THAugs at GermEval 2024 (Shared Task 1: GerMS-Detect): Predicting the Severity of Misogyny/Sexism in Forum Comments with BERT Models (Subtask 1, Closed Track and Additional Experiments)

Corsin Geiss and **Alessandra Zarcone**
Technische Hochschule Augsburg
An der Hochschule 1, 86161 Augsburg, Germany
corsin.geiss@tha.de, alessandra.zarcone@tha.de

Abstract

We present our approach and results for Shared Task 1 of the GermEval2024 competition (GerMS-Detect), in particular Subtask 1, aimed at predicting the severity of misogyny/sexism in text from Austrian online fora. We start from a German BERT-based baseline and a multilingual BERT-based baseline and compare them with a series of finetuned BERT-based models, in order to assess the contribution of (1) finetuning on further data from a high-quality misogyny detection dataset for a different language (Danish) and (2) finetuning on a more generic hate speech dataset for German. The best results, however, were obtained by adapting the `deepset/gbert-large` model to task-specific data, without finetuning on external data, using a weighted loss function and k-fold cross-validation, which resulted in an F1 score of 0.643 and was our submission for the Closed Track. Our findings highlight the complexity of detecting nuanced forms of hate speech and the importance of models adapted to the specific contexts of use.

1 Introduction

In recent years, social media platforms and online news websites have become central mediums for discussing a wide array of topics with a global audience. Various entities, including companies, shops, and TV shows, use these platforms to present content and interact with followers. However, the anonymity afforded by the internet often leads to various forms of harmful content, including sexist and misogynistic expressions, ranging from subtle biases to toxic comments directed at individuals or groups (Van Royen et al., 2017). This can lead to a normalization of misogynistic anti-minority speech, which can in turn perpetuate discrimination (Beukeboom and Burgers, 2019) and even increase the incidence of hate crime and sexual violence (Müller and Schwarz, 2023).

A possible way to address these issues is through automated detection of sexist and misogynistic content, which can support moderation efforts across the spectrum of harmful expressions. A series of GermEval shared tasks evaluation has focused on offensive language detection for the German language in Twitter data (Wiegand et al., 2018; Struß et al., 2019) and Facebook user comments (Risch et al., 2021). The Shared Task 1 at GermEval 2024 poses the challenge of detecting sexism in Austrian news comment as well as the majority grading (Subtask 1) and grading distribution (Subtask 2).

Detecting misogyny and sexism in text from online platforms in German is inherently challenging. Hate speech detection in online platforms for a specific language requires adapting existing models to the specific language, domain, and task, considering the full range of sexist and misogynistic expressions (Karan and Šnajder, 2018). Additionally, as sexist content can range from subtle implications to extremely toxic and violent expressions, annotators typically diverge in their opinions and perceptions of what constitutes misogynistic or sexist language (Stappen et al., 2021).

Furthermore, biases in datasets, such as those created by focused sampling instead of random sampling, can further complicate detection and result in lower classification scores under realistic settings (Wiegand et al., 2019). Unintended biases in misogyny detection models, such as those caused by identity terms, can lead to the misclassification of non-misogynistic content as misogynistic, highlighting the complexity of creating fair and effective detection systems (Nozza et al., 2019).

We present our approach and results for the 2024 GerMS-Detect Competition (GermEval2024, Shared Task 1, Subtask 1). Our starting point are pre-trained encoder-only transformed models (BERT, Devlin et al., 2019) which have shown to be successful in various NLP challenges (Min et al., 2023). The first question we address is whether

a multilingual BERT model can leverage existing task-specific data in a different language (Danish) to bring an advantage over a language-specific German BERT model. The second question is whether language-specific data for finetuning can be extracted from a more generic German hate speech dataset.

Our best-performing model, a finetuned German BERT model, achieved an F1 Score of 0.643 and was submitted to the Closed Track, as it did not use any additional training data. This model was trained using a weighted loss function to handle class imbalance and evaluated through a k-fold cross-validation approach (with $k = 5$). To further enhance the robustness of our predictions, we employed an ensemble method, combining the five best models from cross-validation.

The multilingual BERT model with additional training showed some improvements compared to its corresponding baseline, but still performed worse than the basic version of the German BERT model `deepset/gbert-large` with simple language-modeling finetuning.

We additionally evaluated the contribution of a German hate speech dataset, which was filtered using cosine similarity of sentence embeddings (Reimers and Gurevych, 2019), aiming to select the most relevant training data for misogyny/sexism detection. This experiment did not yield improvements in model performance either.

These experiments provided insights that simply filtering for misogyny is insufficient for capturing the specific nuances of sexism and the differing opinions of annotators. This highlighted the complexity of the task, where understanding the context and subjective interpretations of sexism is crucial.

Our code is released on Github for further research and development.[1]

2 Related Work

The detection of hate speech on social media platforms, as well as the detection of more subtle forms of toxicity and prejudice, including sexist and misogynistic content (ranging from subtle biases to overt violence), has been a significant research area due to its societal impact. Various studies have utilized different methodologies and datasets to address these issues. Poletto et al. (2021) provide a systematic review of resources and benchmark corpora for hate speech detection, which highlights the variety of datasets available for training and evaluating hate speech detection models.

When it comes to misogyny detection, previous work has typically focused on Twitter (Anzovino et al., 2018; Jha and Mamidi, 2017) or Reddit (Farrell et al., 2019; Guest et al., 2021). The need for annotated datasets in multiple languages is underscored by Arango Monnar et al. (2022), who highlights the limitations of existing resources and emphasize the importance of cross-lingual and cross-cultural perspectives in developing hate speech detection models.

Transformer-based models, particularly BERT, have revolutionized NLP with their ability to capture contextual information bidirectionally, significantly improving performance across various NLP tasks (Devlin et al., 2019). The success of these models has prompted their application for hate speech detection, including misogyny and sexism (Pamungkas et al., 2020; Kalra and Zubiaga, 2021; Safi Samghabadi et al., 2020). When it comes to pretrained encoders, multilingual models may be suitable for leveraging cross-lingual information and for exploiting existing datasets in a language different than the target language (Muller et al., 2021). It seems however that language-specific models outperform multilingual ones for tasks involving nuanced language understanding (Zeinert et al., 2021; Rust et al., 2021).

Handling class imbalance is a critical aspect of developing robust models for hate speech detection. Younes and Mathiak (2022) explored the use of pre-trained language models and weighted loss functions to address class imbalance, showing significant improvements in model performance for underrepresented classes. These techniques are essential for ensuring that models do not overlook minority classes, which is a common issue in hate speech detection (Kwarteng et al., 2022).

Ensemble learning techniques have been explored to enhance the performance of hate speech detection models. Mazari et al. (2024) demonstrated the effectiveness of combining BERT with other models through ensemble methods, achieving significant improvements in detecting multiple aspects of hate speech. This approach leverages the strengths of different models to provide more robust and accurate predictions.

[1] https://github.com/tha-atlas/GermEval2024-THAugs/

3 Data, Tasks and Evaluation

3.1 GERMS-AT dataset

The data for the GermEval2024 Shared Task 1 (GerMS-Detect) came from the GERMS-AT dataset (Krenn et al., 2024). It consists of comments from an Austrian online newspaper. The comments were annotated by multiple annotators with regards to the severity of misogyny/sexism present in text, on a scale from 0 to 4, where 0 represents no sexism/misogyny, and 4 indicates extreme sexism/misogyny. The labels capture the distribution and variability of annotators' opinions.

The datasets were released in three phases: Trial, Development, and Competition[2]. Each phase provided training and testing data, with the test data from one phase becoming part of the labeled training data in the next phase, along with the previous phase's training data. In the Competition Phase, the training data included 5,598 comments (including the labeled test data from the Development Phase and the training data from both the Trial and Development phases), with 4,221 labeled as non-sexist and 1,377 as sexist (majority label). The testing data for the Competition Phase comprised 1,986 unlabeled comments.

Content warning: We report examples of sexist and misogynous language to illustrate the data.

Below are two examples that reflect the range of severity captured in the dataset:

- **Original:** "Ja - weils politisch-medial nicht gewünscht ist gemischte Ranglisten zu zeigen bei denen Frauen auf den vorderen Plätzen nicht auftauchen."
  Translation: "Yes - because it is not politically or media-wise desired to show mixed rankings where women do not appear in the top positions."
  Majority Label: 1 (Low Severity)

- **Original:** "wegen der paar frauen die es aus der küche heraus schaffen eine eigene karriere zu starten ist es eigentlich ein overkill eine eigene genderform zu schaffen. der überwiegende teil der frauen freut sich wenn man(n) ! sie regelmäßig ob der hervorzüglich gekochten speisen lobt."
  Translation: "Because of the few women who manage to start their own careers outside of the kitchen, it is actually overkill to create a separate gender form. The majority of women are happy when (a man) regularly praises them for the excellently cooked meals."
  Majority Label: 4 (Extreme Severity)

The dataset is presented in JSONL format, each entry contains:

- **id**: A unique identifier for the comment.
- **text**: The text of the comment.
- **annotations**: An array of dictionaries (only in the labeled dataset), each containing:
  - **user**: An anonymized ID for the annotator (e.g., "A003").
  - **label**: The label assigned by the annotator.
- **annotators**: An array of annotator IDs who labeled the example (only in the unlabeled dataset).

Figure 1 shows the distribution of the misogyny/sexism severity labels in the dataset, considering all labels provided by the annotators. As illustrated, there is a significant class imbalance, with the majority of comments labeled as "None" (label 0), and fewer comments labeled with higher severity levels.

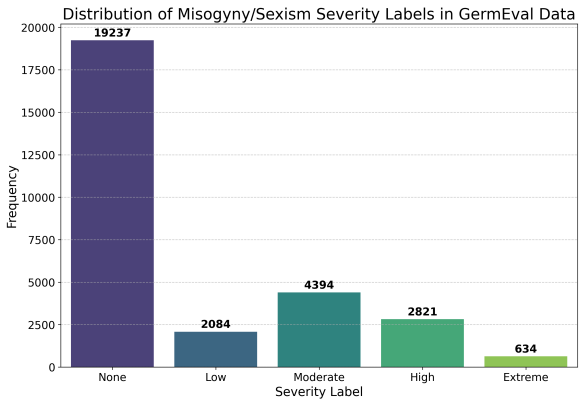

Figure 1: Distribution of misogyny/sexism severity labels in the GermEval2024/GerMS-Detect Data, considering all annotator labels.

3.2 Additional Datasets

To explore potential improvements brought by fine-tuning on more data, we incorporated two additional datasets containing annotated examples of hate speech:

[2] https://ofai.github.io/GermEval2024-GerMS/, Accessed: 2024-07-02

- **Bajer Dataset**: This dataset contains annotated Danish social media posts labeled for misogyny. It includes high-quality annotations of online misogynistic content, providing insights into cross-lingual and cross-cultural aspects of misogyny detection (Zeinert et al., 2021).
- **GAHD Dataset**: The German Adversarial Hate Speech Dataset (GAHD) contains adversarial examples aimed at improving model robustness in detecting hate speech (Goldzycher et al., 2024). This dataset includes texts labeled as hate speech or not.

These datasets were not used in our submission to the Closed Track.

3.3 Data Preprocessing

The preprocessing steps were consistently applied across all stages of our methodology to ensure clean input text. These steps included removing HTML tags, URLs, emojis, and extra whitespaces (Glazkova, 2023). This preprocessing was performed on all datasets used, including the GermEval2024/GerMS-Detect Data, the Danish sexism dataset, and the German adversarial hate speech dataset (GAHD).

3.4 Task Description

In Subtask 1, the goal is to predict the severity of misogyny/sexism for each text based on the labels assigned by multiple annotators. The labels reflect different strategies for combining multiple annotations into a single target label:

- **bin_maj**: Predict 1 if a majority of annotators assigned a label other than 0, otherwise predict 0. Both 1 and 0 are correct if there's no majority.
- **bin_one**: Predict 1 if at least one annotator assigned a label other than 0, otherwise predict 0.
- **bin_all**: Predict 1 if all annotators assigned labels other than 0, otherwise predict 0.
- **multi_maj**: Predict the majority label if there is one; if no majority, any of the labels assigned is counted as correct.
- **disagree_bin**: Predict 1 if there is disagreement on 0 versus all other labels, otherwise predict 0.

3.5 Evaluation

System performance on all five predicted labels (bin_maj, bin_one, bin_all, multi_maj, disagree_bin) is evaluated using the F1 macro score over all classes. The final score (which is used for ranking submissions in the leaderboard) is calculated as the unweighted average over all five scores.

4 Methodology

4.1 Model Architecture

The model architecture used for the final finetuning on the training data is consistent across all our BERT-based models and is designed to handle the specific requirements of the classification tasks. The architecture includes the following components:

- **Input Layer**: Handles tokenized input text, including input IDs and attention masks.
- **BERT Layer**: Utilizes a pretrained BERT model to extract contextual embeddings from the input text. We used the following BERT models in this layer (more details on the models in 4.6-4.7):
  - **deepset/gbert-large**: A German BERT model finetuned for specific language tasks.
  - **google-bert/bert-base-cased**: A base BERT model suitable for general language understanding tasks in German.
  - **bert-base-multilingual-cased**: A multilingual BERT model capable of handling multiple languages, including German.
- **Fully Connected Layers**: Consists of multiple fully connected layers with batch normalization and activation functions.
- **Classifiers**: Comprises several task-specific classifiers for binary and multi-class classification.

The architecture is illustrated in Figure 2.

4.2 Cross-Validation

To ensure model robustness, we employed k-fold cross-validation (with $k = 5$), partitioning the dataset into five subsets, training on four, and validating on one. This process was repeated five times with different validation subsets, and performance metrics were averaged. Stratified splitting ensured

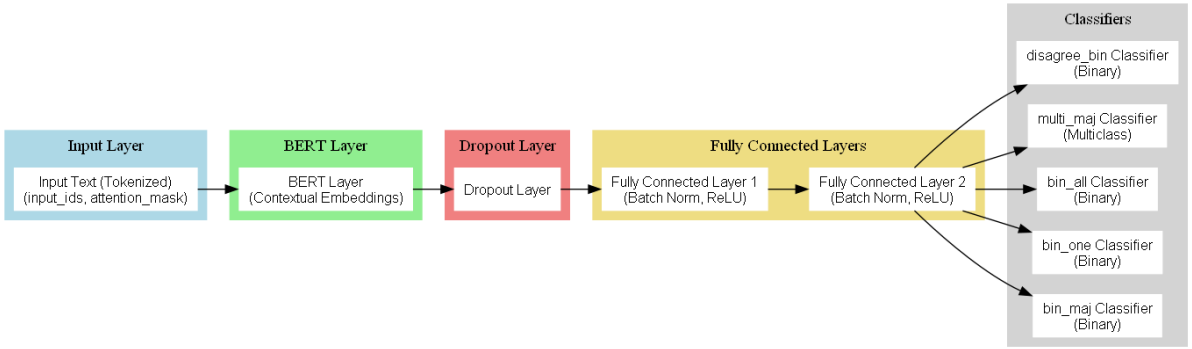

Figure 2: Model architecture used for the finetuning process.

balanced class representation across folds. Except for the final submission, we used the GermEval development data and test data for training and testing. The primary evaluation metric was the F1 score, chosen for its balance between precision and recall, suitable for imbalanced datasets (Sokolova and Lapalme, 2009).

4.3 Ensemble Learning

To enhance prediction robustness, we employed an ensemble learning approach. Our ensemble consisted of five models, each trained on a different fold of the dataset using 5-fold cross-validation. This approach ensures that each model is exposed to a slightly different subset of the training data, potentially capturing different aspects of the problem.

The models in the ensemble shared the same architecture (described in Section 4.1) but differed in their learned parameters due to being trained on different data folds. For prediction, we used the following process:

1. Each of the five models made predictions on the test data independently.

2. For binary classification tasks (bin_maj, bin_one, bin_all, disagree_bin), the raw logits were transformed using a sigmoid function to obtain probability scores.

3. For the multi-class task (multi_maj), a softmax function was applied to the logits to obtain class probabilities.

4. The predictions from all five models were aggregated by averaging the probability scores for each task.

5. For binary tasks, the final prediction was determined by rounding the average probability (threshold of 0.5).

6. For the multi-class task, the class with the highest average probability was selected as the final prediction.

This ensemble approach mitigates individual model variability and improves overall performance by leveraging the collective wisdom of multiple models. The use of probability averaging allows for a more nuanced final prediction, potentially capturing uncertainties that a single model might miss (Mazari et al., 2024).

4.4 Handling Class Imbalance

In order to address class imbalance, we employed a weighted loss function (Younes and Mathiak, 2022), assigning higher weights to underrepresented classes to prevent model bias towards frequent classes.

4.5 Hyperparameter Tuning

For hyperparameter tuning and architecture building, we initially used the `google-bert/bert-base-cased` model[3], due to its lower resource requirements compared to `deepset/gbert-large` (Chan et al., 2020)[4]. We employed Optuna (Akiba et al., 2019), a hyperparameter optimization framework, to efficiently search for optimal hyperparameters. Our search focused on learning rate, batch size, weight decay, hidden layer dimensions, dropout rate, and number of epochs. The search ranges were informed by previous experience with similar

[3] https://huggingface.co/google-bert/bert-base-german-cased, Accessed: 2024-06-25

[4] https://huggingface.co/deepset/gbert-large, Accessed: 2024-06-25

tasks and models. Approximately 50 trials were conducted using Optuna's Bayesian optimization approach, as shown in Table 1.

The model was trained using the AdamW optimizer (Loshchilov and Hutter, 2019), with a ReduceLROnPlateau scheduler (PyTorch Documentation, 2021) to adjust the learning rate during training. For the final `deepset/gbert-large` model used in the competition, we had to adjust some parameters due to GPU memory constraints and the larger model size. Specifically, we reduced the batch size to 16 and adjusted the learning rate to 1×10^{-5}. We used the development training data for hyperparameter tuning

4.6 Closed Track Approaches

4.6.1 Baseline Models

German Baseline We utilized the `german bert cased` model as a baseline to compare the performance of our finetuned models. This model was chosen due to its lower resource requirements and effectiveness in handling German language tasks.

Multilingual Baseline The `bert multilingual` model was used as a baseline for assessing cross-lingual transfer learning capabilities (Devlin et al., 2019). It provided a benchmark for evaluating improvements from finetuning on a dataset in a different language than the target language.

4.6.2 Finetuning German Models

Finetuning `gbert-large` We performed language model finetuning (LM finetuning) on the `deepset/gbert-large` model (Chan et al., 2020) using the GermEval2024 dataset. This step adapted the model to the specific language and context of the GermEval data, enhancing its understanding of the linguistic characteristics of the domain.

Finetuning `german bert cased` We also performed LM-finetuning on the `german bert cased` model using the same dataset. This model served as a point of comparison to evaluate the performance gains achieved from the LM-finetuning process itself.

4.6.3 Finetuning Process Details

For both models, we used a custom dataset class and `BertTokenizer` to tokenize the preprocessed texts, ensuring consistent input size by truncating and padding them to a maximum length. We initialized the `BertForMaskedLM` model, setting up the training environment with specific arguments such as epochs, batch size, and learning rate. A data collator dynamically created masked language modeling data during training. Using the `Trainer` class from the Transformers library, we managed the training loop, including forward and backward passes, optimization, and checkpointing. Post-training, we saved the finetuned models and tokenizers for the classification task.

4.7 Additional Experiments

4.7.1 Harvesting Data for Task-Based Finetuning

To prepare the filtered GAHD dataset for finetuning, we performed a detailed data harvesting process, aimed at finding datapoints which were not just examples of hate speech, but specifically examples of sexism/misogyny:

1. **Data Preparation**: We read the GermEval data, labeled each entry as sexist or non-sexist based on the annotations, and split the dataset into training and test sets to find an optimal threshold for identifying misogynistic content, aiming for a precision value of about 75%.

2. **Model Setup and Finetuning**: We used sentence embeddings (Reimers and Gurevych, 2019) from the T-Systems-onsite/cross-en-de-roberta-sentence-transformer model[5] and finetuned it on the GermEval data to improve the filtering process. The known sexist sentences were all the sexist sentences from GermEval.

3. **Similarity Calculation**: We computed the cosine similarity between the embeddings of the GAHD dataset texts and the known sexist sentences from GermEval.

4. **Filtering**: Using the determined threshold of 0.14, we filtered the GAHD dataset to retain texts with high cosine similarity scores, indicating misogynistic content.

It's important to note that the filtering process may sometimes incorrectly classify certain sentences. For example:

Content warning: We report examples of sexist and misogynous language to analyse model errors.

- **Original:** "Sollten wir nicht zumindest versuchen, diesen Thread in etwas Positives

[5] `https://huggingface.co/T-Systems-onsite/cross-en-de-roberta-sentence-transformer`, Accessed: 2024-06-25

Hyperparameter	Search Range	Optimal Value
Batch size	[16, 32, 64]	64
Learning rate	$[1 \times 10^{-5}, 5 \times 10^{-5}]$	4.053×10^{-5}
Weight decay	$[0.1, 0.3]$	0.2019
Hidden layer dimension 1	[512, 768, 1024]	768
Hidden layer dimension 2	[128, 256, 512]	256
Dropout rate	[0.1, 0.5]	0.3606
Number of epochs	[8, 12]	12

Table 1: Hyperparameter search ranges and optimal values

umzuwandeln, etwa Möglichkeiten, weiße Männer und Frauen vor gemischten Beziehungen zu retten?"
Translation: "Shouldn't we at least try to turn this thread into something positive, such as ways to save white men and women from mixed relationships?"
(This sentence may be incorrectly classified due to the word "Frauen" (women) despite the context not being inherently misogynistic.)

- **Original:** "Ich habe 2019 auf TIK TOK einige SEHR SCHÖNE UND HEISSE SEXY MÄDCHEN VON MUSICALLY gefunden, schade, dass sie alle wie Huren wirken"
  Translation: "In 2019, I found some VERY BEAUTIFUL AND HOT SEXY GIRLS from Musical.ly on TikTok, it's a shame that they all seem like whores"
  (This sentence is correctly identified as sexist due to explicit objectification and derogatory language towards women.)

The final filtered GAHD subset, balanced to include an equal number of sexist and non-sexist examples, contained 769 sentences of each type and was saved for further finetuning.

4.7.2 Finetuning on Filtered GAHD Dataset

To improve its ability to detect misogynistic content, we finetuned the `german bert cased` model on the filtered GAHD dataset prepared in the previous step. The finetuning process involved tokenizing the text, computing class weights to address class imbalance, and training the model using the AdamW optimizer with a set of hyperparameters tailored for this specific task.

4.8 Finetuning Multilingual Models

To assess the multilingual BERT model's cross-lingual transfer learning capability, we performed task-based finetuning on a Danish sexism dataset (Zeinert et al., 2021). The multilingual BERT model was finetuned on the Danish dataset with the following setup:

- Learning rate: 2×10^{-5}
- Batch size: 16
- Number of epochs: 1
- Dropout rates: 0.3 for hidden and attention layers
- Optimizer: AdamW
- Evaluation metrics: Accuracy, F1 score, precision, and recall.

This evaluation aimed to determine if finetuning on Danish data could lead to performance improvements on German language tasks.

5 Results

Table 2 summarizes the F1 scores achieved by different models and configurations during our exploration and experimentation. These models were trained on the GermEval Development data and tested on the GermEval Development test data.

5.1 Closed Track Results

5.1.1 german bert cased (baseline)

The `german bert cased` model, without any finetuning, served as our primary baseline. It achieved an average F1 score of 0.5825 across all tasks, providing a solid starting point for comparison.

5.1.2 bert multilingual (baseline)

Our multilingual baseline, using the `bert-base-multilingual-cased` model without finetuning, achieved an average F1 score of 0.5679. This performance was slightly lower than the German-specific baseline, highlighting the potential benefits of language-specific models for this task.

Model	bin_maj	bin_one	bin_all	multi_maj	disagree_bin	**Avg. F1 Score**
gbert-large (LM-finetuned)	0.7347	0.7707	0.7132	0.2886	0.6132	0.6241
german bert cased (LM-finetuned)	0.7069	0.7532	0.6459	0.2778	0.6085	0.5985
german bert cased (baseline)	0.6865	0.7299	0.6632	0.2669	0.5660	0.5825
german bert cased (task-finetuned on GAHD)	0.6834	0.7180	0.6477	0.2693	0.5700	0.5777
bert multilingual (baseline)	0.6769	0.7301	0.6000	0.2467	0.5858	0.5679
bert multilingual (task-finetuned on Danish)	0.6776	0.7328	0.6185	0.2518	0.6061	0.5773

Table 2: F1 scores achieved by different models and configurations during exploration and experimentation. These models were trained on the GermEval Development data and tested on the GermEval Development test data.

5.1.3 german bert cased (LM-finetuned)

After language model finetuning, the `german bert cased` model showed improved performance, with an average F1 score of 0.5985. This improvement demonstrates the effectiveness of adapting the model to the specific language and context of the task.

5.1.4 gbert-large (LM-finetuned)

The `gbert-large` model, finetuned with language model finetuning, achieved the best performance among all tested configurations. The final evaluation on the Competition test set, which included texts without labels, involved generating predictions using the ensemble models and submitting them for the GermEval contest. The final submission to the Closed Track achieved an average F1 score of 0.643 across all tasks, confirming the robustness of the finetuned `gbert-large` model.

The training procedure for the `deepset/gbert-large` model involved monitoring the training loss and validation F1 score across epochs to ensure proper convergence and avoid overfitting. Figure 3 shows the training loss and validation F1 score across 5 folds, providing insights into the training dynamics and model performance over time. The figure reveals that the training loss consistently decreases across all folds, indicating effective learning and reduction of error on the training data. Concurrently, the validation F1 score initially increases, reflecting improved model performance on the validation set. However, the validation F1 score plateaus after a few epochs, suggesting that further training does not significantly enhance validation performance and helps identify the point of diminishing returns.

5.2 Results of the Additional Experiments

5.2.1 german bert cased (task-finetuned on GAHD)

The `german bert cased` model, when task-finetuned on the filtered GAHD dataset, achieved an average F1 score of 0.5777. This performance did not surpass its LM-finetuned counterpart, indicating that additional task-specific finetuning with filtered data did not provide the expected benefits.

5.2.2 bert multilingual (task-finetuned on Danish)

The multilingual BERT model, after task-specific finetuning on Danish data (Zeinert et al., 2021), showed improved performance with an average F1 score of 0.5773. This improvement over the baseline multilingual model suggests that the model can adapt to new languages and transfer learning across them. However, it still did not outperform the German-specific models.

6 Discussion and Conclusion

This paper presents our approach to Shared Task 1 of the GermEval2024 GerMS-Detect competition (Subtask 1), focusing on predicting the severity of misogyny/sexism in text based on annotations from multiple human annotators. Our methodology primarily utilized a finetuned `deepset/gbert-large` model, which proved effective in understanding and detecting nuanced language features associated with misogyny and sexism.

Our ensemble approach, consisting of five `deepset/gbert-large` models each trained on a different fold of the dataset achieved an average F1 score of 0.643 across all tasks in the competition. Key to this success was the use of weighted loss functions to address class imbalance and an ensemble learning approach to enhance prediction robustness. Hyperparameter tuning using Optuna further optimized performance, ensuring the chosen hyperparameters were well-suited for the task (Akiba et al., 2019).

To explore cross-lingual transfer learning, we finetuned a multilingual BERT model on a Danish sexism dataset (Zeinert et al., 2021). While this model showed slight improvements in certain F1 scores after finetuning, it did not outperform

Figure 3: Training loss (solid lines) and validation F1 score (dashed lines) across epochs for 5 folds of the `gbert-large` model.

the standard `deepset/gbert-large` model. The marginal gains highlight the challenges inherent in cross-lingual transfer learning and the complexity of the task due to the subjective nature of the annotations.

We also investigated filtering the German adversarial hate speech dataset (GAHD) for misogynistic content using cosine similarity, aiming to improve the `german bert cased` model through additional pre-finetuning. This approach did not yield significant improvements, indicating that simply increasing misogynistic content in the training data is not sufficient to capture nuanced perceptions of sexism and misogyny, highlighting the complexity of developing effective models for detecting nuanced and subjective forms of hate speech (Fortuna and Nunes, 2018).

Our findings emphasize the importance of using specialized models tailored to specific linguistic contexts. Although the multilingual BERT model demonstrated some cross-lingual capabilities, the `deepset/gbert-large` model remained more effective for this task. Future research could explore more sophisticated methods for integrating multilingual data and better techniques for handling subjective annotations.

In summary, our work highlights the challenges and potential solutions for detecting misogyny and sexism in text. The advanced transformer models, ensemble learning, and careful handling of class imbalance were effective in achieving robust performance. However, the nuanced and subjective nature of this task requires further exploration and innovation to develop comprehensive and fair detection models.

Acknowledgements

This research was funded by the Bavarian State Ministry for Science and the Arts (StMWK: Bayerische Staatsministerium für Wissenschaft und Kunst - StMWK) as part of the Project "CHIASM" (Changenreiche industrielle Anwendungen für vortrainierte Sprachmodelle).

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 2623–2631, New York, NY, USA. Association for Computing Machinery.

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *Natural Language Processing and Information Systems: 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings 23*, pages 57–64. Springer.

Ayme Arango Monnar, Jorge Perez, Barbara Poblete, Magdalena Saldaña, and Valentina Proust. 2022. Resources for multilingual hate speech detection. In

Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), pages 122–130.

Camiel J Beukeboom and Christian Burgers. 2019. How stereotypes are shared through language: a review and introduction of the aocial categories and stereotypes communication (SCSC) framework. *Review of Communication Research*, 7:1–37.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tracie Farrell, Miriam Fernandez, Jakub Novotny, and Harith Alani. 2019. Exploring misogyny across the manosphere in reddit. In *Proceedings of the 10th ACM conference on web science*, pages 87–96.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

Anna Glazkova. 2023. A comparison of text preprocessing techniques for hate and offensive speech detection in twitter. *Social Network Analysis and Mining*, 13(1):155.

Janis Goldzycher, Paul Röttger, and Gerold Schneider. 2024. Improving adversarial data collection by supporting annotators: Lessons from GAHD, a German hate speech dataset. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4405–4424, Mexico City, Mexico. Association for Computational Linguistics.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, Online. Association for Computational Linguistics.

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? Analysis and classification of ambivalent sexism using Twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 7–16, Vancouver, Canada. Association for Computational Linguistics.

Amikul Kalra and Arkaitz Zubiaga. 2021. Sexism identification in tweets and gabs using deep neural networks. *Preprint*, arXiv:2111.03612.

Mladen Karan and Jan Šnajder. 2018. Cross-domain detection of abusive language online. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 132–137, Brussels, Belgium. Association for Computational Linguistics.

Brigitte Krenn, Johann Petrak, Marina Kubina, and Christian Burger. 2024. GERMS-AT: A sexism/misogyny dataset of forum comments from an Austrian online newspaper. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7728–7739, Torino, Italia. ELRA and ICCL.

Joseph Kwarteng, Serena Coppolino Perfumi, Tracie Farrell, Aisling Third, and Miriam Fernandez. 2022. Misogynoir: challenges in detecting intersectional hate. *Social Network Analysis and Mining*, 12(1):166.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

A.C. Mazari, N. Boudoukhani, and A. Djeffal. 2024. Bert-based ensemble learning for multi-aspect hate speech detection. *Cluster Computing*, 27:325–339.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Comput. Surv.*, 56(2).

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. First align, then predict: Understanding the cross-lingual ability of multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2214–2231, Online. Association for Computational Linguistics.

Karsten Müller and Carlo Schwarz. 2023. From hashtag to hate crime: Twitter and antiminority sentiment. *American Economic Journal: Applied Economics*, 15(3):270–312.

Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended bias in misogyny detection. In *IEEE/WIC/ACM International Conference on Web Intelligence*, WI '19, pages 149–155, New York, NY, USA. Association for Computing Machinery.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Misogyny detection in twitter: a multilingual and cross-domain study. *Information Processing & Management*, 57(6):102360.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(3):477–523.

PyTorch Documentation. 2021. ReduceLROnPlateau.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 1–12, Duesseldorf, Germany. Association for Computational Linguistics.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Niloofar Safi Samghabadi, Parth Patwa, Srinivas PYKL, Prerana Mukherjee, Amitava Das, and Thamar Solorio. 2020. Aggression and misogyny detection using BERT: A multi-task approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 126–131, Marseille, France. European Language Resources Association (ELRA).

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437.

Lukas Stappen, Lea Schumann, Anton Batliner, and Bjorn W. Schuller. 2021. Embracing and exploiting annotator emotional subjectivity: An affective rater ensemble model. In *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 01–08.

Julia Maria Struß, Melanie Siegel, Josep Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of GermEval Task 2, 2019 Shared Task on the Identification of Offensive Language. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, pages 354–365, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

Kathleen Van Royen, Karolien Poels, Heidi Vandebosch, and Philippe Adam. 2017. "Thinking before posting?" Reducing cyber harassment on social networking sites through a reflective message. *Computers in Human Behavior*, 66:345–352.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608. Association for Computational Linguistics.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In *Proceedings of the 14th Conference on Natural Language Processing (KONVENS 2018)*.

Yousef Younes and Brigitte Mathiak. 2022. Handling class imbalance when detecting dataset mentions with pre-trained language models. In *Proceedings of the 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022)*, pages 79–88.

Philine Zeinert, Nanna Inie, and Leon Derczynski. 2021. Annotating online misogyny. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3181–3197, Online. Association for Computational Linguistics.