Detecting Sexism and Its Severity in German Online Comments: Modeling Annotation Subjectivity with BERT and mBERT

Melanie Woodrow and Margot Mieskes
University of Applied Sciences, Darmstadt, Germany
melanie.woodrow@stud.h-da.de
margot.mieskes@h-da.de

Abstract

This study is based on the GermEval 2024 GerMS-Detect Shared Task[1], which collected user comments from an Austrian news platform and manually annotated them to reflect the presence and intensity of sexism. Its aim is to improve the detection of misogyny and sexism in German-speaking online forums. To address this task, we employed multiple classification strategies, including majority voting, the presence of annotator disagreement, and predicting the most frequent label. We fine-tuned BERT and its multilingual variant, mBERT, on the annotated dataset and evaluated their performance using precision, recall, and F1-score. In addition, we explored label distribution prediction to model annotator disagreement, using metrics such as Mean Squared Error, Kullback-Leibler divergence, and Jensen-Shannon divergence. This paper discusses challenges such as class imbalance and overfitting. Despite these, our results show that mBERT slightly outperforms BERT in binary classification tasks, though it faces difficulties with more complex distributional predictions. The findings highlight both the potential and limitations of pre-trained language models for detecting sexism in online discourse.

Content Warning: This paper includes examples of harmful and offensive language that do not reflect our opinions in any way.

1 Introduction

Sexist statements are found both in real life (Pingree et al., 1976) and in online environments (Plakoyiannaki et al., 2008). The European Institute for Gender Equality[2] defines sexism as a "hierarchical [way of thinking] that can be conscious and hostile, or it can be unconscious, manifesting itself as unconscious bias. Sexism can touch everyone, but women are particularly affected". The anonymity afforded by online platforms such as forums often facilitates and amplifies sexist behavior (Fox et al., 2015), leading to consequences that extend beyond the digital space (Skinner, 2022).

As the prevalence of online communication increases, so does the spread of discriminatory content. Identifying and moderating such content is critical to fostering respectful digital interactions. However, research shows that even recent large language models exhibit gender bias and are not reliably equipped to detect sexism (Mehrabi et al., 2021; Sheng et al., 2019). This emphasizes the need for dedicated models and datasets tailored for sexism detection.

In our work, we investigate the prediction of sexist content and its severity in German comments from an Austrian online news portal. We base our work on the GermEval 2024 GerMS-Detect Shared Task[3], provided by the Austrian Research Institute for Artificial Intelligence (OFAI), which includes two classification subtasks and an annotated dataset (GERMS-AT) (Krenn et al., 2024). The main goal of the annotations was to identify comments that make it difficult for women to participate in forum discussions and thus feel unwelcome.

We fine-tune the pre-trained BERT (Devlin et al., 2019) and mBERT models, which have not been explicitly fine-tuned for detecting sexism, on a set of binary, multi-class, and distributional classification tasks. To achieve this, we define multiple labeling strategies, such as majority voting, annotator disagreement, and distribution-based supervision. These strategies are detailed in Section 3 and reflect different ways to interpret subjective annotations. The trained models are evaluated on a test set without ground-truth labels, and results are reported us-

[1] https://ofai.github.io/GermEval2024-GerMS/

[2] https://eige.europa.eu/publications-resources/toolkits-guides/sexism-at-work-handbook/part-1-understand/what-sexism?language_content_entity=en

[3] https://ofai.github.io/GermEval2024-GerMS/

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 1: Long and Short Papers, pages 360–370, 2025

ing a range of classification and divergence-based metrics.

In addition, the distribution of labels should be predicted in order to reflect the diversity of opinions among the annotators. This study highlights the strengths and weaknesses of pre-trained language models in dealing with subjective and culturally nuanced social issues such as sexism.

The main contributions of this work are:

- We explore a variety of labeling strategies that capture different levels of subjectivity and disagreement in human annotations.
- We fine-tune and evaluate BERT and mBERT for both classification and label distribution prediction on German comments.
- We compare the performance of binary, ordinal, and distributional prediction tasks and analyze the impact of model choice and task complexity.
- We highlight the potential and limitations of pre-trained transformer models for detecting sexism and subjective language in online discourse.

The remainder of this paper is structured as follows: Section 2 presents relevant related work on sexism detection and subjective annotation strategies. Section 3 describes the dataset, labeling strategies, and model setup. In Section 4, we report experimental results across the multiple subtasks and strategies. Section 5 discusses the findings in the context of subjectivity and model performance, followed by Section 6 which concludes the paper and outlines directions for future research. Finally, we also outline the limitations and ethical considerations of this work at the very end of the paper.

2 Related Work

Early approaches to text classification relied on manual features such as bag-of-words, TF-IDF, and n-grams, and commonly used algorithms like Support Vector Machines, Decision Trees, or k-Nearest Neighbors (Joachims, 1998). While effective to a degree, these methods often struggle to capture contextual or implicit meaning in language. With the rise of transformer-based models such as BERT (Devlin et al., 2019), classification performance significantly improved across various NLP tasks. The following presented works highlight the success of using BERT and its variants for classifying different types of hate speech.

Das et al. (2023) demonstrated the effectiveness of fine-tuning RoBERTa for explainable sexism detection across multiple subtasks, including binary and fine-grained classification, as part of the SemEval-2023 EDOS task. Furthermore, Belbachir et al. (2024) explored integrating sentiment analysis with transformer models such as RoBERTa and LSTM for the SemEval-2023 Task, achieving higher F1-scores in sexism classification compared to other tested models. Additionally, Saleh et al. (2023) evaluated the performance of BERT against domain-specific embeddings for hate speech detection, further confirming BERT's effectiveness for binary classification in socially sensitive domains such as social media.

Sexism detection is inherently subjective, and different annotators may interpret content in varying ways (Almanea and Poesio, 2022). Several studies have therefore explored how to model disagreement and perspective in annotations (Almanea and Poesio, 2022; Akhtar et al., 2020). Label aggregation strategies, such as majority vote, union, or distribution modeling, are commonly used to approximate ground truth. Jiang et al. (2024) investigated how annotator attitudes influence automated classification of sexist content and showed that incorporating annotator metadata can improve the performance of large language models on Gender-Based Violence (GBV) related tasks.

Furthermore, sexism detection presents challenges distinct from general hate speech detection, as previous work shows higher false positive rates and less consistent performance when models are trained on counterfactual augmented data (Sen et al., 2022). Sexist language often manifests implicitly, through sarcasm or cultural slang, making it difficult for models to detect without contextual understanding and further complication classification (Belbachir et al., 2024).

These challenges highlight the need for models and evaluation frameworks that account for ambiguity, subjectivity, and disagreement, factors which are especially critical in the detection of nuanced social biases such as sexism.

3 Methods and Data

The following sections describe the dataset and annotation framework used in this study, along with

the labeling strategies applied for both classification and distributional modeling (Section 3.1). In Section 3.2 we then describe the pre-processing steps and model architectures (Section 3.3) used for fine-tuning BERT and mBERT, followed by our training setup, hyperparameter choices, and methods for addressing class imbalance (Section 3.4). Finally, in Section 3.5, we explain the evaluation metrics used to assess model performance across both subtasks.

3.1 Dataset and Labeling Strategies

As mentioned in Section 1, this work is based on the dataset and task definitions provided for the GermEval 2024 GerMS-Detect Shared Task (Krenn et al., 2024), which focuses on detecting sexism in user-generated comments from an Austrian online news portal. Each comment was annotated by up to ten annotators with an average of five, who assigned a label on a scale ranging from 0 ("no sexism") to 4 ("extreme sexism"). The training dataset contains approximately 4,500 comments and includes the annotations, specifically the labels given by annotators. The dataset also includes a unique ID for each comment and each annotator involved in the labeling process.

Due to the subjective nature of the task, no single "ground truth" label is provided. Instead, the dataset supports multiple labeling strategies derived from the annotation distributions.[4] We use the following five strategies:

bin_one : Label is 1 if at least one annotator selected a value greater than 0.

bin_all : Label is 1 only if all annotators assigned a value greater than 0.

bin_maj : Label is 1 if the majority of annotators selected any value above 0; 0 if the majority selected 0.

disagree_bin : Label is 1 if annotators disagreed on whether the comment was sexist (i.e., mix of 0 and >0).

multi_maj : Predict the majority label among the five ordinal values. In case of ties, any of the tied labels is considered correct.

The use of multiple labeling strategies enables a deeper understanding of how models handle different types of subjectivity in sexist content detection.

Strategies such as `bin_one`, which labels a comment as sexist if any annotator perceives it as such, are highly sensitive but may increase false positives. Conversely, `bin_all` demands full agreement among annotators, resulting in a stricter but potentially more conservative model.

Majority-based strategies like `bin_maj` aim to balance sensitivity and precision by reflecting the predominant perception. The `disagree_bin` strategy explicitly models cases of annotator disagreement, capturing instances where the presence of sexism is ambiguous or culturally interpreted differently.

Finally, `multi_maj` preserves ordinal information about the severity of sexism but poses challenges due to class imbalance and the subtle gradations between label values.

Our results show that models trained with looser strategies (e.g., `bin_one`) generally achieve higher recall, whereas stricter or distributional strategies (e.g., `multi_maj`) are more difficult for the models to optimize (see Table 3 and Table 4).

In addition to these single-label tasks, the shared task includes a second subtask: predicting the full distribution of labels assigned by annotators. Instead of training the model to predict a single "correct" label, this approach requires the model to output a probability distribution that matches the normalized annotation frequencies for each comment. This subtask captures the nuances and subjectivity of annotator perspectives more accurately. This approach better reflects the inherent subjectivity of sexism detection. We consider two distributional strategies: `dist_bin` collapses the ordinal annotations into a binary distribution (sexist vs. non-sexist), simplifying the task and reducing sparsity. In contrast, `dist_multi` preserves the full five-point ordinal scale, allowing the model to learn finer distinctions in perceived severity. We implement two strategies for this subtask:

dist_bin : Predict the normalized binary distribution by collapsing all values >0 into class 1. For example, [0, 0, 1, 1, 1] becomes [0.4, 0.6].

dist_multi : Predict the full five-class ordinal distribution, preserving all individual annotations.

The GERMS-AT training dataset consists of 4,486 user-generated comments from an Austrian online news platform. Each comment is annotated by an average of five annotators, with an average comment length of 33 tokens. Most comments are relatively short and show typical signs of informal forum communication.

[4] For details on the annotation see https://ofai.github.io/GermEval2024-GerMS/guidelines.html

Examples 1–3

Example 1
Comment: "Was soll das depperte Sternderl? *?"
Translation: "What's with the dumb little star? *?"
Explanation: The asterisk (*) is often used in German to signal gender inclusivity (e.g., "Leser*innen" for "readers of all genders"). The comment expresses ridicule toward this usage.
Label distribution: 0: 4, 1: 3, 2: 3, 3: 0, 4: 0
Derived labels: `bin_maj = 1`, `bin_one = 1`, `bin_all = 0`, `disagree_bin = 1`, `multi_maj = 0`

Example 2
Comment: "Im Chaos zuzeln sie heimlich den Staat aus..."
Translation: "In the chaos, they're secretly sucking the state dry..."
Explanation: "Turquoise" refers to the Austrian People's Party (VP), implying political critique rather than sexist content.
Label distribution: 0: 5, 1–4: 0
Derived labels: `bin_maj = 0`, `bin_one = 0`, `bin_all = 0`, `disagree_bin = 0`, `multi_maj = 0`

Example 3
Comment: "Mit der Fotze hat er sich keinen Gefallen getan..."
Translation: "He did himself no favor with that cunt..."
Explanation: The comment uses a strongly offensive gendered slur, contributing to its classification as severe sexism.
Label distribution: 0: 0, 1: 0, 2: 2, 3: 3, 4: 6
Derived labels: `bin_maj = 1`, `bin_one = 1`, `bin_all = 1`, `disagree_bin = 0`, `multi_maj = 4`

Figure 1: Illustrative examples (1–3) showing comment content, interpretation, and derived labels.

Statistic	Value
Total comments	4,486
Average comment length	33.0 tokens
Longest comment	366 tokens
Average annotations per comment	5.0

Table 1: Statistics of the GERMS-AT training dataset.

Statistic	Value
Total comments	1,512
Average comment length	32.51
Longest comment	146 tokens

Table 2: Statistics of the GERMS-AT test dataset.

In addition to the training data, the test set contains a total of 1,512 comments. The average token length per comment is similar to the training set at approximately 32.5 tokens. However, the longest comment in the test set is notably shorter, with a maximum of 146 tokens compared to 366 in the training data.

The examples in Figure 1 illustrate how different annotation strategies affect the interpretation of sexist content. These comments were taken from the provided training data set. A rough English translation and explanation (where appropriate) are provided for each example.[5]

3.2 Pre-processing and Tokenization

Minimal pre-processing is applied to preserve the original structure and expressions in the comments,

[5]All comment translations were produced by the authors for the benefit of non-German-speaking readers. The views expressed in the comments are those of the comment authors and do not reflect the views of the authors of this paper.

including colloquial language and informal punctuation. All text is lowercased and tokenization is performed using the standard BERT WordPiece tokenizer from the Hugging Face Transformers library (Wolf et al., 2020)[6]. Stopword removal or stemming is not used, based on previous findings that such steps may harm performance in contextual models (Alzahrani and Jololian, 2021).

To determine an appropriate maximum sequence length for input truncation, we analyzed the token length distribution of all comments in the training data. As shown in Figure 2, most comments are relatively short, with the majority containing fewer than 100 tokens. However, some comments are significantly longer. To preserve as much information as possible while avoiding excessive padding and memory usage, we set the maximum token length to 276, which covers about 99% of all comments in the training dataset.

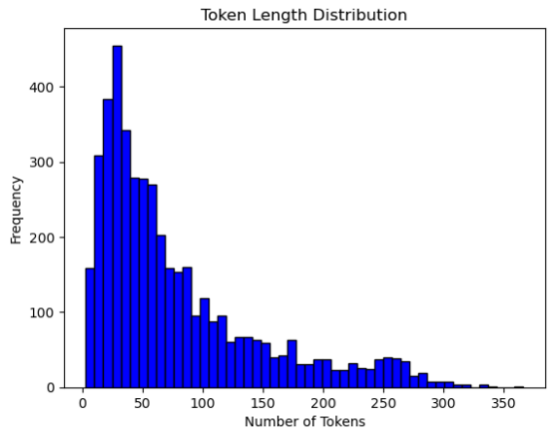

Figure 2: Distribution of comment lengths measured in BERT WordPiece tokens.

3.3 Model Architectures and Training

Although the dataset consists of German texts, `bert-base-uncased` was included as a baseline because uncased models can be more robust to inconsistent capitalization, which frequently occurs in user-generated content such as online comments. This helps mitigate noise from typos, informal writing, or inconsistent use of uppercase letters. Conversely, `bert-base-multilingual-cased` was chosen to preserve capitalization information in the multilingual setting, as German capitalization (especially for nouns) can carry important syntactic and semantic cues. This combination allows us to investigate whether retaining case information in multilingual training benefits performance compared to an uncased approach. For each classification task, a task-specific linear head is added to the [CLS] token output, followed by a softmax activation in the binary and multi-class cases.

For the distribution prediction task (Subtask 2), the model also outputs a five-class softmax probability distribution. This is compared to the gold label distribution using Kullback-Leibler (KL) divergence. This design allows the model to learn both the primary perception of sexism and the uncertainty or disagreement present in the annotations.

We decided to train using 5-fold stratified cross-validation. In each fold, 80% of the data is used for training and 20% for validation. The final evaluation is performed on a test set. All models are trained using the AdamW optimizer[7] with a linear learning rate schedule and warm-up steps.

3.4 Hyperparameters and Class Imbalance

We perform initial hyperparameter tuning using a validation split from one fold and keep the settings consistent across tasks to ensure comparability. We set the learning rate to $2e^{-5}$, batch size to 16, and use early stopping based on validation loss. Dropout is applied with a rate of 0.3.

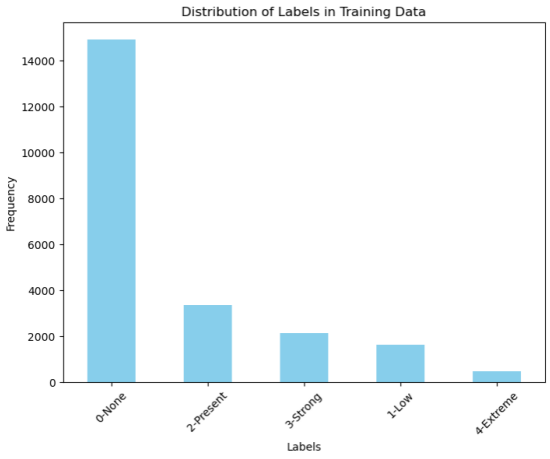

Figure 3: Distribution of the labels in the training dataset.

Some of the binary classification strategies (especially `bin_maj` and `bin_all`) exhibit class imbalance. Figure 3 shows the extent of the label distribution imbalance in the training data set which ultimately accounts for the class imbalance. This

[6] https://huggingface.co/docs/transformers/en/model_doc/bert

[7] https://pytorch.org/docs/stable/generated/torch.optim.AdamW.html

imbalance poses challenges for model training, particularly in the multi-class and distribution prediction tasks, where rarer categories may be underrepresented during learning. For this reason, we apply class weighting in the loss function or oversample the minority class during training. In the case of `multi_maj`, where the class distribution is skewed toward label 0, oversampling of rare labels is applied to each training fold individually.

In Subtask 2, distributional diversity is also limited by the frequency of certain label combinations. To address this, we increase the sampling rate for comments with rare or highly varied annotation patterns, improving the model's ability to generalize across different types of subjectivity.

3.5 Evaluation Metrics

Binary and multi-class tasks are evaluated using precision, recall, and macro-averaged F1-score. For the distributional prediction task, we compute the Kullback-Leibler divergence, Jensen-Shannon divergence and Mean-Squared-Error between the predicted and true annotation distributions. In the case of `multi_maj`, evaluation accounts for ties by treating any of the tied majority labels as correct.

4 Experiments and Results

We conducted experiments using the GERMS-AT dataset to evaluate the performance of BERT and mBERT on two of the subtasks introduced in Section 3.1: binary/multi-class classification and label distribution prediction. Hyperparameters were kept consistent for each subtask, as described in Section 3.3.

For the first subtask, we evaluated the models across five different labeling strategies: `bin_one`, `bin_all`, `bin_maj`, `disagree_bin`, and `multi_maj`. The models were assessed using precision, recall, and F1-score.

Strategy	Precision	Recall	F1-Score
bin_one	0.71	0.71	0.71
bin_all	0.74	0.83	0.77
bin_maj	0.69	0.67	0.67
disagree_bin	0.64	0.61	0.62
multi_maj	0.59	0.59	0.58

Table 3: Evaluation results for BERT (Subtask 1)

Overall, mBERT slightly outperforms BERT in binary classification tasks, particularly for `bin_one` and `bin_maj`. However, BERT achieves higher F1-scores in more complex multi-class tasks like `multi_maj`. Confusion matrices for all strategies of Subtask 1 can be found in Section A of the Appendix.

Strategy	Precision	Recall	F1-Score
bin_one	0.74	0.74	0.74
bin_all	0.77	0.67	0.71
bin_maj	0.74	0.73	0.73
disagree_bin	0.64	0.64	0.64
multi_maj	0.67	0.45	0.52

Table 4: Evaluation results for mBERT (Subtask 1)

In the second subtask, the goal was to predict the full label distribution of annotations using `dist_bin` and `dist_multi`. Table 5 presents the results using Mean Squared Error (MSE), Kullback-Leibler (KL) divergence, and Jensen-Shannon divergence (JSD).

Model	Type	MSE	KL	JSD
BERT	dist_bin	0.14	0.4	0.11
BERT	dist_multi	0.05	0.59	0.15
mBERT	dist_bin	0.12	0.36	0.09
mBERT	dist_multi	0.04	0.56	0.14

Table 5: Evaluation results for distribution prediction (Subtask 2)

It is important to note that Subtask 1 and Subtask 2 use different evaluation paradigms: while higher values of F1-score and accuracy in Subtask 1 reflect better performance, Subtask 2 employs distance- and divergence-based metrics (KL, JSD, MSE), where lower values correspond to better alignment between predicted and true distributions. The results indicate that mBERT consistently achieves lower error and divergence values across both distribution types. While the improvements over BERT are modest, they suggest mBERT is better suited to capture the subtle differences in perceived sexism among annotators. Modeling full distributions (`dist_multi`) is more challenging due to increased label complexity and class imbalance, but it offers a richer representation of annotator disagreement. Our results confirm that predicting multi-class distributions yields higher divergence values than binary ones, indicating greater difficulty for the models.

Despite these challenges, distribution prediction provides a valuable framework for capturing uncertainty and subjective variance in socially sensitive

classification tasks.

5 Discussion

The experimental results demonstrate that both BERT and mBERT can be effectively fine-tuned to detect sexist content in German-language online comments. mBERT generally achieves slightly higher performance in binary classification tasks, particularly under strategies like `bin_one` and `bin_maj`, which focus on detecting the presence of any sexist expression. However, when faced with more complex tasks such as predicting label distributions or modeling fine-grained severity (`multi_maj`), BERT shows more stable and balanced performance.

The difficulty of the distribution prediction task is reflected in higher divergence scores for both models, especially for `dist_multi`. This suggests that capturing nuanced, subjective differences between annotators remains a substantial challenge for current pre-trained models. Our findings support previous observations in the literature that socially sensitive language tasks, particularly those involving subjective judgments, are inherently harder to model.

Moreover, the choice of labeling strategy has a notable impact on model behavior. Strategies prioritizing sensitivity (`bin_one`) lead to higher recall but lower precision, while more conservative strategies (`bin_all`) yield the opposite trend. This trade-off is important when considering real-world applications: systems prioritizing inclusivity might prefer high-recall models, while moderation systems aiming to minimize false positives might favor high-precision models.

Although mBERT performs slightly better in general, the differences are not dramatic. Expanding the dataset, incorporating additional contextual features (e.g., thread structure or user metadata), or using newer architectures like RoBERTa could further improve performance.

To better illustrate the model behavior and the role of annotation subjectivity, we selected three representative examples from the test dataset and their respective true labels from the targets dataset for the `bin_maj` strategy.

For the example in Figure 4 the model correctly identifies the absence of sexism, aligning it with the majority annotation. Figure 5 shows a false positive. The sarcastic tone may have led the model to misclassify the comment, despite the absence of explicitly sexist content. The statement shown in Figure 6 contains a gender-related stereotype but may not be universally interpreted as sexist. The disagreement among annotators illustrates the subjectivity of labeling and highlights the potential value of modeling label distributions instead of relying solely on single-label predictions.

Successful Prediction

Kämpferischer . . . das heisst erzwingen . . .
Das bedeutet, ihr wollt Leute zwingen euch ein sorgenfreies Leben zu finanzieren.
Alles klar . . .

(Translation: Militant... that means to force...
That means you want to force people to finance you a carefree life.
All right then...)

Prediction: 0 **True Label:** 0

Figure 4: Example of a successful prediction

Failed Prediction

Wieso “nett”? Ist Dummheit ident mit “nett”?

(Translation: Why “nice”? Is stupidity identical with “nice”?)

Prediction: 1 **True Label:** 0

Figure 5: Example of a failed prediction (false positive).

Ambiguous Case

Das sind Schätzungen, denn Männer sind ja zu feig, sowas anzuzeigen, nicht wahr?

(Translation: Those are just estimates, because men are too cowardly to report things like that, right?)

Prediction: 0 **True Label:** 1

Figure 6: Example of an ambiguous case with subjective interpretation.

The statement shown in Figure 6 contains a gender-related stereotype but may not be universally interpreted as sexist. The disagreement among annotators illustrates the subjectivity of labeling and highlights the potential value of modeling label distributions instead of relying solely on single-label predictions.

To contextualize our results, we compare them to the scores reported by teams participating in the GermEval 2024 GerMS-Detect Shared Task[8]. Table 6 lists the average F1-scores and Jensen-Shannon divergences (JSD) of all teams who submitted a paper, along with our average scores which represent the average performance across all evaluated labeling strategies and both model variants (BERT and mBERT).

Team	F1-Score	JSD
THAugs	0.642	–
FICODE	0.641	0.354
Quabynar	0.611	0.292
GDA	0.597	0.301
pd2904	0.483	0.388
Ours	**0.669**	**0.123**

Table 6: Comparison with submitted systems from the GermEval 2024 GerMS-Detect Shared Task. Our models outperform all other teams in both classification and distribution tasks.

Our models achieved an F1-score of 0.669 and a JSD of 0.123, outperforming all teams that submitted a publication. While several teams explored advanced architectures and ensembles, such as training a separate model on each annotator, our results show that precise fine-tuning can yield strong performance using standard BERT-based models.

While our model achieved higher F1-scores and lower divergence values than all teams who submitted their publication, these comparisons are based on single evaluations on the test set.

6 Conclusion

In this paper, we investigated the detection of sexist content in German-language online comments using pre-trained transformer models. Based on the GermEval 2024 GerMS-Detect Shared Task, we explored various labeling strategies to account for the subjectivity of human annotations and fine-tuned BERT and mBERT for both classification and distributional prediction tasks.

Our results show that mBERT slightly outperforms the original BERT model in binary classification tasks, while BERT demonstrates more stable performance in distribution tasks. Additionally, we found that modeling annotation disagreement explicitly, using both binary and full distribution predictions, allows models to better capture the nuances of perceived sexism.

These findings highlight both the potential and limitations of using large language models for socially sensitive tasks. Future work could explore the use of more advanced transformer architectures, interpretability methods to explain predictions, or active learning strategies to better deal with subjective annotations.

Limitations

While our findings show that pre-trained language models such as BERT and mBERT can be fine-tuned to detect sexist content in online comments, this study has several limitations.

First, the dataset used is relatively small compared to other large-scale NLP benchmarks such as GLUE (Wang et al., 2018) or XNLI (Conneau et al., 2018). Although we used cross-validation and oversampling to improve robustness, the limited size and source diversity of the data may affect generalizability to other platforms, languages, or domains.

Second, the annotation process is inherently subjective. Although we modeled disagreement and used distributional labels, the presence of annotator bias or cultural interpretation of sexism may still affect the quality of the labels.

Third, we fine-tuned existing models without applying extensive hyperparameter optimization or exploring more recent architectures such as RoBERTa or DeBERTa. More advanced models or training strategies might improve performance, especially in the distributional prediction task.

Finally, our approach is not immediately interpretable. Understanding why a comment was classified as sexist or how the model interprets annotator disagreement remains an open question, which is especially relevant for sensitive applications like moderation. For this purpose, a label for the type of sexist content is required.

[8]Results are published in the proceedings section at https://ofai.github.io/GermEval2024-GerMS/workshop

Ethics Statement

This paper addresses the detection of sexism in online comments, a task that involves sensitive and subjective social judgments. We recognize that both the annotations and the model predictions reflect human interpretations of what constitutes sexist language, which may vary across cultures, identities, and personal experiences. To account for this, we explicitly model annotator disagreement and examine multiple labeling strategies, including distributional predictions.

The dataset used in this study and its annotations were provided as part of the GermEval 2024 Shared Task, which we used as is.

Our models are trained and evaluated for research purposes and are not intended for immediate deployment in real-world moderation systems without further validation. While we aim to contribute toward safer and more inclusive online environments, we also advise against the blind application of such models, especially without considering fairness, interpretability, and the potential for unintended consequences in automated content moderation.

References

S. Akhtar, V. Basile, and V. Patti. 2020. Modeling annotator perspective and polarized opinions to improve hate speech detection. In *In Proceedings of the AAAI conference on human computation and crowdsourcing*, pages 151–154.

D. Almanea and M. Poesio. 2022. Armis-the arabic misogyny and sexism corpus with annotator subjective disagreements. *In Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2282–2291.

E. Alzahrani and L. Jololian. 2021. How different text-preprocessing techniques using the BERT model affect the gender profiling of authors. *arXiv preprint arXiv:2109.13890*.

F. Belbachir, T. Roustan, and A. Soukane. 2024. Detecting online sexism: Integrating sentiment analysis with contextual language models. *AI*, 5(4):2852–2863.

A. Conneau, R. Rinott, G. Lample, A. Williams, S. Bowman, H. Schwenk, and V. Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on EMNLP*.

A. Das, N. Raychawdhary, T. Bhattacharya, Ge. Dozier, and C. D. Seals. 2023. Au_nlp at semeval-2023 task 10: Explainable detection of online sexism using fine-tuned roberta. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 707–717.

J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*.

J. Fox, C. Cruz, and J. Y. Lee. 2015. Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media. *Computers in Human Behavior*.

A. Jiang, N. Vitsakis, T. Dinkar, G. Abercrombie, and I. Konstas. 2024. Re-examining sexism and misogyny classification with annotator attitudes. *arXiv preprint arXiv:2410.03543*.

T. Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Machine Learning: ECML-98*, pages 137–142. Springer.

B. Krenn, J. Petrak, M. Kubina, and C. Burger. 2024. Germs-at: A sexism/misogyny dataset. In *Proceedings of LREC-COLING 2024*.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.

S. Pingree, R. Hawkins, M. Butler, and W. Paisley. 1976. A scale for sexism. *Journal of Communication*.

E. Plakoyiannaki, P. Mathioudaki, K.and Dimitratos, and Y. Zotos. 2008. Images of women in online advertisements of global products: Does sexism exist? *Journal of Business Ethics*.

H. Saleh, A. Alhothali, and K. Moria. 2023. Detection of hate speech using bert and hate speech word embedding with deep model. *AAI*, 37(1):2166719.

I. Sen, M. Samory, C. Wagner, and I. Augenstein. 2022. Counterfactually augmented data and unintended bias: The case of sexism and hate speech detection. *arXiv preprint arXiv:2205.04238*.

E. Sheng, K. Chang, P. Natarajan, and N. Peng. 2019. The woman worked as a babysitter: On biases in language generation. *EMNLP*.

C. Skinner. 2022. The consequences of online misogyny. *Feminist Media Studies*.

A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP*.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Davison. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on EMNLP: system demonstrations*, pages 38–45.

Appendix

A Confusion Matrices

To provide deeper insight into the classification behavior of our models, we include the confusion matrices for both BERT and mBERT on all labeling strategies in Subtask 1.

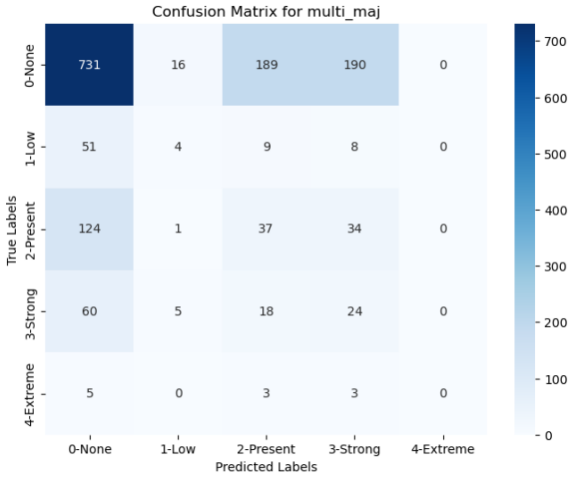

Figure 7: Confusion matrix for multi_maj (BERT)

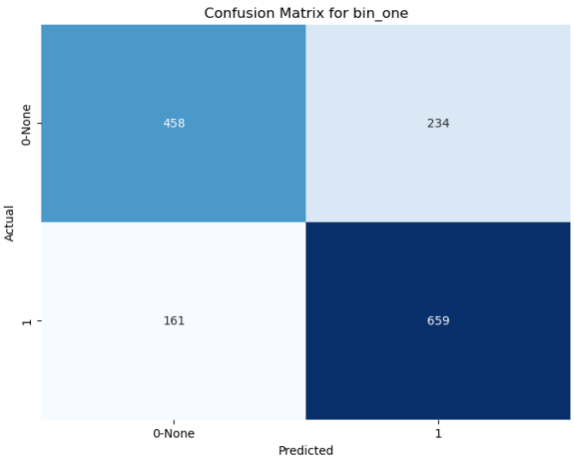

Figure 8: Confusion matrix for bin_one (BERT)

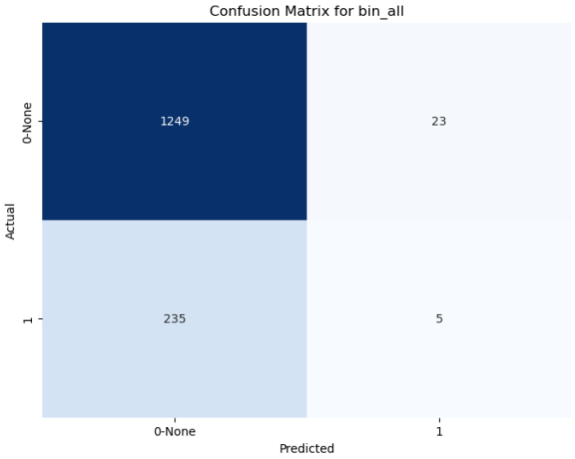

Figure 9: Confusion matrix for bin_all (BERT)

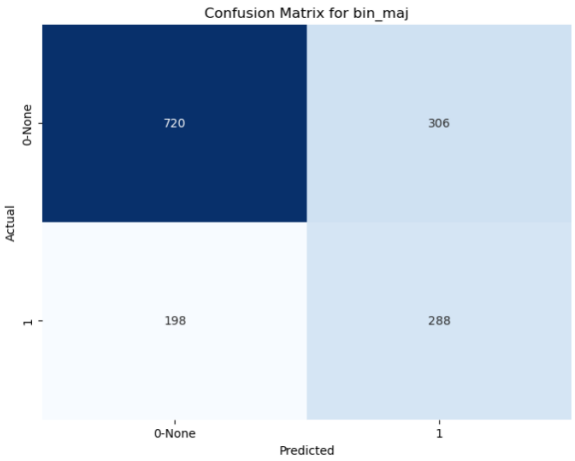

Figure 10: Confusion matrix for bin_maj (BERT)

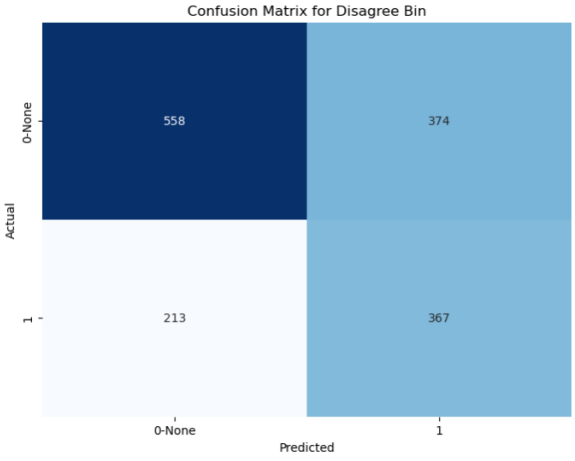

Figure 11: Confusion matrix for disagree_bin (BERT)

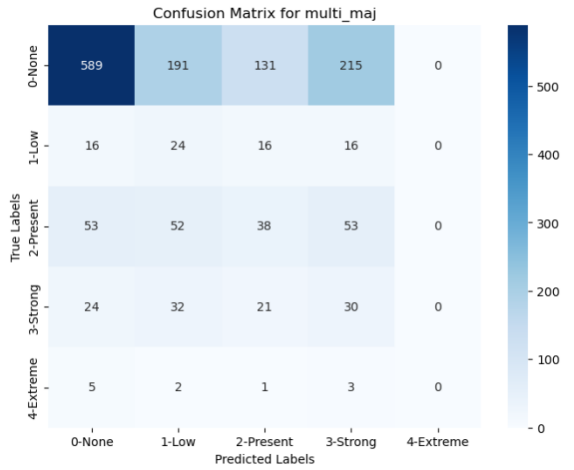

Figure 12: Confusion matrix for multi_maj (mBERT)

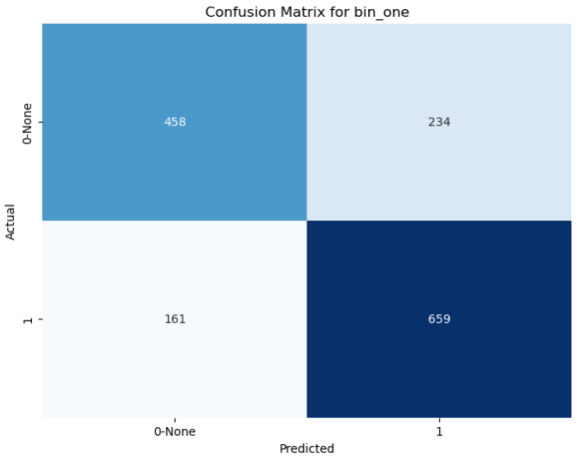

Figure 13: Confusion matrix for bin_one (mBERT)

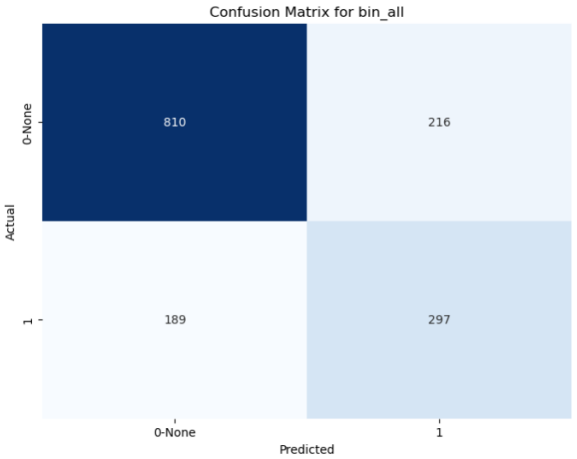

Figure 14: Confusion matrix for bin_all (mBERT)

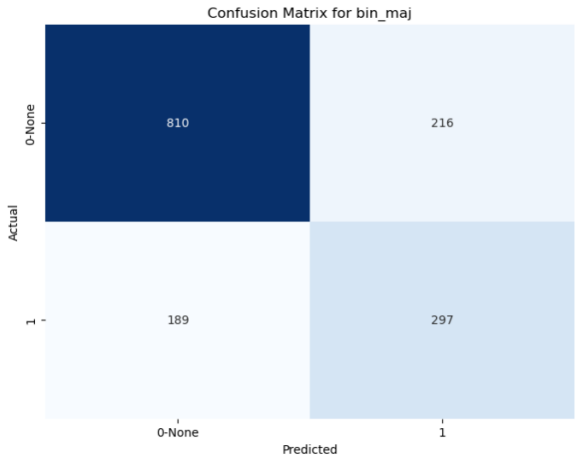

Figure 15: Confusion matrix for bin_maj (mBERT)

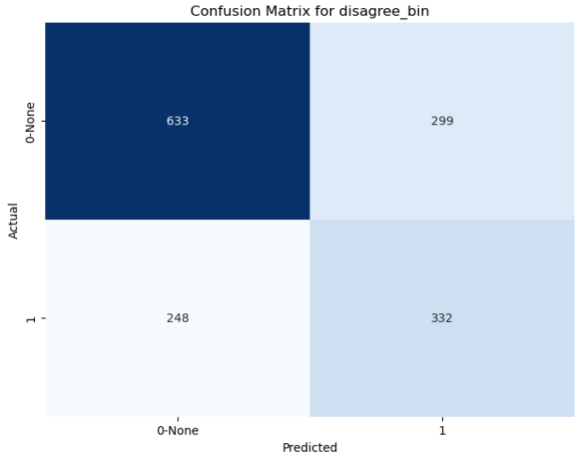

Figure 16: Confusion matrix for disagree_bin (mBERT)