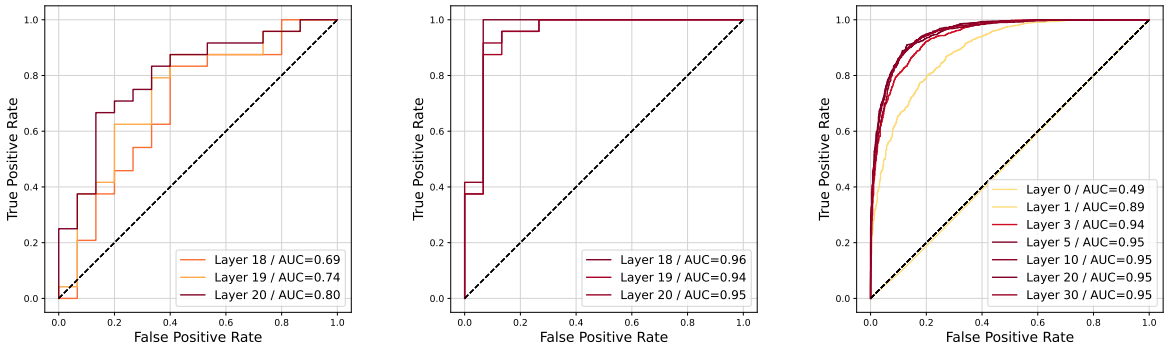
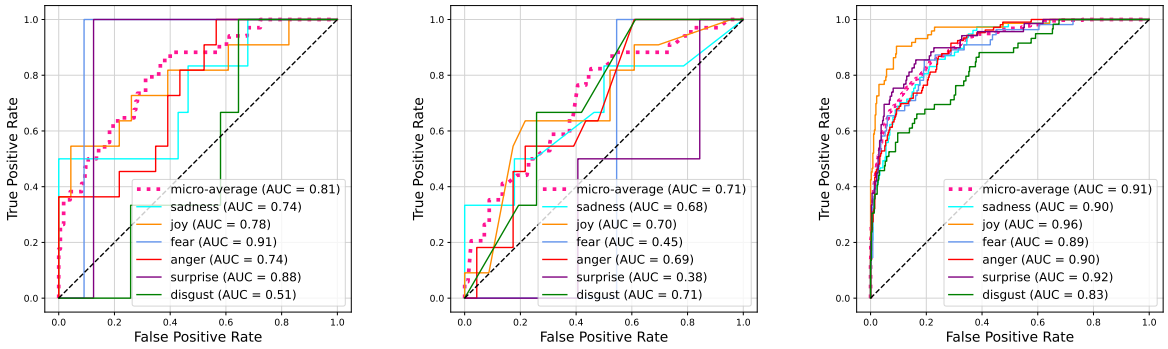



(a) Trained steering vectors (b) Corresponding activation vectors (c) Activation vectors of 17k sentences

Figure 6: Comparison between the classification results on the Shakespeare dataset: Using (a) only the trained steering vectors, (b) the corresponding activation vectors, and (c) activation vectors of 17k sentences for selected layers.



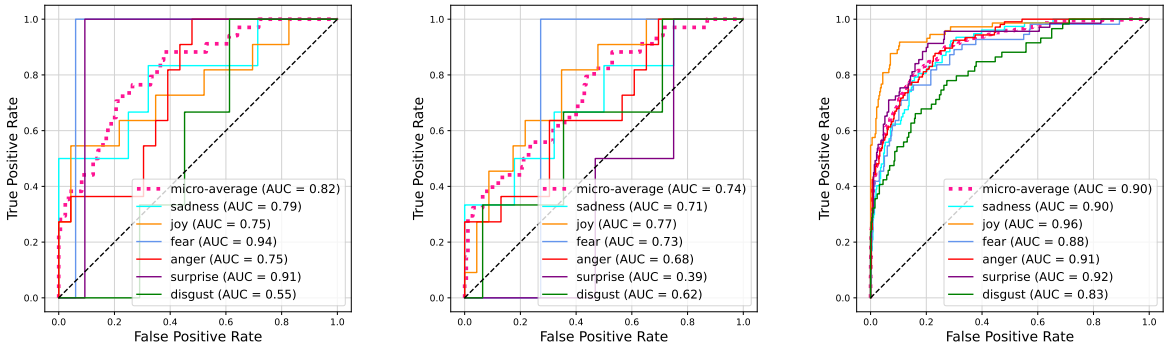

(a) Trained steering vectors (b) Corresponding activation vectors (c) Activation vectors of 2k sentences

Figure 7: Classification results of vectors from layer 18 on the GoEmotions dataset: Using (a) only the trained steering vectors, (b) the corresponding activation vectors, and (c) activation vectors of 2k sentences. The activation vectors only show superior performance if we include more sentences than we have trained steering vectors.

## D Further classification-based evaluation results for output steering

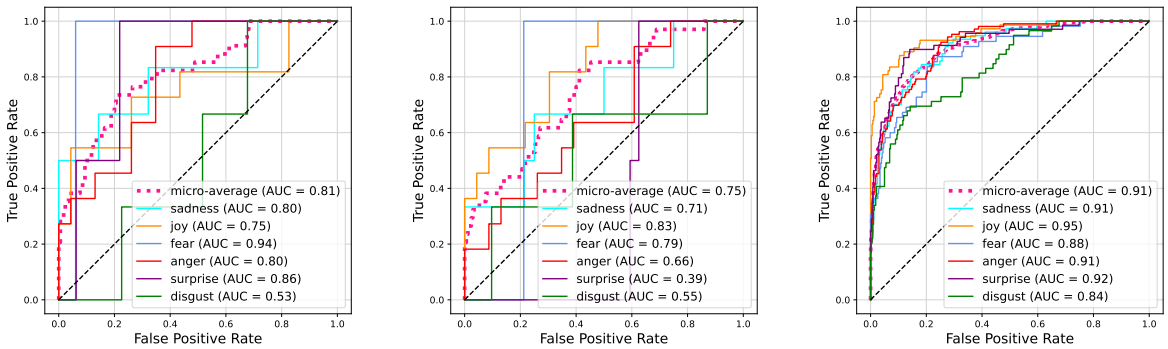
This section compares the training-based style vectors with their corresponding activation-based style vectors. We do this to ensure fairness in the comparison since the number of activation-based style vectors is significantly higher than the number of training-based vectors. In the evaluation of the factual (Fig. 10) and subjective (Fig. 12) prompts using the training-based style vectors on the GoEmotions dataset, we saw that the steering seems to work for all emotions, except disgust and surprise. However, during a closer examination, it became evident that the model's output with $\lambda \geq 0.75$ did not represent proper sentences anymore and were mainly repetitions of keywords related to the emotion, e.g., "sadly" for sadness. For the Yelp dataset, this happened as well, but only for higher $\lambda$. A reason for this unstable behavior in GoEmotions is probably the small number of trained steering vectors that were found, which was especially low for the classes *disgust* and *surprise*.

The steering is much more stable for the activation-based style vectors for factual prompts (Fig. 11), while the subjective are not steered well (Fig. 13) prompts. The generated sentences seem to be biased towards *joy*. Especially, *disgust* does not seem to be steered. These results, especially in comparison to the steering with all activation-based style vectors (5), are, again, the result of the small number of trained steering vectors, which limits the amount of available activation-based style vectors. This, furthermore, highlights the superiority of the activation-based style vectors, which can be just extracted and do not require a computationally expensive learning procedure.



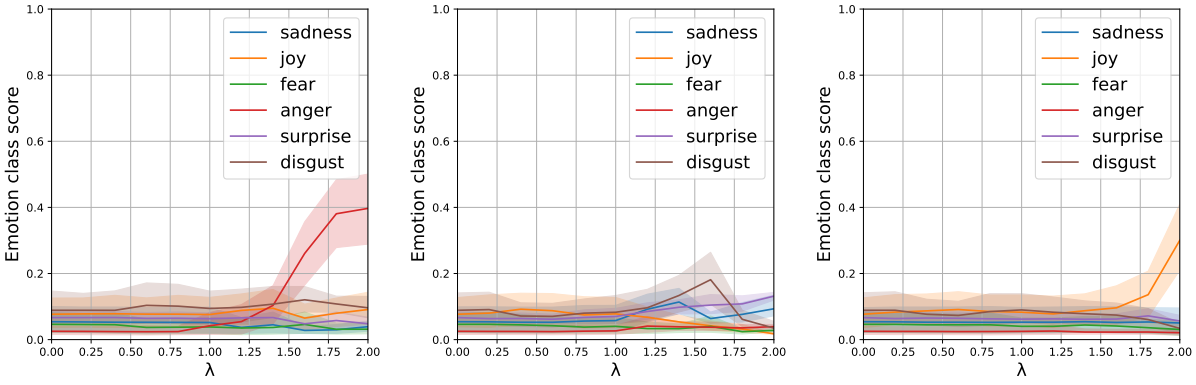

(a) Trained steering vectors (b) Corresponding activation vectors (c) Activation vectors of 2k sentences

Figure 8: Classification results of vectors from layer 19 on the GoEmotions dataset: Using (a) only the trained steering vectors, (b) the corresponding activation vectors, and (c) activation vectors of 2k sentences. The activation vectors only show superior performance if we include more sentences than we have trained steering vectors.




(a) Trained steering vectors (b) Corresponding activation vectors (c) Activation vectors of 2k sentences

Figure 9: Classification results of vectors from layer 20 on the GoEmotions dataset: Using (a) only the trained steering vectors, (b) the corresponding activation vectors, and (c) activation vectors of 2k sentences. The activation vectors only show superior performance if we include more sentences than we have trained steering vectors.




(a) Steering to anger, factual prompts

(b) Steering to disgust, factual prompts

(c) Steering to joy, factual prompts

(d) Steering to fear, factual prompts

(e) Steering to sadness, factual prompts

(f) Steering to surprise, factual prompts

Figure 10: Training-based style vectors: Evaluation of generated texts for *factual* prompts using GoEmotions' style vectors.