BBN-U.Oregon's ALERT system at GenAI Content Detection Task 3:

Robust Authorship Style Representations for Cross-Domain Machine-Generated Text Detection

Detecting AI-Generated Content" Workshop @ COLING 2025

Brian Ulicny, PhD

Date: January 19, 2025

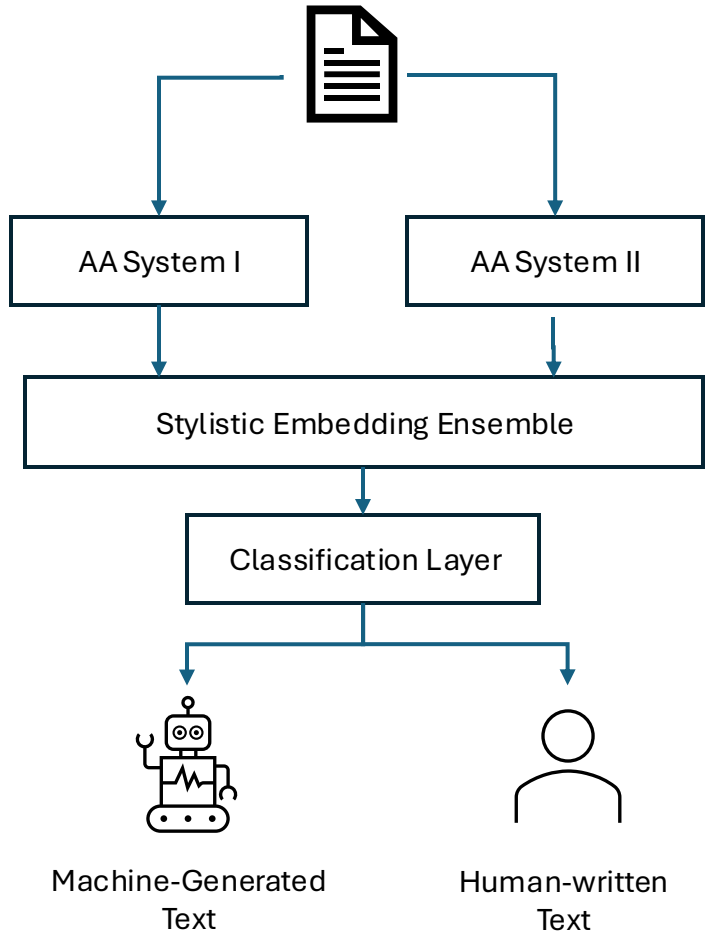
Hemanth Kandula[1] **Chak Fai Li**[1] **Haoling Qiu**[1] **Damianos Karakos**[1]
Hieu Man Duc Trong[2] **Thien Huu Nguyen**[2] **Brian Ulicny**[1]
[1] RTX BBN Technologies [2]University of Oregon

Core Idea

Purpose: Detect cross-domain Machine-Generated Text (MGT) with robust authorship-style representations.

Key Components:

- Authorship Attribution (AA) Systems: (inspired from LUAR [1])
  - System I: Focuses on cross-genre robustness using hard-positive and hard-negative mining.
  - System II: Leverages semantic and lexical clustering for nuanced stylistic contrasts.
- Ensemble Approach: Combines stylistic embeddings from both systems for improved accuracy and domain generalization.



Machines generate *stylistically consistent text* that is *stylistically different from human styles.*

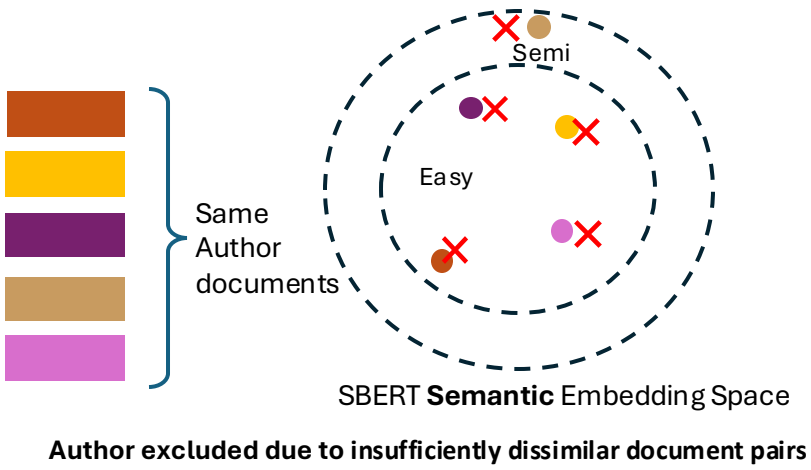
[1] Rivera-Soto, Rafael A., et al. "Learning universal authorship representations." Proceedings of the 2021 conference on empirical methods in natural language processing. 2021.

AA System I

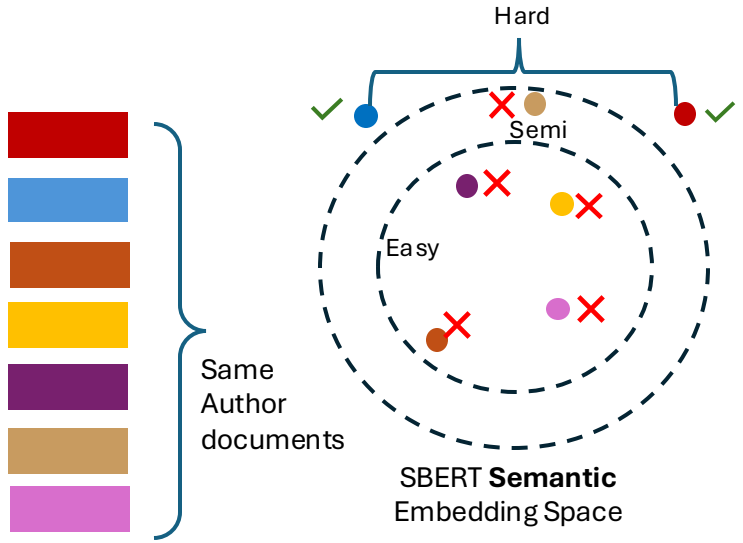
- **Focus:** Cross-genre robustness through hard-positive filtering and hard-negative mining strategy that relies on topically distant documents.
- This approach encourages the model to learn stylistic consistency that is not conflated with topic similarity.

Hard-positive Filtering

- Use the two most topically distant documents available per author
- Focus on learning stylistic similarity rather than topical similarity.
- Exclude authors with insufficiently dissimilar document pairs



Author excluded due to insufficiently dissimilar document pairs

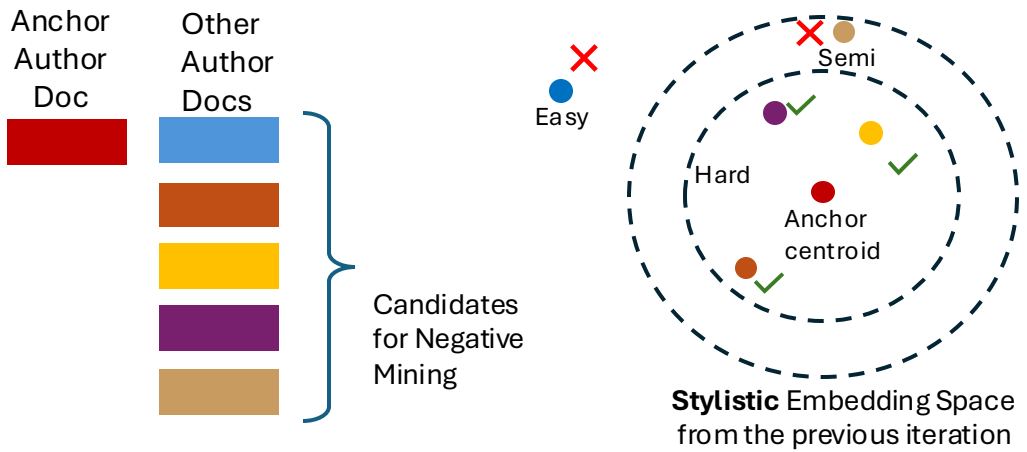


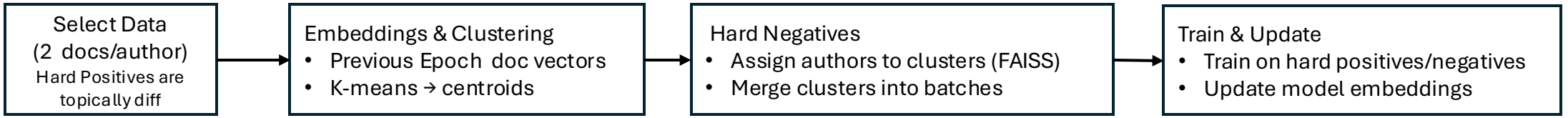
Fincke, S., & Boschee, E. (2024). Separating Style from Substance: Enhancing Cross-Genre Authorship Attribution through Data Selection and Presentation. arXiv preprint arXiv:2408.05192.

AA System I

Hard-negative mining

- 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
- 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.





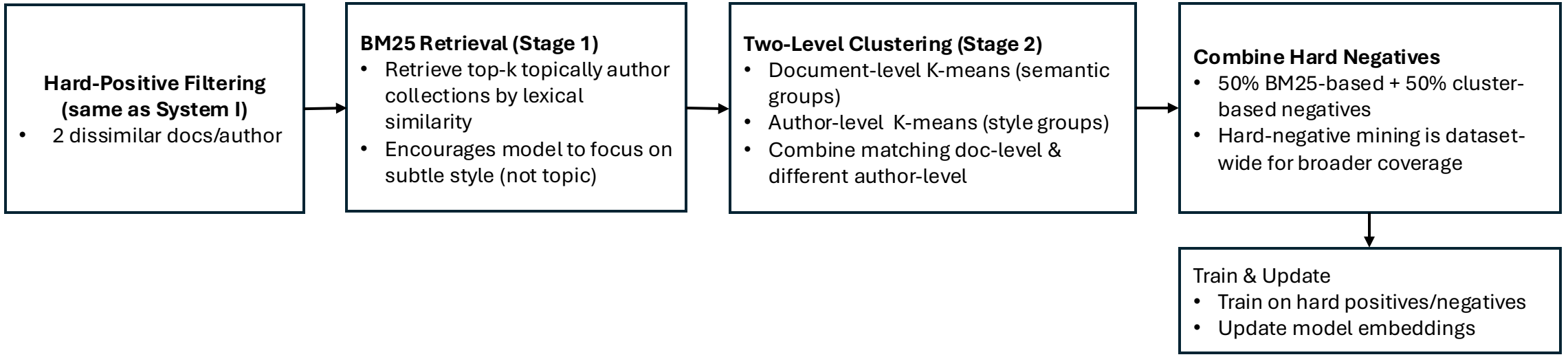
Fincke, S., & Boschee, E. (2024). Separating Style from Substance: Enhancing Cross-Genre Authorship Attribution through Data Selection and Presentation. arXiv preprint arXiv:2408.05192.

AA System II

System II is designed to capture nuanced stylistic differences across authors through hard-positive filtering and a dual-strategy hard-negative mining approach.

Hard-positive filtering: Same as AA System I

Dual-strategy hard-negative mining



Results

Model	Development Set (20% RAID Train)				Evaluation Set (RAID Test)
	Abstracts	Books	News	Average	
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 (RAID Test)
	Abstracts	Books	News	Average	
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)

The ensemble system achieves the higher TPR at FPR=5%, demonstrating high performance and robustness across domains and adversarial settings.

Conclusions

- 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. Thanks to capturing nuanced authorship-style representations, it achieves reliable MGT detection across various genres, LLMs, and adversarial attacks.
- 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.