CNLP-NITS at SemEval-2023 Task 10: Online sexism prediction, PREDHATE!

Advaitha Vetagiri
NIT Silchar, Assam, India
advaithavetagiri@gmail.com

Prottay Kumar Adhikary
NIT Silchar, Assam, India
prottay71@gmail.com

Partha Pakray
NIT Silchar, Assam, India
pakraypartha@gmail.com

Amitava Das
AIISC, South Carolina, USA
Wipro AI Lab, India
amitava.santu@gmail.com

Abstract

Online sexism is a rising issue that threatens women's safety, fosters hostile situations, and upholds social inequities. We describe a task SemEval-2023 Task 10 for creating English-language models that can precisely identify and categorize sexist content on internet forums and social platforms like Gab and Reddit as well to provide an explainability in order to address this problem. The problem is divided into three hierarchically organized subtasks: binary sexism detection, sexism by category, and sexism by fine-grained vector. The dataset consists of 20,000 labelled entries. For Task A, pertained models like Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM), which is called CNN-BiLSTM and Generative Pretrained Transformer 2 (GPT-2) models were used, as well as the GPT-2 model for Task B and C, and have provided experimental configurations. According to our findings, the GPT-2 model performs better than the CNN-BiLSTM model for Task A, while GPT-2 is highly accurate for Tasks B and C on the training, validation and testing splits of the training data provided in the task. Our proposed models allow researchers to create more precise and understandable models for identifying and categorizing sexist content in online forums, thereby empowering users and moderators.

1 Introduction

Sexism is a pervasive issue in society that shows up in a variety of linguistic contexts, including online discourse (Paciello et al., 2021). The topic of automatically recognising sexist language in text has several real-world applications, including bias detection in computer-generated material and social media platform moderation (Verge, 2022). Sexism is a major issue in society today, and it has been shown to have negative effects on both men and women. The prevalence of sexism in online communication makes it difficult for people to identify and address it (Leaper and Brown, 2008). Automatically detecting sexist language in text is an important problem that has many practical applications. For example, social media platforms can use automated systems to detect sexist language and take appropriate actions against users who post offensive content (Rodríguez-Sánchez et al., 2020), this could help reduce the amount of sexism present on these platforms and create a more inclusive environment for all users.

The shared task SemEval-2023 Task 10 (Kirk et al., 2023) was about identifying sexism and proving it by identifing sexist content and explaining why it is sexist enhances the generalization ability, trust, and comprehension of the choices made by automated systems, giving users and moderators more control. This will bring more confidence on the automated system which detects sexist content by providing an explainability for that content. The shared task incorporated with three hierarchical subtasks that each builds on the previous one makes up our main task. The first Subtask, Task A, requires systems to anticipate whether or not a post is sexist and it is a binary classification task. In the second subtask, Task B, systems must predict the category of sexism for posts that are labeled as sexist in Task A. Task B is a four-class classification task. *'Threats', 'Derogation', 'Animosity', and 'Prejudiced Discussions'* based on prejudice are the categories. The third and final subtask, Task C, requires systems to predict one of 11 fine-grained vectors for postings that are rated as sexist in Task A and categorized in Task B. Task C is an 11-class classification task.

In this shared task SemEval-2023 Task 10, the problem of detecting sexist language and explainability was addressed using two models, one is **Convolutional Neural Network** (CNN) (Kim, 2014) and **Bidirectional Long Short-Term Memory** (BiLSTM) (Liu and Guo, 2019) which is interchangeably called throughout the paper CNN-

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 815–822

BiLSTM (Ma et al., 2020) and the other is **Generative Pretrained Transformer 2** (GPT-2) (Radford et al., 2019). A common type of neural network used in image analysis is the CNN. But, by treating each word as a "channel" in the input, they can also be employed in tasks involving natural language processing. In order to capture particular features, such as n-grams and word embeddings, that are important for the classification task, the CNN layer can then learn various sorts of filters. On the other hand, BiLSTM models are a kind of Recurrent Neural Network (Sherstinsky, 2020) created specifically to handle sequential data, like text. They can recognise long-range relationships in the text because they can keep track of the words that came before them in a phrase. GPT-2 is a Transformer-Based Language Model developed by OpenAI (Radford et al., 2019) that has shown remarkable performance in various natural language processing tasks, such as language generation (Ko and Li, 2020), machine translation (He et al., 2020), and question answering (Puri et al., 2020). To deploy our models a web application called 'PRED-HATE!' has been created that enables users to input any text they desire. The application employs our model's forecast to indicate whether the text contains sexist content or not.

The paper has been arranged as follows, Section 2 will be the background work and relative work done in the field, Section 3 will be about the data sampling and training, development and test data splits. Section 4 the system overview of the models and were demonstrated. Section 5 gives the experimental setup of our models for the Tasks A,B & C. Section 6 will provide the results how our models performed on these tasks.

2 Background

In recent years, there has been growing interest in the task of detecting sexism and misogyny (Parikh et al., 2021) in natural language text. This is exemplified by research efforts such as SemEval Task 10, which provides a dataset (Kirk et al., 2023) for training and evaluating machine learning models for this task (Zampieri et al., 2019).

One approach to tackling this challenge is to use classic machine learning techniques, such as n-grams, as demonstrated inn (Anzovino et al., 2018). This study presents a dataset for identifying and classifying misogynistic language on Twitter in both Spanish and English. Meanwhile, (Frenda et al., 2019) applies a similar approach to detect online hate speech against women by combining two datasets. Likewise, we are combining the SemEval Task 10 dataset with other similar ones, which we will be explaining in detail in the section 3.

However, more recent studies have explored the use of advanced deep learning techniques to achieve state-of-the-art results in sexism detection. For instance, (Grosz and Céspedes, 2020) uses GloVe embeddings (Pennington et al., 2014) and modified LSTMs with attention mechanisms (Bahdanau et al., 2014) to detect sexist statements commonly used in the workplace automatically. These studies demonstrate the potential of using deep learning techniques to address the challenge of detecting sexism and misogyny in natural language text.

GPT-2 (Radford et al., 2019) has been used in several related natural language processing (NLP) tasks, including text classification (Anaby-Tavor et al., 2020), sentiment analysis (Alexandridis et al., 2021), and language modeling (Budzianowski and Vulić, 2019).

For example, in a recent study by (Zhang et al., 2022), the authors used a pre-trained GPT-2 model to perform binary classification on a dataset of news articles to predict whether the article is sexist or not sexist. The results showed that the GPT-2 model outperformed several other machine learning models in terms of accuracy.

Another study (Mathew and Bindu, 2020) used a pre-trained GPT-2 model to perform sentiment analysis on a dataset of customer reviews. The authors found that the GPT-2 model was able to achieve state-of-the-art performance in sentiment analysis tasks.

While these studies did not focus specifically on sexism classification, they demonstrate the potential of pre-trained language models like GPT-2 for natural language processing tasks. Similar approaches using GPT-2 could be effective for the sexism classification task (Vaca-Serrano, 2022) as they have done a decent job, for classification of sexism in English and Spanish languages, they have mostly used Encoder-Decoder models for their approach.

3 Data

The organisers have provided the data in the task SemEval-2023 Task 10, which we will be giving a brief overview of the data, how the data has been

Task A	Task B	Task C
Sexist	1. threats, plans to harm and incitement	1.1 threats of harm
		1.2 incitement and encouragement of harm
	2. derogation	2.1 descriptive attacks
		2.2 aggressive and emotive attacks
		2.3 dehumanising attacks & overt sexual objectification
	3. animosity	3.1 casual use of gendered slurs, profanities, and insults
		3.2 immutable gender differences and gender stereotypes
		3.3 backhanded gendered compliments
		3.4 condescending explanations or unwelcome advice
	4. prejudiced discussions	4.1 supporting mistreatment of individual women
		4.2 supporting systemic discrimination against women as a group
Non Sexist	None	None

Table 1: Classes of the given Dataset

split into train, development and test splits as well as the data sampling and annotation in this section.

3.1 Data Sampling and Annotation

(Kirk et al., 2023) have 20,000 entries in their labeled data-set, with 10,000 coming from Gab[1] and 10,000 from Reddit[2]. Three professional annotators first label the entries, then one of two experts decides on disagreements after that. The gold label for Task A is decided by the annotators in unanimity or, in the event of a tie, by an expert review. The gold label for Tasks B and C is chosen based on agreement by at least two annotators or, in the event of a 3-way tie, an expert review. They annotate their data using the "prescriptive paradigm," which includes precise annotation rules, frequent feedback, and training. All annotators and experts are self-identifying women.

3.2 Training, Development, and Test Data

3,398 of the 14,000 entries in their training set, which has a 70% split, are identified as sexist. For Tasks A, B, and C, there is a single CSV file with the following columns: text (the input text), label sexist (the label for Task A), label category (the label for Task B), and label vector (the label for Task C). Label category and Label vector are set to the string "none" for entries that are not sexist.

	Train	Valid	Test
Task A	11200	1400	1400
Task B	2718	340	340
Task C	2718	340	340

Table 2: Dataset Split Statistics

To promote cutting-edge training methods, they also made available two unlabeled datasets from Gab and Reddit, each with one million entries. Development data is used to grade submissions during the development phase and consists of 2,000 entries with a 10% split. Test data is used to grade submissions during the test phase and consists of 4,000 entries with a 20% split. The development and test datasets contain distinct CSV files for Tasks A, B, and C, with Tasks B and C's release being spaced out to prevent task-gaming. Table 2 shows the data-set statistics which has been used to train our models.

4 System Overview

For Task A, we use a binary classification approach where we classify whether a post is sexist or not sexist. For this, we use the CNN-BiLSTM model and the GPT-2 model.

The CNN-BiLSTM model uses a combination of convolutional and recurrent layers for feature extraction and sequence modeling, respectively. The input to the model is a sequence of words, which are first embedded into a dense vector representation. The embedded words are then passed through multiple convolutional layers to extract local features, which are then fed to a Bidirectional LSTM layer for sequence modeling. The output of the LSTM layer is then passed through a dense layer for binary classification.

On the other hand, the GPT-2 model is a transformer-based language model that uses many layers to extract contextual information from a given sequence. The input to the GPT-2 model is also a sequence of words, which are embedded into dense vector representations. The embedded words are then passed through multiple transformer layers to extract contextual information. The output of the last transformer layer is then passed through a dense layer for binary classification.

[1] https://gab.com/
[2] https://www.reddit.com/

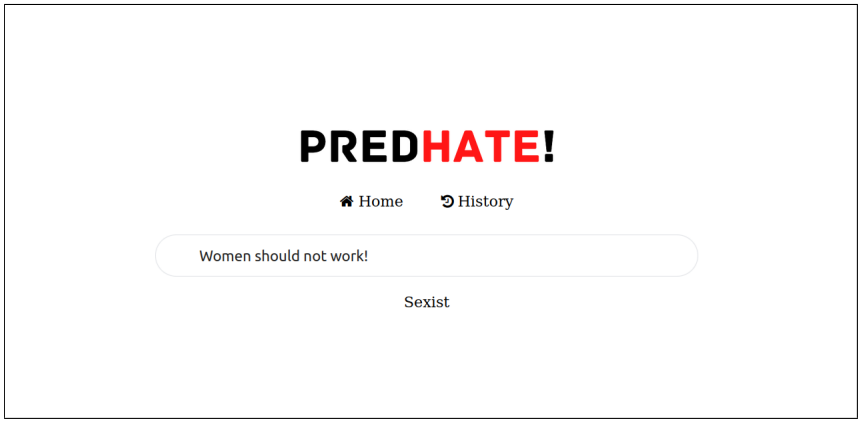

Figure 1: An overview of the GUI for classifying online sexism

For Task B, we classify the posts that are labeled as sexist by Task A into one of four categories: (1) Threats, (2) Derogation, (3) Animosity, (4) Prejudiced Discussions. For Task C, we classify the posts labeled as sexist by Task A into one of eleven fine-grained vectors. We use the GPT-2 models as in Task A for this classification task for both Task B & C. The only difference is that the output layer is modified to predict one of the four categories instead of binary classification for Task B and again modified to predict one of the eleven fine-grained vectors for Task C. Table 1 shows all the dataset's classes.

Towards GUI: We have designed a web app, where users can type any text they want. Based on our model's prediction it will show whether the text is sexist or not. Figure 1 shows an overview of our website 'PREDHATE!' which also provided a few features like checking the history and can classify multiple texts at a time. Code is openly available at: https://github.com/human71/predhate.

5 Experimental Setup

The SemEval Task 10 has been divided into 3 subtasks, for each subtask we have created individual models with respect to their respective classifications. Now the experimental setup has been divided into 2 parts as follows.

5.1 Task A

Again for Task A the experimental setup has been divided into 2 parts. one is using CNN-BiLSTM and the other is GPT-2.

CNN-BiLSTM: Convolutional neural networks (CNN) and bidirectional long short-term memory (BiLSTM) models are used in this paper to demonstrate an experimental setting for Task A of the Sexism Classification Challenge. Task A's objective is to determine if a particular post is sexist or not sexist.

We feed pre-trained word embedding into our CNN-BiLSTM architecture to build our model. Using the labeled training data supplied by the challenge organizers, we train our model. We divide the data into a 70/30 train-test split and use the training set for model development and verification.

Two primary parts make up the architecture of our CNN-BiLSTM system. The first is a CNN layer, which is intended to identify specific correlations and patterns in the input text. A BiLSTM layer, the second, is created to identify long-term dependencies in the input text. To create a binary classification output, the outputs of these two layers are merged and fed through a fully linked layer with a sigmoid activation function.

The Adam optimiser is used to train our models, while binary cross-entropy serves as our loss function. To avoid overfitting, we additionally apply dropout regularization to our fully connected layer. We use a grid search to fine-tune our hyperparameters over a range of values, and we choose the top-performing model based on validation accuracy.

GPT-2: We explore the use of GPT-2, a pre-trained language model, for the task of binary sexism detection (Task A). The use of pre-trained models has been shown to be highly effective in various natural language processing tasks, including text classification.

For this task, we use the pre-trained GPT-2 model with 768 parameters. We fine-tune the model on a given labeled dataset of 14,000 entries, which consists of 10,602 non-sexist posts and 3,398 sexist posts. We split the dataset into 70%

training data, 15% validation data and 15% testing data. We use the Adam optimizer with a learning rate of 1e-5 and a batch size of 8. We train the model for 5 epochs and save the model weights that perform the best on the validation set.

To use the fine-tuned model for predicting sexism in new text, we first tokenize the input text and pass it through the model. We evaluate the performance of the model on our heldout dev dataset of 2,000 entries, which were not labeled either as sexist or not sexist and produced F1-score. In addition, we compare the performance of our GPT-2 model with the previously mentioned CNN-BiLSTM model.

5.2 Task B and C

For Task B and C, we use GPT-2 with fine-tuning for multi-class classification. The pre-trained GPT-2 model will be fine-tuned on our training data to predict the category of sexism and fine-grained vector of sexism.

We will follow a similar preprocessing pipeline as we did for Task A, with the addition of one-hot encoding for the output labels. For Task B, we will use a four-class classification with the following categories: threats, derogation, animosity, and prejudiced discussions. For Task C, we will use an 11-class classification with the following vectors: objectification, sexual violence, physical violence, exploitation, gender discrimination, occupation discrimination, racist discrimination, stereotyping, body shaming, gendered insult, and identity-based insult. We will use the same training, development, and test sets as in Task A. We will fine-tune the pre-trained GPT-2 model on the training set, validate on the development set, and test on the test set. We will use the Adam optimizer with a learning rate of 1e-5, a batch size of 8, and a maximum sequence length of 128 tokens.

We will evaluate the performance of the GPT-2 model for Task B and C using micro-averaged F1 scores, as in Task A. We will compare the performance of the GPT-2 model to the baseline CNN-BiLSTM model, as well as to other state-of-the-art models for multi-class text classification.

We will also perform ablation studies to investigate the impact of different model architectures and hyperparameters on the performance of the GPT-2 model. Specifically, have experimented with different learning rates, batch sizes, sequence lengths, and pre-processing techniques to determine the optimal configuration for each task.

6 Experimental Results

In this part, we present GPT2 model's performance throughout training and testing compared to our other approach. We also compare the result with the other existing system submitted on SemEval-2023 Task 10.

6.1 Evaluation Method

The Macro F1 score is a single score that strikes a balance between Precision and Recall in a single number. Confusion matrices are often used metrics in classification tasks, and F1 offers a single score that strikes a balance between Precision and Recall in a single number. Values for precision and recall are computed, and there are respectively TP, FN scenarios that are True Positive and FN circumstances that are False Negative. Precision, recall, and F1 score are all split by class and each are given a macro average. The Macro F1 we typically observe is called the macro average. To calculate, we have to find out (1) Macro-average precision by calculating (P1+P2)/2 and (2) Macro-average recall by calculating (R1+R2)/2. Finally, just the harmonic mean of these two will constitute the Macro-average F1.

6.2 Evaluation of Test Data

For Task A we have tested on 1400 lines of data, as the data-set split was 80-10-10%, while, for task B and C the number was 340. We have fine-tuned the classification models for all three using GPT-2 as our base model. Additionally, we tested these classification models with three epoch values and three learning rate values. For the classification models with the best detection performance, the confusion matrix in Figure [3, 4, 5] shows the number of correctly and wrongly classified samples that were acquired. The number of true negatives (TN) in the confusion matrix refers to correctly identified instructions that are benign, whereas the number of true positives (TP) refers to correctly identified instructions that are harmful. The percentage of false positives shows innocent instructions mistaken for malevolent ones. The number of false negatives, on the other hand, shows dangerous instructions mistaken for benign ones. Table [4, 5, 6] shows the precision, recall and the Macro-F1 of the models. From that, we can see the macro average decreased along with the decrease of data-set from A to C.

The most important finding from Table [3, 4] is that the GPT-2 model specifically outperforms our CNN-BiLSTM model for Task A, hence we decided not using this approach for tasks B and C. For Task A, GPT-2 scored 75%, while for B and C, it went down to 40% and 20%.

label_sexist	precision	recall	**f1-score**
non-sexist	0.87	0.90	0.89
sexist	0.62	0.53	0.57
Macro F1	0.74	0.72	**0.73**

Table 3: Scores for Task A using CNN-BiLSTM

label_sexist	precision	recall	**f1-score**
non-sexist	0.88	0.89	0.88
sexist	0.63	0.60	0.62
Macro F1	0.75	0.75	**0.75**

Table 4: Scores for Task A using GPT-2

label_category	precision	recall	**f1-score**
1	0.69	0.30	0.42
2	0.51	0.77	0.61
3	0.47	0.28	0.35
4	0.30	0.12	0.17
Macro F1	0.49	0.37	**0.40**

Table 5: Precision, Recall, and F1 score for Task B

label_vector	precision	recall	**f1-score**
1.1	0.00	0.00	0.00
1.2	0.83	0.36	0.50
1.3	0.37	0.42	0.40
1.4	0.35	0.86	0.50
1.5	0.00	0.00	0.00
1.6	0.47	0.13	0.20
1.7	0.50	0.33	0.39
1.8	0.00	0.00	0.00
1.9	0.00	0.00	0.00
1.10	0.00	0.00	0.00
1.11	0.23	0.25	0.24
Macro F1	0.25	0.21	**0.20**

Table 6: Precision, Recall, and F1 score for Task C

If there are no predicted samples for a label, precision and F-score are considered ineffective and are set to 0. Because some of our labels appear in our y_pred but not all, our classifier is unable to predict some labels in our y_test. Here, y_test is the actual values that the dataset has, y_pred is the values our model predicted.

	Top Scorer	CNLP-NITS
Task A	0.87	0.74
Task B	0.73	0.41
Task C	0.56	0.20

Table 7: Semeval-23 Task 10 Leaderboard Macro F1

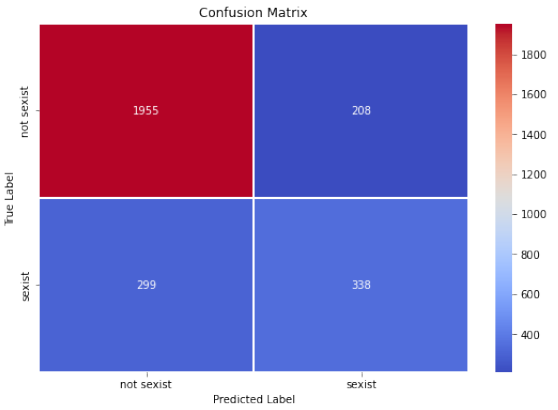

Figure 2: Confusion Matrix of Task A using CNN-BiLSTM

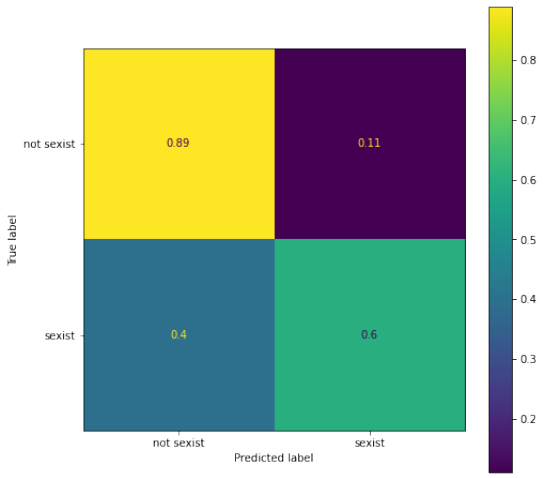

Figure 3: Confusion Matrix of Task A using GPT-2

6.3 Decisions and Problems

For evaluation, we use precision, recall, F1-score as metrics. The section 6.2 shows performance of our models on the development and test sets for all three tasks. We also conduct ablation tests to assess how different model elements affect performance as a whole. In the end, we compared our models with the ones that performed the best. The classification model built on GPT-2 that was adjusted

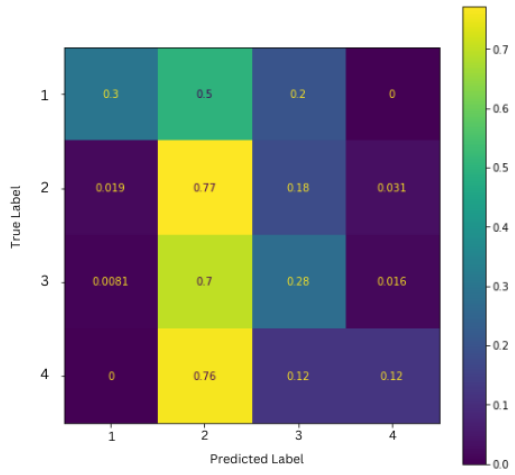

Figure 4: Confusion Matrix of Task B

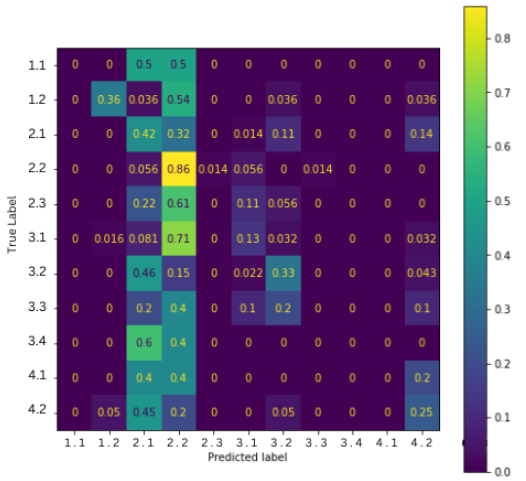

Figure 5: Confusion Matrix of Task C

functioned as expected, as indicated in Table 7. Yet, the results show that the model is biased and does not produce the right answers for a certain group, which may be why they perform worse than other groups.

7 Conclusion

As a conclusion, the identification and categorization of sexism in online content is a significant issue in modern culture, where social interactions take place more and more in online settings. Making online places more welcoming and secure for everyone can be achieved with the use of automated techniques for sexism detection and classification especially for women.

In this study, we provided an overview of the 20,000 sexism-labeled entries from the Gab and Reddit dataset as well as a hierarchy of three sexism detection and classification tasks, namely Task A (Binary Sexism Detection), Task B (Category of Sexism), and Task C (Fine-grained Vector of Sexism). The CNN-BiLSTM and GPT-2 models for Task A as well as the GPT-2 model for Tasks B and C were also given, along with their respective experimental settings.

Our findings demonstrated that the GPT-2 model performed better on Task A than the CNN-BiLSTM model, achieving greater accuracy and F1 scores. Also, the use of GPT-2 for Tasks B and C enabled the creation of more thorough and educational classifications, offering greater insights on the kind and degree of the sexism found in online content.

In the future, we hope that our work will stimulate additional investigation into the automated identification and categorization of sexism and that researchers in this field will find our dataset and experimental setups to be a helpful resource. We also hope that the knowledge we learn from this effort will contribute to the development of more welcoming and secure online environments for everyone.

8 Acknowledgments

We would like to thank National Institute of Technology Silchar for providing us the opportunity to participate in the SemEval-2023 Task 10, financial assistance, CNLP & AI laboratorys for the resources and research environment for smooth conduction of research and experiments.

References

Georgios Alexandridis, Iraklis Varlamis, Konstantinos Korovesis, George Caridakis, and Panagiotis Tsantilas. 2021. A survey on sentiment analysis and opinion mining in greek social media. *Information*, 12(8):331.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7383–7390.

Mary E. Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *International Conference on Applications of Natural Language to Data Bases*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv*.

Paweł Budzianowski and Ivan Vulić. 2019. Hello, it's gpt-2–how can i help you? towards the use of pretrained language models for task-oriented dialogue systems. *arXiv preprint arXiv:1907.05774*.

Simona Frenda, Bilal Ghanem, Manuel Montes y Gómez, and Paolo Rosso. 2019. Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *J. Intell. Fuzzy Syst.*, 36:4743–4752.

Dylan Grosz and Patricia Conde Céspedes. 2020. Automatic detection of sexist statements commonly used at the workplace. *ArXiv*, abs/2007.04181.

Jie He, Tao Wang, Deyi Xiong, and Qun Liu. 2020. The box is in the pen: Evaluating commonsense reasoning in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3662–3672.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Hannah Rose Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 Task 10: Explainable Detection of Online Sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.

Wei-Jen Ko and Junyi Jessy Li. 2020. Assessing discourse relations in language generation from gpt-2. *arXiv preprint arXiv:2004.12506*.

Campbell Leaper and Christia Spears Brown. 2008. Perceived experiences with sexism among adolescent girls. *Child development*, 79(3):685–704.

Gang Liu and Jiabao Guo. 2019. Bidirectional lstm with attention mechanism and convolutional layer for text classification. *Neurocomputing*, 337:325–338.

Ruixin Ma, Shoryu Teragawa, and Zhanjun Fu. 2020. Text sentiment classification based on improved bilstm-cnn. In *2020 Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, pages 1–4.

Leeja Mathew and VR Bindu. 2020. A review of natural language processing techniques for sentiment analysis using pre-trained models. In *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, pages 340–345. IEEE.

Marinella Paciello, Francesca D'Errico, Giorgia Saleri, and Ernestina Lamponi. 2021. Online sexist meme and its effects on moral and emotional processes in social media. *Computers in human behavior*, 116:106655.

Pulkit Parikh, Harika Abburi, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. 2021. Categorizing sexism and misogyny through neural approaches. *ACM Transactions on the Web (TWEB)*, 15(4):1–31.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Raul Puri, Ryan Spring, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2020. Training question answering models from synthetic data. *arXiv preprint arXiv:2002.09599*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, and Laura Plaza. 2020. Automatic classification of sexism in social networks: An empirical study on twitter data. *IEEE Access*, 8:219563–219576.

Alex Sherstinsky. 2020. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306.

Alejandro Vaca-Serrano. 2022. Detecting and classifying sexism by ensembling transformers models. *language*, 2:1.

Tania Verge. 2022. Too few, too little: Parliaments' response to sexism and sexual harassment. *Parliamentary Affairs*, 75(1):94–112.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.