GerAV: Towards New Heights in German Authorship Verification using Fine-Tuned LLMs on a New Benchmark

Lotta Kiefer[1]*, Christoph Leiter[2]*, Sotaro Takeshita[3], Elena Schmidt, Steffen Eger[4]

[1,2,3,4]University of Technology Nuremberg (UTN)

[1]lotta.kiefer@utn.de [2]christoph.leiter@utn.de [4]steffen.eger@utn.de

Abstract

Authorship verification (AV) is the task of determining whether two texts were written by the same author and has been studied extensively, predominantly for English data. In contrast, large-scale benchmarks and systematic evaluations for other languages remain scarce. We address this gap by introducing GerAV, a comprehensive benchmark for German AV comprising over 400k labeled text pairs. GerAV is built from Twitter and Reddit data, with the Reddit part further divided into in-domain and cross-domain message-based subsets, as well as a profile-based subset. This design enables controlled analysis of the effects of data source, topical domain, and text length. Using the provided training splits, we conduct a systematic evaluation of strong baselines and state-of-the-art models and find that our best approach, a fine-tuned large language model, outperforms recent baselines by up to 0.09 absolute F1 score and surpasses GPT-5 in a zero-shot setting by 0.08. We further observe a trade-off between specialization and generalization: models trained on specific data types perform best under matching conditions but generalize less well across data regimes, a limitation that can be mitigated by combining training sources. Overall, GerAV provides a challenging and versatile benchmark for advancing research on German and cross-domain AV.[1]

1 Introduction

Authorship analysis aims to assess the likelihood that a text was produced by a particular individual based on stylistic characteristics. It has important applications in a range of domains, including plagiarism detection, the analysis of misinformation spread, and forensic investigations (Ramnath et al., 2025). In particular, the German Federal Criminal Police Office (Bundeskriminalamt) reports the use of author recognition as a forensic tool when written texts are relevant to criminal offenses (BKA, 2025). Given the sensitive and high-stakes nature of these applications, the development of reliable and high-performing methods for authorship analysis is of critical importance.

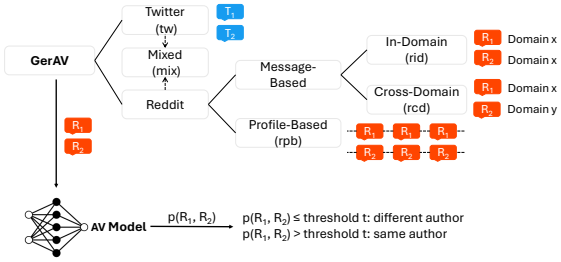

Figure 1: Overview of our approach. The figure presents the GerAV benchmark and its structure, illustrating how the datasets are used as input for AV models during training and testing. Based on a decision threshold, the models predict whether a given pair of texts was written by the same or by different authors. Each pair consists of one or several posts from a single author, contrasted with post(s) from either the same or another author. The figure represents these post pairs as T_1 and T_2 (Twitter, in blue) and R_1 and R_2 (Reddit, in orange).

Authorship analysis is typically divided into two closely related tasks: authorship attribution (AA) and authorship verification (AV). While AA identifies the author of a text from a closed set of candidates, AV determines whether two given texts were written by the same author. This work focuses on AV, which better reflects real-world forensic scenarios where candidate authors are unknown.

While authorship analysis has received considerable attention in the Natural Language Processing (NLP) literature, particularly through feature-based and neural approaches (Tyo et al., 2023), the use of large language models (LLMs) remains underexplored (Huang et al., 2025). Moreover, most existing methods are evaluated primarily on English benchmarks, a limitation that is especially

*Equal contribution

[1]Our code and information about data access are available on GitHub.

Findings of the Association for Computational Linguistics: ACL 2026, pages 40050–40069

problematic given that authorship analysis relies on language-specific stylistic and linguistic dependencies.

To address these gaps, we make the following contributions:

- ✓ ***GerAV* benchmark**: We introduce GerAV, one of the largest benchmarks for the evaluation of **Ger**man **A**uthorship **V**erification. Comprising a substantial number of German-language online posts (over 400k message pairs), it enables robust and diverse evaluation scenarios. The benchmark includes subsets from different data sources and domains, allowing us to analyze their effects on the evaluation models. Additionally, we provide a profile-based subset in which messages from each author's profile are concatenated to evaluate model performance when more data is available.

- ✓ **Broad comparison of baseline models**: We perform a broad evaluation of established AV methods, recent approaches, and zero-shot LLMs, which, to our knowledge, is the most comprehensive AV evaluation to date that includes recent LLMs.

- ✓ **State-of-the-art models for German AV**: Using distinct train splits of the GerAV datasets, we fine-tune several open-source LLMs for the task of German AV. By doing so, several model versions strongly outperform all existing baselines. Most notably, Gemma-3-12b, tuned on a mix of train splits from all GerAV subsets, achieves an average F1-Score of 0.83 across all test splits. This is 0.09 higher than the best baseline at 0.74. On a small test subset, this model also beats GPT-5 by 0.08 points F1-Score while being significantly smaller in size.

- ✓ **Exploring cross-domain and message length**: We explore performance differences when tuning and evaluating on in-domain vs. cross-domain data and on single messages vs. whole profiles. The baseline methods' performance (including LLMs) increases with input length, from an F1-score of approx. 0.78 at 25 words to up to almost 1.0 at 900 words. For our tuned models, this is only the case if the training data contains long examples. Also, the average performance across all models on the GerAV cross-domain subset is 0.06 points lower than on the in-domain subset. Tuning on cross-domain data solves this issue, but leads to weaker performance on in-domain data and other datasets.

- ✓ **Investigating low-resource and cross-language scenarios**: We evaluate the robustness of our approach under reduced training data conditions and find that decreasing the number of training samples to one quarter results in only a 0.04 drop in F1. Furthermore, we highlight the importance of language-specific AV benchmarks and models by applying a model trained on English Twitter data to our GerAV Twitter test set, where we observe a strong decrease of 0.12 in F1.

Figure 1 displays an overview of our approach.

2 Related Work

Feature-Based and Neural Approaches In the NLP literature, AA and AV methods can be broadly categorized into three classes: feature-based approaches, neural models, and methods leveraging LLMs. Feature-based methods rely on handcrafted representations derived from morphological, syntactic, and semantic properties, such as character n-gram frequencies, part-of-speech variability, and semantic dependency features (Stamatatos, 2009), with short character n-grams being especially effective (Grieve, 2007; Stamatatos, 2006). Although these approaches are highly interpretable and offer transparent explanations, they typically underperform compared to more complex models (Zeng et al., 2025).

Neural approaches alleviate the need for explicit feature engineering by learning representations directly from data. Prior work has explored recurrent neural networks (RNNs) (Gupta et al., 2019), long short-term memory networks (LSTMs) (Qian et al., 2017), convolutional neural networks (CNNs) (Hossain et al., 2021), and attention-based siamese architectures (Boenninghoff et al., 2019). More recently, pretrained transformer models such as BERT (Devlin et al., 2019), its variants, and Sentence-BERT (Reimers and Gurevych, 2019) have achieved strong performance in AA and AV (Fabien et al., 2020; Manolache et al., 2021; Rivera-Soto et al., 2021; Schlicht and de Paula, 2021). While these models offer superior performance and

domain adaptability, they generally lack the interpretability of feature-based methods.

Large Language Models Authorship analysis, like many other NLP tasks, is increasingly influenced by advances in LLMs. Huang et al. (2024) investigate zero- and few-shot prompting strategies with GPT-3.5 (OpenAI, 2022) and GPT-4 (OpenAI, 2023), showing that linguistically informed prompting (LIP) can outperform traditional baselines without task-specific training. However, reliance on online APIs raises concerns about reliability, reproducibility, and data privacy (Ramnath et al., 2025), limiting use in sensitive domains such as forensic analysis. To overcome this, Ramnath et al. (2025) employ offline LLMs in zero- and few-shot settings and further improve performance through fine-tuning. Generating rationales through carefully designed prompts enhances explainability, reduces output variability, and surpasses both GPT-4 zero-shot and feature-based baselines in their work. Similarly, Hu et al. (2024) explore instruction-tuning for AV using OPT (Zhang et al., 2022) and LLaMA (Touvron et al., 2023b,a) models, integrating explanations directly into the model output and benchmarking against BERT-based and zero-shot approaches.

Despite these advances, the potential of LLMs for authorship analysis remains underexplored (Huang et al., 2025). Existing studies are limited in scope, often evaluating only a small number of models or baselines. For example, Ramnath et al. (2025) and Hu et al. (2024) each compare their approaches against only a single non-LLM baseline method, which is insufficient to establish superiority over alternative baselines that have demonstrated strong performance in prior work. Although several surveys and shared tasks exist for AA and AV (Stamatatos, 2009; Kestemont et al., 2020; Bevendorff et al., 2021; He et al., 2024; Tyo et al., 2023), LLMs are largely absent from these evaluations. Moreover, comparisons across studies are hindered by inconsistent datasets, splits, and evaluation metrics (Tyo et al., 2023), underscoring the need for systematic and comprehensive benchmarking including LLMs. We address this gap by conducting a detailed evaluation of zero-shot and tuned LLMs against a wide range of high-performing baselines.

Multilingual Settings Authorial style is inherently shaped by language-specific grammar, morphology, and writing conventions, making language-aligned evaluation essential. Without it, models risk capturing language-dependent artifacts rather than true stylistic signals. While some work has explored multilingual modeling (Qiu et al., 2025; Kim et al., 2025), non-English resources remain limited. Few datasets address this gap: Israeli et al. (2025) introduce the One Million Author Corpus with Wikipedia articles across dozens of languages, including over 100,000 German authors, with cross-domain samples from Wikipedia namespaces; Murauer and Specht (2019) construct a multilingual AV dataset from Reddit spanning many languages, including 80,000 German posts, using subreddit membership to induce cross-topic variation; and Halvani et al. (2016) evaluate AV across five languages, including a small German corpus from news, reviews, forums, and novels. Closest to our work, Boenninghoff et al. (2024) present a German AV dataset derived from a newspaper discussion forum. However, their work focuses on author diarization in chronologically ordered text streams, whereas our benchmark emphasizes diverse and realistic evaluation settings tailored to forensic applications.

This work contributes to language-specific AV research by introducing a comprehensive and challenging German benchmark from two major social media platforms that reflects a range of real-world use cases and enables extensive evaluation under language-specific conditions, thereby addressing the lack of diverse non-English benchmarks.

3 Data Curation

For this study, we introduce two large German-language source corpora derived from Reddit and Twitter. From these corpora, we construct five benchmark datasets for AV, covering in-domain, cross-domain, profile-based, and mixed-source settings (see Appendix A for examples of each dataset and their English translations). The datasets are derived from Twitter and Reddit posts and therefore consist of relatively short texts compared to existing benchmark sources, e.g., derived from novels (Bogdanova and Lazaridou, 2014), Wikipedia articles (Israeli et al., 2025), or news (Liu, 2006). This characteristic makes the datasets particularly suitable for training systems intended for forensic applications, in which AV is applied to online forums (e.g., incriminated darknet forums or Telegram groups) to identify an active author within a given forum. In the following, the curation and

preprocessing steps of each dataset are described.

Reddit GerAV We collect a new German Reddit corpus aiming at covering German-language subreddits as comprehensively as possible. Starting from a list of German seed subreddits, we retrieved all posts and iteratively expanded the dataset by following linked usernames to other subreddits, assuming that users active in German communities are likely to participate in additional German-language subreddits.

Using the *langid* [2] Python package, we retained only posts classified as German, yielding 787,372 posts from 120,538 users across 182 subreddits, being nearly ten times bigger than the German part of the Reddit dataset by Murauer and Specht (2019). Bot accounts are filtered by excluding usernames containing certain keywords and users with unusually high post counts. To enable pair construction, users with fewer than two posts are removed. To focus on meaningful stylistic features, only posts with a minimum of 25 words are retained. Posts exceeding 1,000 words, containing URLs or user mentions are excluded to reduce noise.

From this corpus, we derive three benchmark AV datasets: in-domain, cross-domain, and profile-based. For all datasets, authors are partitioned into training (60%), validation (20%), and test (20%) splits with no author overlap. For each author, up to two positive pairs (two posts by the same author) and two negative pairs (a post paired with one from a different author in the same split) are sampled. Pair limits keep dataset sizes manageable for training large models, reducing computational cost and training time. We enforce class balance by ensuring an equal number of positive and negative pairs in each split, resulting in balanced binary classification datasets.

To construct the **reddit cross-domain dataset (rcd)**, we leverage topic separation across subreddits by prompting GPT4o (Hurst et al., 2024), resulting in 14 distinct topical domains that were checked manually for plausibility (see Appendix B for the prompt and C for the resulting clusters). In the **reddit in-domain dataset (rid)**, paired posts always come from the same domain, while in the cross-domain dataset, pairs are drawn from different domains. This cross-domain benchmark creates a more challenging evaluation, which both increases AV difficulty (Kestemont et al., 2020) and reflects real-world applications. Furthermore, this setting enables the investigation of content bias, assessing whether models rely on topical similarity or genuinely capture stylistic signals. Strong performance in topically unrelated settings provides evidence that the models are learning stylistic characteristics rather than merely exploiting content similarities.

Finally, the **reddit profile-based dataset (rpb)** is created by concatenating all posts from each user in random order into a post list. For each user, a random split point is selected at the end of a post, falling between 30% and 70% of the total words in their post list. Negative pairs are sampled to ensure that the length difference distributions of positive and negative pairs is similar, thereby minimizing potential length-based biases that a model could exploit. This dataset enables evaluation at the user-profile level, reflecting a realistic scenario in which all available texts of an author are leveraged to make more reliable predictions.

Twitter GerAV The **Twitter dataset (tw)** is created from Kratzke (2023), which contains monthly German Twitter updates from April 2019 to December 2022, and is accessible to researchers upon request. To further ensure the selection of German-language content, *langid* is applied to all texts, resulting in a dataset of 606,588 posts from 54,544 unique authors. In line with the Reddit dataset, users with fewer than two posts are removed, and tweets containing fewer than 25 words are excluded. To mitigate the effect of outliers, tweets exceeding 300 characters and users with more than 200 tweets in total are excluded.

Consistent with the Reddit GerAV dataset, the final Twitter GerAV dataset is constructed by creating training, validation, and test splits with no author overlap. For each author, up to two positive and two negative samples are generated, resulting in a balanced dataset.

Mixed GerAV As a final dataset, we combine the three Reddit datasets (rid, rcd, and rpb) and the Twitter dataset by sampling 20,000 pairs for the training set and 4,000 each for validation and test sets from each source while maintaining a balanced distribution of labels, resulting in a **mixed dataset (mix)**. For an evaluation of GPT-5, we further create a small version of the mixed dataset that covers 480 samples (120 per dataset).

Table 1 summarizes dataset statistics, including the number of samples, unique posts and users, and mean sample length for each dataset. The

[2] https://pypi.org/project/langid/

Dataset	Sample Number	Unique Posts	Unique Users	Mean Sample Len (in Words)
Reddit in-domain	62,328	87,979	23,083	69
Reddit cross-domain	62,328	82,902	22,577	67
Reddit profile-based	46,166	50,775	23,083	326
Twitter	120,858	133,329	35,181	34
Mixed	112,000	158,561	47,816	124
Mixed (small)	480	952	692	112

Table 1: Statistical comparison of all GerAV datasets.

post length distributions for the Reddit and Twitter datasets before preprocessing are provided in Appendix D. Both datasets are dominated by short posts, with frequencies decreasing rapidly as length increases. Twitter posts are capped at roughly 75 words, reflecting its maximum post length, while Reddit exhibits a long-tailed distribution.

4 Experiment Setup

In this section, we describe the evaluation measures we use and the methods that we benchmark for German AV. This includes (1) established and recent baselines, (2) zero-shot LLMs, and (3) LLMs with LoRAs (Hu et al., 2022) fine-tuned on our training sets.

Evaluation Measures To evaluate the AV methods on GerAV, we measure accuracy in correctly identifying which messages are written by the same or different authors. Further, we measure the F1-score to identify the methods with the highest precision and recall. As a third measure commonly used for AV evaluation, we choose ROC AUC[3] that quantifies how well similarity/distance scores returned by the metric can separate samples from same and different authors with arbitrary thresholds. Note that the first two measures use binary labels, while ROC AUC requires numeric scores. To measure statistical significance, we perform paired non-parametric bootstrap tests for model comparison (Efron, 1979). We generate 10,000 bootstrap resamples per dataset and compute F1 scores for each model on identical resampled labels, yielding paired F1 distributions. For each model pair, the p-value was estimated as the proportion of samples in which the higher-scoring model performed no better than its competitor, providing a direct measure of whether observed differences can be attributed to sampling variability.

AV Baselines To cover a diverse set of approaches, we start by following Tyo et al. (2023), who identify four high-performing AV methods with proven effectiveness (Fabien et al., 2020; Murauer and Specht, 2021; Kestemont et al., 2020; Neal et al., 2017). The first method *ngram* is a feature-based approach by Weerasinghe et al. (2021), which represents texts as n-gram character feature vectors and feeds them into a logistic regression classifier. The second baseline, *Prediction by Partial Matching (ppm)* Teahan and Harper (2003), builds a character-level language model for one text and computes cross-entropy on the second text using a ppm compression model. The third method, *hlstm*, introduced by Boenninghoff et al. (2019), employs a hierarchical Bi-LSTM to process text at the character, word, and sentence levels, producing unified style representations for AV. The final baseline *sbert* of this set uses BERT as the backbone for a siamese network trained with contrastive loss. We adapt the implementations to suit German inputs, e.g, BERT has been substituted with its German version (Chan et al., 2019). We train each baseline on all five newly created GerAV datasets, resulting in a total of 20 models.

Furthermore, we evaluate the recent style embedding models *mStyleDistance* (Qiu et al., 2025) that includes German in its training data and *Multilingual Style Representation* (we abbreviate it as *msr*) (Kim et al., 2025), which is multilingual but was not trained on German data. For AV, the cosine similarity between styles is considered.

To evaluate the baseline models using accuracy and F1 score, we first determine a decision threshold t.[4] The threshold is tuned on the respective validation sets using Youden's J statistic (Youden, 1950): we compute the ROC curve over the validation data and choose the threshold that maximizes the sum of sensitivity and specificity.

Zero-Shot LLMs We also benchmark the zero-shot performance of four recent LLMs: *gemma_3_12b_it* (Team, 2025), *llama-3.1-8b-instruct* (Meta AI, 2024a), *llama-3.2-3b-instruct* (Meta AI, 2024b), and *qwen-2.5-7b-instruct* (Yang et al., 2025) using the following prompt template:

> *Are the following two texts written by the same author?*
> *Text A:* {text_a}
> *Text B:* {text_b}
> *Please answer with "Yes" or "No".*
> *Answer:*

[3] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html

[4] Scores above the threshold indicate "same author" while scores below the threshold indicate "different author".

During inference, we generate tokens until the first "yes"-token (Yes, yes, YES) or "no"-token (No, no, NO) is generated. Then, to get a more accurate confidence estimate, we calculate the prediction score as $s = p(\text{“}yes\text{”} - token) - p(\text{“}no\text{”} - token)$ where we consider the 10 most likely tokens. We use s as the input score for the ROC AUC measure. Positive values of s mean that the probability of two texts being written by the same author is higher than by different authors. Again, we tune the threshold on the validation set.

As an additional zero-shot model, we evaluate and compare the closed-source model GPT-5 (OpenAI, 2025) on 120 subsamples of each dataset besides mixed. For GPT-5, we (1) use the same zero-shot prompt as above and (2) also evaluate a prompt using the LIP prompt component (Huang et al., 2024) that instructs the model to first analyze stylistic features of both texts, as well as (3) a version of LIP translated to German (see Appendix E for prompt details).

Tuned LLMs We also train LoRAs on each of the four LLMs (*gemma_3_12b_it*, *llama-3.1-8b-instruct*, *llama-3.2-3b-instruct*, and *qwen-2.5-7b-instruct*). For each of our five datasets (Twitter, Reddit in-domain, Reddit cross-domain, Reddit profile-based and Mixed), we train one LoRA per LLM, yielding 20 LoRAs in total. The input prompt follows the zero-shot format but omits "Please answer with "Yes" or "No.", as the required output format will be learned during tuning. Information about the training hardware, hyperparameters, runtime and licences can be found in Appendix F. Because the models are tuned on the training sets, we do not perform threshold tuning and use 0 as a threshold for accuracy and F1-score.

5 Experiment Results

Figure 2 shows the F1-scores of the baselines and tuned LLMs on our datasets (see Appendix G for the full results). We first consider the overall performance of the models. We then evaluate the effect of different data sources by comparing results on Twitter and Reddit data. Subsequently, we compare results on Reddit in-domain data with cross-domain data. We also evaluate the effect of text length by comparing message-based with profile-based evaluation. Finally, we present the comparison with closed-source models on a separate subset and analyze the effects of reduced training data and cross-lingual evaluation settings.

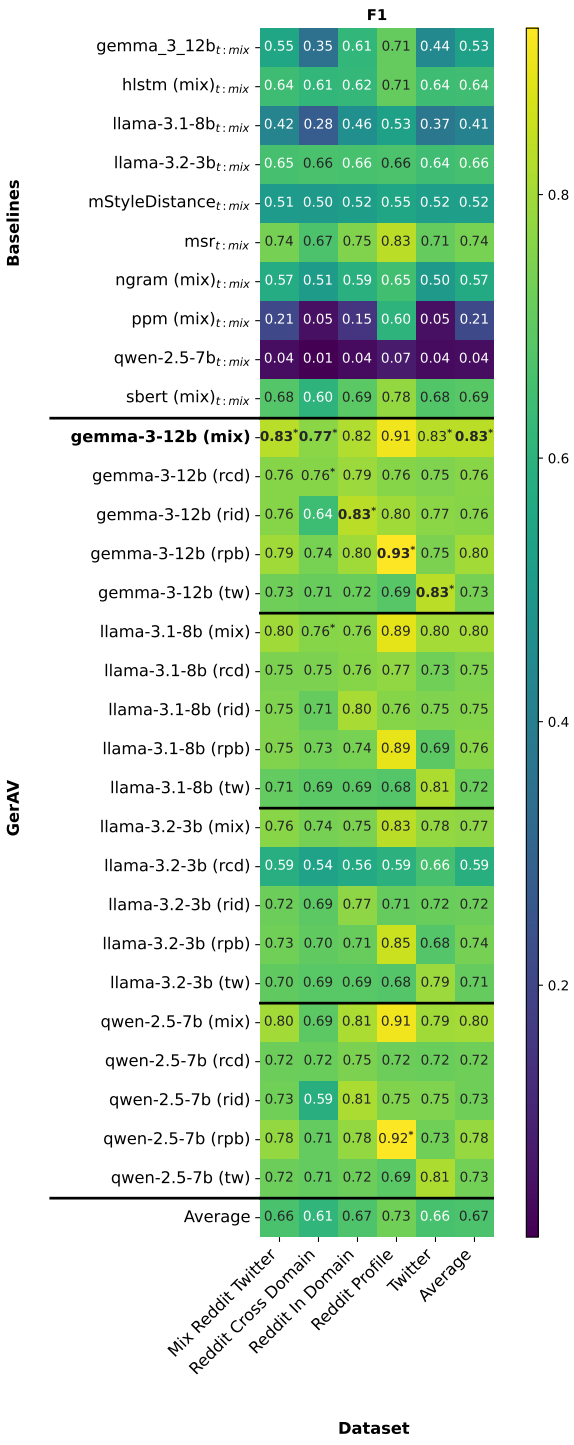

Figure 2: F1-Scores for the baselines and our GerAV models. The y-axis shows the model names and the x-axis shows the evaluated test set. Brackets indicate the training set used: **Mix** Reddit Twitter (mix), **R**eddit **C**ross **D**omain (rcd), **R**eddit **I**n **D**omain (rid), **R**eddit **P**rofile **B**ased (rpb) and **T**witter (tw). The subscript t denotes the validation set used to tune decision thresholds for the baselines. Column-wise best results are shown in bold. The best-performing model groups for each dataset (defined as those that statistically outperform all other models ($p < 0.01$) and show no statistically significant differences within the group) are marked with an asterisk next to the score.

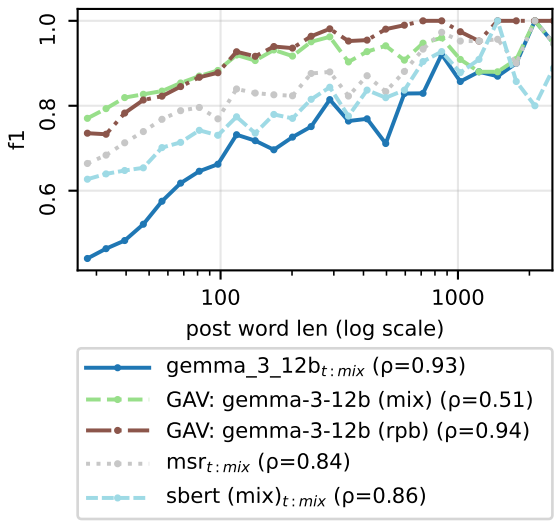

Figure 3: F1-Score per message length in words (up to 2500) on the mixed test set. ρ is the Spearman correlation between message length and F1 scores. We display the results with bucket sizes increasing on a log-scale, to account for less examples existing for longer messages.

Overall Performance We first consider the average (6th) column, finding that models trained on the mixed train set achieve the highest performance. Specifically, Gemma-3 (mix) scores the highest overall F1 of 0.83, 0.03 points above Llama-3.1 (mix) and Qwen-2.5 (mix) ($p<0.01$). The second-best results come from models trained on the profile-based dataset, with Gemma-3 (rpb) scoring 0.80 and Qwen-2.5 (rpb) 0.02 points lower. For Gemma-3 and Llama-3.1, the lowest average performance occurs when trained on the Twitter dataset, while for Llama-3.2 and Qwen-2.3, it occurs with Reddit cross-domain.

Among the baselines, Multilingual Style Representation (msr), with its threshold tuned on the mixed validation set, achieves the highest average F1 of 0.74, outperforming the second-best baseline, sbert, by 0.05. This also exceeds the best zero-shot baseline, Llama-3.2, which attains an average F1 of 0.66, while hlstm performs nearly on par, trailing by 0.02. Compared to our tuned LLMs, msr matches the average F1 of Llama-3.2 (rpb). The weakest models are Qwen-2.5 in zero-shot, with F1 = 0.04, followed by ppm (mix) with F1 = 0.21. Since the mixed test set contains a stratified subset of the other test sets, its results closely mirror the average column, differing by at most 0.02.

These results show that our fine-tuned LLMs clearly outperform existing AV baselines, and their performance predictably correlates with model size, with larger models generally performing better. This suggests that AV benefits from greater model capacity and that even stronger results may be achievable with larger-scale models. However, parameter efficiency remains important, making msr as the best-performing baseline an attractive alternative when computational resources are limited.

Overall, the strong performance of models trained on the mixed dataset suggests that exposure to stylistic variability across domains enables learning robust and generalizable representations of authorial style, challenging the assumption that models must be highly data-specific. This is particularly relevant in real-world applications where test data may differ from any single training source.

Reddit vs. Twitter Here, we consider the evaluation on the Reddit in-domain dataset (3rd column) and Twitter dataset (5th column). Models tuned on the Twitter training set perform better on the Twitter test set (Gemma: +0.11 F1), while models trained on Reddit in-domain perform better on the Reddit in-domain test set (Gemma: +0.06 F1), indicating that models capture dataset-specific characteristics such as message style, topic, or length. The larger gap for Reddit suggests that Reddit data promotes greater generalization than Twitter. The strongest model on Reddit in-domain is Gemma-3 (rid), and the strongest model on Twitter is Gemma-3 (tw), with Gemma-3 (mix) consistently ranking second (non-significant difference on Twitter, $p>=0.01$), again highlighting the benefit of combining training data from multiple sources. The average performance across all models for Reddit in-domain is 0.01 points higher than for Twitter.

In-Domain vs. Cross-Domain Next, we compare the performance on the Reddit in-domain test set (3rd column) with the cross-domain test set (4th column). In most cases, models trained on the respective training set (in-domain and cross-domain) perform stronger on the respective test set than vice versa. However, for Llama-3.2, training on cross-domain leads to a much lower performance than with in-domain for all dataset (0.13 lower F1-score on average). Perhaps, because of its smaller parameter count, the model is not able to handle the more complex case of cross-domain evaluation well. Also, the mixed data models outperform or perform on par with the cross-domain models on the cross-domain dataset. This means that mixing data from many sources has a higher benefit for cross-domain evaluation than just training on

cross-domain data acquired in the same manner as the test set. The average across all models shows that cross-domain evaluation is more difficult than in-domain evaluation, with the average F1-score being 0.06 higher for in-domain. Overall, the slight performance drop suggests that the models do rely to some extent on content cues. However, the modest decrease indicates that they still capture meaningful stylistic signals and remain effective under content-controlled conditions.

Message-Based vs. Profile-Based Here, we compare the performance on the Reddit in-domain test set (3rd column) with the profile-based test set (5th column). As before, models trained on the respective datasets show the strongest performance, with Gemma-3 (rpb) and Qwen-2.5 (rpb) performing best. The mixed models perform almost on par for both datasets. For example, for profile-based evaluation, Gemma-3 (mix) has only 0.02 difference in F1-Score. Overall, profile-based evaluation yields higher metrics than other scenarios (e.g., 0.09 higher accuracy), likely because concatenated messages provide more information on an author's style than single messages. The best-performing model remains a tuned Gemma variant.

Figure 3 shows F1-scores for the two best-performing baselines, msr and sbert (trained on mixed), as well as Gemma-3-12b in zero-shot and fine-tuned (mixed and profile-based) settings across varying word counts (x-axis), alongside corresponding Spearman correlations. To this end, samples from the mixed test set were grouped into word-length buckets on a logarithmic scale (see Appendix I for details). Gemma-3 (mix) performs best up to about 100 words, after which it is surpassed by Gemma-3 (rpb). Around 500 words, Gemma-3 (mix) reaches a performance plateau, and at about 1000 words, the performance of other models is almost on par with it. Gemma-3 (mix) is decreasing from about 500 words. Gemma-3 zero-shot shows the steepest gain, starting with the lowest F1 and gradually catching up as word count increases. The baseline models exhibit similar curves, with msr generally outperforming sbert except at around 1500 words, where sbert peaks before declining. The high Spearman correlations between message length and F1-Score (up to ρ=0.94) indicate that AV is highly sensitive to text length, with short texts often lacking sufficient stylistic information. This highlights the importance of minimum length requirements in practical use.

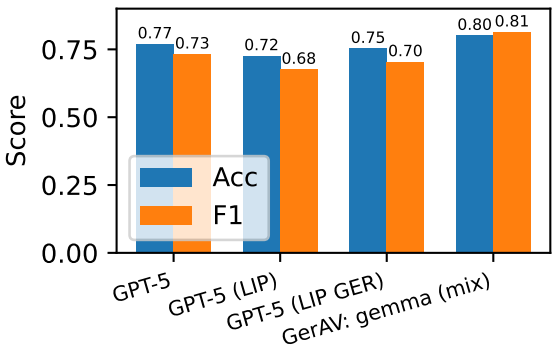

Figure 4: Accuracy and F1-scores of GPT-5 and GerAV: gemma-3-12 (mix). They are evaluated on a stratified 480 sample subset of the mixed test set.

Closed-Source Evaluation Figure 4 shows the accuracy and F1-scores of Gemma-3 (mix) on a stratified subset of the mixed test set. The GerAV model significantly outperforms GPT-5 by 0.031 accuracy and 0.081 F1-score ($p<0.01$). ROC scores are not reported, as GPT-5's log probabilities are inaccessible. This result is especially significant for applications, as fine-tuned open-source LLMs can be deployed locally, providing full control over data and reproducibility without sacrificing performance. Interestingly, GPT-5 with the short zero-shot prompt achieves a better performance than with both LIP prompts, with the German LIP prompt performing slightly better than the English one.

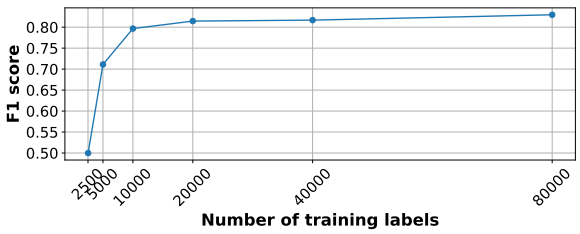

Figure 5: F1 scores of Gemma-3-12b plotted against number of training samples.

Effect of Number of Training Labels While GerAV comprises a large number of authors and texts, in realistic scenarios, training data is often scarce, particularly for low-resource languages. To investigate the impact of limited training data, we analyze the effect of progressively reducing the number of training samples during fine-tuning of our best-performing Gemma model. Specifically, we iteratively halved the size of the mixed training set five times and retrained the model on each subset.

model	Twitter German	Twitter English
gemma-3-12b (tw_en)	0.70	0.73
gemma-3-12b (tw_ger)	0.82	0.68

Table 2: F1-Scores for models trained on English (tw_en) and German (tw_ger) twitter datasets and tested on respective test datasets.

Full results are provided in Appendix I. Figure 5 presents the F1 scores when tested on the mixed test set. Reducing the dataset from 80,000 to 70,000 samples has a negligible impact (0.83 to 0.82, $p < 0.01$). Performance declines gradually to 0.79 with one-eighth of the data, but drops more sharply at smaller scales (0.71 at 5,000; 0.52 at 2,500), indicating that this range approaches a lower limit for effective adaptation. Overall, these findings indicate that training data can be substantially reduced while maintaining competitive performance, with performance gains largely saturating at around one-fourth of the full dataset. This highlights the potential of the proposed LoRA approach for application in low-resource settings.

Language-Specific Investigation To further examine the value of language-specific AV corpora and models, we use an English Twitter dataset from Cheng et al. (2010), following the splits defined by Tyo et al. (2023). For comparability, we applied our preprocessing pipeline, resulting in 22,352 training samples. We then sampled an equally sized subset from GerAV Twitter.

We trained Gemma-3-12b on both datasets and evaluated each model on both test sets (see Table 2). While the German model reaches an F1 of 0.82 when applied to the German test set, the English model only reaches 0.70 F1 when tested on the English test set, suggesting that the English dataset is more challenging, potentially due to shorter messages (mean length: 26.9 in English vs. 34.1 words in German). Cross-lingual evaluation reveals performance drops: the German model reveals a moderate drop, scoring 0.68 on English, while the English model drops more substantially to an F1 of 0.70 on German test data. Overall, these results indicate that less dominant languages benefit more from language-specific fine-tuning, underscoring the importance of dedicated AV datasets. The dominance of English pretraining appears to partially compensate for task adaptation on a different language, though a notable performance gap remains.

We further evaluated the German LLM LLäMmlein-7b-chat (Pfister et al., 2025), to assess whether language-specific pretraining and embeddings provide measurable advantages for AV. We train and test the model on all GerAV subsets but did not observe any performance gains over the English LLMs, achieving e.g. an F1 of 0.78 when trained and tested on the mixed dataset(see Appendix G for full results). This suggests that our fine-tuning approach can sufficiently adapt models to language-specific phenomena relevant to AV even when the pretraining data is dominated by another language. However, this finding should be confirmed by evaluating more models and languages in future work.

6 Conclusion

Overall, this work provides a comprehensive evaluation of fine-tuned LLMs alongside a diverse set of established baselines representing different methodological approaches on our newly introduced GerAV benchmark. In doing so, we address two major gaps in the AV literature: the lack of systematic evaluations including multiple LLMs and multiple baselines and the scarcity of robust benchmarks for non-English languages. Our results provide several insights into the behavior of LLMs for AV in German with implications for both real-world applications and future research. (1) We clearly demonstrate the superiority of fine-tuned LLMs over traditional baselines. (2) We show that the best-performing model, Gemma-3-12b, is robust to domain shifts and handles both very short and very long input texts well. Moreover, strong performance can be maintained even when the amount of training data is substantially reduced. (3) Our results show the importance of training on language-specific corpora, revealing significant performance drops in cross-lingual settings. (4) GerAV demonstrates that training on combined datasets enhances generalization abilities, alleviating the need for highly data-specific models, and provides a solid foundation for future work on improving performance in cross-domain scenarios as well as handling varying text lengths.

Overall, our results provide a solid foundation for future research aimed at improving performance and paving the way toward robust, high-performing AV models suitable for practical applications across languages.

Limitations

Key limitations of this work include the need for a critical assessment of potential biases in the datasets and the absence of interpretable features in the model outputs. Future work should focus on incorporating explanations into the framework to address the crucial requirement of interpretability in AV systems. In addition, the influence of content has been only partially addressed and should be examined in more detail to ensure that model predictions are driven by authorial style rather than topical cues. The GerAV cross-domain dataset should further be extended to more challenging scenarios, such as cross-platform author identification or varying text genres produced by the same author.

In addition, the evaluation of our fine-tuning strategy is limited, and future work should explore a wider range of architectures, prompt designs, hyperparameter settings, and adaptation methods. The experimental setup should be extended to additional non-English languages beyond German with a specific focus on low-resource languages.

Finally, this work does not include human evaluation or direct comparisons between model and human performance, which are necessary steps toward reliably replacing expert judgment with automated AV systems. The absence of human evaluation in AV tasks is a general issue in the literature and should be addressed in future work.

Ethical Considerations

While authorship analysis generally aims to support socially beneficial applications, such as forensic investigations or plagiarism detection, these methods may also be misused: awareness of stylistic markers could enable deliberate style obfuscation, and the analysis of anonymous texts may infringe on privacy or undermine legitimate anonymity. We therefore emphasize that authorship verification results are probabilistic and should be applied cautiously, particularly in sensitive or high-stakes contexts.

We report all licenses and configurations of models used in this work (see Appendix F), and all data used in this work is publicly available for research purposes. Information about data licences and access can be found in Appendix J. We did not extract nor try to predict any personally identifiable information, such as e.g, names, birth dates, addresses, or gender, of any author included in the datasets used for this work. While we assume that our developed AV systems do not systematically predict two texts to originate from the same author for reasons unrelated to stylistic similarity, such as shared social or demographic characteristics, we do not investigate this behavior further. Consequently, we cannot guarantee the fairness of our models.

Acknowledgements

We gratefully acknowledge support from the German Federal Ministry of Research, Technology and Space (BMFTR) through the research grant ALiAS (13N17272 and 13N17273) in the security research program, and from the German Research Foundation (DFG) through the Heisenberg Grant EG 375/5-1.

References

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, and 30 others. 2024. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. ACM.

Janek Bevendorff, Berta Chulvi, Gretel Liz De La Peña Sarracén, Mike Kestemont, Enrique Manjavacas, Ilia Markov, Maximilian Mayerl, Martin Potthast, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, Benno Stein, Matti Wiegmann, Magdalena Wolska, and Eva Zangerle. 2021. Overview of pan 2021: authorship verification, profiling hate speech spreaders on twitter, and style change detection. In *International conference of the cross-language evaluation forum for european languages*, pages 419–431. Springer.

BKA. 2025. Autorenerkennung. Accessed: 2025-12-28.

Benedikt Boenninghoff, Steffen Hessler, Dorothea Kolossa, and Robert M Nickel. 2019. Explainable authorship verification in social media via attention-based similarity learning. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 36–45. IEEE.

Benedikt Boenninghoff, Henry Hosseini, Robert M. Nickel, and Dorothea Kolossa. 2024. Who wrote when? author diarization in social media discussions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15721–15734, Miami, Florida, USA. Association for Computational Linguistics.

Dasha Bogdanova and Angeliki Lazaridou. 2014. Cross-language authorship attribution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2015–2020, Reykjavik, Iceland. European Language Resources Association (ELRA).

Branden Chan, Timo Möller, Malte Pietsch, and Tanay Soni. 2019. German BERT (bert-base-german-cased). *Hugging Face repository*. Accessed on 2025-31-12.

Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, page 759–768, New York, NY, USA. Association for Computing Machinery.

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.

Bradley. Efron. 1979. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1 – 26.

Maël Fabien, Esau Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. 2020. BertAA : BERT fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 127–137, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

Jack Grieve. 2007. Quantitative authorship attribution: An evaluation of techniques. *Literary and linguistic computing*, 22(3):251–270.

Shriya TP Gupta, Jajati Keshari Sahoo, and Rajendra Kumar Roul. 2019. Authorship identification using recurrent neural networks. In *Proceedings of the 2019 3rd International Conference on Information System and Data Mining*, ICISDM '19, page 133–137, New York, NY, USA. Association for Computing Machinery.

Oren Halvani, Christian Winter, and Anika Pflug. 2016. Authorship verification for different languages, genres and topics. *Digital Investigation*, 16:S33–S43. DFRWS 2016 Europe.

Xie He, Arash Habibi Lashkari, Nikhill Vombatkere, and Dilli Prasad Sharma. 2024. Authorship attribution methods, challenges, and future research directions: A comprehensive survey. *Information*, 15(3).

Md. Rajib Hossain, Mohammed Moshiul Hoque, M. Ali Akber Dewan, Nazmul Siddique, Md. Nazmul Islam, and Iqbal H. Sarker. 2021. Authorship classification in a resource constraint language using convolutional neural networks.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yujia Hu, Zhiqiang Hu, Chun-Wei Seah, and Roy Ka-Wei Lee. 2024. Instructav: Instruction fine-tuning large language models for authorship verification. *Preprint*, arXiv:2407.12882.

Baixiang Huang, Canyu Chen, and Kai Shu. 2024. Can large language models identify authorship? pages 445–460.

Baixiang Huang, Canyu Chen, and Kai Shu. 2025. Authorship attribution in the era of llms: Problems, methodologies, and challenges. 26(2):21–43.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, and 1 others. 2024. Gpt-4o system card.

Abraham Israeli, Shuai Liu, Jonathan May, and David Jurgens. 2025. The million authors corpus: A cross-lingual and cross-domain Wikipedia dataset for authorship verification. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25997–26017, Vienna, Austria. Association for Computational Linguistics.

Mike Kestemont, Enrique Manjavacas, Ilia Markov, Janek Bevendorff, Matti Wiegmann, Efstathios Stamatatos, Martin Potthast, and Benno Stein. 2020. Overview of the cross-domain authorship verification task at pan 2020.

Junghwan Kim, Haotian Zhang, and David Jurgens. 2025. Leveraging multilingual training for authorship representation: Enhancing generalization across languages and domains. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34855–34880, Suzhou, China. Association for Computational Linguistics.

Nane Kratzke. 2023. Monthly samples of german tweets (2019 - 2022).

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Zhi Liu. 2006. Reuter_50_50. *UCI Machine Learning Repository*.

Andrei Manolache, Florin Brad, Antonio Barbalau, Radu Tudor Ionescu, and Marius Popescu. 2021. Transferring bert-like transformers' knowledge for authorship verification.

Meta AI. 2024a. Introducing llama 3.1: Our most capable models to date.

Meta AI. 2024b. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models.

Benjamin Murauer and Günther Specht. 2019. Generating cross-domain text classification corpora from social media comments. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 114–125, Cham. Springer International Publishing.

Benjamin Murauer and Günther Specht. 2021. Developing a benchmark for reducing data bias in authorship attribution. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 179–188, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. 2017. Surveying stylometry techniques and applications. *ACM Computing Surveys (CSuR)*, 50(6):1–36.

OpenAI. 2022. Gpt-3.5. Technical report, OpenAI.

OpenAI. 2023. Gpt-4. Technical report, OpenAI.

OpenAI. 2025. Gpt-5. Technical report, OpenAI.

Jan Pfister, Julia Wunderle, and Andreas Hotho. 2025. LLäMmlein: Transparent, compact and competitive German-only language models from scratch. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2227–2246, Vienna, Austria. Association for Computational Linguistics.

Chen Qian, Tianchang He, and Rao Zhang. 2017. Deep learning based authorship identification. *Report, Stanford University*, pages 1–9.

Justin Qiu, Jiacheng Zhu, Ajay Patel, Marianna Apidianaki, and Chris Callison-Burch. 2025. mStyleDistance: Multilingual style embeddings and their evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16917–16931, Vienna, Austria. Association for Computational Linguistics.

Sahana Ramnath, Kartik Pandey, Elizabeth Boschee, and Xiang Ren. 2025. CAVE: Controllable authorship verification explanations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8939–8961, Albuquerque, New Mexico. Association for Computational Linguistics.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Rafael A. Rivera-Soto, Olivia Elizabeth Miano, Juanita Ordonez, Barry Y. Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. 2021. Learning universal authorship representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 913–919, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ipek Baris Schlicht and Angel Felipe Magnossão de Paula. 2021. Unified and multilingual author profiling for detecting haters. *Preprint*, arXiv:2109.09233.

Efstathios Stamatatos. 2006. Ensemble-based author identification using character n-grams. In *Proceedings of the 3rd International Workshop on Text-based Information Retrieval*, volume 36, pages 41–46.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

William J Teahan and David J Harper. 2003. Using compression-based language models for text categorization. In *Language modeling for information retrieval*, pages 141–165. Springer.

Gemma Team. 2025. Gemma 3.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Jacob Tyo, Bhuwan Dhingra, and Zachary C. Lipton. 2023. Valla: Standardizing and benchmarking authorship attribution and verification through empirical evaluation and comparative analysis. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 649–660, Nusa Dua, Bali. Association for Computational Linguistics.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. TRL: Transformers Reinforcement Learning.

Janith Weerasinghe, Rhia Singh, and Rachel Greenstadt. 2021. Feature vector difference based authorship verification for open-world settings. In *CLEF (Working Notes)*, pages 2201–2207.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Perric Cistac, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. pages 38–45. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

William J Youden. 1950. Index for rating diagnostic tests. *Cancer*, 3(1):32–35.

Peter Zeng, Pegah Alipoormolabashi, Jihu Mun, Gourab Dey, Nikita Soni, Niranjan Balasubramanian, Owen Rambow, and H. Schwartz. 2025. Residualized similarity for faithfully explainable authorship verification. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15824–15837, Suzhou, China. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *Preprint*, arXiv:2205.01068.

A GerAV Examples

Table 3 shows examples for each dataset in GerAV and their English translation.

B Domain Clustering Prompt

Figure 6 shows the prompt used to prompt GPT-4o for clustering German subreddits into topical domains.

C Topical Domains

Table 4 displays the names of all topical domains, a description of each domain and all subreddits belonging to this domain as outputted by GPT-4o.

> I will provide you with a list of German subreddits. Your task is to cluster these subreddits into semantically coherent topical groups. Create clusters that are as distinct from each other as possible. You may ignore subreddits that don't fit any cluster.
> For each cluster, provide:
> - `A cluster name`
> - `A short description of the theme`
> - `The list of subreddits belonging to it`
>
> If a subreddit could fit into multiple topics, place it in the one that fits best and explain your choice briefly. Do not create clusters just to use all subreddits; fewer, cleaner clusters are preferred.
> Here is the subreddit list:
> [...]

Figure 6: Prompt used for clustering German Subreddits into topical domains.

D GerAV Post Lengths

Figure 7 shows the distributions in the German Reddit and Twitter datasets before preprocessing.

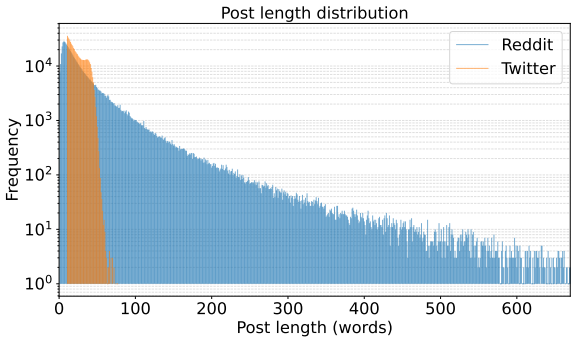

Figure 7: Log frequencies of post word lengths for the Reddit (blue) and Twitter (orange) dataset before preprocessing. The long tail of the Reddit distribution is truncated at 0.05% for visualization purposes; the true maximum post length is 4,321 words.

Dataset	Text A	Text B
Original examples		
Twitter	Das sieht man z.B. an München, hier verkauft die Stadt eine Umweltspur als Paradigmenwechsel, doch in Wirklichkeit müssen hier Radler die Spur mit Bussen und E-SUV teilen - kein Fortschritt	Ist hier irgendwo ein*e Arzt*in aus #Stuttgart? Das ist ein tolles Projekt, das sich zu unterstützen lohnt. Es gibt Malteser Medizin übrigens auch in #Nürnberg
Reddit In-Domain	**Subreddit: wasistdas** | Sieht nach Weihnachtsbeleuchtung aus Da gibt es recht viele verschiedene Lämpchen Die "Watt Zahl " hängt unter anderem davon ab wie viele Lampen an der Lichterkette sind	**Subreddit: wasistdas** | Das müsste zum Testen von Batterien sein Zumindest erinnere ich mich an so einen Widerstands- Draht in dem Zusammenhang Und 2 volt passt für die einzelnen Zellen einer Batterie vom Gabelstapler usw
Reddit Cross-Domain	**Subreddit: Ratschlag** | Es gibt aus dem Material und damit aus genau der Stärke vom Stoff auch Mützen. Buff hat welche, die sind extra fürs Tragen unterm Helm gemacht, also ultra dünn und trotzdem halten sie den Wind fern.	**Subreddit: kannmandasnochessen** | Hab in dem Monat so eine Dose mit 2013 drauf im Keller gefunden. Aufmachen, riechen, kosten. Meine war genauso wie sie auch vor 13 Jahren geschmeckt hätte.
Reddit Profile-Based	[...]Finde ich top, aber da gibts auch viele andere gute Alternativen. Sonst finde ich irgendeine Art v[...]	[...]en, nach meinem (rein subjektiven) Gefühl ist das von Landkreis zu Landkreis anders. Grundsätzlich h[...]
Translated examples		
Twitter	This can be seen in Munich, for example, where the city is promoting a bike lane as a paradigm shift, but in reality cyclists have to share the lane with buses and electric SUVs—no progress at all.	Is there a doctor from #Stuttgart here? This is a great project that is worth supporting. Incidentally, Malteser medical services are also available in #Nuremberg.
Reddit In-Domain	**Subreddit: whatisthat** | Looks like Christmas lights. There are quite a few different lights. The “wattage” depends, among other things, on how many lights are on the string.	**Subreddit: whatisthat** | That must be for testing batteries. At least, I remember seeing a resistance wire like that in that context. And 2 volts is suitable for the individual cells of a forklift battery, etc.
Reddit Cross-Domain	**Subreddit: Advice** | There are also hats made from this material, which means they have the same thickness as the fabric. Buff has some that are specially designed to be worn under a helmet, so they are ultra-thin but still keep the wind out.	**Subreddit: canyoustilleatthat** | I found a can with 2013 on it in the basement this month. Open it, smell it, taste it. Mine tasted exactly as it would have 13 years ago.
Reddit Profile-Based	[...]I think it’s great, but there are also many other good alternatives. Otherwise, I find some kind o[...]	[...]en, in my (purely subjective) opinion, this varies from county to county. Basically, h[...]

Table 3: Examples of our four sub-datasets with same-author pairs and their English translations. The mixed dataset is excluded, as is built from these four. For profile-based, we only show a small part of the concatenated messages. We changed the involved entities.

Cluster	Description	Subreddits	Counts
General German Community & Discussion	Broad discussion spaces about life in Germany, general questions, culture, and mixed topics.	de, AskGermany, FragReddit, FragenUndAntworten, KeineDummenFragen, Ratschlag, wirklichgutefrage, tja, WerWieWas, de_IAmA, Nachrichten, GuteNachrichten, DErwachsen, einfach_posten, wasistdas, serviervorschlag, duschgedanken, tragedeigh	110,453
Humor, Memes & Satire	German memes, ironic content, absurd humor, and satire.	ich_iel, hessich_iel, LTB_iel, deutschememes, Berliner_memes, aberBitteLaminiert, asozialesnetzwerk, OkBrudiMongo, schkreckl, wortwitzkasse, wasletztestern, wasletzterezension, NichtDieTagespresse, satire_de_en, GermansGoneWild, ichbin14unddasisttief	47,302
Politics, Society & Ideology	German politics, political parties, activism, and social debates.	politik, ich_politik, PolitikBRD, DIE_LINKE, Kommunismus, antiarbeit, Klimawandel, Klimagerechtigkeit, umwelt_de, Verkehrswende, OeffentlicherDienst, GoldenerAluhut	44,741
Finance, Careers & Economics	Personal finance, investing, macroeconomics, taxes, and professional life.	Finanzen, Aktien, Immobilieninvestments, Normalverdiener, Wirtschaftsweise, Energiewirtschaft, Steuern, arbeitsleben, Azubis, InformatikKarriere, selbststaendig, mauerstrassenwetten, CapitolVersicherungAG	92,212
Technology, IT & Engineering	Computing, hardware, electronics, IT knowledge, and advanced technologies.	de_EDV, PCGamingDE, PCBaumeister, Elektroinstallation, technologie, KI_Welt, informatik	19,824
Transportation, Mobility & Vehicles	Cars, bikes, trains, aviation, and everyday mobility topics.	Fahrrad, automobil, autobloed, TuningVerbrechen, RoestetMeinAuto, Elektroautos, MotorradDeutschland, deutschebahn, bahn, LuftRaum, Fuehrerschein, Falschparker	46,116
Food, Cooking & Household	Everyday cooking, recipes, food culture, organization, and home life.	Kochen, Backen, keinstresskochen, vegetarischDE, VeganDE, kantine, kannmandasnochessen, SchnitzelVerbrechen, Leberkasverbrechen, Doener, Doenerverbrechen, SpeziVerbrechenDE, Kleiderschrank, Einrichtungstipps, wohnkultur, wohnen, Hausbau, Canbau, Garten, Balkonkraftwerk	84,548
Regional & Local Communities	City-based, state-based, and regional identity communities in German-speaking areas.	frankfurt, Leipzig, dresden, karlsruhe, bavaria, wien, VfBStuttgart, BayernMunich, borussiadortmund, eintracht, MannausSachsen, RentnerfahreninDinge, rentnerzeigenaufdinge	36,169
Science, Knowledge & Education	Academic learning, research, history, science, documentaries, and reading.	Studium, abitur, WissenIstMacht, Wissenschaft, Weltraum, Geschichte, GeschichtsMaimais, Dokumentationen, buecher	34,202
Lifestyle, Relationships & Personal Life	Relationships, parenting, gender discussions, fitness, self-expression.	beziehungen, FragtMaenner, FragNeFrau, Eltern, Weibsvolk, FitnessDE, BeautyDE, Beichtstuhl, BinIchDasArschloch	84,948
Pets, Animals & Nature	Pets, wildlife, nature exploration, and animal-related humor.	Katzengruppe, Gittertiere, naturfreunde, PferdeSindKacke	7,404
Media, Entertainment & Pop Culture	Movies, music, podcasts, creators, streaming, and entertainment content.	Filme, musik, YouTubeDE, Twitch_DE, FestundFlauschig, WolfgangMSchmitt, Augenschmaus, Augenbleiche, zocken, de_punk	23,979
Work, Trades & Practical Skills	Craftsmanship, DIY skills, emergency services, manual professions.	Handwerker, Handarbeiten, selbermachen, feuerwehr	10,887
Law, Administration & Institutions	Legal issues, police, military, emergency services, and administrative institutions.	recht, polizei, pozilei, blaulicht, bundeswehr, dhl_deutsche_post	20,476

Table 4: Table showing clusters of German subreddits, including a brief description, subreddit membership, and post counts.

E AV Prompts

Below, we list all prompts that were used for the AV experiments in this paper:

1. For LLM baselines (including GPT-5):

   Are the following two texts written by the same author?
   Text A: {text_a}
   Text B: {text_b}
   Please answer with "Yes" or "No".
   Answer:

2. For fine-tuned LLMs:

   Are the following two texts written by the same author?
   Text A: {text_a}
   Text B: {text_b}
   Answer:

3. For the English LIP prompt, slightly adjusted to prevent the LLM from refusing the task:

   For a scientific experiment, given two texts determine if they are written by the same author. Analyze the writing styles of the input texts, disregarding the differences in topic and content. Focus on linguistic features such as phrasal verbs, modal verbs, punctuation, rare words, affixes, quantities, humor, sarcasm, typographical errors, and misspellings.
   Text A: {text_a}
   Text B: {text_b}
   Please think step-by-step, then answer with 'Yes' or 'No' as your last word.
   Answer:

4. For the German translation of the LIP prompt:

   Für ein wissenschaftliches Experiment möchten wir feststellen, ob zwei Texte von demselben Autor verfasst wurden. Analysiere die Schreibstile der Eingabetexte und ignoriere dabei Unterschiede im Thema und Inhalt. Konzentriere dich auf linguistische Merkmale wie Phonologie, Morphologie, Wortbildung, Syntax, Wortstellung, Kasussystem, Genus, Tempus, Modus, Passiv, Kongruenz, Valenz, Artikelgebrauch, Pronomen, Negation, Modalverben, Semantik, Pragmatik, Informationsstruktur, Prosodie, Wortarten, Nebensätze, Konjunktiv, Modalpartikeln, seltene Wörter und Idiomatik.
   Text A: {text_a}
   Text B: {text_b}
   Bitte denke Schritt für Schritt nach und antworte am Ende mit 'Ja' oder 'Nein' als letztes Wort.
   Answer:

F Training Setup, Hyperparameters and Model Licences

We train the LoRAs on a Slurm cluster with H100 GPUs. Training times range from under an hour for Llama3.2-3B to 4:40h for Gemma-12B on the mixed dataset. We use the following hyperparameters (evaluated in experiments on the Twitter validation set): batch_size=2, num_train_epochs=1, tuning_seed=10, lang="en", learning_rates=3e-4, weight_decays=0.01, seeds=42, do_lora=true, lora_rs=128, lora_alphas=32, lora_dropouts=0.0. We use the following library versions: **torch 2.9.0** (Ansel et al., 2024), **transformers 4.57.3** (Wolf et al., 2020), **vllm 0.12.0** (Kwon et al., 2023), **trl 0.21.0** (von Werra et al., 2020).

Our use of all pretrained language models is consistent with their intended use as specified in the respective licenses that are listed in Table 5. The models were used solely for research purposes, including evaluation and fine-tuning, and in compliance with all stated access and usage conditions. Any derived artifacts created in this work are explicitly intended for research use only, and their intended use remains compatible with the original license terms and downstream restrictions of the source models.

G Accuracy, F1 and RoC_AuC on GerDATA

Figure 8 show the results of all baseline model variations (with thresholds tuned on different GerAV subsets), and Figure 9 shows the results of all tuned variants. For some baselines, we only tune the threshold on the mixed validation set, because mixed shows the strongest results for most models. The German-pretrained LLäMmlein-7B-Chat model does not outperform the English-pretrained models. Compared to the best-performing model,

Model	License / Terms	Link
Gemma-3-12B-it (Team, 2025)	Gemma Terms of Use	https://ai.google.dev/gemma/terms
LLaMA-3.1-8B-instruct (Meta AI, 2024a)	LLAMA 3.1 Community License Agreement	https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/LICENSE
LLaMA-3.2-3B-instruct (Meta AI, 2024b)	LLAMA 3.2 Community License Agreement	https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct
LLäMmlein-7B-chat (Pfister et al., 2025)	LLäMmlein RESEARCH-ONLY RAIL-M	https://huggingface.co/LSX-UniWue/LLaMmlein_7B_chat/blob/main/license.md
Qwen-2.5-7B-instruct (Yang et al., 2025)	Apache License	https://huggingface.co/Qwen/Qwen2.5-7B-Instruct/blob/main/LICENSE

Table 5: License names and links to all models used for fine-tuning in this work.

Gemma-3, it consistently underperforms across fine-tuned variants and evaluation datasets, with an average deficit of 0.04 points F1 in the mixed setting. When compared to Qwen-2.5-7B, which has a the same parameter count, results are more similar: LLäMmlein outperforms Qwen in two of the five fine-tuning configurations (reddit cross-domain and reddit in-domain) on average, while Qwen achieves better results in the remaining three. Overall, these findings suggest that AV fine-tuning does not significantly benefit from German pre-training. This difference may also reflect the overall lower baseline performance of the LLäMmlein model compared to the other models, rather than an effect of its German pretraining, and should be tested on further models and lanugages.

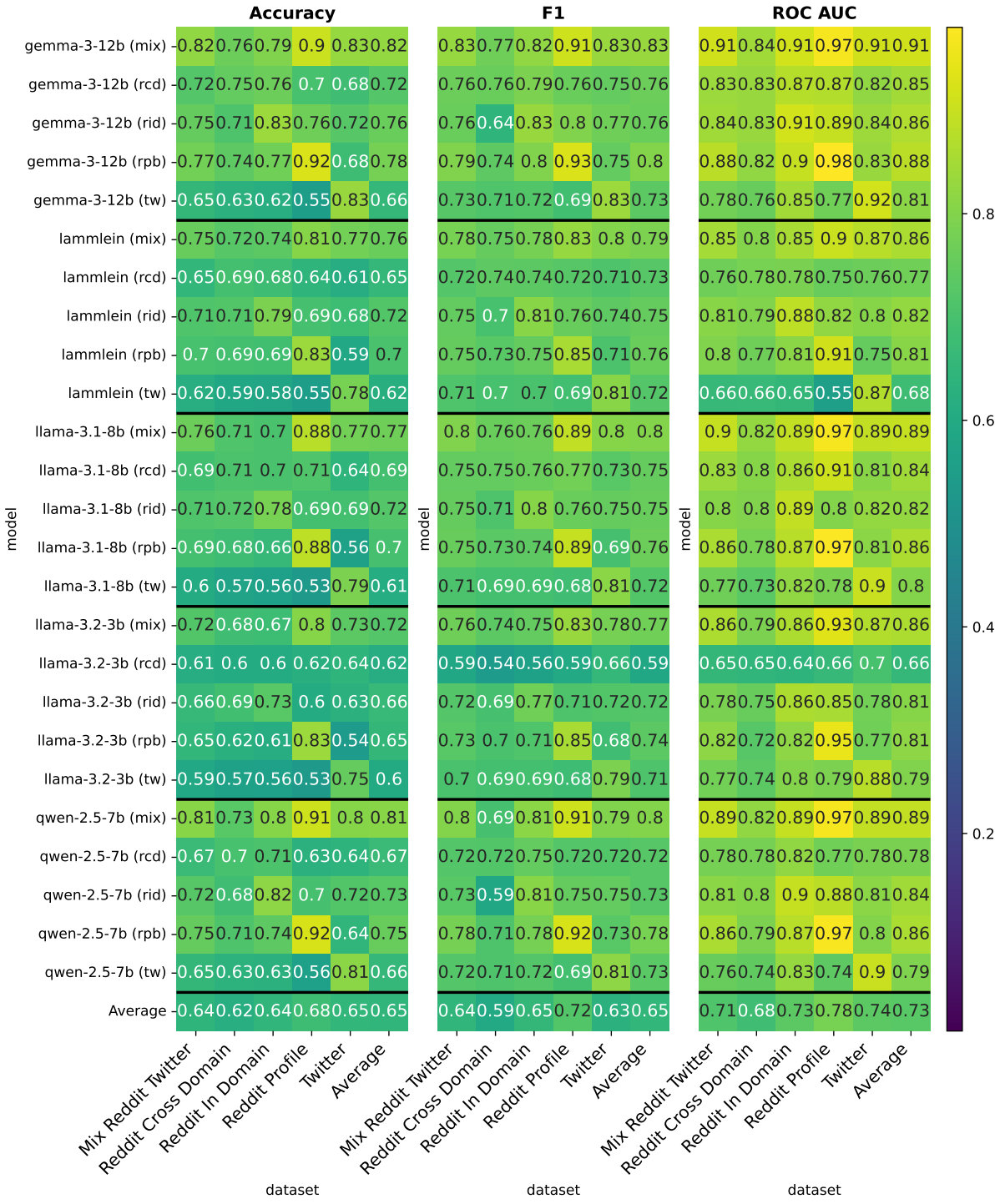

Figure 8: Accuracy, F1-Score and RoC_AuC Score for the baselines. The y-axis shows the model names and the x-axis shows the evaluated test set. The subscript indicates the validation set that was used to tune the decision threshold: **Mix** Reddit Twitter (t:mix), **R**eddit **C**ross **D**omain (t:rcd), **R**eddit **I**n **D**omain (t:rid), **R**eddit **P**rofile **B**ased (t:rpb) and **Tw**itter (t:tw). The highest value in every column is written in bold.

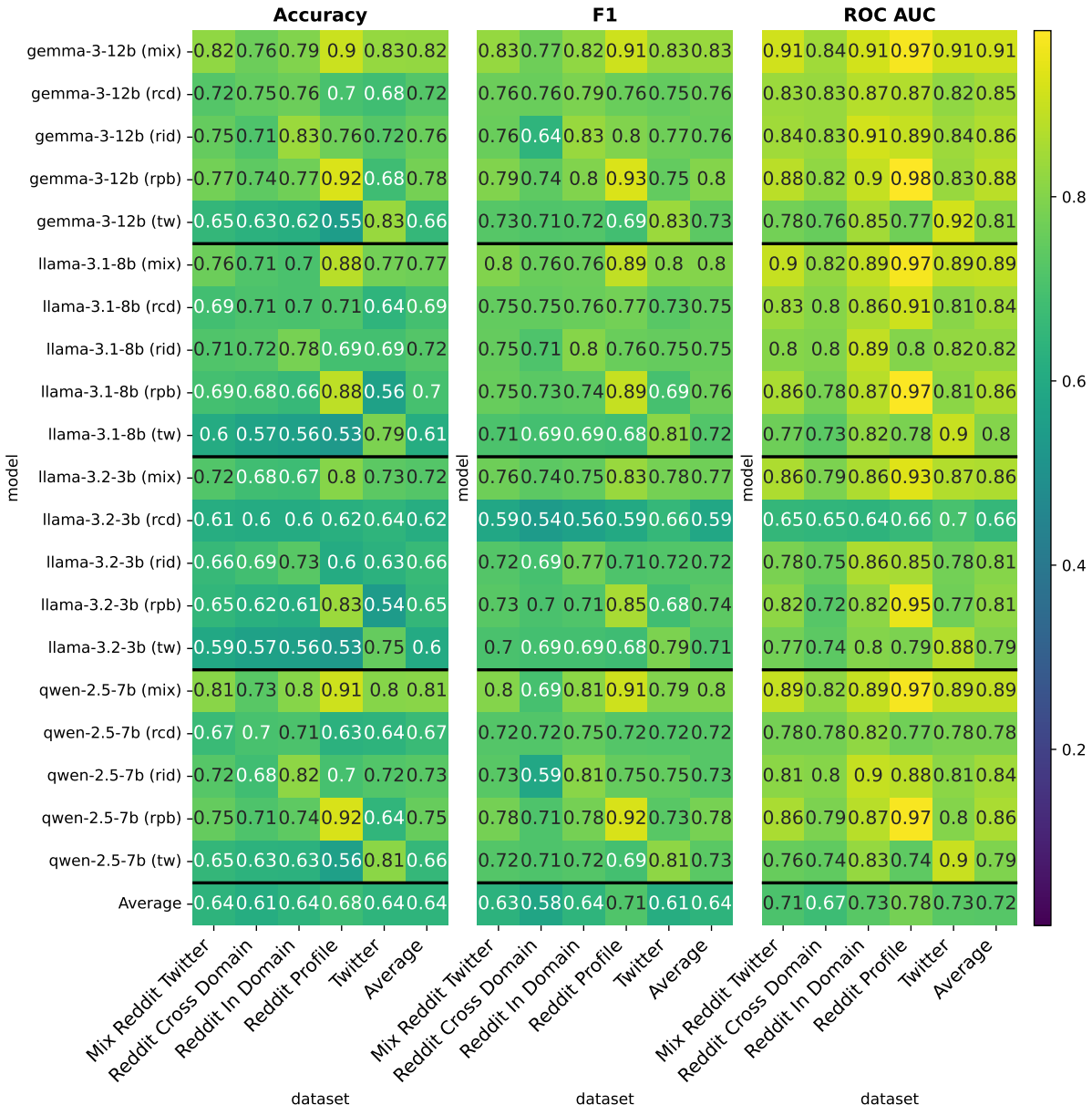

Figure 9: Accuracy, F1-Score and RoC_AuC Score for our GerAV models. The y-axis shows the model names and the x-axis shows the evaluated test set. The brackets behind the model indicate its training set: **Mix** Reddit Twitter (mix), **R**eddit **C**ross **D**omain (rcd), **R**eddit **I**n **D**omain (rid), **R**eddit **P**rofile **B**ased (rpb) and **Tw**itter (tw). The highest value in every column is written in bold.

H Word Length buckets

Table 6 displays the distribution of samples across log-spaced length bins on the mixed test set. Bins with less than 2 samples were excluded, resulting in 27 distinct bins.

I Varying Number of Training Samples

Bin	Range	Samples
0	[25.00, 29.96]	3604
1	[29.96, 35.90]	3340
2	[35.90, 43.01]	2310
3	[43.01, 51.54]	1053
4	[51.54, 61.76]	896
5	[61.76, 74.00]	784
6	[74.00, 88.67]	583
7	[88.67, 106.25]	516
8	[106.25, 127.32]	413
9	[127.32, 152.56]	356
10	[152.56, 182.80]	348
11	[182.80, 219.04]	295
12	[219.04, 262.47]	210
13	[262.47, 314.51]	190
14	[314.51, 376.86]	145
15	[376.86, 451.57]	113
16	[451.57, 541.10]	97
17	[541.10, 648.38]	73
18	[648.38, 776.92]	55
19	[776.92, 930.95]	41
20	[930.95, 1115.52]	25
21	[1115.52, 1336.67]	14
22	[1336.67, 1601.67]	14
23	[1601.67, 1919.21]	12
24	[1919.21, 2299.71]	3
25	[2299.71, 2755.64]	10
26	[2755.64, 3301.96]	3
27	[3301.96, 3956.59]	1
28	[3956.59, 4741.00]	1

Table 6: Distribution of samples across log-spaced length bins on the mixed test set.

Figure 10 shows the full results of F1 scores for all training set sizes of GerAV mixed tested on all 5 test subsets. Among these, the Reddit cross-domain test set is most affected by reductions in training data, showing the largest performance decline. This suggests that larger training corpora are particularly important for this challenging setting, likely because increased data diversity helps the model better capture cross-topic stylistic variation required for AV.

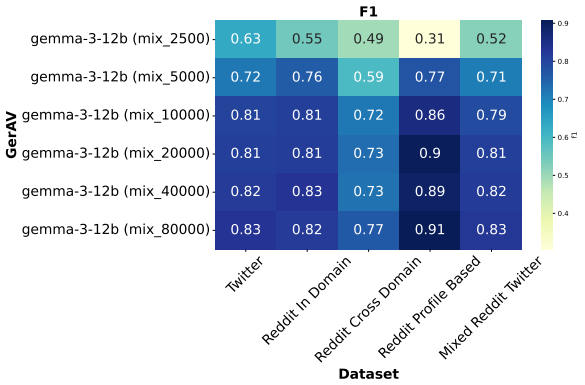

Figure 10: F1-scores for Gemma-3-12b trained on varying numbers of training samples from Mixed Gerav and tested on all GerAV test subsets.

J Data licence and reproducibility

Since GerAV is collected from two distinct sources, different access methods and licences apply:

1. Twitter: The original Twitter data is publicly accessible with restricted access for academic research via Zenodo[5]. The licence of the original data source applies. After gaining access, the GerAV preprocessing and data splits can be reproduced using our code published on GitHub[6].

2. Reddit: The Reddit data was collected via the reddit4researchers API [7] that prohibits direct redistribution of the data. To support the reproducibility of GerAV, we provide access to the post URLs for academic research purposes only. Further information about access and the preprocessing pipeline can be found on our GitHub repository. After accessing the Reddit posts, the original Reddit licence applies.

[5] https://zenodo.org/records/6421331
[6] https://github.com/NL2G/GerAV/
[7] https://support.reddithelp.com/hc/en-us/articles/14945211791892-Developer-Platform-Accessing-Reddit-Data