UzLitAttr: A Benchmark for Closed-Set Authorship Attribution in Uzbek Literary Texts

Anonymous ACL submission

Abstract

Author attribution is a core task in stylometry and computational text analysis, with applications in digital humanities, literary studies, and forensic linguistics. While extensively studied for high-resource languages, it remains largely unexplored for low-resource languages, including Uzbek. We introduce a new benchmark dataset for closed-set author attribution in Uzbek literary prose, comprising 54 authors and 269 works. To prevent information leakage, we adopt a work-level train–test split, ensuring that all text segments from a given literary work appear exclusively in either the training or test set. We benchmark classical stylometric approaches based on function words and character n -grams, as well as five multilingual transformer-based models including state-of-the-art mmBERT. We also conduct an ablation study identifying (6–8) as the optimal character n -gram configuration. Furthermore, we analyze the effect of text length by evaluating all models across chunk sizes ranging from 100 to 800 words. Our results reveal that while mmBERT achieves the strongest overall performance (0.91 Macro-F1 at 800 words), classical character n -gram representations remain highly competitive (0.87 Macro-F1), substantially outperforming standard multilingual encoders such as mBERT and XLM-R across most chunk sizes. These findings highlight both the practical strength of classical stylometry and the promise of massively multilingual pretraining for low-resource author attribution. This benchmark provides a reliable foundation for future research on Uzbek and other low-resource languages.

1 Introduction

Authorship attribution seeks to identify the author of a text using only linguistic evidence. The task has a long tradition in stylometry, where early studies demonstrated that seemingly unobtrusive stylistic cues—such as function word distributions and character-level patterns—can reliably capture authorial signal (Mosteller and Wallace, 1963; Stamatatos, 2009). Building on this foundation, recent works have increasingly adopted neural language models for authorship attribution, often reporting strong performance, particularly in high-resource settings (Ge et al., 2016; Fabien et al., 2020).

Despite methodological advances, research on authorship attribution remains heavily skewed toward a small set of high-resource languages, most notably English. This imbalance mirrors a broader trend in NLP, where language coverage is dominated by a limited number of Indo-European languages (Joshi et al., 2020; Ponti et al., 2019). As a result, low-resource languages are underrepresented not only in modeling approaches but also in the availability of carefully curated, publicly accessible benchmark datasets. Uzbek, a Turkic language characterized by rich morphology and a complex orthographic history, exemplifies this gap. While Uzbek has a substantial literary tradition, it has received little attention in authorship attribution research, and no standardized benchmark currently exists that supports rigorous and reproducible evaluation.

The absence of standardized datasets is particularly problematic for author attribution, where experimental design choices can significantly affect reported performance. Prior work has shown that naïve document-level splitting can lead to substantial data leakage when texts from the same literary work appear in both training and test sets, artificially inflating results (Eder et al., 2016; Kestemont, 2019). Recent evaluation campaigns have therefore emphasized the importance of work-level or cross-domain splits to assess true generalization across unseen texts (Kestemont et al., 2018). However, such evaluation protocols are rarely applied in low-resource settings due to the lack of suitable datasets.

In this work, we address this gap by introduc-

ing a benchmark dataset for closed-set authorship attribution in Uzbek literary prose. The dataset comprises works by 54 authors and spans novels, short stories, and other narrative forms. All texts undergo rigorous preprocessing, including the removal of non-content material, orthographic normalization, and manual verification of author-work associations. Crucially, we adopt a work-level train–test split, ensuring that no literary work appears in both splits and thereby preventing leakage across evaluation phases.

Beyond dataset construction, we provide a comprehensive evaluation framework. We benchmark classical stylometric baselines alongside multilingual transformer-based models and systematically study the effect of text length by varying chunk sizes from 100 to 800 words, following established findings on the sensitivity of authorship attribution to text length and data availability (Luyckx and Daelemans, 2011; Sapkota et al., 2015). To facilitate reproducibility and future research, we publicly release the dataset, standardized splits, and evaluation scripts, aligning with recent recommendations for transparent and reproducible NLP benchmarking (Dodge et al., 2019).

Contributions. Our main contributions are as follows:

- We introduce the first standardized benchmark for closed-set author attribution in Uzbek literary texts, covering 54 authors and 269 works, with publicly documented preprocessing and evaluation protocols.
- We propose a rigorous work-level train-test splitting strategy that prevents textual leakage and ensures realistic attribution evaluation, aligning with best practices in stylometric research.
- We provide a comprehensive comparison of classical stylometric features and multilingual transformer-based models across multiple chunk sizes (100–800 words), enabling systematic analysis of the impact of text length on attribution performance.
- We conduct an ablation study on character n-gram configurations, demonstrating robustness across different feature ranges and identifying (6–8) as the optimal configuration.
- We empirically demonstrate that character n-gram representations remain highly competitive under leakage-free evaluation, substantially outperforming standard multilingual encoders (mBERT, XLM-R) across all chunk sizes, while massively multilingual pretraining (mmBERT) achieves the strongest overall performance — highlighting both the practical strength of classical stylometry and the promise of broad multilingual coverage for low-resource author attribution.

2 Related Work

2.1 Authorship Attribution and Stylometry

Authorship attribution has a long history in stylometry, where classical approaches aim to capture stable aspects of an author's writing style using carefully designed linguistic features. Prominent examples include function word frequencies, character n-grams, and syntactic patterns (Stamatatos, 2009). Among these, character n-gram representations have consistently been shown to be robust across languages and genres, as they implicitly encode morphological, orthographic, and stylistic regularities while remaining relatively insensitive to topic variation (Sapkota et al., 2015).

More recent works have explored neural approaches to authorship attribution, including recurrent neural networks and transformer-based language models (Gupta et al., 2019; Ge et al., 2016; Huang et al., 2024; Fabien et al., 2020). These models often achieve strong performance in high-resource settings, particularly when large amounts of training data are available. However, prior studies have also noted that neural models can be sensitive to data scarcity and evaluation design, raising questions about their robustness under realistic generalization settings (Kestemont et al., 2018).

2.2 Stylometry in Low-Resource Settings

While authorship attribution has been extensively studied for high-resource languages, substantially less attention has been paid to low-resource settings. This imbalance reflects a broader trend in NLP, where research and benchmark datasets are dominated by a small number of languages (Joshi et al., 2020; Ponti et al., 2019). In low-resource scenarios, limited data availability exacerbates challenges related to overfitting and evaluation leakage, making experimental design choices particularly critical.

Recent evaluation campaigns in authorship attribution have therefore emphasized the importance of work-level or cross-domain splits, which prevent texts from the same literary work from appearing in both training and test sets (Kestemont et al., 2018). These protocols aim to assess true stylistic generalization rather than memorization of work-specific patterns. However, such rigorous evaluation setups are rarely adopted in low-resource language studies due to the lack of suitable, carefully curated datasets.

2.3 Uzbek Natural Language Processing

Recent progress in Uzbek natural language processing has been largely driven by the release of task-specific datasets and resources targeting core NLP problems. Existing works have primarily focused on text classification, sentiment analysis, topic categorization, and named entity recognition, often motivated by practical downstream applications and the availability of annotated data (Kuriyozov et al., 2022, 2023; Mengliev et al., 2025). In addition, recent studies have explored multilingual and cross-lingual approaches for Uzbek, particularly in the context of machine translation and representation learning (Allaberdiev et al., 2024; Rabbimov and Kobilov, 2020).

Despite these advances, stylometric analysis and authorship attribution remain largely unexplored for Uzbek. Existing datasets are not designed to support stylistic evaluation, and they typically lack the metadata and experimental controls required for rigorous authorship attribution, such as work-level separation to prevent data leakage. To the best of our knowledge, no publicly available benchmark currently exists for Uzbek author attribution that enables leakage-free evaluation and systematic comparison across modeling approaches. This work addresses this gap by introducing a standardized benchmark and evaluation framework for authorship attribution in Uzbek literary prose.

3 Dataset

3.1 Data Collection and Sources

We collect Uzbek literary prose from publicly available digital archives, focusing on narrative texts such as novels, short stories, and novellas. The dataset includes works by 54 authors representing modern and classical Uzbek literature. To ensure authorial consistency, we manually inspected author directories and individual works, removing misattributed texts, duplicated files, and non-literary material. Only authors with at least two distinct works were retained to support closed-set attribution under a work-level split.

3.2 Data Cleaning and Normalization

Raw texts often contain substantial noise, including cover pages, tables of contents, page headers, footers, and OCR artifacts. We applied a multi-stage cleaning pipeline to remove non-content material and normalize the text. This process included removing page markers, publisher watermarks, and isolated page numbers, as well as correcting hyphenated line breaks introduced during OCR. Given the coexistence of Cyrillic and Latin scripts in Uzbek texts, all content was normalized to a unified Latin-based orthography. Apostrophe variants were standardized, Unicode normalization was applied, and all text was lowercased. These steps were designed to reduce orthographic variation while preserving stylistic signals relevant to authorship.

3.3 Dataset Overview

After preprocessing and author-level filtering, the resulting dataset comprises 54 authors and 269 literary works. Table 1 summarizes the key statistics of the dataset. The majority of authors contribute between 2 and 6 works, with a median of 4.0 and a mean of 4.98 works per author. At 400 words, the dataset yields 35,091 chunks in total, split into 28,073 training and 7,018 test chunks at an approximately 4:1 ratio.

Figure 1 shows the distribution of works per author. The distribution is right-skewed, with most authors contributing 3–5 works and a small number contributing up to 12. All authors have at least two distinct works, satisfying the minimum requirement for work-level train–test splitting. This imbalance in author coverage is reflected in the per-author performance variance observed in the error analysis, and motivates our use of macro-averaged F1 as the primary evaluation metric.

3.4 Chunking Strategy

To enable controlled analysis of text length effects, each literary work was segmented into fixed-length chunks based on word counts. We generate multiple versions of the dataset with chunk sizes ranging from 100 to 800 words. Chunks shorter than the specified length are discarded to ensure consistency

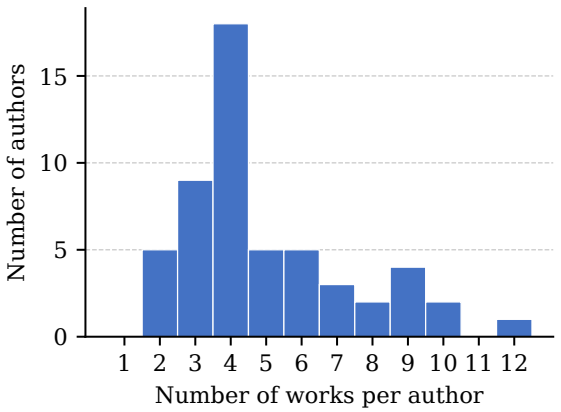

Figure 1: Distribution of works per author in the benchmark dataset. The majority of authors contribute 3–5 works (median: 4.0), while a small number contribute up to 12. All authors have at least two works, satisfying the requirement for work-level train–test splitting.

Statistic	Value
# Authors	54
# Works	269
Min works per author	2
Max works per author	12
Median works per author	4.0
Mean works per author	4.98
Total chunks (400 words)	35,091
Train chunks (400 words)	28,073
Test chunks (400 words)	7,018
Train/test ratio	4.0:1

Table 1: Overview of the Uzbek literary author attribution dataset after cleaning and work-level filtering.

across samples. This setup allows systematic analysis of how the amount of textual context influences attribution performance.

3.5 Train-Test Split

To prevent information leakage, we adopt a work-level train–test split. All chunks derived from a given literary work are assigned exclusively to either the training or test set. Furthermore, we restrict the dataset to authors with at least two distinct works, ensuring that each author is represented in both splits. At 400 words, this yields 28,073 training and 7,018 test chunks at an approximately 4:1 ratio. This design enforces a strict closed-set attribution scenario and reflects realistic conditions under which models must generalize across works rather than memorizing individual texts. The full distribution of chunks per author across both splits is provided in Appendix A.1 (Figure 5).

4 Task Definition

We formulate author attribution as a multi-class classification problem. Given a text chunk x, the goal is to predict its author y from a fixed set of candidate authors. Each chunk is treated as an independent instance, and no metadata beyond the text itself is provided to the model. Evaluation is performed using the macro-averaged F1 score to account for class imbalance across authors.

5 Models and Baselines

5.1 Classical Stylometric Baselines

We evaluate two established stylometric representations that are commonly used in authorship attribution. The first baseline models relative frequencies of selected function words, capturing habitual grammatical preferences that are weakly correlated with topical content. The second baseline employs TF–IDF weighted character n-grams. This representation encodes subword patterns and orthographic regularities that have been shown to be robust indicators of authorial style across languages and writing domains.

To determine the optimal n-gram range, we conduct an ablation study comparing seven configurations: (2,4), (3,5), (4,6), (5,7), (6,8), (7,9), and (8,10). Based on this analysis (see Section 7.2), we select the (6–8) configuration for all main experiments. Both classical representations are paired with linear Support Vector Machine classifiers, providing a strong and computationally efficient baseline for comparison with neural models.

5.2 Neural Models

We evaluate both generic multilingual transformer encoders and a stylometry-oriented neural architecture. For transformer-based modeling, we fine-tune three widely adopted multilingual pretrained models: mBERT (Pires et al., 2019), XLM-R (Conneau et al., 2020), and mmBERT (Marone et al., 2024). These models provide strong semantic representations and are commonly used as baselines for multilingual and low-resource text classification tasks. Notably, mmBERT represents the current state-of-the-art multilingual encoder, trained on over 1,800 languages and showing substantial improvements over previous multilingual models.

To incorporate inductive biases more closely aligned with stylometric analysis, we additionally evaluate a character-level convolutional neural network (CharCNN) (Zhang et al., 2015). Unlike

transformer encoders, CharCNN operates directly on character sequences and is designed to capture subword and orthographic patterns that are known to be informative for authorship attribution.

All neural models are trained independently for each chunk-size configuration under identical data splits and evaluation protocols, enabling a controlled comparison of their sensitivity to input length and their ability to capture authorial style. To ensure statistical robustness, all experiments are repeated across three random seeds (42, 43, 44), and we report mean ± standard deviation for all neural models.

6 Experimental Setup

All experiments follow a unified evaluation pipeline to ensure comparability between classical and neural approaches. For every chunk-size configuration, the same work-level train–test splits are applied to all models, and evaluation is performed using macro-averaged F1 score as the primary metric.

Neural inputs are tokenized using the default pretrained tokenizers associated with each model, with fast tokenizers used where available. To handle texts exceeding the 512-token limit without discarding content, a sliding window strategy is applied: each document is split into overlapping chunks of up to 512 tokens with a stride of 64 tokens, producing multiple partially overlapping subsequences per document. Shorter sequences are padded to the maximum length to maintain consistent batch structure. Models are trained on these chunk-level inputs, and at inference time, softmax probability scores across all chunks belonging to the same document are averaged and the argmax is taken as the final document-level prediction. Each chunk-size dataset is processed and trained independently under identical optimization settings: a learning rate of 2e-5, weight decay of 0.01, batch size of 16, and 3 training epochs, with the best checkpoint selected by macro-F1 on the validation set.

Training is conducted on a single workstation equipped with an Intel® Core™ Ultra 9 275HK CPU, 32 GB RAM, and an NVIDIA GeForce RTX 5080 GPU with 16 GB VRAM. To ensure fairness across experimental conditions, hyperparameters are held constant across models and chunk sizes. Complete training configurations and optimization parameters are provided in Appendix A (Tables 5–7) to support reproducibility.

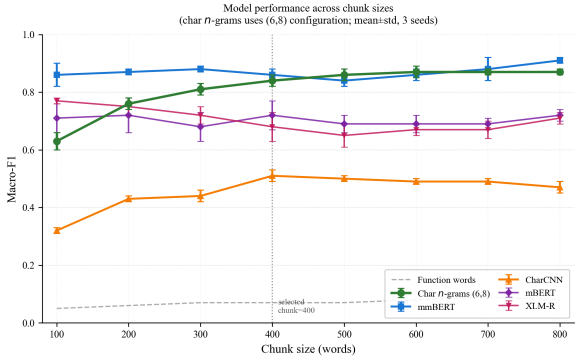

Figure 2: Model performance comparison across chunk sizes. mmBERT achieves the strongest overall performance, while character n-grams (6–8) remain highly competitive and substantially outperform standard multilingual encoders. Error bars indicate standard deviation across 3 random seeds for neural models.

7 Results

7.1 Overall Performance

Figure 2 summarizes Macro-F1 performance across chunk sizes ranging from 100 to 800 words for all evaluated models. Several consistent patterns emerge.

Among all evaluated approaches, mmBERT achieves the strongest overall performance, increasing steadily from 0.86±0.04 at 100 words to 0.91±0.01 at 800 words. This result demonstrates that massively multilingual pretraining across 1,833 languages translates into robust stylometric generalization, even under a strict work-level evaluation protocol.

TF–IDF character n-grams (6–8) represent the strongest classical baseline, increasing from 0.63±0.03 at 100 words to 0.87±0.01 at 600–800 words and exhibiting only minor variation thereafter. This trend aligns with established stylometry findings that character-level patterns capture orthographic and morphological regularities that remain informative even under topic variation and leakage-free evaluation. Notably, character n-grams remain competitive with mmBERT across most chunk sizes, trailing by only 0.04 Macro-F1 points at 800 words (0.87 vs. 0.91), and substantially outperform all other evaluated models.

In contrast, the function-word baseline remains uniformly weak (0.05–0.09) and shows only marginal gains with longer chunks, indicating that function-word distributions alone are insufficient

to separate authors in this dataset under a strict work-level split.

Among the remaining multilingual transformer baselines, mBERT and XLM-R achieve moderate performance. mBERT maintains relatively stable results across chunk sizes (0.71±0.05 at 100 words; 0.72±0.02 at 800 words), suggesting that its representations are reasonably consistent across input lengths but lack the stylistic sensitivity needed for strong attribution. XLM-R performs comparably at shorter chunks (0.77±0.01 at 100 words) but exhibits a gradual decline at longer chunks, reaching 0.71±0.02 at 800 words. Both models fall substantially below character n-grams and mmBERT across all evaluated settings, indicating that intermediate multilingual coverage without massively broad pretraining is insufficient to bridge the gap with classical stylometric representations.

To assess whether a neural architecture with a more explicit stylistic inductive bias narrows this gap, we additionally evaluate CharCNN. CharCNN consistently underperforms both the transformer baselines and character n-grams across all chunk sizes, peaking at 0.51±0.02 at 400 words and declining to 0.47±0.02 at 800 words. This result suggests that a character-level convolutional architecture alone, without large-scale pretraining, is insufficient to capture the stylistic signals required for reliable author attribution in this setting.

Table 2 reports the corresponding Macro-F1 values for each model and chunk-size configuration. Overall, the results reveal two complementary findings: classical character n-gram features remain remarkably competitive given their simplicity, while the broad multilingual coverage of mmBERT yields the strongest attribution performance, demonstrating that both carefully engineered representations and massively multilingual pretraining offer viable paths toward robust author attribution in low-resource settings.

7.2 Character N-gram Configuration Ablation

To identify the optimal character n-gram range and assess the sensitivity of classical stylometric features to hyperparameter choices, we evaluate seven configurations: (2,4), (3,5), (4,6), (5,7), (6,8), (7,9), and (8,10). Table 3 presents results at 400 words, where all configurations perform competitively against neural baselines, with the best configuration approaching mmBERT performance (0.842±0.022 vs. 0.86±0.02).

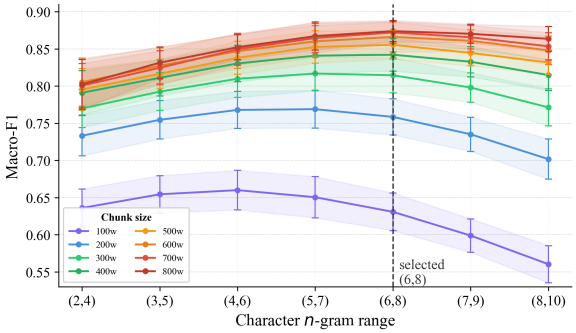

Figure 3: Character n-gram configuration ablation across chunk sizes. All configurations remain competitive with or superior to most neural baselines, with (6–8) achieving optimal performance. mmBERT is shown as a reference line for the strongest neural baseline.

The (6,8) configuration achieves the highest performance (0.842±0.022), representing the optimal balance between pattern specificity and generalization. Shorter n-grams (2,4) capture frequent but less discriminative patterns (0.791±0.030), while very long n-grams (8,10) become overly specific, reducing performance (0.815±0.021). The (5,7) and (6,8) configurations perform comparably, with only marginal differences between them.

Notably, all evaluated n-gram configurations substantially outperform standard multilingual encoders such as mBERT (0.72±0.05) and XLM-R (0.68±0.05) at 400 words, as well as CharCNN (0.51±0.02), demonstrating the robustness of character-level features across hyperparameter choices. While mmBERT remains the strongest individual model at this chunk size, the margin over the best n-gram configuration is narrow (0.86 vs. 0.84), and character n-grams achieve this without any pretraining or GPU-intensive fine-tuning. Figure 3 illustrates performance across all chunk sizes for different configurations, showing that all n-gram variants maintain consistent competitiveness, with performance gaps between configurations remaining small and stable.

7.3 Error Analysis

To better understand model behavior under a strict work-level split, we conduct a qualitative error analysis focusing on the character n-gram (6–8) model at 400 words, as it represents the strongest classical baseline and enables interpretable analysis of stylistic failure modes. Table 4 presents representative cases illustrating common sources of error. In particular, authors with very limited test-set support exhibit highly unstable performance,

Table 2: Macro-F1 scores for closed-set authorship attribution across different chunk sizes. All models report mean ± standard deviation across 3 random seeds. **Bold** indicates best result per chunk size.

Model	Chunk Size (words)							
	100	200	300	400	500	600	700	800
Function Words	0.05	0.06	0.07	0.07	0.07	0.08	0.09	0.08
Char n-grams (6–8)	0.63 ±.03	0.76 ±.02	0.81 ±.02	0.84 ±.02	**0.86 ±.02**	**0.87 ±.02**	0.87 ±.01	0.87 ±.01
CharCNN	0.32 ±.01	0.43 ±.01	0.44 ±.02	0.51 ±.02	0.50 ±.01	0.49 ±.01	0.49 ±.01	0.47 ±.02
mBERT	0.71 ±.05	0.72 ±.06	0.68 ±.05	0.72 ±.05	0.69 ±.03	0.69 ±.03	0.69 ±.02	0.72 ±.02
XLM-R	0.77 ±.01	0.75 ±.03	0.72 ±.03	0.68 ±.05	0.65 ±.04	0.67 ±.02	0.67 ±.03	0.71 ±.02
mmBERT	**0.86 ±.04**	**0.87 ±.01**	**0.88 ±.01**	**0.86 ±.02**	0.84 ±.02	0.86 ±.02	**0.88 ±.04**	**0.91 ±.01**

Table 3: Macro-F1 of character n-gram configurations at chunk size 400 (mean ± std, 3 seeds). **Bold** marks the selected configuration.

Config.	Macro-F1	Description
$(2, 4)$	0.791 ± 0.030	Short n-grams
$(3, 5)$	0.811 ± 0.024	
$(4, 6)$	0.830 ± 0.025	
$(5, 7)$	0.841 ± 0.024	
$\mathbf{(6, 8)}$	$\mathbf{0.842 \pm 0.022}$	**Selected**
$(7, 9)$	0.833 ± 0.020	
$(8, 10)$	0.815 ± 0.021	Long n-grams

Table 4: Per-author performance for the five best and five worst performing authors using character n-grams (6,8) at 400-word chunks. "Test inst." indicates the number of test chunks available for that author.

Author	Test inst.	P	R	F1
Top 5 Authors				
Omon Muxtor	277	0.982	0.982	0.982
Togay Murod	116	0.959	1.000	0.979
Normurod Norqobilov	216	0.968	0.972	0.970
Murod Muhammad Dost	109	0.990	0.945	0.967
Mirmuhsin	240	0.926	0.992	0.958
Bottom 5 Authors				
Xudoyberdi Toxtaboyev	22	0.457	0.727	0.561
Sadulla Karomatov	52	0.354	0.538	0.427
Nusrat Rahmat	6	0.208	0.833	0.333
Abdulla Qodiriy	8	0.214	0.375	0.273
Ibrohim Rahim	80	0.189	0.125	0.150

where a small number of misclassifications disproportionately affects precision and recall. Additionally, stylistic variation across an author's works further challenges generalization, as models are required to transfer stylistic cues across distinct narratives rather than relying on memorization.

These observations reflect a known trade-off in leakage-free author attribution, where methodological rigor increases performance variance for low-resource authors. Figure 4 visualizes the confusion patterns among the most frequently confused authors, revealing that errors tend to occur between authors with similar writing styles or overlapping time periods. We expect these patterns to persist across model families, as they reflect structural properties of the dataset rather than model-specific limitations.

Overall, the error analysis highlights the importance of work-level evaluation for realistic author attribution benchmarks, particularly in low-resource literary settings.

8 Discussion

Our results reveal two complementary findings. First, massively multilingual pretraining yields the strongest overall performance, with mmBERT achieving 0.91 Macro-F1 at 800 words and maintaining stable results across all chunk sizes. Second, classical character n-gram representations remain remarkably competitive despite their simplicity, reaching 0.87 Macro-F1 at 800 words and substantially outperforming standard multilingual encoders across all settings.

The strong performance of mmBERT demonstrates that broad multilingual pretraining translates meaningfully to stylometric tasks. Unlike mBERT and XLM-R, which show inconsistent or declining performance as chunk size increases, mmBERT maintains stable and improving results, suggesting that sufficiently broad multilingual coverage enables the model to capture stylistic regularities that more narrowly pretrained encoders fail to generalize. Nevertheless, the narrow margin over character n-grams (0.91 vs. 0.87) warrants consideration: character n-grams require no pretraining, no GPU-intensive fine-tuning, and introduce no sensitivity to tokenization or sequence length constraints, yet capture stable authorial signals that neural encoders

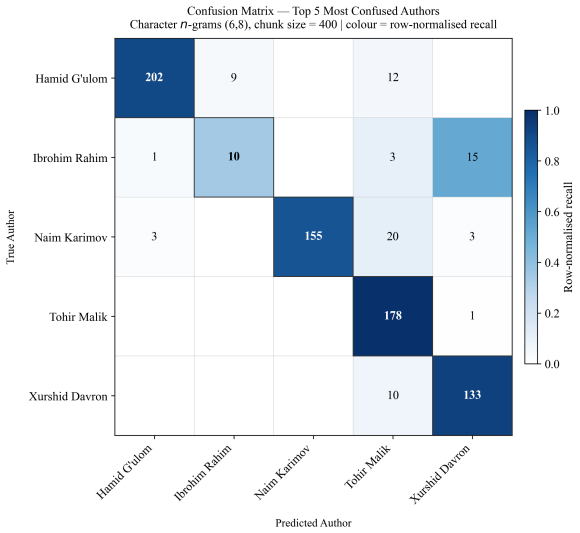

Figure 4: Confusion matrix for the top-5 most confused author pairs using character n-grams (6–8) at 400 words. Errors primarily occur between stylistically similar authors.

only marginally surpass (Stamatatos, 2009; Sapkota et al., 2015).

The underperformance of mBERT and XLM-R likely reflects their optimization for semantic understanding rather than stylistic signals; as chunk length increases, these models may increasingly attend to topical content rather than the subtle markers crucial for authorship attribution (Fabien et al., 2020). CharCNN further illustrates that character-level architectural bias alone is insufficient without large-scale pretraining. Our ablation study confirms that classical robustness holds across all seven n-gram configurations, each outperforming mBERT, XLM-R, and CharCNN, while our error analysis shows that low per-author performance stems from structural dataset properties rather than model-specific failures.

Taken together, these findings present a meaningful trade-off: character n-grams offer strong, computationally efficient attribution without pretraining infrastructure, while mmBERT provides a modest but consistent advantage at significantly greater computational cost. Hybrid approaches combining classical features with multilingual pretraining remain a promising direction for future work.

9 Limitations

Several limitations should be acknowledged. First, the dataset is restricted to literary prose, and findings may not generalize to other genres such as journalism or social media. Second, author-level imbalance in available text can lead to high performance variance for authors with few test works. Third, the benchmark covers only closed-set attribution; open-set attribution and authorship verification remain directions for future research. Fourth, neural baselines were evaluated under fixed hyperparameters without task-specific optimization; strategies such as domain-adaptive pretraining or hybrid classical-neural approaches may yield further improvements, though the computational cost of mmBERT-scale models raises practical accessibility concerns for low-resource settings. Fifth, conclusions may not directly transfer to typologically different languages without further validation. Finally, rapid advances in multilingual modeling mean that newer models may yield different results, though the benchmark provides a stable evaluation framework for future comparisons.

10 Conclusion

We introduce the first benchmark for closed-set author attribution in Uzbek literary texts, combining rigorous preprocessing, work-level evaluation, and a comprehensive comparison of classical and neural approaches. Our results reveal a meaningful trade-off: mmBERT achieves the strongest overall performance (0.91 Macro-F1 at 800 words), while classical character n-grams remain highly competitive (0.87 Macro-F1) without requiring pretraining or GPU-intensive fine-tuning, substantially outperforming standard multilingual encoders across all settings. Our ablation study confirms that this classical advantage is robust across all seven evaluated n-gram configurations. These findings suggest that future work may benefit from hybrid approaches combining classical stylometric features with multilingual pretraining. The benchmark provides a reliable foundation for authorship attribution research on Uzbek and other under-studied languages.

References

Bobur Allaberdiev, Gayrat Matlatipov, Elmurod Kuriyozov, and Zafar Rakhmonov. 2024. Parallel texts dataset for uzbek-kazakh machine translation. *Data in Brief*, 53:110194.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of EMNLP-IJCNLP*, pages 2185–2194.

Maciej Eder, Jan Rybicki, and Mike Kestemont. 2016. *Stylometry with R: A Package for Computational Text Analysis*. R Foundation for Statistical Computing.

Maël Fabien, Esaú Villatoro-Tello, and Petr Motlíček. 2020. Bertaa: Bert fine-tuning for authorship attribution. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 4377–4387.

Zhenhao Ge, Yufang Sun, and Mark Smith. 2016. Authorship attribution using a neural network language model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

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*, pages 133–137.

Baixiang Huang, Canyu Chen, and Kai Shu. 2024. Can large language models identify authorship? *arXiv preprint arXiv:2403.08213*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Mike Kestemont. 2019. Cross-domain authorship attribution: What is easy and what is hard? In *Proceedings of the Workshop on Computational Linguistics for Authorship Analysis*, pages 1–10.

Mike Kestemont, Michael Tschuggnall, Efstathios Stamatatos, Walter Daelemans, Günther Specht, Benno Stein, and Martin Potthast. 2018. Overview of the author identification task at pan-2018: cross-domain authorship attribution and style change detection. In *Working Notes Papers of the CLEF 2018 Evaluation Labs*, pages 1–25.

Elmurod Kuriyozov, Sanatbek Matlatipov, Miguel A. Alonso, and Carlos Gómez-Rodríguez. 2022. Construction and evaluation of sentiment datasets for low-resource languages: The case of uzbek. In *Human Language Technology. Challenges for Computer Science and Linguistics*, pages 232–243. Springer International Publishing.

Elmurod Kuriyozov, Ulugbek Salaev, Sanatbek Matlatipov, and Gayrat Matlatipov. 2023. Text classification dataset and analysis for uzbek language. *arXiv preprint arXiv:2302.14494*.

Kim Luyckx and Walter Daelemans. 2011. The effect of author set size and data size in authorship attribution. *Literary and linguistic Computing*, 26(1):35–55.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2024. mmbert: Scaling multilingual encoders to 1800+ languages. *arXiv preprint arXiv:2509.06888*.

Davlatyor Mengliev, Vladimir Barakhnin, Mukhriddin Eshkulov, Bahodir Ibragimov, and Shohrux Madirimov. 2025. A comprehensive dataset and neural network approach for named entity recognition in the uzbek language. *Data in Brief*, 58:111249.

Frederick Mosteller and David L. Wallace. 1963. Inference in an authorship problem. *Journal of the American Statistical Association*, 58(302):275–309.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

I M Rabbimov and S S Kobilov. 2020. Multi-class text classification of uzbek news articles using machine learning. *Journal of Physics: Conference Series*, 1546:012097.

Upendra Sapkota, Thamar Solorio, Manuel Montes-y Gómez, and Paolo Rosso. 2015. An empirical evaluation of character-based model for authorship attribution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 130–139.

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

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

A Reproducibility Details

A.1 Data Distribution

Figure 5 shows the distribution of chunks per author in training and test sets, illustrating the class imbalance inherent in the dataset.

A.2 Classical Models

A.3 Neural Models

A.4 CharCNN Model

Component	Setting
Classifier	Linear Support Vector Machine (LinearSVC)
Train/test split	Work-level split; test ratio 0.2; seed 42
Function word list	Curated Uzbek function word set (30 tokens)
Function word normalization	Relative frequency per document
Character n-gram range	6–8
Feature weighting	TF–IDF
Vectorizer analyzer	Character-level
Minimum document frequency	5
Maximum feature dimension	50,000
Class weighting (function words)	Balanced
Lowercasing	Enabled
Orthographic normalization	Unified Uzbek Latin script
Evaluation metric	Macro-F1
Hardware	Intel Core Ultra 9 275HK; 32GB RAM; NVIDIA RTX 5080 (16GB)

Table 5: Configuration of classical stylometric baselines used in all experiments. The same work-level data splits and evaluation protocol were applied as for neural models.

Parameter	Value
Learning rate	2×10^{-5}
Batch size	16
Eval batch size	16
Epochs	3
Weight decay	0.01
Max sequence length	512
Optimizer	AdamW
FP16 training	Yes
Random seeds	42, 43, 44

Table 6: Hyperparameters for neural model training.

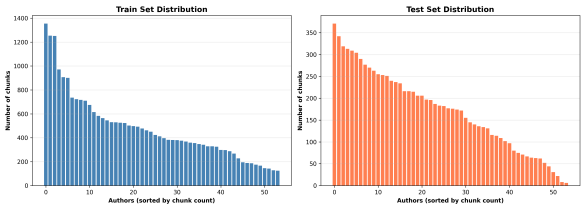

Figure 5: Distribution of chunks per author in training and test sets at 400 words. The imbalance justifies the use of macro-averaged F1 as the primary evaluation metric.

Component	Setting
Model	Character-level Convolutional Neural Network (CharCNN)
Input representation	Raw character sequences
Character vocabulary	Built from training data; includes <PAD> and <UNK> tokens
Maximum input length	3000 characters (right-side truncation)
Padding	Fixed-length character padding
Embedding dimension	128
Convolution kernels	1D convolutions with kernel sizes 3, 4, and 5
Channels per kernel	128
Pooling	Global max-pooling over time
Classifier	Fully connected linear layer
Task	Multi-class author classification (cross-entropy loss)
Train/test split	Work-level split; test ratio 0.2; seed 42
Train batch size	32
Epochs	10
Optimizer	Adam
Learning rate	1×10^{-3}
Regularization	Implicit via early stopping and max pooling
Metrics	Accuracy, Macro-F1
Hardware	Intel Core Ultra 9 275HK; 32GB RAM; NVIDIA GeForce RTX 5080 (16GB)

Table 7: Configuration of the stylometry-oriented CharCNN neural baseline. The model operates directly on character sequences and is trained independently for each chunk-size configuration using the same evaluation protocol as other models.