pd2904 at GermEval2024 (Shared Task 1: GerMS-Detect): Exploring the Effectiveness of Multi-Task Transformers vs. Traditional Models for Sexism Detection (Closed Tracks of Subtasks 1 and 2)

Pia Donabauer
Information Science
University of Regensburg
pia.donabauer@stud.uni-regensburg.de

Abstract

The rise of social platforms has led to an increase in hateful, racist and sexist comments, impacting mental health and well-being. Detecting sexist texts automatically is a crucial first step to addressing this issue. This paper describes two approaches for the GermEval2024 GerMS-Detect Shared Task 1 on identifying sexist and misogynistic multi-annotated comments. Given the challenge of imbalanced data, the effectiveness of a multi-task transformer BERT model with TF-IDF weights is compared against traditional machine learning models. After training each model with individually optimized hyperparameters, 5-fold cross-validation showed that the traditional approach appears to perform better than the transformer model in several metrics. Given these results, the solution based on traditional models was submitted, achieving an F_1 score of 0.483 for subtask 1 and a Jensen-Shannon distance of 0.338 for subtask 2 in the final submission. The code is publicly available on GitHub [1].

1 Introduction

Although the rapid development of technology and social network sites has facilitated global communication, the anonymity online has enabled the unpunished expression of hateful, racist and sexist discourses (Rodríguez-Sánchez et al., 2020). This leads users to engage in behaviours they would avoid in face-to-face interactions, known as the *online disinhibition effect* (Wright et al., 2019). As a result, insults and harassment, such as sexism and misogyny, are prevalent in social media and online fora. Sexism is defined as prejudice, stereotyping, or discrimination based on sex, while misogyny refers to the hatred or dislike of women (Rodríguez-Sánchez et al., 2020). The variety and volume of language used in online platforms make it challenging to manage these issues (Bellmore et al., 2015). As victims of online sexist insults suffer from low self-esteem, emotional distress, and other negative emotions (Felmlee et al., 2020), it is crucial to develop language-specific models for sexism detection to foster a safer online environment. Given this real-world problem, GermEval2024 GerMS-Detect aims to identify sexism and misogyny in German-language comments from an Austrian online newspaper. The texts have been labeled by multiple human annotators, often with differing opinions. The submissions described in this paper are limited to the competition's closed track, which prohibits the use of additional data labeled for sexism, models or embeddings trained on data labeled for sexism, and Large Language Models (LLMs). This constraint requires the exploration of alternative solutions. Therefore, this paper elaborates on two approaches for detecting sexism in online fora: 1) several conventional machine learning classifiers, including Random Forest, Extreme Gradient Boosting (XGBoost), Light Gradient-Boosting (LightGBM), Support Vector Machines (SVM), and CatBoost, and 2) a deep learning transformer-based method, specifically a multi-task model using Bidirectional Encoder Representations from Transformers (Devlin et al., 2018) (BERT) with the integration of term frequency–inverse document frequency (TF-IDF). Multiple traditional models are experimented with, as performance across tasks may differ (Panwar and Mamidi, 2023). Evaluation shows that while the transformer-based approach yields promising results, hyperparameter-tuned conventional models tailored to each annotator turn out to perform better on predicting sexism in these experiments. This paper describes the implemented models for the GermEval2024 shared task, discusses possible reasons for the performance differences, and highlights the importance of developing effective detection methods to mitigate sexism and misogyny in online spaces.

[1] https://github.com/piadonabauer/GermEval2024

2 Background

GermEval2024's shared task focuses on detecting sexism and misogyny in texts posted in German-language to the comment section of an Austrian online newspaper.

2.1 Task Description

The shared task is divided into two subtasks:

Subtask 1: Predict a binary label indicating the presence or absence of sexism in four different ways, based on the original grading of the texts by several annotators; also predict the majority grading assigned by annotators. Evaluation is based on the macro-averaged F_1 score.

Subtask 2: Predict binary soft labels, based on the different opinions of annotators about the text; predict the distribution of the original gradings by annotators. Evaluation uses Jensen-Shannon distance to compare predicted and actual distributions.

Both subtasks are organized into closed tracks, where only the provided dataset may be used and advanced approaches such as LLMs are prohibited, and open tracks, where all materials and methods are allowed. Participation in this paper is limited to the closed track.

2.2 Annotations

The dataset is annotated by a varying subset of ten annotators using numeric classes ranging from 0 to 4, with 0 = not sexist, 1 = mildly sexist, 2 = sexist, 3 = strongly sexist, and 4 = extremely sexist. However, while the annotation guidelines[2] define what types of sexism and misogyny should be annotated, there are no rules about the severity, resulting in annotations reflecting personal judgments.

2.3 Dataset Exploration

GermEval 2024's labeled dataset in German-language consists of 5998 entries, with an unlabeled dataset of 1986 entries for competition submission. One data example, along with its annotations, is displayed in Table 1.

Corpus statistics show variation in the length of data points, ranging from 1 to 173 words. On average, each data point contains approximately 32.9 words, with a median length of 23.0 words. Figure 1 and Figure 2 provide insights into the distributions of annotators and labels. The imbalance in label distribution is apparent, with label 0 (non-sexist) being the most prevalent category. In Figure 1, the leftmost red bar represents 65% of all data points, indicating missing annotations, as not every annotator labeled every data point. Additionally, a few annotators made limited contributions by providing fewer than 2000 annotations, as shown in Figure 2.

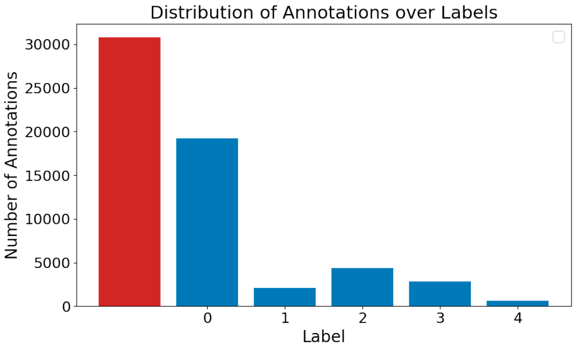

Figure 1: Distribution of labels given by all annotators collectively. The red bar visualizes missing annotations, since not all annotators labeled every data point.

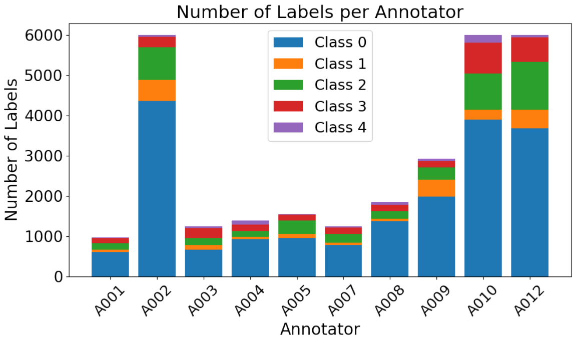

Figure 2: Distribution of labels given by each annotator individually.

Lastly, pairwise agreement among annotators was assessed using Krippendorff's Alpha. The highest agreement between two annotators was 0.043, suggesting highly diverse labeling strategies. Therefore, clustering annotators based on their agreements was not feasible.

3 Related Work

Extensive research has been conducted in the field of sexism prediction, multi-task frameworks, and data augmentation. Therefore, this section will primarily focus on recent concepts closely related to the competition.

[2] https://ofai.github.io/GermEval2024-GerMS/guidelines.html

German	Mit der Fox×÷e [sic] hat er sich keinen Gefallen getan. Ja, ich weiß der Ausdruck ist eigentlich nicht forums tauglich.									
English	He didn’t do himself any favours with that c×÷t [sic]. Yes, I know the expression is not really suitable for a forum.									
Annotations										
Annotator ID	01	02	03	04	05	07	08	09	10	12
Label	4	4	4	4	3	4	2	3	4	2

Table 1: An example comment in German and English-language with its corresponding labels.

3.1 Multi-Task Learning

Advances in deep neural networks have enabled multi-task models to learn multiple tasks simultaneously, sharing parameters across tasks to improve training efficiency and performance (Xu et al., 2022). Zhou (Zhou, 2023) used multi-task learning to approach the hierarchical classification of sexism by pre-training RoBERTa and DeBERTa models on 2 million data points, resulting in boosted performance of their models. In other works, multi-task models were used to address the issue of disagreement among annotators for multi-labeled text. Davani et al. (Davani et al., 2022) proposed a multi-annotator architecture in order to preserve the internal consistency of each annotator’s labels. Their multi-task model includes a fully connected layer explicitly fine-tuned for each annotator, predicting each annotator’s judgments as separate subtasks after being trained on 27k and 50k data points, respectively.

3.2 Data Augmentation

Augmenting new data is the synthesis of existing training data, aiming to improve the performance of a downstream model (Wong et al., 2016). Due to participating in the closed track, the focus of this work will be on traditional augmentation methods (Schmidhuber and Kruschwitz, 2024). For instance, Butt et al. (Butt et al., 2021) applied *Back Translation* for data augmentation using the *deep-translator* library. By translating Spanish and English data into German and then back to their respective languages, the identification of sexism could be enhanced. Furthermore, to address the issue of class imbalance within their dataset, Martinez et al. (Martinez et al., 2023) employed *Random Oversampling* to replicate minority classes with slight variations. Other research applied multiple strategies (Mohammadi et al., 2023), such as performing *Synonym Replacement*, the replacement of words with their synonyms, *Random Word Swapping*, randomly swapping pairs of words in the text, and *Random Character Insertion*, randomly inserting characters into words.

3.3 Summary

Drawing inspiration from these studies, the approach in this work adopts a multi-task learning framework inspired by Zhou’s model (Zhou, 2023), where the different tasks correspond to the different subtasks of the competition, but with fewer data points. The multi-annotator architecture proposed by Davani et al. (Davani et al., 2022) is integrated, leveraging the sharing of knowledge in initial layers to enhance generalization. Additionally, data augmentation methods are employed, specifically *Back Translation* and *Synonym Replacement* inspired by previous research, as augmentation has shown performance improvements. This combined approach is designed to take advantage of the strengths of these previous models.

4 Experimental Setup

As annotator disagreement may capture important nuances, all annotators’ judgements were treated as separate tasks within the multi-annotator architecture. The described approach encompassed two main strategies: a multi-task transformer fine-tuning framework, where each task corresponds to predicting labels from individual annotators, and a baseline comparison involving the individual training of conventional machine learning models tailored to each annotator.

To optimize model performance, a hierarchical classification was adopted, initially predicting binary labels followed by multi-class prediction on texts categorized as sexist.

4.1 Materials & Methods

Due to significant imbalance in label distribution, methods for data balancing[3] were explored, such as the integration of class weights for each annotator's labels, and implementation of Focal Loss. The latter approach incorporates disagreement among annotators into the loss function during training, inspired by Plank et al. (Plank et al., 2014). While in these experiments class weights enhanced model performance, utilizing Focal Loss did not yield improved results in this setting, thus it was not employed.

Further experimentation involved feature engineering of text vectorization, incorporating lexical features, and testing various transformer models. Preprocessing steps such as lemmatization, stemming, and stop word removal harmed model performance, confirming previous work (Xu, 2022), hence the data was preserved in its original form.

In order to address the little amount of data to train a transformer model with multiple heads, especially since the number of data points for some annotators was less than 2000, data augmentation was performed. Additionally, the training of the conventional models benefited from the availability of more data. The intention of augmenting data was not to improve class balance, thus downsampling of most frequent classes was not performed.

Data Augmentation

The provided dataset was expanded from 5998 to 17'913 entries using two augmentation techniques, namely *Back Translation* and *Synonym Replacement*, which were mainly found in recent works (Butt et al., 2021; Mohammadi et al., 2023).

- **Back Translation**: Utilizing Helsinki NLP models (de-en[4] and en-de[5]), sentences were translated to English and then back-translated to German. Duplicates resulting from translations were removed.

- **Synonym Replacement**: Replacing tokens with synonyms using vectors for the German language from fasttext[6], filtering synonyms based on original POS tags. The words *woman*, *women*, *man* and *men* were kept for contextual relevance.

Example augmentations for both techniques are displayed in Table 2.

4.2 Classification Models

For model training, the shuffled dataset was split into a training (85%, N=15'226) and testing split (15%, N=2'687). Other than the methods described, we did not apply any techniques to account for class, annotator, or augmented data balance. Final model training was performed on 100% of the data. The code was developed using the PyTorch framework.

4.2.1 Transformer for Multi-Task Learning

Despite having less training data for transformers compared to previous research, BERT was fine-tuned for multi-task classification, expecting this approach to benefit from sharing knowledge among layers and thus enhancing robustness and generalization (Hashimoto et al., 2016; Davani et al., 2022).

Architecture: Training on various transformer architectures, such as *bert-base-german-cased*[7], *bert-base-multilingual-cased*[8] and *xlm-roberta-large-finetuned-conll03-german*[9] from HuggingFace, resulted in *bert-base-german-cased* showing the best performance as the backbone. The model architecture of BERT was modified to process CLS token embeddings through newly introduced shared dense layers, facilitating dimensionality reduction and feature extraction via ReLU activations, dropout regularization (0.2), and batch normalization. Following the implementation of ten annotator-specific output heads, utilizing sigmoid activation for binary tasks and softmax for multi-class tasks, TF-IDF scores were integrated. Inspired by Chen et al. (Chen et al., 2020), during the forward pass, the CLS token output from BERT was multiplied by TF-IDF weights specific to the training data, which were precomputed and stored in dictionaries. This approach allows the model to benefit from BERT's contextual embeddings and the importance of individual terms as captured by

[3] https://datascientest.com/en/management-of-unbalanced-classification-problems-ii

[4] https://huggingface.co/Helsinki-NLP/opus-mt-de-en

[5] https://huggingface.co/Helsinki-NLP/opus-mt-en-de

[6] https://fasttext.cc/docs/en/crawl-vectors.html

[7] https://huggingface.co/google-bert/bert-base-german-cased

[8] https://huggingface.co/google-bert/bert-base-multilingual-cased

[9] https://huggingface.co/FacebookAI/xlm-roberta-large-finetuned-conll03-german

Back Translation	**German**	**English**
Original	"Ich **habe wahnsinnige** Kopfschmerzen!" Mädchen - neue generation - Angst vor **eh** fast allem....	"I **have** a **terrible** headache!" Girls - new generation - afraid of almost everything...
Augmented	"Ich **hatte verrückte** Kopfschmerzen!" Mädchen - neue Generation - Angst vor fast allem **sowieso**...	"I **had** a **crazy** headache!" Girls - new generation - afraid of almost everything **anyway**...
Synonym Replacement	**German**	**English**
Original	Das **schöne Gesicht der** Frauenquote	The **beautiful face** of the women's quota
Augmented	Das **wunderschöne Antlitz die** Frauenquote	The **wonderful countenance** of the women's quota

Table 2: Examples of comments in their original form and their variations through the data augmentation techniques of *Back Translation* and *Synonym Replacement*.

TF-IDF. The loss function accounts only for available labels provided by annotators, ignoring any missing values.

Training: The multi-task learning approach uses a shared BERT backbone and dense layers trained collectively across all tasks. Each annotator has a specific output head for predicting their annotations, trained simultaneously. Loss is calculated separately for each annotator's head using the appropriate loss function. The total loss for each training step is the sum of the losses from all heads, used to update both the shared BERT backbone and annotator-specific heads.

Hyperparameter tuning was conducted to identify optimal values, including the learning rates and number of epochs. Stochastic Gradient Descent optimization with 10% warm-up steps, a cosine weight decay of 1e-4, a batch size of 16, and a maximum sequence length of 64 was used. For binary classification, the tuning process determined a learning rate of 5e-3 for 6 epochs, while for multi-task classification, it identified a learning rate of 1e-2 for 7 epochs.

Feature Engineering: Beyond TF-IDF scores, additional features (e.g. sentiment analysis, token length, and punctuation ratios) did not improve performance and were therefore excluded from the final solution.

4.2.2 Conventional Machine Learning Models

The baseline comparison involves an intuitive approach for multi-annotator models, where several conventional classifiers are trained, each one individually on the labels provided by a single annotator. Given performance variations among models in sexism detection observed by Panwar et al. (Panwar and Mamidi, 2023), multiple traditional model architectures including Random Forest, SVM, XGBoost, LightGBM, and CatBoost were explored. Hyperparameter tuning and feature engineering using CountVectorizer, TfidfVectorizer, and transformer methods were conducted for each annotator, with training enhanced by class weights.

5 Results

During training of the traditional models for binary prediction, the choice of model for each annotator was varied with all models (Random Forest, LightGBM, XGBoost, SVM, and CatBoost) being employed. The most frequently used one was XGBoost, selected four out of ten times. Vectorization using the *bert-based-german-cased* model showed the best results seven out of ten times. For multi-class labeling, only the models Random Forest, XGBoost, and LightGBM were deployed, with Random Forest and XGBoost being the most common, each selected four out of ten times. The vectorization techniques used most frequently this time were both Transformer and CountVectorizer, each used four out of ten times. Detailed assignments and hyperparameter tuning results can be found in the code.

Due to time constraints, model evaluation was based on accuracy, precision, recall, and F_1 score, rather than using the specified evaluation metrics for the subtasks. Both the BERT model and conventional models were evaluated on the test split after each training epoch and separately using 5-fold cross-validation, as shown in Table 3. For the traditional approach, an individual model was trained for each annotator, hence evaluation results were averaged across all ten models for both binary and multi-class classifications. The displayed metrics are solely for performance evaluation and do not refer to any submitted outcome. Evaluation results on the test split indicate that the BERT model seems to perform better than traditional models in binary classification. However, in multi-class classification, the performance is more variable, with traditional models achieving higher accuracy and F_1 scores. When assessed using 5-fold cross-validation, traditional models consistently perform better than BERT across most metrics for both classification tasks, except in binary classification where BERT shows higher precision.

Final submission results show that the traditional models achieved a lower Jensen-Shannon distance and thus better values compared to BERT, as visualized in Table 5. Therefore, the traditional models were chosen as the final submission approach for subtasks 1 and 2, resulting in the metrics shown in Tables 4 and 5. Details of the evaluation of the submission can be found on the original competition website for subtasks 1[10] and 2[11].

6 Discussion

This section elaborates on two multi-annotator frameworks designed to predict individual labels corresponding to different annotators.

6.1 Model Architectures

The baseline approach showed that hyperparameter tuning played a crucial role in optimizing model performance, with diverse model selection underscoring a rigorous approach to achieving the best results. Especially XGBoost proved to be a suitable choice due to its effective handling of sparse data. Its ability to automatically learn imputation strategies and its incorporation of L1 and L2 regularization techniques help prevent overfitting by penalizing complex models (Nielsen, 2016). These attributes may have contributed to XGBoost being the top-performing traditional model in this multi-annotator scenario.

The multi-task approach initially demonstrated promising results on the 15% test split. TF-IDF scores emphasized term importance, while additional features such as sentiment analysis, token length, or punctuation ratios did not enhance performance, possibly due to the model's difficulty in extracting meaningful patterns. However, during 5-fold cross-validation, traditional models showed better performance in all metrics except for precision in binary classification. Given that these findings contrast with previous research, such as the work by Davani et al. (Davani et al., 2022), which reported that the multi-task architecture obtained better results than the baseline models, possible reasons for this outcome are discussed.

6.2 Occurrence of Overfitting

To evaluate potential overfitting in the multi-task model, training loss and accuracy were plotted, as shown in Figure 3. For binary classification, loss steadily decreased and accuracy increased, but training was stopped after 7 epochs due to stagnation in evaluation metrics. For multi-class classification, training continued beyond optimal performance, achieving minimal loss and maximum accuracy after 4 epochs. This suggests a high likelihood of overfitting, particularly in the multi-class setting. This unintended overfitting is further observed in 5-fold cross-validation, which shows worse performance compared to the test split evaluation, likely due to the model's overfitting to the training data and resulting in less generalizable performance across different splits.

6.3 Data Augmentation and Leakage

The initial assumption that general data augmentation would be the most effective strategy led to neglecting the downsampling of frequently occurring classes, which might have improved performance. Furthermore, augmenting data, shuffling, and then splitting it into training and test sets caused data leakage. This overlap of transformed data points between training and testing phases led to misleadingly high performance metrics on the test set, as the model encountered familiar data points during testing. This increased the chances of overfitting and may explain the discrepancy between high evaluation results and lower final submission scores. This issue affects both models.

Furthermore, the distribution of annotations,

[10] https://ofai.github.io/GermEval2024-GerMS/subtask1.html

[11] https://ofai.github.io/GermEval2024-GerMS/subtask2.html

Model Evaluation	Metric	Binary		Multi-class	
		Test Split	5-fold CV	Test Split	5-fold CV
Multi-task BERT + TF-IDF	Accuracy	**0.807**	0.659	0.445	0.247
	Precision	**0.818**	**0.830**	**0.762**	0.446
	Recall	**0.796**	0.549	**0.540**	0.494
	F_1	**0.750**	0.661	0.337	0.470
Traditional ML Models	Accuracy	0.737	**0.768**	**0.561**	**0.586**
	Precision	0.782	0.780	0.738	**0.640**
	Recall	0.628	**0.768**	0.474	**0.586**
	F_1	0.690	**0.742**	**0.559**	**0.546**

Table 3: Evaluation results on the test split (15%) and 5-fold cross-validation after training both the fine-tuned multi-task BERT and the traditional models. For the traditional models, metrics from all ten models tailored to individual annotators were averaged, each for binary and multi-class classification. Predictions were made using a hierarchical approach, starting with binary and followed by multi-class predictions.

Subtask 1	
Model	**Traditional ML**
bin_maj_f1	0.543
bin_one_f1	0.633
bin_all_f1	0.458
multi_maj_f1	0.223
disagree_bin_f1	0.560
Total Score	**0.483**

Table 4: The final submission scores for subtask 1 are measured using F_1 scores. Specifically: **bin_maj** represents if most annotators' label are non-zero; **bin_one** indicates if any annotator labeled it as non-zero; **bin_all** shows if all annotators label it as non-zero; **multi_maj** refers to the majority label; **disagree_bin** captures cases where there is disagreement among annotators on zero versus non-zero labels. The final score is the unweighted average of these five F_1 scores. Given that traditional models showed better results in subtask 2, only this approach was submitted for subtask 1.

Subtask 2		
Model	**MT BERT**	**Trad. ML**
js_dist_bin	0.433	**0.306**
js_dist_multi	0.540	**0.371**
Total Score	0.487	**0.338**

Table 5: The final submission scores for subtask 2 are measured using the Jensen-Shannon Distance. Here, **dist_bin_0** refers to the portion of annotators labeling the text as 'not-sexist', while **dist_bin_1** refers to the portion of annotators labeling the text as 'sexist'. Lower scores indicate smaller distances and thus better performance. The final score is the unweighted average of the two distances.

classes, and other factors was not consistent across training and test sets, potentially leading to unbalanced data and performance issues during training and thus harming performance.

6.4 Dataset Size and Model Complexity

Despite the benefits of data augmentation, tripling the dataset size may still have been insufficient for fine-tuning ten separate heads in a transformer model, particularly due to the limited number of sexist instances for each annotator. Traditional models, which require less data, demonstrated better performance, likely due to their better feature representation handling in multi-class tasks with numerous annotators and skewed label distributions.

The performance decrease of the multi-task model could also be related to the choice of backbone architecture. To keep the approach simple, only BERT was used. Future research should explore enhanced versions such as RoBERTa and DistilBERT trained for the German language, as they may be crucial for performance improvement.

Ethical Statement: The annotated dataset from the GermEval2024 competition, gathered in accordance with ethical standards, was used. The dataset contained sexist remarks, posing risks to those targeted. The described classification algorithms were designed not to exacerbate harm; they address online sexism and foster inclusivity and equity. This

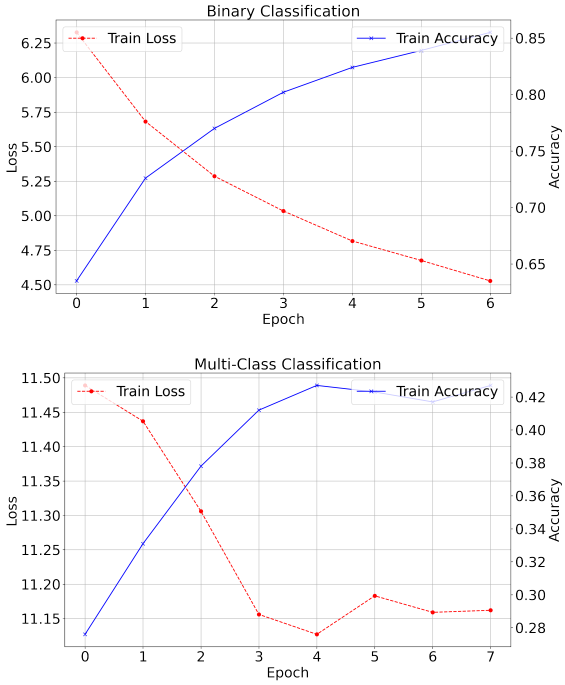

Figure 3: Plotted accuracy (blue) and loss (red) during training of the multi-task transformer for binary and multi-class classification.

study aims to contribute to automated technologies for analyzing sexism, enhancing awareness to combat oppression. This work represents a modest step towards a more equitable online environment.

7 Conclusion

This paper presents different multi-annotator methods for detecting sexism and misogyny in German-language comments, addressing challenges arising from a highly imbalanced dataset and diverse annotations provided by ten annotators. The study evaluates the effectiveness of two primary approaches: conventional machine learning models and a multi-task transformer with BERT architecture. Extensive experiments with various feature combinations and hyperparameter tuning were conducted. Results demonstrate that hyperparameter-tuned traditional models achieved better performance metrics than the multi-task transformer in detecting sexism. Furthermore, the importance of ensuring a consistent distribution of annotations and classes across dataset splits and avoiding data leakage by augmenting only the training data is emphasized. These findings highlight the difficulty of achieving reliable results in multi-task learning with limited data, especially in contexts where annotator opinions vary widely. Future research should validate these observations and explore new methods for multi-task learning frameworks, as well as hybrid models that leverage the strengths of both traditional and deep learning approaches.

References

Amy Bellmore, Angela J Calvin, Jun-Ming Xu, and Xiaojin Zhu. 2015. The five w's of "bullying" on twitter: Who, what, why, where, and when. *Computers in human behavior*, 44:305–314.

Sabur Butt, Noman Ashraf, Grigori Sidorov, and Alexander F Gelbukh. 2021. Sexism identification using bert and data augmentation-exist2021. In *IberLEF@ SEPLN*, pages 381–389.

Weilong Chen, Xin Yuan, Sai Zhang, Jiehui Wu, Yanru Zhang, and Yan Wang. 2020. Ferryman at semeval-2020 task 3: bert with tfidf-weighting for predicting the effect of context in word similarity. In *Proceedings of the fourteenth workshop on semantic evaluation*, pages 281–285.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Diane Felmlee, Paulina Inara Rodis, and Amy Zhang. 2020. Sexist slurs: Reinforcing feminine stereotypes online. *Sex roles*, 83(1):16–28.

Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. 2016. A joint many-task model: Growing a neural network for multiple nlp tasks. *arXiv preprint arXiv:1611.01587*.

Elizabeth Martinez, Juan Cuadrado, Juan Carlos Martinez-Santos, and Edwin Puertas. 2023. Detection of online sexism using lexical features and transformer. In *2023 IEEE Colombian Caribbean Conference (C3)*, pages 1–5. IEEE.

Hadi Mohammadi, Anastasia Giachanou, Ayoub Bagheri, et al. 2023. 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)*, volume 3497, pages 1000–1011. CEUR Workshop Proceedings.

Didrik Nielsen. 2016. Tree boosting with xgboost-why does xgboost win" every" machine learning competition? Master's thesis, NTNU.

Jayant Panwar and Radhika Mamidi. 2023. Panwarjayant at semeval-2023 task 10: Exploring the effectiveness of conventional machine learning techniques for online sexism detection. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1531–1536.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751.

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.

Maximilian Schmidhuber and Udo Kruschwitz. 2024. Llm-based synthetic datasets: Applications and limitations in toxicity detection. *LREC-COLING 2024*, page 37.

Sebastien C Wong, Adam Gatt, Victor Stamatescu, and Mark D McDonnell. 2016. Understanding data augmentation for classification: when to warp? In *2016 international conference on digital image computing: techniques and applications (DICTA)*, pages 1–6. IEEE.

Michelle F Wright, Bridgette D Harper, and Sebastian Wachs. 2019. The associations between cyberbullying and callous-unemotional traits among adolescents: The moderating effect of online disinhibition. *Personality and individual differences*, 140:41–45.

Rayden Xu. 2022. Jigsaw rate severity of toxic comments. https://www.kaggle.com/competitions/jigsaw-toxic-severity-rating/discussion/308938. Last accessed on 2024-05-02.

Xiaogang Xu, Hengshuang Zhao, Vibhav Vineet, Ser-Nam Lim, and Antonio Torralba. 2022. Mtformer: Multi-task learning via transformer and cross-task reasoning. In *European Conference on Computer Vision*, pages 304–321. Springer.

Mengyuan Zhou. 2023. Pinganlifeinsurance at semeval-2023 task 10: using multi-task learning to better detect online sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2188–2192.