Argument Mining: Sexism Detection using Master of Experts

Clara Wicharz
University of Potsdam
wicharz@uni-potsdam.de

Jatin Karthik Tripathy
University of Potsdam
tripathy@uni-potsdam.de

Abstract

Online sexism is an increasing issue on online social media platforms. To improve sexism detection and provide further explainability, Kirk et al. (2023) curated a dataset with finer-grained sexism annotations. We developed two ensemble approaches to perform multi-class classification for sexism detection. In the first approach "Master of Domain Experts" (MoDE) we ensemble two domain-experts via a fully connected neural network. As domain-experts BERT and HateBERT were chosen and differently combined over three experiments. In the second approach "Master of Class Experts" (MoCE) we ensemble binary classifiers, one for each class, via a fully connected neural network. And once again HateBERT was chosen as the model for class-experts. We found that MoDE with two HateBERTs outperformed MoCE with a macro-averaged F1 score of 0.5288.

1 Introduction

Sexism is a phenomenon that comprises any abuse or negative sentiment directed towards women based on their gender. It can also occur in combination with further identity attributes, such as skin colour, religion or sexual orientation (Kirk et al., 2023). It has become a growing problem on online social media platforms. Beyond causing immediate harm, the occurrence makes online spaces less welcoming to women. This can push women out of their respective spaces, solidifying social asymmetries and injustices.

With the rise of sexism on social media, solutions for its automatic detection have surged. However, these solutions often assess sexism only on a high level, therefore lacking explainability. This study is regarded with automatic sexism detection in Semeval 2023's task 10 "Explainable Detection of Online Sexism". In this task, the organisers propose a more differentiated look at sexism. Therefore, they provide an English language dataset of Gab and Reddit content. The instances are annotated on three levels with each level breaking down sexism into more fine-grained sub-classes (Kirk et al., 2023).

This study [1] identifies sexism on the second annotation level, which differentiates all sexist instances into four sub-classes: Threats, Derogation, Animosity and Prejudice as seen in Section 3. Research on abusive language and sexism detection suggests that pre-trained transformers such as BERT outperform older approaches such as linear classifiers or neural networks (Caselli et al., 2021; Davies et al., 2021). Moreover, Caselli et al. (2021); Lin et al. (2022) suggest the benefits of using transformers that have been pre-trained on data that is semantically related to the target task. Davies et al. (2021); Allen-Zhu and Li (2020); Lin et al. (2022); Janatdoust et al. (2022); Portelli et al. (2022) investigated how the performance of standalone pre-trained transformers can be increased by ensembling them.

In Section 2 we first investigate these two lines of research: domain-specifically pre-trained transformers and transformer-ensembles. Subsequently, we present our two approaches towards multi-class sexism detection that combine the previous approaches. Finally, we will present and discuss the performance of our models and propose ideas to further improve them.

2 Related Work

We observed two main themes in current research on abusive language and sexism detection: pre-training transformers on semantically related data and ensemble transformers.

Further pre-training of transformers on semantically related data has received much attention across domains. It has shown to be a viable ap-

[1] Link to Code

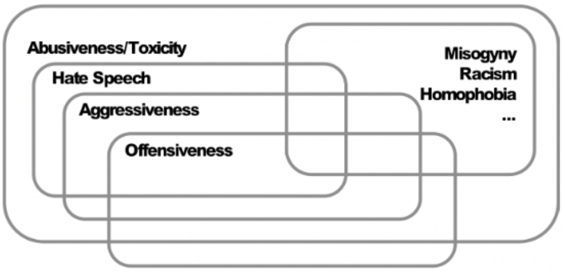

Figure 1: Relationship of Abusive language phenomena

proach to obtain domain-specific models cheaply and quickly (Caselli et al., 2021). This advantage resulted in various domain-specific BERT spin-offs across domains. Caselli et al. (2021) investigated this technique to improve the detection of abusive language, a broader phenomenon entailing sexism as they show in Figure 1.

Caselli et al. (2021) first curated the large-scale dataset RAL-E, consisting of English comments from Reddit communities banned for being offensive, abusive or hateful. Secondly, Caselli et al. (2021) retrained BERT base-uncased on around 1.5 Mio messages from the RAL-E dataset by applying the Masked Language Model objective. The resulting model, named HateBERT, was then tested against its vanilla version, BERT base-uncased. Therefore, the models were fine-tuned and tested on three datasets: OffensEval2019 (14,000 tweets), AbusEval (14,000 tweets) and HatEval (13,000 tweets). All datasets had been binarily annotated for offensive, abusive and hate speech.

For testing, both an in-dataset and a cross-dataset evaluation were conducted. For the in-dataset evaluation, HateBERT outperformed its general-purpose counterpart on each dataset. For example, on the OffensEval2019, it reached a macro-F1-score of 0.809, exceeding the general-purpose version by 0.74%. On AbusEval, it scored 0.765, exceeding BERT by 5.22%, and on HateEval, it reached 0.516, outperforming BERT by 7.5%.

In the cross-dataset evaluation, HateBert outperformed BERT when trained on OffensEval2019 or AbusEval and tested on other datasets. However, HateBERT was outperformed by BERT when trained on HatEval and tested on the different datasets. Caselli et al. (2021) interpret this as HateBERT providing better portability over the general-purpose counterpart when being introduced to a language phenomenon that is larger and entails the language phenomenon that is supposed to be detected in testing. When fine-tuning a more specific language phenomenon, Caselli et al. (2021) conclude that a general-purpose transformer such as BERT will allow higher portability to capture wider language phenomena at testing.

The EXIST21 workshop has inspired many projects investigating sexism detection in online social media platforms. Participants were provided with a Twitter and Gab content dataset with sexist expressions. 3436 samples were in English, and 3541 samples were in Spanish. The task was to do binary or multi-class classification.

The participants (Schütz et al., 2021) worked on the workshop's binary and multi-class classification task. They used multilingual BERT versions in 3 different approaches, which they referred to as "fine-tuning", "pre-training", and "late-fusion strategy". For their fine-tuning strategy, the English language part of the dataset was translated into Spanish and the other way around. Subsequently, they chose mBERT as the multilingual model and fine-tuned it on the augmented dataset. They experimented with fine-tuning over the entire model vs. only on the classification head. Without reporting, performance metrics Schütz et al. (2021) state a higher performance for fine-tuning over the whole model.

The "pre-training strategy" (Schütz et al., 2021) used XML-RoBERTa as a multilingual model. First, the model was further pre-trained on the target dataset and two semantically related datasets, HatEval2019 and MeTwo. MeTwo consists of 3600 Spanish tweets, annotated as either sexist, doubtful or not sexist. HatEval2019 consists of 13000 English and 6000 Spanish language instances, including hate speech against women and immigrants. After this additional pre-training, the model was fine-tuned on the EXIST data.

For the "late-fusion strategy", the two classifiers obtained from the two previous approaches were ensembled. The logits of the fine-tuned mBERT and the pre-trained and fine-tuned XLM-RoBERTa were first summed up per class and then argmaxed.

The mere pre-training strategy outperformed both the fine-tuning and late-fusion strategy on the binary and multi-class tasks and showed minor signs of overfitting. For example, on the binary task, the pre-training strategy reached a macro-averaged F1 score of 77.52, while the "late-fusion strategy" reached 76.56, and the "fine-tuning strategy" scored 71.21. In the multi-class task, the pre-training strategy got a score of 55.89, followed by the late-fusion strategy with 55.59 and the fine-tuning strategy with 51.95. The results suggest the advantage of pre-training pre-trained transformers with additional semantically related data to boost performance on the target task.

Davies et al. (2021) also participated in the previously mentioned EXIST21 workshop. Davies et al. (2021) investigated the approach of ensembling pre-trained transformers to perform binary and multi-class sexism detection. Their specific ensemble approach was based on Allen-Zhu and Li (2020)'s findings. According to Allen-Zhu and Li (2020), ensembles of deep learning models with the same architecture, trained with the same algorithm, the same loss function and on the same dataset show superior performance over their standalone counterparts when the constituent models differ by their initial weight configurations. Furthermore, while Schütz et al. (2021) chose multilingual models, Davies et al. (2021) exploited the fact that data came with a language tag. Therefore Davies et al. (2021) created one ensemble per language, arguing that language-specific BERT versions better capture language subtleties. They used BERT base-uncased for their English language ensemble and Spanish language BETO. Each ensemble consisted of three of the respective models. For the ensemble, first, each model was fine-tuned separately and with an individual weight initialization on the respective dataset. Whether all layers or only a classification head was fine-tuned was not specified. Subsequently, the fine-tuned models were ensemble via a majority vote. This ensemble approach was applied to both classification tasks and performed with a macro-averaged F1 score of 0.766 on the binary task and 0.535 in the multi-class classification tasks. Error analysis revealed that the models were often fooled by words that commonly appear in sexist environments, creating false positives.

An ensemble of pre-trained transformers has not just been used for sexism detection, Janatdoust et al. (2022) investigated the approach of ensembling pre-trained transformers for detecting signs of depression in social media texts. For Google's LANGUAGE TECHNOLOGY FOR EQUALITY, DIVERSITY, INCLUSION workshop 2022 (Janatdoust et al., 2022), they developed a ternary classifier to predict severe, moderate or no signs of depression in English social media comments. Without reporting on the decision criteria, Janatdoust et al. (2022) chose BERT, ALBERT, DistilBERT and RoBERTa for an ensemble classifier. Each model was fine-tuned separately on the social media comments and ensembled via majority vote. The ensemble outperformed each standalone model with a macro-averaged F1 score of 0.54. Standalone RoBERTa followed with a score of 0.52, ALBERT with 0.51 and BERT with 0.50. The ensemble ranked 5th in the competition and provided further evidence for the general approach of ensembling general-purpose pre-trained transformers.

Portelli et al. (2022) also used an ensemble of pre-trained transformers to solve two tasks for the Social Media Mining for Health 2022 workshop; in the first task, models needed to predict whether Spanish tweets contain literature/news reports, personal reports or reports on somebody else's COVID-19 symptoms. In the second task, English tweets needed to be classified as containing general COVID-19 vaccine chatter or personal reports confirming the vaccination status.

For the ensemble Portelli et al. (2022) decided to use three standalone models. The choice of candidate models was narrowed down to GPT-2, BERT, mBERT, PubMedBERT, Spambot, Twitter-roBERTa-base for Sentiment Analysis and XLM-Roberta. Each model was fine-tuned and tested five times for each task. The best model for the job was used twice in the respective ensemble, and the second best model once.

For the ternary classification task, two mBERTs and one XLM-RoBERTa were used. For the binary classification task, two mBERTs and one BERT were used. In both ensembles, models were coordinated via a majority vote. The main difference between the two resulting ensembles was that the ternary classification task used models with different architectures, while the ensemble for the binary classification task used models with the same.

In the ternary classification task, the ensemble slightly outperformed its standalone models. The ensemble scored a macro-averaged F1 score of 0.838, while the two BERTs scored 0.826 and 0.833, and XLM-RoBERTa reached 0.8223. Furthermore, Portelli et al. (2022) calculated the agreement between the ensemble's constituent models via Cohen's Kappa Coefficient. The coefficient revealed that the constituent models with different architecture agree less with one another than models of the same type.

In the binary classification task, a standalone constituent model outperformed the ensemble. The two mBERTs reached a macro-averaged F1 score of 0.835 and 0.799, while the ensemble only scored 0.834. BERT got a score of 0.807. Cohen's Kappa Alpha revealed that all models had a higher agreement than those with a different architecture from the other ensemble. Portelli et al. (2022) conclude that ensembles of models with different architectures complement each other better, leading to higher ensemble performance.

Lin et al. (2022) combined previous approaches of using transformers pre-trained on semantically related data and ensemble for their contribution to the BioCreative VII challenge. Their task was to perform multi-label topic classification on COVID-19 literature to facilitate the search in fast-growing libraries such as PubMed. Various BERT models were tested in 10-fold cross-validation experiments to select suitable pre-trained models for the ensemble. The best performers, BioBERT, PubMedBERT, Sultan and Bioformer, were used in the choir. The models were fine-tuned and ensemble by averaging the logits of the constituent models and argmaxing over classes. The ensemble outperformed its constituent models with a macro-averaged F1 score of 0.9392, followed by its standalone models Bioformer which scored 0.9334 and BioBERT, 0.9346. Combining the approaches of pre-training transformers on semantically related data and ensembling transformers has led to state-of-the-art results for topic detection in COVID-19 literature, according to Lin et al. (2022).

3 Dataset

We use the data provided by Kirk et al. (2023) at the second level, broken down into four conceptually separate groups for sexist material.

Table 1: Class Sizes of Data

	Train	Dev	Test
Harm	310	44	89
Derogation	1590	227	454
Animosity	1165	167	333
Prejudice	333	48	94

On the second annotation level Kirk et al. (2023) split sexism into four categories Table 1: Threats, Derogation, Animosity and Prejudice. Threats entail any threats and plans to harm women directly or incite others to harm women via strategies for damage and stimulation. This sexism category is often marked with overt linguistic cues, e.g. with words like "kill" or "rape".

Derogatory language is intended to disparage, dehumanize, belittle, or disrespect women. It contains disparaging remarks and gender tropes, objectification of women's bodies, vehemently negative emotional statements, and dehumanizing parallels. It includes disparaging remarks about particular women and women in general. This category is also often expressed with linguistic cues, such as mentioning females together with negative attributes such as "slow", "clumsy" or "incompetent".

In the category of animosity misogyny, stereotypes and descriptive assertions are conveyed both subtly such as in backhanded compliments and overtly such as in gender slurs.

Prejudice uses wording that rationalizes sexism and refutes the presence of prejudice. It includes the rejection of gender inequality and its explanation, the justification of women's abuse, and the idea of masculine victimhood (Kirk et al., 2023).

4 Methodology

In Section 2, we discussed three main findings: First, pre-trained transformer-based language models are the state-of-the-art method for sexism detection. Secondly, re-training transformers with semantically related data improves performance on downstream tasks, if the downstream task is of the same semantic field or a smaller subfield of the pre-training data. Thirdly, ensembles of pre-trained transformer-based outperform their constituting standalone models.

Inspired by the third finding, we also chose an ensemble approach. Portelli et al. (2022) indicate that ensemble performance increases with a higher disagreement between constituent models. Following this idea, we investigated different approaches to obtain models that are suitable to the classification task and, at the same time, disagree enough to give other "perspectives" on the data. We identified three levels in which models could differ in an ensemble: model architecture, pre-training data and fine-tuning data. Since Portelli et al. (2022) has already investigated the benefits of models with different architectures within an ensemble, we wanted to explore ensembles with constituent models that differ regarding their pre-training or fine-tuning data.

We hypothesized that it can be beneficial to ensemble models that are not necessarily the best, as long as their domain knowledge is task-relevant and the models complement each other. We will refer to the constituting models of our ensembles as "Experts" since they do not need to be good at the entire task, but just especially good at one aspect of the task. Furthermore, we hypothesized that the "perspective" each standalone model has on the data needs to be coordinated efficiently. Therefore, we chose a neural network to coordinate the Experts' output. We will refer to this neural network as "Master". Based on these hypotheses we developed two ensemble approaches: "Master of Domain Experts" (MoDE) and "Master of Class Experts" (MoCE).

4.1 Master of Domain Experts (MoDE)

In this approach, experts gain their "expertise" through the domain of their pre-training data. Given limited computational power, we opted to test our underlying hypotheses rather than develop state-of-the-art classifiers. Therefore, we chose the minimum number of experts for this approach, which is two. The sexist classes pose different challenges to a classifier, from picking up overt linguistic patterns such as in the class "threats" to more implicit patterns such as in the class "Prejudice". Since we only have two experts, we wanted to use one expert that models general language phenomena and another model that models language phenomena more related to sexism. To avoid further computational costs when re-training on semantically related data, we chose HateBERT as one expert, since it proved to outperform its general-purpose counterpart on abusive language or its sub-phenomena. To avoid the model architecture becoming a confounding variable, we chose HateBERT's general-purpose counterpart, BERT-base-uncased, as the second expert.

Both Experts were fine-tuned on the train and validation data according to the organizer's split since we did not perform hyperparameter tuning for this approach. Since Schütz et al. (2021) results suggest a better performance when fine-tuning the whole model instead of only the classification head, we fine-tuned the experts on all layers with hyperparameters shown in Table 2. Furthermore, we fine-tuned two HateBERTs and two BERTs, each with different weight initialization to later not only create our ensemble of interest, HateBERT-BERT, but also HateBERT-HateBERT and BERT-BERT for further comparison.

After fine-tuning the Experts, their classification head was removed. To train the master, the training data was passed on to the Experts. The Experts outputted embeddings with a size of 512 each. The expert's embeddings were concatenated and further passed on to the master model, one layer deep NN, that finally outputs the class prediction, as shown in Figure 2. The master was trained with hyperparameters shown in Table 2

4.2 Master of Class Experts (MoCE)

In this approach, we tried to improve the individual learning of the experts by breaking down the multi-class classification problem into multiple binary classification problems. Since the data tends to be quite complicated, with classes posing different linguistic challenges in their detection, this approach aimed to create four classifiers that are each optimized for the challenges the respective class poses.

To reduce the overall computational costs of training the experts and its performance advantage on abuse language over its general-purpose counterpart, we went with HateBERT for each expert. Then, based on how the task organizer had broken up the training and test instances, the experts were fine-tuned over the entire model and evaluated on the validation set. After a rough hyperparameter-tuning, we obtained the hyperparameters as shown in Table 3 with which we retrained the models on train and validation data.

After training the Experts, their classification head is again removed. To train the Master the train and validation data is passed to the Experts.

Table 2: Training Details for MoDE

Expert Batch size	1
Master Batch Size	1
Expert Epochs	50
Master Epochs	50
Learning Rate	1.5e-6

Table 3: Training Details for MoCE

Expert Batch size	1
Master Batch Size	8
Expert Epochs	10
Master Epochs	200
Learning Rate	3e-4

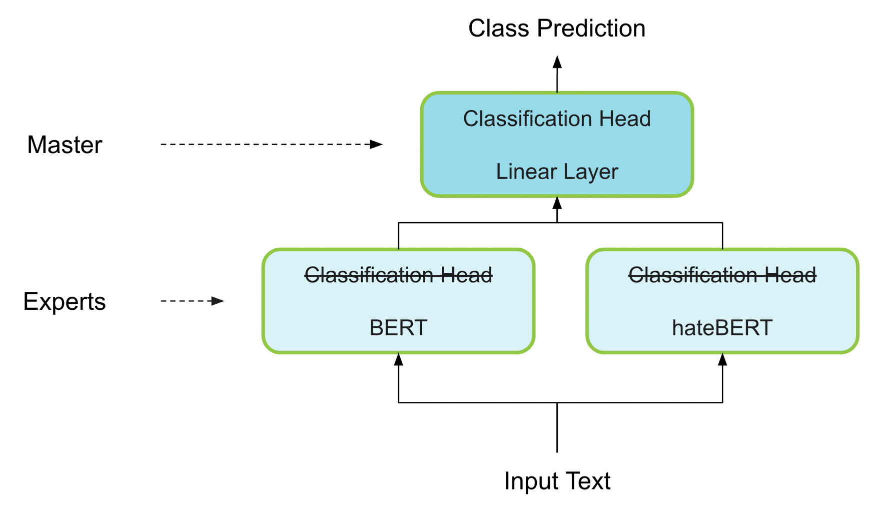

Figure 2: Master of Domain Experts Architecture

Their embeddings with a dimension of 512 are concatenated and passed on to a four-layer deep, fully connected network, see Figure 3. The master's classification head finally gives the classification.

5 Results and Discussion

We chose the macro-averaged F1 score as an evaluation metric to give equal importance to all classes independent of future class-balancing changes and ensure that the results are as consistent as possible. The results from both the baselines and our two approaches are shown in Table 4

As the first baseline, the experts of MoDE were tested. HateBERT outperformed BERT in every class, reaching a macro-averaged F1 score of 0.5197 instead of BERT's score of 0.4526. This finding is consistent with Caselli et al. (2021) 's conclusion that HateBERT outperforms BERT in detecting abusive language or its sub-phenomena. Furthermore, BERT and HateBERT show the same pattern in performance over classes: "Harm" and "Derogation" were better detected than "Animosity" and "Prejudice". The reason for that could be that these classes are often characterized with solid linguistic cues such as "kill" or "rape" in the category "Harm" or words describing females co-occurring with negative attributes such as "slow" or "clumsy" in the category "Derogation". The lower performance for "Animosity" and "Prejudice" is consistent with our assumption that these classes are more difficult to classify due to less overt linguistic marking and more dependent on context. As an example, "backhanded compliment" is a sub-phenomenon of "Animosity" and occurs in utterances such as "Women are delicate flowers who need to be cherished". Statements like this are marked with generally positively attributed words, but the belittlement and implication of inferiority only reveal themselves in the context. "Prejudice" a priori seems to be the hardest to capture since it requires assessing whether the utterance author denies, understates, or seeks to justify the discrimination of women. As expected, the standalone models perform the worst in this class.

For the MoDE approach, all three ensemble configurations were evaluated: BERT-BERT, HateBERT-HateBERT and our ensemble of interest, HateBERT-BERT. As expected BERT-BERT was outperformed by HateBERT-HateBERT, which scored 0.5288. Against our hypothesis, our ensemble of interest HateBERT-BERT only reached a score of 0.4936, therefore being outperformed by HateBERT-HateBERT and even standalone HateBERT. The results suggest that standalone HateBERT has such a performance advantage in sexism detection that the strength of the standalone model outweighs any potential advantage of having disagreeing constituent models in our ensemble. All ensemble's performance over classes shows the same trend as in the performance of the standalone models: "Harm" and "Derogation" were detected better than "Animosity and Prejudice".

In contrast, the MoCE scored a much lower macro-averaged F1 score of 37.62 and an extremely low score of only 0.1537 for the class "Prejudice". Moreover, the performance pattern over classes was inconsistent with the pattern appearing in the standalone models and all three ensembles for the

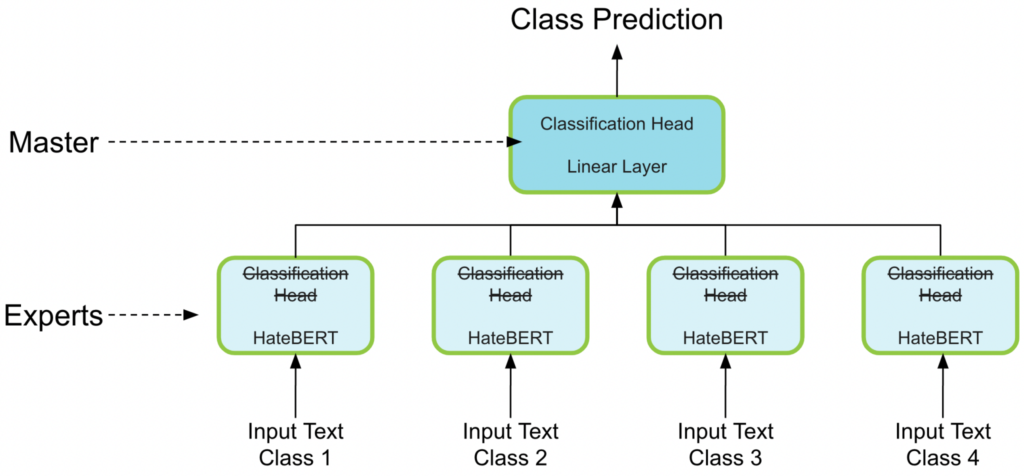

Figure 3: Master of Class Experts Architecture

Table 4: Results for MoDE and MoCE

	Expert Baselines		**MoDE**			**MoCE**
	BERT	HateBERT	BERT/HateBERT	BERT/BERT	HateBERT/HateBERT	HateBERT
Harm	0.5212	0.5793	0.5534	0.5398	0.5815	0.2701
Derogation	0.5075	0.5354	0.5777	0.5641	0.5381	0.5543
Animosity	0.3971	0.4957	0.4054	0.4519	0.4902	0.4205
Prejudice	0.3846	0.4685	0.4378	0.4042	0.5054	0.1537
Macro F1	0.4526	0.5197	0.4936	0.4900	0.5288	0.3762

MoDE approach. Despite the overt linguistic cues of the class "Harm" and the relatively high scores of all other models, suggesting that it is easy to learn, MoCE only scored 0.2701. In contrast, "Animosity" scored relatively high with 0.4205, despite our assumption that learning should be exceptionally intricate due to sometimes misleading linguistic cues. The performance was relatively high for classes with rather big class sizes at training and relatively low for classes with small class sizes. The MoCE was trained on balanced data. However, this approach means that the class expert for "Harm" was trained on a dataset of only 620 instances, while the class expert for "Derogation" was trained on a dataset of 3180 cases. Furthermore, the Master's first linear layer received twice as much input, since it received the 512-dimensional embeddings of four as opposed to just two experts. Additionally, the Master is four layers deep, which makes it a quite complex model given the relatively scarce training data. This challenge might be another factor that leads to the low, seemingly unstable performance of MoCE.

6 Future Work

Multiple aspects could be explored to improve both approaches. For the MoCE approach, we suggest experimenting with data-balancing. Instead of perfect class balancing by downsampling, one could experiment with upsampling and balancing with different ratios. For the MoDE approach, different models could be chosen as Experts and one could experiment with the number of “copies” of each domain expert, each copy just being different in their weight initialization. We hypothesize that the marginal benefit of having one more copy of a domain expert is gonna decrease while the marginal benefit of one more qualitatively different Expert should increase. Furthermore, an optimal depth of the MoDE Master should be identified.

7 Conclusion

In this work, we investigated the effectiveness of different ensemble approaches for sexism detection.

To this end, we use Kirk et al. (2023)'s dataset, breaking sexism into four distinct categories. We developed two different ensembles called the Master of Domain Expert (MoDE) and Master of Class Expert (MoCE). Both of our ensemble approaches revolve around using large language models, in this work HateBERT, which allows the models to focus on different aspects of the data, thus allowing the whole ensemble to gain better classification results. The primary idea behind MoDE was to see if having multiple experts trained on the entire dataset performs better than using a single model. While the concept behind MoCE was to see if breaking down the data into a binary classification problem would improve the experts' classification. While MoDE with HateBERT-HateBERT performs decently, even beating the standalone HateBERT baseline, MoCE is quite unstable and performs worse than the standalone BERT baseline. MoCE obtains a macro F1 score of 0.3762 whereas MoDE got a score of 0.5288.

References

Zeyuan Allen-Zhu and Yuanzhi Li. 2020. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *ArXiv*, abs/2012.09816.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Lily Davies, Marta Baldracchi, Carlo Alessandro Borella, and Konstantinos Perifanos. 2021. Transformer ensembles for sexism detection. *CoRR*, abs/2110.15905.

Morteza Janatdoust, Fatemeh Ehsani-Besheli, and Hossein Zeinali. 2022. KADO@LT-EDI-ACL2022: BERT-based ensembles for detecting signs of depression from social media text. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 265–269, Dublin, Ireland. Association for Computational Linguistics.

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

Sheng-Jie Lin, Wen-Chao Yeh, Yu-Wen Chiu, Yung-Chun Chang, Min-Huei Hsu, Yi-Shin Chen, and Wen-Lian Hsu. 2022. A BERT-based ensemble learning approach for the BioCreative VII challenges: full-text chemical identification and multi-label classification in PubMed articles. *Database*, 2022. Baac056.

Beatrice Portelli, Simone Scaboro, Emmanuele Chersoni, Enrico Santus, and Giuseppe Serra. 2022. AILAB-Udine@SMM4H'22: Limits of transformers and BERT ensembles. In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pages 130–134, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Mina Schütz, Jaqueline Boeck, Daria Liakhovets, Djordje Slijepcevic, Armin Kirchknopf, Manuel Hecht, Johannes Bogensperger, Sven Schlarb, Alexander Schindler, and Matthias Zeppelzauer. 2021. Automatic sexism detection with multilingual transformer models ait fhstp@exist2021. In *IberLEF@SEPLN*.