Understanding writing style in social media with a supervised contrastively pre-trained transformer

Javier Huertas-Tato[1], Alejandro Martín[1], and David Camacho[1]

Department of Computer Systems Engineering,
Universidad Politécnica de Madrid,
St. Ramiro de Maeztu, 28040, Madrid, Spain
{javier.huertas.tato, alejandro.martin, david.camacho}@upm.es

Abstract. We introduce STAR[1][2] a pre-trained transformer designed for robust detection and analysis of writing style in social media. The model, trained on a heterogeneous corpus, has seen $4.5{\cdot}10^6$ texts from 70k authors. Supervised Contrastive Learning and massive batch sizes grant competitive performance in challenges in zero- and many-shot scenarios.

Keywords: Authorship Attribution, Transformers, Natural Language Processing, Representation Learning, Contrastive Learning

1 Introduction

Malicious actors at online social networks (OSNs) frequently create multiple accounts to evade disciplinary measures. Early detection of these tactics must rely on authorship analysis to detect the nuances of speech produced by participants. Thus, our aim is to characterize an author using exclusively the textual content they produce.

Authorship attribution (AA) requires detecting the author of a text. Recent advances in this field perform Contrastive Learning (CL) to align representations into an authorship embedding space [2, 3]. Novel models like these give way to STAR [1], which leverages a large-scale AA supervised contrastive learning objective to effectively generate an authorial embedding space.

STAR stands as a powerful open-source and open-weight model able to extract state of the art authorship embeddings, which requires no fine-tune to offer competitive performance under distribution shift. Our contributions include: *1)* Methodology for effective AA pre-training at scale, *2)* Generation of authorship representations for arbitrary chunks of text. *3)* Applicability to a breadth of demonstrated tasks including AA and verification, among others.

* This work has been funded by the project PCI2022-134990-2 (MARTINI) of the CHISTERA IV Cofund 2021 program, funded by MCIN/AEI/10.13039/501100011033 and by the "European Union NextGenerationEU/PRTR"; by European Comission under IBERIFIER Plus - Iberian Digital Media Observatory (DIGITAL-2023-DEPLOY-04-EDMO-HUBS 101158511); and by EMIF managed by the Calouste Gulbenkian Foundation, in the project MuseAI.

[1] We share our pre-trained model at huggingface AIDA-UPM/star and our code is available at jahuerta92/star

[2] This Keywork summarizes the original STAR publication at [1].

2 Supervised contrastive Training for Authorship Representations

The summary of **STAR** can be seen in Fig. 1. The STAR model is trained with an AA objective. In layman terms, our idea consists on choosing a learning objective that requires that two text created by the same author share the same representation, regardless of contents. Formally, given an author set $A = \{A^{(i)} : i \in [1, N]\}$, where each Author has M documents as in $A^{(i)} = \{D_j^{(i)} : j \in [1, M]\}$, we want to parametrize a function f_θ such that the representation vectors produced for the exact same author are similar via some distance measure (usually cosine similarity). An embedding set can be computed as $e = \{f_\theta(D_j^{(i)}) : i \in [1, N], j \in [1, M]\}$ This is performed via Supervised Contrastive Learning (SCL) [4] as expressed in function1.

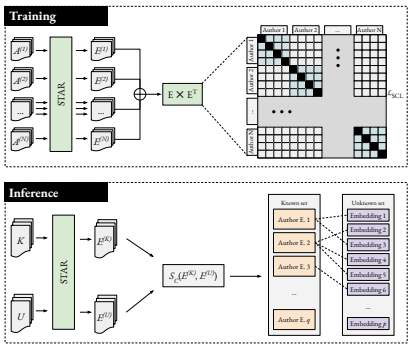


Fig. 1. Summary of training and inference for STAR

$$\mathcal{L}_{SCL}(e) = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(e_i \cdot e_p / \tau)}{\sum_{a \in A(i)} \exp(e_i \cdot e_a / \tau)} \tag{1}$$

where e is the total embedding set, $P(i)$ is the positives set for embedding at index i.

The proposed methodology requires several positive same-author documents as well as further negatives. While other previous work uses single-example positives (via InfoNCE [5]), STAR uses several in-batch positives.

Data: Another key to this work was the breadth of data used. We train on four datasets, curated for writing style heterogeneity. Plus, as our focus is on social media, we prioritized texts from different OSN sources. We use the Blog authorship corpus [6], the Reddit TLDR dataset [7], the twitter author corpus [8][3], and the Standardized Gutenberg Corpus [9]. Each batch is balanced, each $\frac{1}{4}$th belonging to each dataset. The final composite dataset has more than 4 million documents from 70 thousand authors.

[3] Due to changes in Twitter/X policy, now this dataset is inaccessible.

Effects of batch size: One key to effective contrastive learning is high batch sizes. Large batches enforce high numbers of comparisons, usually increasing performance. Thus, we consider important to note the huge batch size employed in this work, using >16k total size with 16 documents per author and 1k authors.

Usage: After training, inference can be performed in a task-to-task basis. STAR is meant to produce embeddings with its last layer; however, it can be adapted for other tasks like Author Verification[4] (see Figure 2)

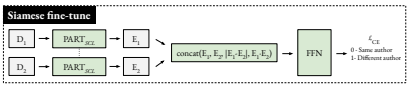


Fig. 2. Example adaption to authorship verification

3 Evaluation summary

STAR has been tested against other state of the art baselines like Style Embeddings [3] and PART [2], as well as standard zero-shot baselines with RoBERTa [10] and DeBERTav3 [11]. We used PAN-CLEF[5] attribution, verification and clustering challenges, as well as a hard partition of our own reddit dataset.

3.1 Challenge performance

STAR excels at AA against embedding authorship models: We find that on all PAN11/12/18/19 challenges STAR dominates results in Accuracy and F1-score, with PART being a second close. This demonstrates fair performance under distribution shift.

Fine-tuning STAR for verification yields top-competition results: This authorship model competes directly on all PAN20/21/22/23 tasks with the best results of the competitions with the sole exception of PAN2022 Task 2 Dataset 3. Compared to PART, the STAR model increasis effectiveness.

Raw representations allow for effective clustering: Finally, we tested on PAN16-17 challenges for clustering, demonstrating top-line performance of STAR at zero-shot. This, again shows robustness to distribution shift.

3.2 Application to Reddit

The aim of this work targeted social media. While PAN challenges contain twitter partitions, it is not the focus. Instead we partition the Reddit dataset, picking authors outside the training set for validation. Our validation includes the number of positives and negatives.

Our results show that STAR outperform other models both in raw performance at any setting, but also much more stable. Under repeated subsampling

[4] *Author Verification Task*: Determine if two texts are authored by the same person.
[5] All semi-public at pan.webis.de/data.html

of the dataset, STAR shows the lowest standard deviation among results from all the chosen baselines. For comparison, when there are 1616 negative authors, PART manages a 72.96% while STAR achieves a 93.74% Top-5 accuracy when there are 8 positive examples for references. For perspective, Style-Embeddings barely achieves 13% and non-authorship models are below 11%. This demonstrates potential application to a social media network like Reddit.

4 Conclusions

In this article, we introduce STAR, our authorship representation Model focused on Online Social Network. Our key findings are: usage of multiple positives, massive dataset upscaling and large batch size. Operationalizing these keys, we manage to produce an open-source open-weight model capable of representing numerically authorship in social media. Our evaluation shows the robustness under distribution shift, state of the art capabilities of fine-tuning and in-domain application to several authorship tasks.

References

1. Javier Huertas-Tato, Alejandro Martín, and David Camacho. Understanding writing style in social media with a supervised contrastively pre-trained transformer. *Knowledge-Based Systems*, 296:111867, 2024.
2. Huertas-Tato, Martín, and Camacho. PART: Pre-trained authorship representation transformer. *Hum.-centric Comput. Inf. Sci.*, 14(0):25–45, July 2024.
3. Anna Wegmann, Marijn Schraagen, and Dong Nguyen. Same author or just same topic? towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268, 2022.
4. Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.
5. Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
6. Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205, 2006.
7. Michael V"olske, Martin Potthast, Shahbaz Syed, and Benno Stein. TL;DR: Mining Reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.
8. Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 759–768, 2010.
9. Martin Gerlach and Francesc Font-Clos. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Entropy*, 22(1):126, 2020.
10. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
11. Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.