BBN-U.Oregon's ALERT system at GenAI Content Detection Task 3: Robust Authorship Style Representations for Cross-Domain Machine-Generated Text Detection

Hemanth Kandula[1] **Chak Fai Li**[1] **Haoling Qiu**[1] **Damianos Karakos**[1]
Hieu Man Duc Trong [2] **Thien Huu Nguyen** [2] **Brian Ulicny**[1]
[1] RTX BBN Technologies [2]University of Oregon
{hemanth.kandula, chak.fai.li, haoling.qiu, damianos.karakos, brian.ulicny}@rtx.com
{hieum,thienn}@uoregon.edu

Abstract

This paper presents BBN-U.Oregon's system, ALERT, submitted to the Shared Task 3: Cross-Domain Machine-Generated Text Detection. Our approach uses robust authorship-style representations to distinguish between human-authored and machine-generated text (MGT) across various domains. We employ an ensemble-based authorship attribution (AA) system that integrates stylistic embeddings from two complementary subsystems: one that focuses on cross-genre robustness with hard-positive and negative mining strategies and another that captures nuanced semantic-lexical-authorship contrasts. This combination enhances cross-domain generalization, even under domain shifts and adversarial attacks. Evaluated on the RAID benchmark, our system demonstrates strong performance across genres and decoding strategies, with resilience against adversarial manipulation, achieving 91.8% TPR at FPR=5% on standard test sets and 82.6% on adversarial sets.

1 Introduction

The rapid advancement of large language models (LLMs) has transformed numerous fields, from natural language processing to automated content generation, enabling machines to generate text that is often indistinguishable from human writing. Models are now capable of producing fluent, coherent, and contextually relevant text, sparking widespread adoption across industries for tasks ranging from content creation to customer service. However, alongside these advancements, the potential for misuse has grown, particularly in areas like disinformation, academic plagiarism, automated phishing attacks, and social media manipulation. New challenges arise in distinguishing between human-authored and machine-generated content.

To address these challenges, we developed the ALERT (Authorship through Learnable and Explainable Rich Transformations) system, which was submitted to the Shared Task on Cross-Domain Machine-Generated Text Detection within the COLING Workshop on Detecting AI Generated Content. The focus of this shared task (Dugan et al., 2025) is on evaluating the cross-domain robustness of MGT detectors across diverse domains, generative models, and decoding strategies. This challenge addresses the critical need for detectors that maintain high accuracy and low false positive rates when applied to MGT in varied real-world contexts.

Traditional approaches for detecting MGT typically rely on supervised learning, where detectors are trained on labeled corpora of human-written and machine-generated documents. However, these methods often struggle with generalization, particularly as new, more sophisticated LLMs emerge (Zellers et al., 2019). Furthermore, these models are highly sensitive to distribution shifts, meaning that performance degrades when applied to LLMs or domains not seen during training (Mitchell et al., 2023). As LLMs become more prevalent and diversified, such approaches become increasingly impractical.

To address these challenges, recent research has focused on learning authorship style representations (Soto et al., 2024). Style, a fundamental characteristic of human authorship, varies across individuals and tasks but tends to be more consistent within a specific LLM. By capturing stylistic nuances, detectors can more effectively identify MGT, even in the face of emerging LLMs or previously unseen content domains. This paper expands on the work of (Rivera-Soto et al., 2021) by proposing an improved Authorship Attribution (AA) system for MGT using authorship style representations. Our models, which generate these representations, are trained using advanced techniques such as GradCache (Gao et al., 2021), various hard-negative mining (Robinson et al., 2021) (Fincke and Boschee, 2024) and hard-positive filter-

Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect), pages 358–364
January 19, 2025.

ing strategies, and their embeddings are combined by a fusion module to improve performance. These techniques improve the models' ability to distinguish between the writing styles of humans and MGT, across LLMs and domains. The rest of the paper is organized as follows: Section 2 reviews related work in MGT detection, Section 3 describes our methodology and ensemble-based AA system, Section 4 outlines the experimental setup and results, and Section 5 concludes the paper with future directions.

2 Related Work

Detecting MGT has become an increasingly critical task with the rapid growth of large language models (LLMs) like GPT-3 and -4, LLaMA-2 and -3, which can produce highly fluent and human-like text. Early approaches focused on supervised methods, where models were trained on labeled datasets of humans and MGT. For instance, OpenAI's AI Detector was designed to distinguish between text written by GPT-2 and human authors, but its performance declined with the release of more advanced models like GPT-3 and ChatGPT (Solaiman et al., 2019). These supervised detectors, while effective in their specific settings, often fail to generalize to unseen LLMs due to the constant evolution of model architectures and training paradigms (Zellers et al., 2019). Soto et al. (2024) propose a method based on authorship style representations (Rivera-Soto et al., 2021), which leverages the stylistic features of human-written text to detect machine-generated content in a few-shot scenario, without relying on large amounts of training data from the target LLMs. This method addresses the limitations of supervised learning by focusing on invariant features of writing style. Hans et al. (2024) introduce a zero-shot detection method, "Binoculars", which contrasts the outputs of two closely related LLMs to identify MGT with high accuracy. Recent advances also include adversarial learning approaches like RADAR (Hu et al., 2023), which improve robustness by training detection models to identify adversarially crafted MGT. These methods offer improved performance in challenging scenarios where LLMs are specifically designed to bypass detection systems.

3 Methodology

The core of our detection framework is based on the hypothesis that MGT exhibits consistent stylistic patterns that differ from those of human authors. To capture these stylistic cues effectively, we implement an ensemble-based AA system, combining two complementary subsystems optimized with advanced training techniques for robustness and cross-domain generalization.

Our methodology builds upon the contrastive learning approach used by (Rivera-Soto et al., 2021), with key improvements. The AA system ensemble is designed using a Siamese neural architecture, which captures nuanced stylistic signatures through embeddings that serve as distinctive authorship signatures. The ensemble integrates cues from multiple linguistic and stylistic features, enabling a cohesive detection framework with broad generalization capabilities. While Rivera-Soto et al. (2021) introduced authorship attribution with contrastive learning, our approach extends it by incorporating advanced hard-positive and hard-negative mining strategies (BM25-based and cluster-based), GradCache for larger batch sizes, and fusing embeddings from multiple systems.

The core components of our framework, AA System I and AA System II, each employ unique training strategies. Below, we provide training strategies employed by each system and its specific optimizations.

3.1 AA system I: Cross-Genre Robustness with Hard-Positive and Negative Mining

AA System I employs a training methodology that emphasizes cross-genre robustness through specialized hard-positive filtering and hard-negative mining strategies adopted from (Fincke and Boschee, 2024). For hard-positive examples, the system uses the two most topically distant documents available per author, promoting the learning of stylistic consistency rather than topical similarity. To refine this process, authors with insufficiently dissimilar document pairs are excluded from training, resulting in fewer but more challenging examples that improve performance in both genre-specific and cross-genre contexts. For hard-negatives, the system generates batches containing clusters of authors where each author contributes two documents: one near the cluster center for similarity and the other in the outer reaches for dissimilarity, ensuring stylistic contrast. K-means clustering determines initial centroids, with each centroid representing one author, and documents closest to each centroid are selected to populate clusters. FAISS-based similarity search (Douze et al., 2024)maintains clustering

efficiency by capping retrieval to the nearest 2,024 entries. Once clusters are formed, centroids are grouped to fill each batch with a set number of authors, creating more coherent batches and ensuring that each batch offers challenging stylistic contrasts. Further details on these methods, including clustering and selection criteria, are available in Fincke and Boschee (2024). In summary, AA System I focuses on cross-genre robustness by applying hard-positive filtering and a clustering-driven hard-negative mining strategy that relies on topically distant documents. This approach encourages the model to learn stylistic consistency that is not conflated with topic similarity.

3.2 AA System II: Semantic, Lexical, Clustering based Contrastive Learning

System II is designed to capture nuanced stylistic differences across authors through hard-positive filtering and a dual-strategy hard-negative mining approach. The same hard-positive mining strategy from System I Sec 3.1 is used in this system for dataset filtering. This subsystem, while sharing foundational techniques with AA System I, incorporates distinct selection criteria for training examples to improve the model's ability to distinguish stylistic similarities across diverse topics. For mining hard-negative examples, in the first stage, BM25 (Robertson et al., 2009) is applied to retrieve top-k collections, where each "collection" refers to the set of documents written by a single author. By selecting collections that are lexically similar to the anchor documents yet originate from different authors, the model is encouraged to focus on subtle stylistic patterns rather than topical similarities. This process enables the model to focus on subtle stylistic patterns, reducing the influence of the topic. Subsequently, a two-level clustering approach using K-means is adopted, to capture more nuanced semantic content. The first level performs document-level clustering, grouping documents based on their semantic content, primarily capturing topical similarities. The second level implements author-level clustering, organizing author collections based on aggregated embeddings that reveal patterns in authorship style. Within each author-level cluster, collections from different authors are selected as Hard-negative examples, further refined by retaining only documents that fall within the same document-level clusters as the anchor documents. Hard-negative mining (Robinson et al., 2021) is performed on the complete dataset, not limited to training subsets, to ensure a broader range of potential hard-negatives. Document-level and author-level clusters are set to 512, with a balanced distribution of 50% BM25-mined and 50% cluster-mined negative examples. This comprehensive approach supports a variety of negative examples, challenging training instances, and robust model performance. By combining semantic, lexical, and clustering-based approaches, the aforementioned process makes the model focus on the most important features for authorship style discriminability.

Additionally, the GradCache (Gao et al., 2021) technique allows for larger batch sizes, storing intermediate gradients to reduce memory load. This enables the model to handle a higher volume of examples per batch, improving generalization across diverse domains and effectively distinguishing subtle stylistic differences in authorship.

Overall, AA System II builds upon similar concepts but differs notably in its hard-negative mining strategy. While System I relies on clustering and topically distant pairs, System II adopts a dual-strategy method: first, BM25-based retrieval identifies lexically similar yet differently authored documents; second, a two-level clustering approach (document-level and author-level) further refines these candidates. This combination enables System II to pinpoint subtler stylistic discrepancies that persist even among topically and lexically similar texts.

3.3 Machine Style Detection

The MGT Style Detection system uses learned authorship style representations to accurately distinguish between humans and MGT. In the final classification stage, a fully connected layer processes these stylistic embeddings, followed by a binary classification layer specifically trained to detect MGT. The AA (sub-)system produces domain-invariant style representations, thus making the MGT detection system domain/genre-invariant as well. Furthermore, an ensemble system enhances detection capability by combining style embeddings from both AA systems, achieving a robust and comprehensive understanding of stylistic nuances for greater accuracy across diverse domains.

Model	Development Set (20% RAID Train)				Evaluation Set
	Abstracts	Books	News	Average	(RAID Test)
AA System I (Sec: 3.1)	0.790	0.838	0.927	0.852	-
AA System II (Sec: 3.2)	0.975	0.939	0.982	0.965	0.893
Ensemble System	0.966	0.971	0.982	0.973	0.918

Table 1: Performance of Cross-Domain MGT Detection on RAID Dataset (Subtask-A: No Adversarial Attacks)

Model	Development Set (20% RAID Train)				Evaluation Set
	Abstracts	Books	News	Average	(RAID Test)
AA System I (Sec: 3.1)	0.612	0.650	0.912	0.794	-
AA System II (Sec: 3.2)	0.887	0.866	0.937	0.897	0.788
Ensemble System	0.876	0.934	0.978	0.930	0.826

Table 2: Performance of Cross-Domain MGT Detection on RAID Dataset (Subtask-B: with Adversarial Attacks)

4 Experiments and Results

4.1 Data

Both authorship systems are trained on various datasets (see Appendix Table 5) with authorship labels, employing various author contrastive learning objectives—with a focus on authors who have produced at least 100 documents. To increase sample diversity, longer documents are split into shorter segments, augmenting the training pool. The Cross-domain MGT Detection task (Dugan et al., 2025) uses the RAID benchmark (Dugan et al., 2024) which consists of over 10 million documents spanning 11 LLMs, 11 genres, 4 decoding strategies, and 12 adversarial attacks. To evaluate our models, we utilized the training set from RAID. We divided the RAID dataset into 60% train, 20% validation, and 20% development sets, ensuring an equal representation of genres, LLMs, and adversarial attacks. Document source information was used to prevent overlap between training and test sets.

4.2 Experiment Setup

For the AA systems, we use Qwen2 1.5B [1] and E5-mistral-7b-instruct [2] for text embeddings in Systems I and II, respectively. Model optimization was done using the AdamW Optimizer (Loshchilov, 2017), and training was conducted on 4 NVIDIA RTX A6000 GPUs.

To assess cross-domain generalization, we conducted cross-validation experiments by training on two of the three genres in RAID (Abstracts, Books, and News) and testing on the held-out genre. Final classification layers are trained using the 60% of the train set, 20% validation set to select the classification layer weights and results reported in Sec 4.3 are on 20% of the development set. While initial experiments involved domain-specific splits to guide hyperparameter selection, in the final reported model, the final classification layer is trained on the full 60% using all available training domains for maximum coverage.

We use the official evaluation metric, TPR @ FPR=5%, which measures the model's accuracy in detecting MGT at a fixed false positive rate of 5%.

4.3 Results

Table 1 and Table 2 show results with our MGT detection models showing strong cross-domain performance, particularly highlighting the effectiveness of the ensemble-based approach. The cross-validation on the development set reveals that the ensemble system achieves the highest average TPR at FPR=5%, which is also reflected in the evaluation results. Without adversarial attacks, the ensemble system outperforms individual models by capturing more varied stylistic representations, which enabled it to generalize well even when facing domain shifts. In adversarial settings, the ensemble maintained robustness, showing less performance degradation compared to individual systems.

Our results on the Development and Evaluation sets indicate that while both AA System I and AA System II contribute to performance, System II provides a stronger baseline detection accuracy due to its dual-strategy hard-negative mining, which integrates both lexical and semantic constraints. Although System II alone is highly effective, the ensemble capitalizes on System I's cross-genre ro-

[1] https://huggingface.co/Qwen/Qwen2-1.5B

[2] https://huggingface.co/Linq-AI-Research/Linq-Embed-Mistral

bustness and System II's nuanced stylistic discrimination. As a result, combining them leads to more stable and improved performance, particularly in challenging or previously unseen domains.

The results on the Evaluation set further validate the generalizability of our models. The ensemble model (ALERT v1.1 reported in (Dugan et al., 2025)) resulted in a TPR at FPR=5% of 0.918 without adversarial attacks and 0.826 with such attacks, indicating consistent stability across diverse genres and text styles. Although the model was not specifically fine-tuned for adversarial attacks, these results suggest that capturing nuanced authorship styles enhances detection performance across varied content types and adversarial scenarios.

5 Conclusions

We show that our ensemble-based authorship style representations from two complementary subsystems identify MGT across varied domains and adversarial attacks. By integrating advanced training techniques such as GradCache, contrastive learning, and hard-positive/negative mining, the system demonstrates strong cross-domain generalization, achieving reliable MGT detection across various genres, LLMs, and adversarial attacks, thanks to capturing nuanced authorship-style representations. Future work could extend the framework to handle more sophisticated adversarial attacks and support additional languages and low-resource domains, making it adaptable to a wider range of real-world applications. Exploring domain adaptation techniques could improve robustness in detecting MGT by new or unseen models.

Acknowledgments

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 2022-22072200003. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *Preprint*, arXiv:2401.08281.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12463–12492, Bangkok, Thailand. Association for Computational Linguistics.

Liam Dugan, Andrew Zhu, Firoj Alam, Preslav Nakov, Marianna Apidianaki, and Callison-Burch Chris. 2025. Genai content detection task 3: Cross-domain machine generated text detection challenge. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Steven Fincke and Elizabeth Boschee. 2024. Separating style from substance: Enhancing cross-genre authorship attribution through data selection and presentation. *arXiv preprint arXiv:2408.05192*.

Luyu Gao, Yunyi Zhang, Jiawei Han, and Jamie Callan. 2021. Scaling deep contrastive learning batch size under memory limited setup. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 316–321, Online. Association for Computational Linguistics.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting llms with binoculars: Zero-shot detection of machine-generated text. *arXiv preprint arXiv:2401.12070*.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023. Radar: Robust ai-text detection via adversarial learning. *Advances in Neural Information Processing Systems*, 36:15077–15095.

I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*.

Rafael A Rivera-Soto, Olivia Elizabeth Miano, Juanita Ordonez, Barry Y Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. 2021. Learning universal authorship representations. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 913–919.

Model	Abstracts			Books			News		
	Sys I	Sys II	Ensemble	Sys I	Sys II	Ensemble	Sys I	Sys II	Ensemble
ChatGPT	0.982	1.000	1.000	0.961	0.996	0.999	NA	NA	NA
Cohere	0.103	0.823	0.647	0.355	0.728	0.778	NA	NA	NA
Cohere-Chat	0.780	0.966	0.973	0.667	0.910	0.924	NA	NA	NA
GPT-2	0.710	0.960	0.978	0.843	0.948	0.991	0.916	0.893	0.987
GPT-3	0.773	0.986	0.986	0.879	0.997	0.999	NA	NA	NA
GPT-4	0.966	1.000	0.996	0.527	0.994	0.989	NA	NA	NA
Llama-Chat	0.999	1.000	1.000	0.996	0.996	0.999	0.994	0.996	0.999
Mistral	0.403	0.972	0.953	0.794	0.883	0.967	0.796	0.831	0.926
Mistral-Chat	0.955	0.998	0.999	0.987	0.995	1.000	0.986	0.996	1.000
MPT	0.856	0.967	0.981	0.853	0.869	0.951	0.895	0.942	0.987
MPT-Chat	0.989	0.999	0.999	0.956	0.976	1.000	0.973	0.972	0.993
Aggregate	0.790	0.975	0.966	0.838	0.939	0.971	0.927	0.938	0.982

Table 3: Performance on Various LLMs Detection on Development Set (20% RAID Train). Sys I refers to (Sec: 3.1), Sys II refers to (Sec: 3.2).

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Rafael Alberto Rivera Soto, Kailin Koch, Aleem Khan, Barry Y. Chen, Marcus Bishop, and Nicholas Andrews. 2024. Few-shot detection of machine-generated text using style representations. In *The Twelfth International Conference on Learning Representations*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Advances in neural information processing systems*, 32.

A Detailed Results on MGT Detection

The detailed performance of individual models on the development set (20% of RAID Train) is provided in Table 3, showcasing their performance across specific domains. Similarly, Table 4 highlights the results of each model on the RAID Test set. While the average performance is shown in Table 1 and 2, these results underline the contributions and effectiveness in detecting individual models in various domains.

B Authorship Attribution Model Training, Implementation and Datasets Details

This section provides key implementation details for our Authorship Attribution (AA) systems and the final Machine-Generated Text (MGT) detection classifier.

B.1 Authorship Attribution Models

The architecture leverages pre-trained transformer models, such as `Qwen2-1.5B` and`E5-mistral-7b-instruct` (Sec 4), to process input documents into high-dimensional feature vectors. For longer documents, the text is divided into smaller segments, like paragraphs, to enhance the model's ability to capture local stylistic nuances effectively. These segment embeddings are subsequently aggregated using techniques like self-attention mechanisms and max-pooling to produce a cohesive representation of the author's stylistic signature. The model is optimized with a contrastive loss function, ensuring that embeddings of documents by the same author are positioned closer in the vector space than those of different authors. For a detailed discussion on model training, refer to (Rivera-Soto et al., 2021). Dataset used to train AA System I and System II are shown in Table 5

Hard-Positive and Hard-Negative Mining: For positives, we select each author's two most topically distant documents to emphasize stylistic over topical consistency. For negatives, System I focuses on cluster-based mining, grouping authors via K-means and selecting documents that are both cluster-center and periphery examples. System II integrates BM25 retrieval to find lexically simi-

Model	Subtask-A		Subtask-B	
	ALERT v1.1	ALERT v1.2	ALERT v1.1	ALERT v1.2
ChatGPT	0.976	0.958	0.882	0.854
GPT-4	0.943	0.917	0.834	0.812
GPT-3	0.917	0.932	0.828	0.805
GPT-2	0.919	0.897	0.826	0.787
Mistral	0.862	0.826	0.778	0.740
Mistral-Chat	0.973	0.943	0.874	0.832
Cohere	0.706	0.725	0.629	0.605
Cohere-Chat	0.848	0.823	0.767	0.707
Llama-Chat	0.988	0.952	0.889	0.852
MPT	0.905	0.873	0.825	0.784
MPT-Chat	0.960	0.922	0.859	0.811
Aggregate	0.918	0.893	0.826	0.788

Table 4: Performance of ALERT Detectors for each model on Cross-Domain MGT Detection (RAID Test Dataset) for Subtask-A: no Adversarial Attacks and Subtask-B: with Adversarial Attacks

lar but differently authored documents and then refines these candidates via document- and author-level clustering, ensuring that negative pairs are semantically and lexically close but differ in style.

Optimization: We use the AdamW optimizer (Loshchilov, 2017) with a learning rate of 5×10^{-5}. GradCache (Gao et al., 2021) enables an effective batch size of 2048. Each AA system is trained for about 5 epochs, with model selection based on validation performance.

B.2 MGT Detection Classifier

Once AA models are trained, we apply them to produce embeddings for each RAID document. We concatenate embeddings from System I and System II and feed them into a two-layer feed-forward classifier (hidden size 512, ReLU activation, 0.1 dropout) to predict whether the text is machine-generated. The classifier is optimized with AdamW at a 1×10^{-4} learning rate for 3–5 epochs, using a validation set to select the best checkpoint.

Dataset Name	# authors	# documents
English Reddit Million User Dataset	7.6K	4.7M
English Pushshift Reddit Dataset	28.9K	2.0M
English Twitter	13	1.9K
English Hackernews	12.3K	1.7M
English StackExchange	19.8K	1.4M
Russian stihi	7.9K	1.4M
English Amazon Review	3.6K	827.2K
Russian proza	1.9K	206.6K
English NYT Comment	1.3K	172.5K
English Blog Authorship Corpus	1223	140.3K
Russian Telegram	2.8K	128.9K
English Yelp Review	485	113.9K
Russian KP	313	43.0K
Russian Pushshift Reddit Dataset	247	37.5K
English IMDb1M/IMDb62	253	3.1K
Russian Stackexchange	122	1.1K

Table 5: Datasets for Authorship Attribution Training