

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

Oybek Eraliev Maripjon Ugli*

School of Computer and
Information Engineering
Inha University in Tashkent
Tashkent, Uzbekistan
o.eraliyev@inha.uz

Azizbek Muhsinov

School of Computer and
Information Engineering
Inha University in Tashkent
Tashkent, Uzbekistan
azizbekmuhsinov9@gmail.com

Jumabek Alikhanov

Dept. of Computer Engineering
Gachon University
Seongnam, Republic of Korea
Bee Intelligence Global Co., Ltd.
Incheon, Republic of Korea
jumabek@humblebee.ai

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 stylistic approaches based on function words and character n -grams against four neural baselines, including three multilingual transformer encoders (mBERT, XLM-R, and mmBERT) and a character-level convolutional network. We also conduct an ablation study identifying (6–8) as the optimal character n -gram configuration, and analyze the effect of text length by evaluating all models across chunk sizes ranging from 100 to 800 words. In the course of this analysis we identify and correct an evaluation-unit mismatch under which classical models were scored per text chunk while neural models were scored after aggregating predictions across each work (per document): under this matched evaluation, classical character n -gram representations outperform every neural baseline, including mmBERT, at every chunk size and under both evaluation units, an advantage that a paired bootstrap significance test confirms is statistically robust ($p < .05$) across most of the chunk-size range, though not statistically distinguishable from zero at the largest chunk sizes (700–800 words) where the two models converge. This reverses the paper’s original, mixed-unit headline comparison (mmBERT 0.91 vs. n -grams 0.87 Macro-F1 at 800 words) and demonstrates that classical stylistic features remain a strong, computation-

ally inexpensive foundation for low-resource author attribution. Beyond the Uzbek benchmark itself, we surface this evaluation-unit confound as a concrete, quantified cautionary case for classical-versus-neural comparisons in authorship attribution more broadly. We publicly release the dataset splits, labels, derived stylistic features, and evaluation scripts as a 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). Recent work has increasingly adopted neural language models for the task, often reporting strong performance 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, mirroring 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). Uzbek, a Turkic language with rich morphology and a complex orthographic history, exemplifies this gap: despite 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.

This absence is particularly problematic because experimental design choices can significantly affect reported performance: naïve document-level splitting can leak information when texts from the same work appear in both training and test sets, artificially inflating results (Eder et al., 2016; Keste-

* Corresponding author.

mont, 2019). Work-level or cross-domain splits address this (Kestemont et al., 2018), but are rarely applied in low-resource settings due to the lack of suitable datasets.

In this work, we address this gap by introducing a benchmark dataset for closed-set authorship attribution in Uzbek literary prose, comprising 54 authors and 269 works spanning novels, short stories, and other narrative forms. We deliberately scope the task to *closed-set* attribution over a fixed, known list of 54 authors and to *literary prose* specifically: we make no claim that findings transfer to open-set attribution (where the true author may be outside the candidate list), to authorship verification, or to other genres such as journalism or social media, and we return to this scoping decision in §10. All texts undergo rigorous preprocessing, including removal of non-content material, orthographic normalization, and manual verification of author-work associations, and we adopt a strict work-level train–test split to prevent leakage across evaluation phases.

Beyond dataset construction, 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 (Luyckx and Daelemans, 2011; Sapkota et al., 2015). We publicly release the dataset, standardized splits, and evaluation scripts, aligning with 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 (54 authors, 269 works), with a rigorous, leakage-free work-level train–test split and publicly documented preprocessing and evaluation protocols.
- We provide a comprehensive comparison of classical stylometric features and multilingual transformer-based models across chunk sizes (100–800 words), including an ablation over character n -gram configurations identifying (6–8) as optimal.
- We identify and correct an evaluation-unit mismatch between classical models (scored per text chunk) and neural models (scored per document, after aggregating predictions across a

work’s chunks). Under matched evaluation, character n -grams outperform every evaluated neural baseline, including mmBERT, at every chunk size and unit — reversing the original mixed-unit comparison and offering a concrete illustration of how evaluation-protocol confounds can invert reported conclusions in classical-versus-neural comparisons.

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 function word frequencies, character n -grams, and syntactic patterns (Stamatatos, 2009). Character n -gram representations in particular have consistently been shown to be robust across languages and genres, implicitly encoding morphological, orthographic, and stylistic regularities while remaining relatively insensitive to topic variation (Sapkota et al., 2015). More recent work has explored neural approaches, including recurrent networks and transformer-based language models (Gupta et al., 2019; Ge et al., 2016; Huang et al., 2024; Fabien et al., 2020), which often achieve strong performance in high-resource settings but can be sensitive to data scarcity and evaluation design (Kestemont et al., 2018).

2.2 Stylometry in Low-Resource Settings

Substantially less attention has been paid to low-resource settings, reflecting 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); limited data availability there exacerbates overfitting and evaluation-leakage risks, making experimental design choices particularly critical. Recent evaluation campaigns have accordingly emphasized work-level or cross-domain splits that prevent texts from the same work appearing in both training and test sets (Kestemont et al., 2018), though such protocols are rarely adopted in low-resource studies due to the lack of suitable, curated datasets.

Our finding that classical character n -gram features remain competitive with, or outperform, neural encoders is consistent with a growing body of low-resource stylometry results rather than a novel discovery in itself: concurrent work on Romanian authorship attribution similarly finds

that lightweight character n -gram representations match or exceed more complex learned representations on the ROST benchmark (Lupsă et al., 2025). We view the primary novelty of the present work as twofold: introducing the first leakage-aware benchmark for Uzbek author attribution, and, in the course of building it, surfacing and correcting an evaluation-unit confound (§7) that shows how easily such classical-versus-neural comparisons can be inverted by an evaluation-protocol artifact rather than a genuine capability gap.

2.3 Uzbek Natural Language Processing

Recent progress in Uzbek NLP has been driven largely by task-specific datasets for text classification, sentiment analysis, topic categorization, and named entity recognition (Kuriyozov et al., 2022, 2023; Mengliev et al., 2025), alongside multilingual and cross-lingual work on machine translation and representation learning (Allaberdiev et al., 2024; Rabbimov and Kobilov, 2020). Stylometric analysis and authorship attribution remain largely unexplored for Uzbek: existing datasets lack the metadata and experimental controls (e.g., work-level separation) required for rigorous, leakage-free attribution research. This work addresses that gap.

3 Dataset

3.1 Data Collection and Sources

We collect Uzbek literary prose from the Ziyouz.com digital library (<https://n.ziyouz.com/kutubxona>), a publicly accessible Uzbek-language cultural archive, collected in 2026 and 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. The source archive permits personal, non-commercial use of its materials with attribution; we therefore release derived stylometric feature representations, splits, and labels directly, together with a work-identifier list and a lightweight script enabling each user to fetch their own copy of the source text under the same personal-use terms, rather than redistributing raw text ourselves (see Ethics Statement for details).

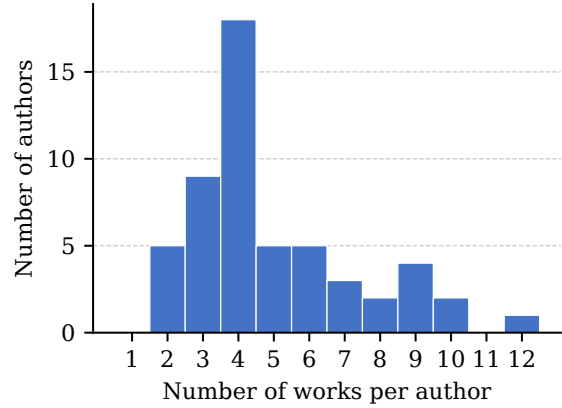


Figure 1: Distribution of works per author in the benchmark dataset. The majority of authors contribute 3–5 works (median: 4.0), while a small number contribute up to 12. All authors have at least two works, satisfying the requirement for work-level train–test splitting.

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 such non-content material, correct hyphenated line breaks introduced by OCR, and normalize orthography: given the co-existence of Cyrillic and Latin scripts in Uzbek texts, all content was normalized to a unified Latin-based orthography, with apostrophe variants standardized, Unicode normalization applied, and text lowercased, to reduce orthographic variation while preserving stylistic signals.

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: authors contribute a median of 4.0 (mean 4.98) works each, and at 400 words the dataset yields 35,091 chunks total (28,073 train / 7,018 test, $\approx 4:1$). Figure 1 shows the right-skewed distribution of works per author (most contribute 3–5, up to 12); all authors have at least two works, satisfying the minimum requirement for work-level splitting. This imbalance motivates our use of macro-averaged F1 as the primary 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, generating dataset versions with chunk sizes from 100 to 800

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.

words; chunks shorter than the target length are discarded. Table 2 reports the total chunk count at each size, ranging from 140,774 (100 words) down to 17,477 (800 words), reflecting the fewer chunks extractable from a fixed amount of text as chunk length grows.

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, and only authors with at least two distinct works are retained, ensuring each author is represented in both splits (28,073 training / 7,018 test chunks at 400 words). This design enforces a strict closed-set attribution scenario reflecting realistic conditions under which models must generalize across works rather than memorize individual texts. The full distribution of chunks per author across both splits is provided in Appendix A.1 (Figure 3).

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. The first models relative frequencies of a manually curated list of 27 Uzbek function words (conjunctions, particles, and other closed-class

items; full list in Appendix B.2), capturing habitual grammatical preferences weakly correlated with topical content. The second employs TF–IDF weighted character n -grams, encoding subword and orthographic regularities shown to be robust indicators of authorial style across languages and domains. To determine the optimal n -gram range, we compare seven configurations: (2,4), (3,5), (4,6), (5,7), (6,8), (7,9), and (8,10), and select (6–8) for all main experiments based on this ablation (§7.3). Both representations are paired with linear SVM classifiers, providing a strong and computationally efficient baseline for comparison with neural models.

5.2 Neural Models

We fine-tune three widely adopted multilingual pre-trained transformer encoders: mBERT (Pires et al., 2019), XLM-R (Conneau et al., 2020), and mmBERT (Marone et al., 2025), a recent massively multilingual encoder trained on over 1,800 languages. We additionally evaluate a character-level convolutional neural network (CharCNN) (Zhang et al., 2015), which operates directly on character sequences to capture subword and orthographic patterns relevant to authorship attribution, as a stylometry-oriented neural architecture.

All neural models are trained independently for each chunk-size configuration under identical data splits and evaluation protocols. To ensure statistical robustness, all experiments are repeated across three random seeds (42, 43, 44), and we report mean $\pm$ standard deviation.

6 Experimental Setup

All experiments follow a unified evaluation pipeline: for every chunk-size configuration, the same work-level train–test split is applied to every model *at a given seed*, and evaluation uses macro-averaged F1.

What the three seeds represent. Each of the three reported seeds (42, 43, 44) corresponds to an independently drawn work-level train–test split, not merely a different model initialization: re-signing which works fall into train versus test is itself seeded. We verified directly from the released per-instance predictions that, within a given seed, the test-work set is identical across classical and neural models (so classical-versus-neural comparisons at a fixed seed remain matched), but the three seeds’ test-work sets are largely disjoint

Table 2: Total chunk count by chunk size (before train–test splitting). Chunk counts fall as chunk size grows because longer chunks consume more of each work’s text per chunk; the 400-word train/test split (28,073 / 7,018) is reported in §3.3.

Words	100	200	300	400	500	600	700	800
Chunks	140,774	70,320	46,828	35,091	28,042	23,342	20,000	17,477

from one another (as low as 5/72 works in common between two seeds at 400 words). Consequently, the mean $\pm$ std figures reported throughout this paper (e.g., Table 4) reflect variance across independently sampled splits, not only across neural training stochasticity: for classical models, whose LinearSVC classifier is otherwise deterministic given a fixed split, essentially all of the reported $\pm$ std is split-composition variance rather than optimization variance. We view this as a stronger, not weaker, robustness result than seed-only variance would be, since it shows the character n -gram advantage (§7.2) holds up under different train/test compositions, not just different random initializations of the same split.

Neural inputs are tokenized with each model’s default fast tokenizer. Because a substantial fraction of Uzbek text chunks exceed the 512-token limit of the pretrained encoders, we apply a sliding-window strategy at the sub-token level: each text chunk is split into overlapping token windows of up to 512 tokens (stride 64), producing multiple partially overlapping token sequences per chunk. Predictions are then aggregated in two stages at inference time: softmax scores across all token windows of the same chunk are averaged and argmaxed to produce a *chunk-level* prediction, and scores across all chunks of the same work are further averaged and argmaxed to produce a *document-level* prediction. Classical models, by contrast, are scored directly at the chunk level with no analogous aggregation step. §7 reports both units for every model and discusses why the original single-unit comparison (chunk-level for classical, document-level for neural) is not directly comparable. Each chunk-size dataset is trained independently under identical settings: learning rate 2e-5, weight decay 0.01, batch size 16, 3 epochs, with the best checkpoint selected by validation macro-F1.

Training is conducted on a single workstation (Intel® Core™ Ultra 9 275HK CPU, 32 GB RAM, NVIDIA RTX 5080 GPU, 16 GB VRAM), with hyperparameters held constant across models and chunk sizes. Complete configurations are in Ap-

pendix A (Tables 6–8).

7 Results

7.1 Overall Performance

As described in §6, our original evaluation design scores classical models (Function Words, Char n -grams, CharCNN) per chunk and transformer models (mBERT, XLM-R, mmBERT) per document (chunk-aggregated) — i.e., two different evaluation units. Table 9 and Figure 4 (Appendix B.1) report this original, mixed-unit comparison in full; §7.2 below identifies why this comparison is not directly interpretable and reports a corrected, matched-unit alternative.

Under this original view, mmBERT achieves the strongest performance, increasing from 0.86 ± 0.04 at 100 words to 0.91 ± 0.01 at 800 words. Character n -grams (6–8) are the strongest classical baseline, increasing from 0.63 ± 0.03 to 0.87 ± 0.01 at 600–800 words and appearing to trail mmBERT by only 0.04 points at 800 words. The function-word baseline remains uniformly weak (0.05–0.09). mBERT and XLM-R achieve moderate, comparatively stable performance (0.68–0.77) but fall well below character n -grams and mmBERT throughout, and CharCNN underperforms all transformer baselines, peaking at 0.51 ± 0.02 at 400 words. §7.2 shows that the apparent 0.04-point mmBERT advantage does not survive matched-unit evaluation.

7.2 Matched-Unit Evaluation

The comparison above evaluates classical and neural models under different units: classical models are scored once per text chunk, whereas neural models are scored after aggregating predictions across every chunk of a work into a document-level prediction (§6). Document-level aggregation can only help, never hurt, Macro-F1 relative to scoring individual chunks, while classical models above are held to the harder, per-chunk standard, so it is not possible to tell whether mmBERT’s apparent 0.04-point edge reflects a genuine representational advantage or a scoring-protocol advantage.

To resolve this, we compute Macro-F1 for every model, at every chunk size, under *both* units, reusing the saved model checkpoints and vectorizers with no retraining: classical predictions are additionally aggregated to the document level by majority vote over each work’s chunks, and neural chunk-level predictions are recovered from the intermediate per-chunk aggregation step (over 512-token windows, §6) that the pipeline already computes before the final document-level step.

Table 4 reports the resulting matched comparison at the two headline chunk sizes (full results in Appendix B.3; Figure 2 visualizes all chunk sizes). Under matched evaluation, character n -grams outperform every neural baseline, including mmBERT, at every chunk size and under both units: at the chunk level, by 0.10–0.33 points, widening with chunk size (0.30 at 400 words: 0.84 vs. 0.54; 0.31 at 800 words: 0.87 vs. 0.56); at the document level, by a smaller but still consistent 0.05–0.14 points, narrowing to 0.05 at 800 words (0.958 vs. 0.910). This reverses the conclusion suggested by Table 9: the original 0.04-point mmBERT advantage was an artifact of comparing a per-chunk classical score against a per-document, multi-chunk-aggregated neural score, not a genuine representational advantage for mmBERT.

Two caveats. First, because neural chunk-level scores are themselves an aggregate over overlapping 512-token windows while classical models score the full chunk as a single feature vector, the neural chunk-level numbers likely already benefit from an internal averaging effect; if anything, this makes the character n -gram advantage at the chunk level an underestimate. Second, XLM-R’s fine-tuning did not converge (near-random accuracy) on roughly one third of seed/chunk-size configurations; we exclude these non-converged runs from all reported XLM-R statistics (see Limitations).

Significance testing. To check whether the document-level margin above is distinguishable from sampling noise rather than assumed from its size alone, we run a paired bootstrap test (10,000 resamples of the shared test documents, two-sided) comparing character n -grams against mmBERT at each chunk size, separately for each seed (§6 explains why splits differ by seed, so we do not pool across seeds). Chunk-level significance testing is not included: per-instance chunk-level predictions were retained for classical models but not for neural models (only the aggregate metric was

Table 3: Paired bootstrap significance test (char n -grams vs. mmBERT, document-level, 10,000 resamples per seed) summarized across chunk sizes. “Gap range” spans the observed (char n -grams – mmBERT) Macro-F1 gap across the available seeds; “Sig.” counts seeds with two-sided $p < 0.05$. mmBERT/800/seed 44 did not complete evaluation (§10), so 800 words has only 2 seeds. Full per-seed results in Appendix B.4.

Chunk size	Seeds	Gap range	Sig. ($p < .05$)
100	3	+0.094 to +0.145	3/3
200	3	+0.081 to +0.119	3/3
300	3	+0.065 to +0.116	2/3
400	3	+0.068 to +0.159	2/3
500	3	+0.109 to +0.157	3/3
600	3	+0.103 to +0.130	3/3
700	3	+0.069 to +0.101	2/3
800	2	+0.014 to +0.067	0/2

saved, §6), and re-generating them would require re-running inference from the saved checkpoints, which we did not have time for (§10). Table 3 summarizes the result (full per-seed table in Appendix B.4): the gap favors character n -grams in all 23 available (seed, chunk-size) combinations with a completed mmBERT run, and is significant at $p < .05$ in 18/23. The 5 non-significant cases cluster at the larger chunk sizes (300, 400, and 700–800 words), consistent with §7.2’s observation that the document-level gap narrows as chunk size grows; at 800 words specifically, the gap is not statistically distinguishable from zero in either of the 2 seeds with a completed mmBERT run (one run, seed 44, never finished evaluation; §10). We therefore state our claim precisely: character n -grams are never worse than mmBERT at the document level under any split or chunk size we tested, and the advantage is statistically robust across most of the range, but is not statistically established at the largest chunk sizes, where the two models converge.

We view this correction as strengthening, not weakening, confidence in the paper’s central claim: the character n -gram advantage is not a narrow, potentially artifactual margin, but positive at every chunk size and unit we tested, and statistically significant across most of that range.

7.3 Character N -gram Configuration Ablation

To identify the optimal character n -gram range, we evaluate seven configurations: (2,4), (3,5), (4,6), (5,7), (6,8), (7,9), and (8,10). Table 5 presents results at 400 words. The (6,8) configuration achieves

Table 4: Macro-F1 under **matched evaluation units** at the two headline chunk sizes. *Chunk*: one prediction per text chunk (sub-chunk for neural models, see §7). *Document*: predictions aggregated across all of a work’s chunks. **Bold** marks the best model per unit/chunk-size. Full 8-chunk-size results in Appendix B.3.

Model	400 words		800 words	
	Chunk	Document	Chunk	Document
Function Words	0.070	0.088 ± 0.017	0.080	0.078 ± 0.004
Char n -grams (6–8)	0.840 ± 0.020	0.967 ± 0.029	0.870 ± 0.010	0.958 ± 0.021
CharCNN	0.510 ± 0.020	0.731 ± 0.047	0.470 ± 0.020	0.695 ± 0.050
mBERT	0.394 ± 0.021	0.720 ± 0.050	0.387 ± 0.013	0.720 ± 0.020
XLM-R	0.372	0.680 ± 0.050	0.384 ± 0.014	0.710 ± 0.020
mmBERT	0.544 ± 0.040	0.860 ± 0.020	0.564 ± 0.037	0.910 ± 0.010

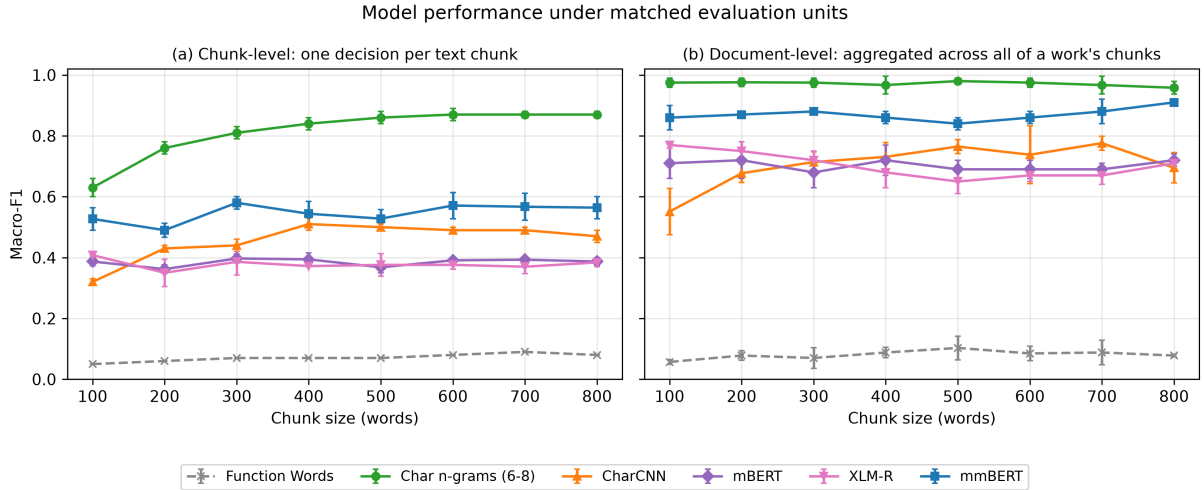


Figure 2: Macro-F1 under matched evaluation units across chunk sizes: (a) chunk-level and (b) document-level, for all six models. Character n -grams outperform every neural baseline at every chunk size under both units. Error bars indicate standard deviation across (up to) 3 random seeds; XLM-R excludes non-converged seeds (§7.2).

the highest performance (0.842 ± 0.022), 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 (0.815 ± 0.021). All seven configurations substantially outperform mBERT (0.72 ± 0.05), XLM-R (0.68 ± 0.05), and CharCNN (0.51 ± 0.02) at 400 words under the original mixed-unit comparison (§7); Figure 5 (Appendix B.1) shows this holds consistently across all chunk sizes, with small, stable gaps between configurations.

7.4 Error Analysis

We conduct a qualitative error analysis on the character n -gram (6–8) model at 400 words. Table 13 (Appendix B.5) reports the five best- and worst-performing authors: authors with very limited test-set support exhibit highly unstable performance, and stylistic variation across an author’s own works further challenges generalization.

Table 5: 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	
(6, 8)	0.842 ± 0.022	Selected
(7, 9)	0.833 ± 0.020	
(8, 10)	0.815 ± 0.021	Long n -grams

To make this instability explicit rather than implicit, Table 13 also reports a bootstrap 95% confidence interval for each author’s F1 (10,000 resamples over the full seed-42 test set at 400 words, recomputing precision/recall/F1 for that author on each resample). The contrast is stark: the top-5 authors’ F1 intervals are narrow (width 0.02–0.05), while the bottom-5 authors’ intervals are wide (width 0.16–0.52) and, for two authors with

only 6–8 test chunks (Nusrat Rahmat, Abdulla Qodiriy), the interval’s lower bound approaches or reaches 0.00. This confirms quantitatively that the low point-estimate F1 scores for these authors are not precise measurements of weak stylistic signal so much as artifacts of having too few test chunks to estimate F1 reliably at all; we return to this in §10. Figure 6 (same appendix section) shows that confusion errors concentrate between stylistically similar authors, a pattern we expect to reflect structural dataset properties rather than model-specific limitations.

To test that expectation, Table 14 (same appendix section) reports the analogous best/worst five authors for mmBERT. Four of its five best-performing authors overlap with the character n -gram model’s top 5, and its worst case is the same author, Ibrohim Rahim (F1=0.150 for character n -grams at the chunk level, 80 test chunks; F1=0.000 for mmBERT at the document level, both test documents misclassified). Because this failure recurs across two architecturally unrelated models and is not explained by low support alone (80 test chunks is mid-sized), we attribute it to unusually high stylistic variation between this author’s own works, rather than a systematic model failure. We also note that document-level evaluation, mmBERT’s native reporting unit, provides far fewer test instances per author than chunk-level evaluation (1–3 works vs. dozens of chunks for most authors), a further, independent reason to prefer chunk-level analysis for interpretability.

8 Discussion

Our results reveal two complementary findings. First, among neural models, massively multilingual pretraining yields the strongest performance: mmBERT reaches 0.91 Macro-F1 at 800 words (document level) and maintains stable results across all chunk sizes, clearly ahead of mBERT and XLM-R, likely because broad multilingual coverage helps it capture stylistic regularities that more narrowly pretrained encoders fail to generalize. Second, and more centrally, classical character n -gram representations outperform every neural baseline we evaluate, including mmBERT, once evaluation units are matched (§7.2): at every chunk size, under both chunk-level and document-level scoring, character n -grams achieve the highest Macro-F1 of any model, requiring no pretraining, no GPU-intensive fine-tuning, and no sensitivity to the 512-token win-

dowing described in §6. This is not a new finding in the abstract — classical character-level features have long been known to be robust authorship signals (Stamatatos, 2009; Sapkota et al., 2015), and a concurrent study reports the same pattern for Romanian (Lupsă et al., 2025) — but it had not previously been shown for Uzbek, and our original, mixed-unit evaluation had in fact obscured it for this dataset until we identified and corrected the underlying evaluation-protocol confound.

The underperformance of mBERT and XLM-R relative to mmBERT likely reflects their optimization for semantic understanding rather than stylistic signals (Fabien et al., 2020); CharCNN further illustrates that character-level architectural bias alone, without large-scale pretraining, is insufficient to match TF-IDF character n -grams. Our ablation study confirms that classical robustness holds across all seven n -gram configurations, each outperforming mBERT, XLM-R, and CharCNN at the chunk level, while our error analysis (§7) shows that the same low-support authors are hardest for both the strongest classical model and mmBERT, suggesting that low per-author performance stems from structural dataset properties rather than model-specific failures.

Taken together, these findings do not describe a trade-off so much as a one-sided practical case: character n -grams are never worse than mmBERT at the document level under any split or chunk size we tested, the advantage is statistically significant across most of that range (§7.2), and character n -grams are the cheapest model to train. We see two qualifications. First, this is an empirical finding about this benchmark and evaluation setup, not a general claim that classical features dominate neural encoders for authorship attribution; different genres, languages, or better-tuned neural training recipes could yield different results (§10). Second, hybrid approaches combining classical features with multilingual pretraining remain a promising direction for future work, particularly if they can recover mmBERT’s cross-lingual transfer benefits without sacrificing the efficiency of character n -gram features.

9 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. In the course of this evaluation we identified and corrected a chunk-level-versus-document-level evaluation-unit mismatch between classical and neural models (§7.2), which reversed the paper’s original headline comparison: under matched evaluation, classical character n -grams outperform every evaluated neural baseline, including mmBERT, at every chunk size and under both evaluation units, without requiring pre-training or GPU-intensive fine-tuning; a paired bootstrap test confirms this advantage is statistically significant across most, though not all, of the chunk-size range (§7.2). Among neural models, mmBERT remains the strongest, and our ablation study confirms that the classical advantage is robust across all seven evaluated n -gram configurations. Beyond the Uzbek benchmark itself, we hope this evaluation-unit correction serves as a concrete, quantified caution for classical-versus-neural comparisons elsewhere in authorship attribution and stylometry. These findings suggest that future work may benefit from hybrid approaches combining classical stylometric features with multilingual pretraining, and from the further controls and extensions identified in Limitations. The benchmark provides a reliable foundation for authorship attribution research on Uzbek and other under-studied languages.

10 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: as Table 13 quantifies directly (§7.4), per-author F1 for authors with only 6–8 test chunks carries a bootstrap 95% confidence interval as wide as 0.52 F1 points, compared to 0.02–0.05 for well-supported authors, so single-point F1 estimates for the lowest-support authors should be read as unreliable rather than as precise measurements. Third, the benchmark covers only closed-set attribution over a fixed, known author list; open-set attribution and authorship verification, where the true author may not be among the candidates, remain directions for future research and would likely require a substantially different evaluation protocol. 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. Sixth, this study is restricted to a single language, Uzbek; while we expect the general methodology (work-level splitting, chunk-size analysis, classical versus multilingual-encoder comparison, matched-unit evaluation) to transfer to other Turkic and low-resource languages, we make no empirical claim about cross-lingual transfer of the specific findings reported here. Seventh, the benchmark is comparatively small in absolute terms (54 authors, 269 works) relative to high-resource authorship attribution corpora; this reflects the limited availability of curated, author-verified digital literary text in Uzbek rather than a design choice, and we view expanding coverage as important future work rather than a solved problem. Eighth, our function-word list (27 items, Appendix B.2) was manually curated rather than derived from corpus frequency statistics, which may explain its weak performance (Macro-F1 0.05–0.10) relative to typical function-word stylometry results; we also do not know what proportion of mmBERT’s pretraining corpus is Uzbek text, which limits causal interpretation of why it outperforms mBERT and XLM-R specifically, and we have not conducted a systematic audit of digitization/OCR-source consistency across authors within our single source archive. Ninth, XLM-R fine-tuning did not converge (near-random Macro-F1) on roughly one third of seed/chunk-size configurations; we exclude these non-converged runs from all reported XLM-R statistics (§7.2) rather than silently averaging them in, which should be kept in mind when interpreting XLM-R’s reported numbers as a lower bound on its achievable performance under more stable optimization. Relatedly, the mmBERT run at 800 words, seed 44, did not complete evaluation (only a training checkpoint was saved); the paper’s mmBERT/800-word statistics and the significance test in §7.2 are therefore based on 2 of the 3 seeds at that one chunk size, not 3, and we report this explicitly rather than silently treating it as a 3-seed result. Tenth, several controls and extensions were identified as valuable during revision but not completed for this submission due to time constraints, and are prioritized future work:

(a) a zero-/few-shot LLM baseline evaluated across all 54 authors, which we did not have time to run and report with confidence and would rather flag explicitly than submit in a rushed, underpowered form; (b) a topic-classification control experiment (fitting the same character n -gram features to predict an unsupervised topic-cluster label instead of author identity) to further rule out topical or source-archive leakage as an explanation for character n -gram performance; and (c) a frequency-derived, expanded function-word list (60–100 items) to test whether the function-word baseline’s weak performance is a list-size artifact rather than a property of the dataset. 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.

Ethics Statement

Data and privacy. All texts in this benchmark are excerpts from previously published literary works by named, publicly recognized authors, obtained from the Ziyouz.com digital library (see §3.1). The dataset does not contain personal, private, or sensitive information about private individuals: author identity is already public information attached to each work (e.g., on the book cover and in library catalogs), not inferred or newly disclosed by this work.

Dual-use risk. Authorship attribution methods carry a general dual-use risk: the same techniques that identify an author of a literary work could, in principle, be adapted to attempt to de-anonymize an anonymous or pseudonymous writer (e.g., a whistleblower, journalist, or dissident) elsewhere. We note two mitigating factors specific to this benchmark: (1) it is a *closed-set* task over a fixed, known list of 54 published authors, which does not directly transfer to open-set real-world de-anonymization where the true author may not be among any known candidate set; and (2) the texts studied are literary prose by authors who are already publicly identified with their work, not anonymous or pseudonymous text. We nonetheless acknowledge that stylometric techniques validated on this benchmark could be repurposed for de-anonymization in other contexts, and we release this resource with the expectation that it be used for research on stylometry, digital humanities, and low-resource NLP, consistent with the ACL Code of Ethics.

Licensing and consent. Texts were collected from the Ziyouz.com digital library (<https://n.ziyouz.com/kutubxona>), a publicly accessible Uzbek-language cultural archive. The site’s terms of use permit materials to be used for personal, non-commercial reading with attribution to the source, while prohibiting commercial use, reproduction, and distribution. To respect these terms while supporting reproducibility, we do not redistribute raw text ourselves. Instead, we publicly release the work-level train/test splits, author and work labels, derived stylometric feature representations (character n -gram and function-word matrices), and evaluation scripts, together with a work-identifier list and a lightweight script that lets each user fetch their own copy of the source text directly from the archive, consistent with the site’s permitted personal-use terms. We plan to additionally request explicit permission from Ziyouz.com for research redistribution of the corpus; should this be granted, a future revision of this resource will include direct raw-text release.

Annotator effort. Author-work associations and text cleaning were manually verified by the authors of this paper as part of dataset construction (§3.1); no external crowdworkers or annotators were employed, so no additional compensation or IRB-style annotator protections apply.

Use of AI writing assistance. Portions of this manuscript’s prose were drafted and/or polished with the assistance of a large language model. The experimental design, implementation, data collection, and analysis are the authors’ own work; all AI-assisted text was reviewed and verified by the authors before submission, consistent with the ACL Publication Ethics Policy on the use of AI writing assistance.

References

- Bobur Allaberdiyev, 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, Petr Motlíček, 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, Patna, India.
- 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*.
- Dana Lupsă, Sanda-Maria Avram, and Radu Lupsă. 2025. Oldies but goldies: The potential of character n-grams for romanian texts. *arXiv preprint arXiv:2506.15650*.
- 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. 2025. mmbert: A modern multilingual encoder with annealed language learning. *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, Steven Bethard, Manuel Montes, and Tamar Solorio. 2015. Not all character n-grams are created equal: A study in authorship attribution. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 93–102, Denver, Colorado.
- 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 3 shows the distribution of chunks per author in training and test sets, illustrating the class imbalance inherent in the dataset.

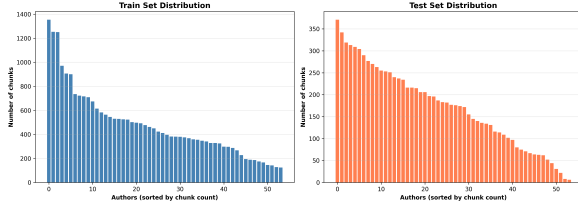


Figure 3: Distribution of chunks per author in training and test sets at 400 words. The imbalance justifies the use of macro-averaged F1 as the primary evaluation metric.

A.2 Classical Models

A.3 Neural Models

A.4 CharCNN Model

B Additional Results and Materials

B.1 Original (Mixed-Unit) Results

Figure 4 and Table 9 report our original evaluation design in full (§7): classical models scored per chunk, transformer models scored per document. Figure 5 shows the character n -gram ablation (§7.3) across all chunk sizes under this same original comparison.

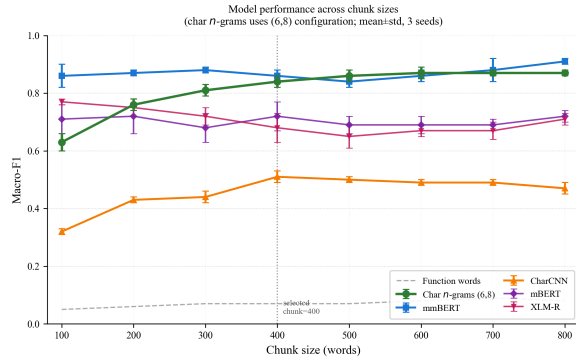


Figure 4: Model performance comparison across chunk sizes under the original, mixed-unit evaluation (classical models per chunk, neural models per document). Error bars indicate standard deviation across 3 random seeds for neural models.

B.2 Function Word List

Table 10 lists all 27 Uzbek function words used for the function-word baseline (§5.1).

B.3 Matched-Unit Evaluation: Full Results

Table 11 reports the full matched-unit evaluation (§7.2) across all eight chunk sizes, for both chunk-level and document-level scoring.

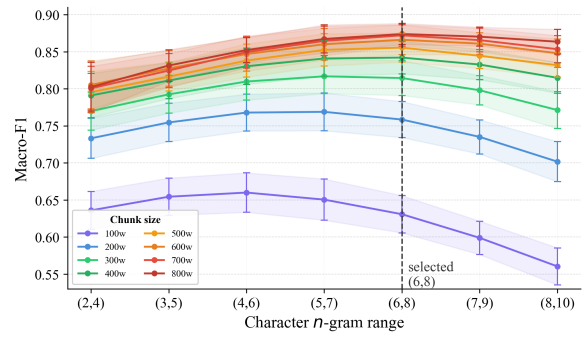


Figure 5: Character n -gram configuration ablation across chunk sizes (original, mixed-unit comparison). All configurations remain competitive with or superior to the neural baselines, with (6–8) achieving optimal performance. mmBERT is shown as a reference line.

B.4 Significance Testing: Full Results

Table 12 reports the full per-seed paired bootstrap significance test (§7.2) comparing character n -grams (6–8) against mmBERT at document level, at every chunk size and seed for which the mmBERT run completed evaluation.

B.5 Per-Author Error Analysis

Table 13 and Figure 6 report the character n -gram (6–8) error analysis at 400 words (§7); Table 14 reports the analogous mmBERT breakdown, discussed together in §7.

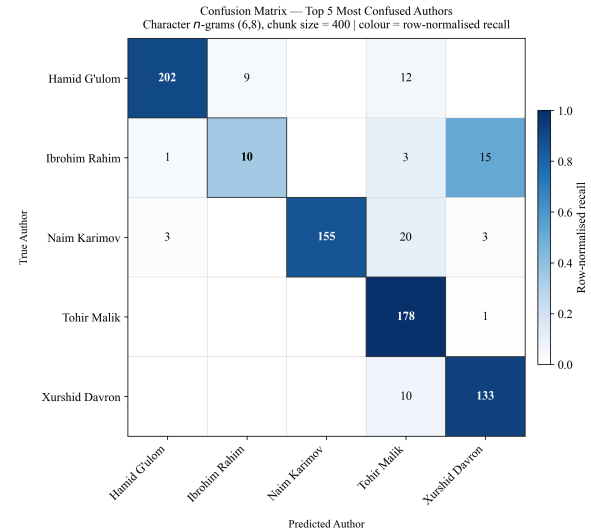


Figure 6: Confusion matrix for the top-5 most confused author pairs using character n -grams (6–8) at 400 words. Errors primarily occur between stylistically similar authors.

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 (27 tokens; full list in Appendix B.2)
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 6: 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 7: Hyperparameters for neural model training.

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 8: 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.

Table 9: Macro-F1 scores for closed-set authorship attribution across different chunk sizes, as originally scored: classical models (Function Words, Char n -grams, CharCNN) are evaluated per text **chunk**, while transformer models (mBERT, XLM-R, mmBERT) are evaluated per **document** (predictions aggregated across all chunks of a work) — i.e., this table mixes two different evaluation units across model families. All models report mean $\pm$ standard deviation across 3 random seeds. **Bold** indicates best result per chunk size. See §7 (Table 4) for a matched-unit comparison under which classical and neural models are scored identically.

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 $\pm$.03	0.76 $\pm$.02	0.81 $\pm$.02	0.84 $\pm$.02	0.86 $\pm$.02	0.87 $\pm$.02	0.87 $\pm$.01	0.87 $\pm$.01
CharCNN	0.32 $\pm$.01	0.43 $\pm$.01	0.44 $\pm$.02	0.51 $\pm$.02	0.50 $\pm$.01	0.49 $\pm$.01	0.49 $\pm$.01	0.47 $\pm$.02
mBERT	0.71 $\pm$.05	0.72 $\pm$.06	0.68 $\pm$.05	0.72 $\pm$.05	0.69 $\pm$.03	0.69 $\pm$.03	0.69 $\pm$.02	0.72 $\pm$.02
XLM-R	0.77 $\pm$.01	0.75 $\pm$.03	0.72 $\pm$.03	0.68 $\pm$.05	0.65 $\pm$.04	0.67 $\pm$.02	0.67 $\pm$.03	0.71 $\pm$.02
mmBERT	0.86 $\pm$.04	0.87 $\pm$.01	0.88 $\pm$.01	0.86 $\pm$.02	0.84 $\pm$.02	0.86 $\pm$.02	0.88 $\pm$.04	0.91 $\pm$.01

Table 10: The 27 Uzbek function words used for the function-word baseline (§B.2), grouped by rough category. The list was manually curated by the authors rather than derived from corpus frequency statistics (Limitations, point 8).

Category	Words
Coordinating conjunctions	va, ham, ammo, lekin, biroq, esa, yoki, hamda
Causal / concessive conjunctions	negaki, chunki, garchi
Postpositions	bilan, uchun, sababli, tufayli, kabi
Restrictive / emphatic particles	faqat, xolos, hatto, yana
Conditional / comparative	agar, go‘yo, xuddi
Discourse / emphatic adverbs	endi, demak, axir, albatta

Table 11: Macro-F1 (mean across seeds) under matched evaluation units, all chunk sizes; full per-seed values with standard deviations are in review/workstream_A/results/official_matched_table.csv in the released code. Top: chunk-level. Bottom: document-level.

<i>Chunk-level</i>								
Model	100	200	300	400	500	600	700	800
Function Words	0.050	0.060	0.070	0.070	0.070	0.080	0.090	0.080
Char n -grams (6–8)	0.630	0.760	0.810	0.840	0.860	0.870	0.870	0.870
CharCNN	0.320	0.430	0.440	0.510	0.500	0.490	0.490	0.470
mBERT	0.387	0.362	0.397	0.394	0.368	0.391	0.393	0.387
XML-R	0.408	0.350	0.386	0.372	0.376	0.376	0.370	0.384
mmBERT	0.527	0.490	0.580	0.544	0.528	0.571	0.567	0.564
<i>Document-level</i>								
Model	100	200	300	400	500	600	700	800
Function Words	0.057	0.078	0.070	0.088	0.103	0.085	0.088	0.078
Char n -grams (6–8)	0.975	0.976	0.975	0.967	0.980	0.975	0.967	0.958
CharCNN	0.551	0.677	0.714	0.731	0.765	0.738	0.776	0.695
mBERT	0.710	0.720	0.680	0.720	0.690	0.690	0.690	0.720
XML-R	0.770	0.750	0.720	0.680	0.650	0.670	0.670	0.710
mmBERT	0.860	0.870	0.880	0.860	0.840	0.860	0.880	0.910

Table 12: Full paired bootstrap significance test results: char n -grams (6–8) vs. mmBERT, document-level Macro-F1, at every (seed, chunk-size) combination with a completed mmBERT run (23 of 24; mmBERT/800/seed 44 did not finish evaluation, §10). Each row: 10,000 bootstrap resamples of the shared 72 test documents for that seed; gap = char n -grams – mmBERT; two-sided p -value from the bootstrap distribution. The gap is positive (favoring char n -grams) in all 23 rows and significant at $p < .05$ in 18/23.

Seed	Chunk	Char n-grams	mmBERT	Gap	95% CI	p-value	Sig.
42	100	0.960	0.815	+0.145	[+0.069, +0.232]	0.0016	Yes
42	200	0.963	0.882	+0.081	[+0.018, +0.160]	0.0184	Yes
42	300	0.960	0.894	+0.065	[+0.000, +0.128]	0.1028	No
42	400	0.935	0.867	+0.068	[+0.000, +0.131]	0.0522	No
42	500	0.975	0.818	+0.157	[+0.082, +0.252]	0.0004	Yes
42	600	0.960	0.830	+0.130	[+0.052, +0.221]	0.0014	Yes
42	700	0.935	0.833	+0.101	[+0.030, +0.178]	0.0098	Yes
42	800	0.935	0.920	+0.014	[−0.014, +0.045]	0.6264	No
43	100	0.990	0.896	+0.094	[+0.030, +0.165]	0.0322	Yes
43	200	0.990	0.872	+0.119	[+0.035, +0.197]	0.0114	Yes
43	300	0.990	0.874	+0.116	[+0.034, +0.192]	0.0126	Yes
43	400	0.990	0.832	+0.159	[+0.069, +0.247]	0.0012	Yes
43	500	0.990	0.849	+0.141	[+0.057, +0.223]	0.0034	Yes
43	600	0.990	0.880	+0.110	[+0.035, +0.180]	0.0124	Yes
43	700	0.990	0.921	+0.069	[+0.000, +0.127]	0.0974	No
43	800	0.965	0.899	+0.067	[−0.019, +0.148]	0.1762	No
44	100	0.975	0.879	+0.096	[+0.031, +0.186]	0.0158	Yes
44	200	0.975	0.857	+0.118	[+0.047, +0.208]	0.0022	Yes
44	300	0.975	0.867	+0.109	[+0.034, +0.192]	0.0104	Yes
44	400	0.975	0.891	+0.084	[+0.016, +0.161]	0.0324	Yes
44	500	0.975	0.867	+0.109	[+0.034, +0.195]	0.0104	Yes
44	600	0.975	0.873	+0.103	[+0.035, +0.180]	0.0112	Yes
44	700	0.975	0.891	+0.084	[+0.022, +0.161]	0.0292	Yes

Table 13: 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. The F1 column also reports a bootstrap 95% CI (10,000 resamples over the full test set; §7.4), which widens sharply for low-support authors.

Author	Test inst.	P	R	F1	F1 95% CI
<i>Top 5 Authors</i>					
Omon Muxtor	277	0.982	0.982	0.982	[0.97, 0.99]
Togay Murod	116	0.959	1.000	0.979	[0.96, 1.00]
Normurod Norqobilov	216	0.968	0.972	0.970	[0.95, 0.99]
Murod Muhammad Dost	109	0.990	0.945	0.967	[0.94, 0.99]
Mirmuhsin	240	0.926	0.992	0.958	[0.94, 0.98]
<i>Bottom 5 Authors</i>					
Xudoyberdi Toxtaboyev	22	0.457	0.727	0.561	[0.39, 0.71]
Sadulla Karomatov	52	0.354	0.538	0.427	[0.32, 0.53]
Nusrat Rahmat	6	0.208	0.833	0.333	[0.09, 0.54]
Abdulla Qodiriy	8	0.214	0.375	0.273	[0.00, 0.52]
Ibrohim Rahim	80	0.189	0.125	0.150	[0.07, 0.23]

Table 14: Per-author performance for the five best and five worst performing authors using mmBERT at 400-word chunks (seed 42), mirroring Table 13. Following mmBERT’s native reporting unit, “Test inst.” here counts *documents* (works), not chunks, so support is far lower than in the character n -gram table; see §7 for discussion.

Author	Test inst.	P	R	F1
<i>Top 5 Authors</i>				
Tohir Malik	3	1.000	1.000	1.000
Omon Muxtor	2	1.000	1.000	1.000
Togay Murod	2	1.000	1.000	1.000
Normurod Norqobilov	2	1.000	1.000	1.000
Mirmuhsin	2	1.000	1.000	1.000
<i>Bottom 5 Authors</i>				
Ibrohim Rahim	2	0.000	0.000	0.000
Abdulla Qodiriy	1	0.000	0.000	0.000
Mirzakalon Ismoiliy	1	0.000	0.000	0.000
Sadulla Karomatov	1	0.000	0.000	0.000
Tilavoldi Jorayev	1	0.000	0.000	0.000