 \cdot 10^{19}$	$1.33 \cdot 10^{-3}$	0.92
die	$1.73 \cdot 10^{19}$	$1.32 \cdot 10^{-3}$	0.92
given	$1.73 \cdot 10^{19}$	$1.32 \cdot 10^{-3}$	0.92
west	$1.69 \cdot 10^{19}$	$1.29 \cdot 10^{-3}$	0.92
believe	$1.67 \cdot 10^{19}$	$1.27 \cdot 10^{-3}$	0.93
super	$1.64 \cdot 10^{19}$	$1.25 \cdot 10^{-3}$	0.93
doe	$1.59 \cdot 10^{19}$	$1.21 \cdot 10^{-3}$	0.93
hold	$1.57 \cdot 10^{19}$	$1.2 \cdot 10^{-3}$	0.93
maga	$1.57 \cdot 10^{19}$	$1.2 \cdot 10^{-3}$	0.93
end	$1.57 \cdot 10^{19}$	$1.2 \cdot 10^{-3}$	0.93
lonely	$1.51 \cdot 10^{19}$	$1.15 \cdot 10^{-3}$	0.93
voter	$1.5 \cdot 10^{19}$	$1.14 \cdot 10^{-3}$	0.93
incel	$1.48 \cdot 10^{19}$	$1.13 \cdot 10^{-3}$	0.94
probably	$1.44 \cdot 10^{19}$	$1.1 \cdot 10^{-3}$	0.94
plot	$1.4 \cdot 10^{19}$	$1.07 \cdot 10^{-3}$	0.94
leg	$1.4 \cdot 10^{19}$	$1.06 \cdot 10^{-3}$	0.94

Token	SHAP Importance	Importance Ratio	Cumulative Importance
loser	$1.37 \cdot 10^{19}$	$1.05 \cdot 10^{-3}$	0.94
watch	$1.37 \cdot 10^{19}$	$1.04 \cdot 10^{-3}$	0.94
sexual	$1.37 \cdot 10^{19}$	$1.04 \cdot 10^{-3}$	0.94
wow	$1.33 \cdot 10^{19}$	$1.01 \cdot 10^{-3}$	0.94
money	$1.32 \cdot 10^{19}$	$1.01 \cdot 10^{-3}$	0.94
killed	$1.27 \cdot 10^{19}$	$9.7 \cdot 10^{-4}$	0.94
lulz	$1.23 \cdot 10^{19}$	$9.34 \cdot 10^{-4}$	0.95
cost	$1.22 \cdot 10^{19}$	$9.26 \cdot 10^{-4}$	0.95
support	$1.21 \cdot 10^{19}$	$9.23 \cdot 10^{-4}$	0.95
reply	$1.21 \cdot 10^{19}$	$9.2 \cdot 10^{-4}$	0.95
willing	$1.19 \cdot 10^{19}$	$9.06 \cdot 10^{-4}$	0.95

## References

- Kurasawa, F.; Rondinelli, E.; Kilicaslan, G. Evidentiary activism in the digital age: On the rise of feminist struggles against gender-based online violence. *Inf. Commun. Soc.* **2021**, *24*, 2174–2194.
- Papaevangelou, C. ‘The non-interference principle’: Debating online platforms’ treatment of editorial content in the European Union’s Digital Services Act. *Eur. J. Commun.* **2023**, *38*, 466–483.
- Ortiz, S.M. “If Something Ever Happened, I’d Have No One to Tell:” how online sexism perpetuates young women’s silence. *Fem. Media Stud.* **2023**, *24*, 119–134.
- Aldana-Bobadilla, E.; Molina-Villegas, A.; Montelongo-Padilla, Y.; Lopez-Arevalo, I.; Sordia, O. A language model for misogyny detection in Latin American Spanish driven by multisource feature extraction and transformers. *Appl. Sci.* **2021**, *11*, 10467.
- Lee, F.L.; Liang, H.; Cheng, E.W.; Tang, G.K.; Yuen, S. Affordances, movement dynamics, and a centralized digital communication platform in a networked movement. *Inf. Commun. Soc.* **2022**, *25*, 1699–1716.
- Feng, C. A simple voting mechanism for online sexist content identification. *arXiv* **2021**, arXiv:2105.14309.
- Schütz, M.; Boeck, J.; Liakhovets, D.; Slijepcevic, D.; Kirchknopf, A.; Hecht, M.; Bogensperger, J.; Schlarb, S.; Schindler, A.; Zeppelzauer, M. Automatic Sexism Detection with Multilingual Transformer Models, CoRR abs/2106.04908. 2021. Available online: <https://arxiv.org/abs/2106.04908> (accessed on).
- Kumar, R.; Pal, S.; Pamula, R. Sexism Detection in English and Spanish Tweets. In Proceedings of the IberLEF@ SEPLN; 2021; pp. 500–505. Available online: [https://ceur-ws.org/Vol-2943/exist\\_paper17.pdf](https://ceur-ws.org/Vol-2943/exist_paper17.pdf) (accessed on).
- de Paula, A.F.M.; da Silva, R.F.; Schlicht, I.B. Sexism prediction in spanish and english tweets using monolingual and multilingual bert and ensemble models. *arXiv* **2021**, arXiv:2111.04551.
- Altin, L.S.M.; Saggion, H. Automatic detection of sexism in social media with a multilingual approach. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), Málaga, Espanya, 21 September 2021; [Málaga]: CEUR Workshop Proceedings Series; CEUR Workshop Proceedings: Aachen, Germany, 2021; pp. 415–419.
- Mehta, H.; Passi, K. Social media hate speech detection using explainable artificial intelligence (XAI). *Algorithms* **2022**, *15*, 291.
- Gil Bermejo, J.L.; Martos Sánchez, C.; Vázquez Aguado, O.; García-Navarro, E.B. Adolescents, ambivalent sexism and social networks, a conditioning factor in the healthcare of women. *Healthcare* **2021**, *9*, 721.
- Hoofnagle, C.J.; Van Der Sloot, B.; Borgesius, F.Z. The European Union general data protection regulation: What it is and what it means. *Inf. Commun. Technol. Law* **2019**, *28*, 65–98.
- Mathew, B.; Saha, P.; Yimam, S.M.; Biemann, C.; Goyal, P.; Mukherjee, A. Hatexplain: A benchmark dataset for explainable hate speech detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 2–9 February 2021; Volume 35, pp. 14867–14875.
- Velankar, A.; Patil, H.; Joshi, R. A review of challenges in machine learning based automated hate speech detection. *arXiv* **2022**, arXiv:2209.05294.
- Jiang, J.A. Identifying and addressing design and policy challenges in online content moderation. In Proceedings of the Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; pp. 1–7.
- Danilevsky, M.; Qian, K.; Aharonov, R.; Katsis, Y.; Kawas, B.; Sen, P. A survey of the state of explainable AI for natural language processing. *arXiv* **2020**, arXiv:2010.00711.
- Søgaard, A. *Explainable Natural Language Processing*; Morgan & Claypool Publishers: San Rafael, CA, USA, 2021.
- Mohammadi, H.; Giachanou, A.; Bagheri, A. Towards robust online sexism detection: A multi-model approach with BERT, XLM-RoBERTa, and DistilBERT for EXIST 2023 Tasks. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023); CEUR Workshop Proceedings: Aachen, Germany*, 2023.
- Böck, J.; Schütz, M.; Liakhovets, D.; Satriani, N.Q.; Babic, A.; Slijepčević, D.; Zeppelzauer, M.; Schindler, A. AIT\_FHSTP at EXIST 2023 benchmark: Sexism detection by transfer learning, sentiment and toxicity embeddings and hand-crafted features. In *Working Notes of CLEF*; 2023.
- Daouadi, K.E.; Boualleg, Y.; Guehairia, O. Deep Random Forest and AraBert for Hate Speech Detection from Arabic Tweets. *J. Univers. Comput. Sci.* **2023**, *29*, 1319–1335.

22. Lopez-Lopez, E.; Carrillo-de Albornoz, J.; Plaza, L. Combining Transformer-Based Models with Traditional Machine Learning Approaches for Sexism Identification in Social Networks at EXIST 2021. In Proceedings of the IberLEF@ SEPLN; 2021; pp. 431–441. Available online: [https://ceur-ws.org/Vol-2943/exist\\_paper10.pdf](https://ceur-ws.org/Vol-2943/exist_paper10.pdf) (accessed on).
23. Samory, M.; Sen, I.; Kohne, J.; Flöck, F.; Wagner, C. “Call me sexist, but...”: Revisiting Sexism Detection Using Psychological Scales and Adversarial Samples. In Proceedings of the International AAAI Conference on Web and sOcial Media, [Online, 7–10 June 2021](#); Volume 15, pp. 573–584.
24. Rodríguez-Sánchez, F.; de Albornoz, J.C.; Plaza, L. Automatic Classification of Sexism in Social Networks: An Empirical Study on Twitter Data. *IEEE Access* **2020**, *8*, 219563–219576. <https://doi.org/10.1109/ACCESS.2020.3042604>.
25. Jha, A.; Mamidi, R. When Does a Compliment Become Sexist? Analysis and Classification of Ambivalent Sexism Using Twitter Data. **2017**; pp. 7–16. Available online: <https://aclanthology.org/W17-2902/> (accessed on). <https://doi.org/10.18653/v1/W17-2902>.
26. Jiang, A.; Yang, X.; Liu, Y.; Zubiaga, A. SWSR: A Chinese dataset and lexicon for online sexism detection. *Online Soc. Netw. Media* **2022**, *27*, 100182.
27. Das, A.; Rahgouy, M.; Zhang, Z.; Bhattacharya, T.; Dozier, G.; Seals, C.D. Online Sexism Detection and Classification by Injecting User Gender Information. In Proceedings of the 2023 IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings), [Mount Pleasant, MI, USA, 16–17 September 2023](#); IEEE: [Piscataway, NJ, USA, 2023](#); pp. 1–5.
28. Kirk, H.R.; Yin, W.; Vidgen, B.; Röttger, P. SemEval-2023 Task 10: Explainable Detection of Online Sexism. In Proceedings of the [17th International Workshop on Semantic Evaluation \(SemEval-2023\)](#). Association for Computational Linguistics, [Toronto, ON, Canada, 9–14 July 2023](#). <https://doi.org/10.48550/arXiv.2303.04222>.
29. Tasneem, F.; Hossain, T.; Naim, J. KingsmanTrio at SemEval-2023 Task 10: Analyzing the Effectiveness of Transfer Learning Models for Explainable Online Sexism Detection. In Proceedings of the Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), [Toronto, ON, Canada, 31 January 2023](#); pp. 1916–1920.
30. Kiritchenko, S.; Nejadgholi, I.; Fraser, K.C. Confronting abusive language online: A survey from the ethical and human rights perspective. *J. Artif. Intell. Res.* **2021**, *71*, 431–478.
31. Lamsiyah, S.; El Mahdaouy, A.; Alami, H.; Berrada, I.; Schommer, C. UL & UM6P at SemEval-2023 Task 10: Semi-Supervised Multi-task Learning for Explainable Detection of Online Sexism. In Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), [Toronto, ON, Canada, 31 January 2023](#); pp. 644–650.
32. Kotapati, G.; Gandhimathi, S.K.; Rao, P.A.; Muppagowni, G.K.; Bindu, K.R.; Reddy, M.S.C. A Natural Language Processing for Sentiment Analysis from Text using Deep Learning Algorithm. In Proceedings of the 2023 2nd International Conference on Edge Computing and Applications (ICECAA), [Namakkal, India, 19–21 July 2023](#); IEEE: [Piscataway, NJ, USA, 2023](#); pp. 1028–1034.
33. Chauhan, R.; Gusain, A.; Kumar, P.; Bhatt, C.; Uniyal, I. Fine Grained Sentiment Analysis using Machine Learning and Deep Learning. In Proceedings of the 2023 International Conference on Sustainable Emerging Innovations in Engineering and Technology (ICSEIET), [Ghaziabad, India, 14–15 September 2023](#); IEEE: [Piscataway, NJ, USA, 2023](#), pp. 423–427.
34. Mariappan, U.; Balakrishnan, D.; Subhashini, S.; Kumar, N.V.A.S.; Rao, S.L.S.M.; Alagusundar, N. Sentiment and Context-Aware Recurrent Convolutional Neural Network for Sentiment Analysis. In Proceedings of the 2023 3rd Asian Conference on Innovation in Technology (ASIANCON), [Pune, India, 25–27 August 2023](#); IEEE: [Piscataway, NJ, USA, 2023](#); pp. 1–6.
35. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4765–4774.
36. Ribeiro, M.T.; Singh, S.; Guestrin, C. “Why should i trust you?” Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, [San Francisco, CA, USA, 13–17 August 2016](#); pp. 1135–1144.
37. Lai, V.; Carton, S.; Bhatnagar, R.; Liao, Q.V.; Zhang, Y.; Tan, C. Human-ai collaboration via conditional delegation: A case study of content moderation. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, [New Orleans, LA, USA, 29 April–5 May 2022](#); pp. 1–18.
38. Molina, M.D.; Sundar, S.S. When AI moderates online content: Effects of human collaboration and interactive transparency on user trust. *J. Comput.-Mediat. Commun.* **2022**, *27*, zmac010.
39. Rallabandi, S.; Kakodkar, I.G.; Avuku, O. Ethical Use of AI in Social Media. In Proceedings of the 2023 International Workshop on Intelligent Systems (IWIS), [Ulsan, Republic of Korea, 9–11 August 2023](#); IEEE: [Piscataway, NJ, USA, 2023](#); pp. 1–9.
40. Kirk, H.R.; Yin, W.; Vidgen, B.; Röttger, P. SemEval-2023 Task 10: Explainable Detection of Online Sexism. *arXiv* **2023**, arXiv:2303.04222.
41. Beddiar, D.R.; Jahan, M.S.; Oussalah, M. Data expansion using back translation and paraphrasing for hate speech detection. *Online Soc. Netw. Media* **2021**, *24*, 100153.
42. Zheng, Z.; Cai, Y.; Li, Y. Oversampling method for imbalanced classification. *Comput. Inform.* **2015**, *34*, 1017–1037.
43. Xu, Y.; Vaziri-Pashkam, M. Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nat. Commun.* **2021**, *12*, 2065.
44. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
45. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised cross-lingual representation learning at scale. *arXiv* **2019**, arXiv:1911.02116.

46. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108.
47. Prabha, M.I.; Srikanth, G.U. Survey of sentiment analysis using deep learning techniques. In Proceedings of the 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), **Chennai, India, 25–26 April 2019**; IEEE: **Piscataway, NJ, USA**, 2019, pp. 1–9.
48. Mohammadi, H.; Giachanou, A.; Bagheri, A. Code for “Towards Robust Online Sexism Detection: A Multi-Model Approach with BERT, XLM-RoBERTa, and DistilBERT for EXIST 2023 Tasks”, 2023. Available online: <https://zenodo.org/records/8144300> (accessed on). <https://doi.org/10.5281/zenodo.8144300>.
49. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
50. Brownlee, J. A gentle introduction to early stopping to avoid overtraining neural networks. *Mach. Learn. Mastery* **2018**, *7*.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.