SYSML: StYlometry with Structure and Multitask Learning: Implications for Darknet Forum Migrant Analysis

Pranav Maneriker, Yuntian He, Srinivasan Parthasarathy

The Ohio State University

Summary

- Stylometry-based **multitask learning approach** for natural language and model interactions.
- Uses **graph embeddings** to construct low-dimensional representations of short episodes of user activity for authorship attribution.
- Lift of up to **2.5X** on **Mean Retrieval Rank** and 2X on Recall@10 compared to baselines across four darknet market forums.
- Analysis of textual stylometric features, **obfuscation strategies**, and **migration** of users across forums.

Motivation

- Recent advancements in using neural networks for character and word-level modeling for authorship attribution have shown promising results on shorter text.
- The identity of individuals on a crypto-market is associated only with a username; building trust on these networks does not follow conventional models prevalent in eCommerce.
- Our goal is to understand how textual style evolves on darknet markets and how users on such markets, to eventually analyze the trust signalling strategies as users transition across markets.
- The base unit of activity that we aim to identify is an episode, i.e., collection of posts by a user, along with the post contexts (subforums) and posting times.

Approach

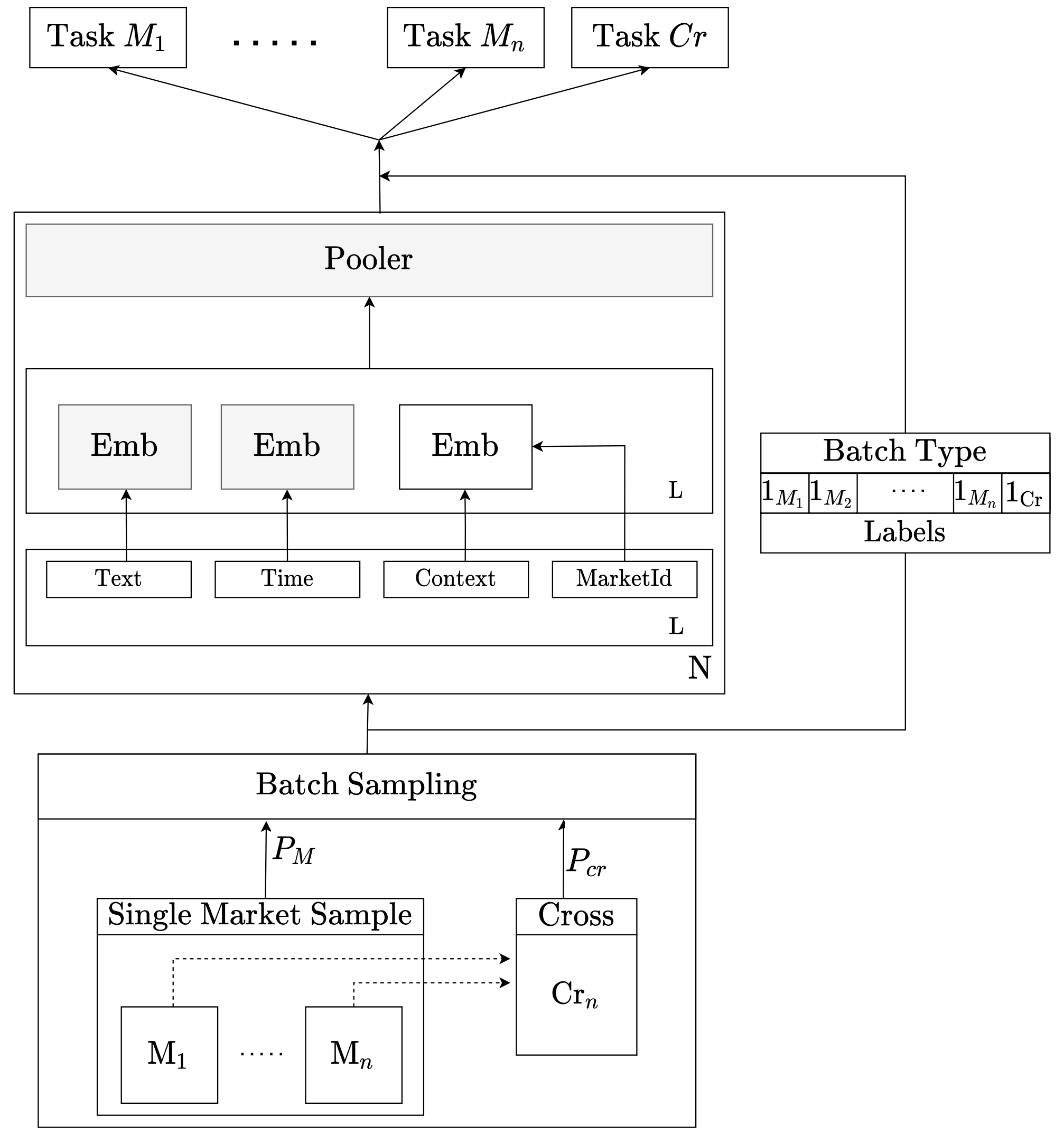

Figure: Multi-task setup for SYSML

We extend the social media user representations built by Andrews and Bishop (2019) by adding heterogeneous graph representations and multi-task learning. Text embeddings are generated using a character and byte pair level CNN, while time embeddings are one-hot day of week embeddings. An episode $e = \{(t_i, \tau_i, c_i) | 1 \leq i \leq L\}$ with t_i, τ_i, c_i representing text, time, and context respectively.

Implementation Details

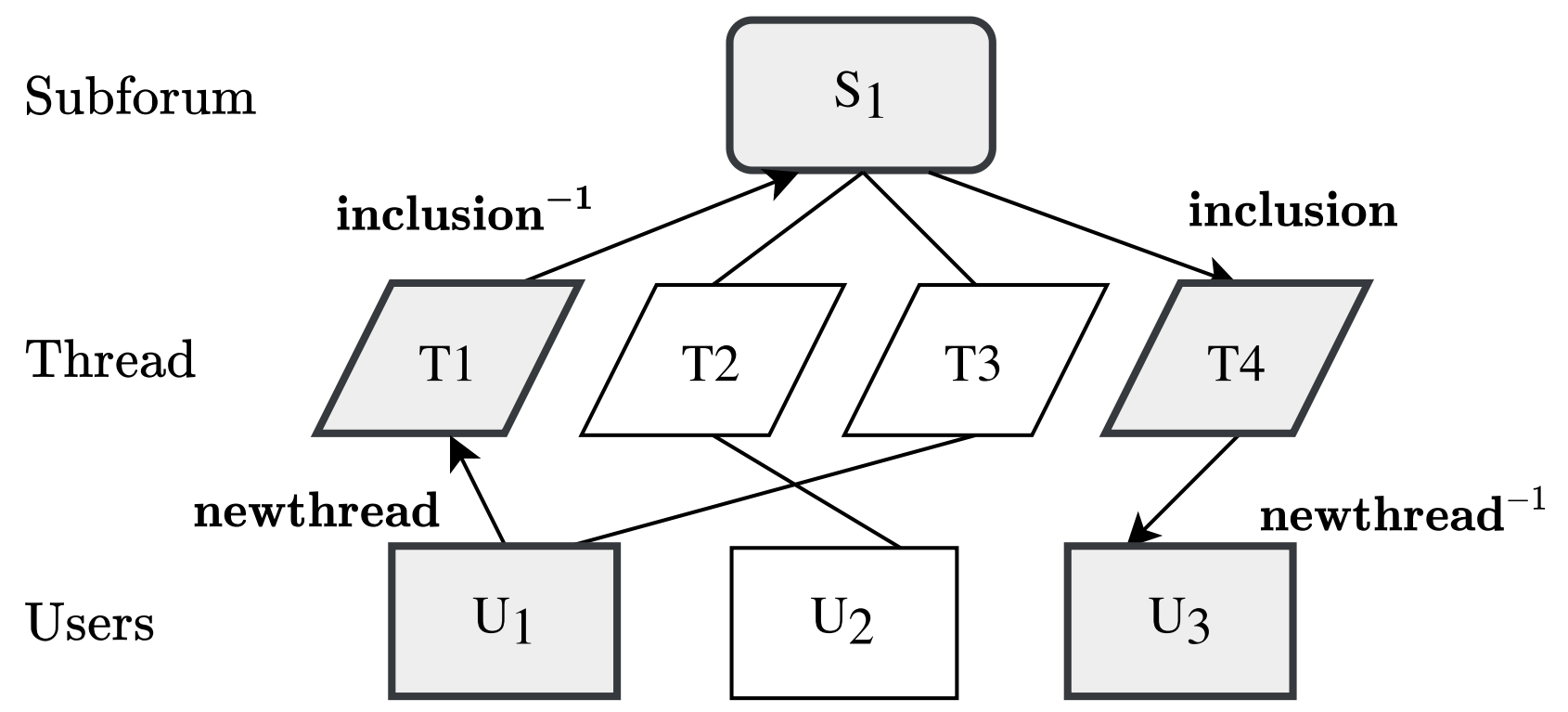

Figure: An instance of meta-path 'UTSTU' in a subgraph of the forum graph.

We use the forum post **heterogeneous graph** to generate conntext embeddings for episodes. Metapath2vec generates embeddings by maximizing the probability of heterogeneous neighbourhoods, normalizing it across typed contexts. To fully capture the semantic relationships in the heterogeneous graph, we use seven meta-path schemes: UPTSTPU, UTSTPU, UPTSTU, UTSTU, UPTPU, UPTU, and UTPU. Thus, the learned embeddings will preserve the semantic relationships between each subforum, included posts as well as relevant users (authors).

$$\mathcal{L} = \underset{\mathcal{T}_i \sim \mathcal{T},\, \mathcal{E} \sim \mathcal{T}_i}{\mathbb{E}} [\mathcal{L}_i(\mathcal{E})]$$

We use authorship attribution as the metric learning task for each market. The first step in the multi-task approach is to choose a market ($\mathcal{T}_M$) or cross-market ($\mathcal{T}_{cr}$) metric learning task $\mathcal{T}_i \sim \mathcal{T} = \{\mathcal{T}_M, \mathcal{T}_{cr}\}$. Following this, a batch of N episodes $\mathcal{E} \sim \mathcal{T}_i$ is sampled from the corresponding task. The embedding module generates the embedding for each episode $f_\theta^N : \mathcal{E} \to \mathbb{R}^{N \times E}$. Finally, the task-specific metric learning layer $g_\phi^{\mathcal{T}_i}$ is selected and a task-specific loss is backpropagated through the network.

Results

Market	Train Posts	Test Posts	#Users train	#Users test
SR (Silkroad)	379382	381959	6585	8865
SR2 (Silkroad 2)	373905	380779	5346	6580
BMR (Black Market Reloaded)	30083	30474	855	931
Agora	175978	179482	3115	4209

Table: Dataset Statistics for Darkweb Markets.

For cross-market users, we manually annotate a small set ($\sim$ 100) of users as migrant/non-migrants. The data is sourced from the work of Munksgaard and Demant (2016) on the analysis of politics on the darknet forums.

We evaluated our method using retrieval-based metrics over the embeddings generated by each approach. With all episodes $E = \{e_1, \ldots e_n\}$ and $Q = \{q_1, q_2, \ldots q_\kappa\} \subset E$ be a sampled subset. We computed the cosine similarity of the query episode embeddings with all episodes. Let $R_i = \langle r_{i1}, r_{i2}, \ldots r_{in} \rangle$ denote the list of episodes in E ordered by their cosine similarity with episode q_i (excluding itself). We compute the Mean Retrieval Rank (MRR) and Recall@k (R@k) for the retrieved episodes.

Method	BMR		Agora		SR2		SR	
	MRR	R@10	MRR	R@10	MRR	R@10	MRR	R@10
Shrestha et al. (2017) (CNN)	0.07	0.165	0.126	0.214	0.082	0.131	0.036	0.073
+ time + context	0.235	0.413	0.152	0.263	0.118	0.21	0.094	0.178
+ time + context + transformer pooling	0.219	0.409	0.146	0.266	0.117	0.207	0.113	0.205
Andrews and Bishop (2019) (IUR)								
mean pooling	0.223	0.408	0.114	0.218	0.126	0.223	0.109	0.19
transformer pooling	0.283	0.477	0.127	0.234	*0.13*	*0.229*	0.118	0.204
SYSML (single)	*0.32*	*0.533*	*0.152*	*0.279*	0.123	0.21	*0.157*	*0.266*
- graph context	0.265	0.454	0.144	0.251	0.089	0.15	0.049	0.094
-graph context - time	0.277	0.477	0.123	0.198	0.079	0.131	0.04	0.08
SYSML (multitask)	**0.438**	**0.642**	**0.303**	**0.466**	**0.304**	**0.464**	**0.227**	**0.363**
- graph context	0.396	0.602	**0.308**	**0.469**	0.293	0.442	0.214	0.347
- graph context - time	0.366	0.575	0.251	0.364	0.236	0.358	0.167	0.28

Table: Best performing results in **bold**. Best performing single-task results in *italics*. All $\sigma_{MRR} < 0.02$, $\sigma_{R@10} < 0.03$, For all metrics, higher is better. Results suggest single-task performance largely outperforms the state-of-the-art Shrestha et al. (2017); Andrews and Bishop (2019), while our novel multi-task cross-market setup offers a substantive lift (**up to 2.5X on MRR and 2X on R@10**) over single-task performance. Results shown for 5 posts per episode

Analysis

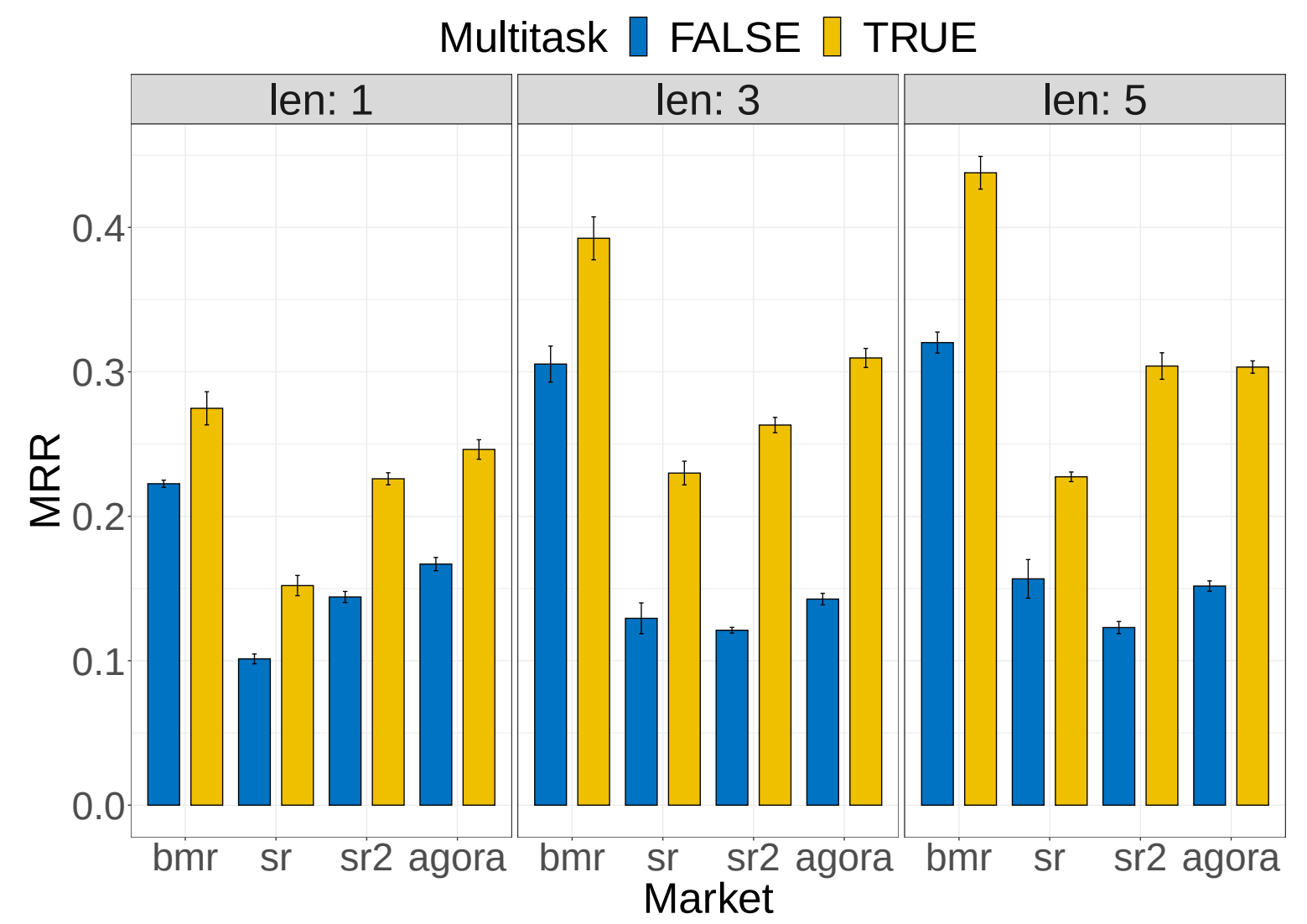

Figure: Multitask learning consistently and significantly ($p < 0.01$) improves performance across all markets and all episode lengths.

Author	Word Importance
HSI-Sec	. . . 2 cents, anyway . . . PGP Key Fingerprint = ... PGP Key Fingerprint . . . . security is NOT retroactive. . . . Is it possible for a gpg key to request that)
LSI-1	Check out the link in my sig . . . [IMAGE alt=8)] Hey dude, just run a search . . . I can not help much . . . Im sure if you ask . . . German , he may be willing to lend a hand. Good luck freind [IMAGE alt=8)]
LSI-2	[QUOTE] From: . . . Just my opinion, I 've done just about everything, . . . IMAGE alt=8)] couldnt agree more [QUOTE] From : . . . strangely enough, when im in . . . I too jabber meaningless jibberish . . .
	Negative 🟥 Neutral □ Positive 🟩

Table: Integrated Gradient based attribution of posts. H=High, L=Low, SI=stylometric identifiablity. Authors with low stylometric identifiability used classical adversarial strategies (imitation-based and obfuscation based) and avoid detection.

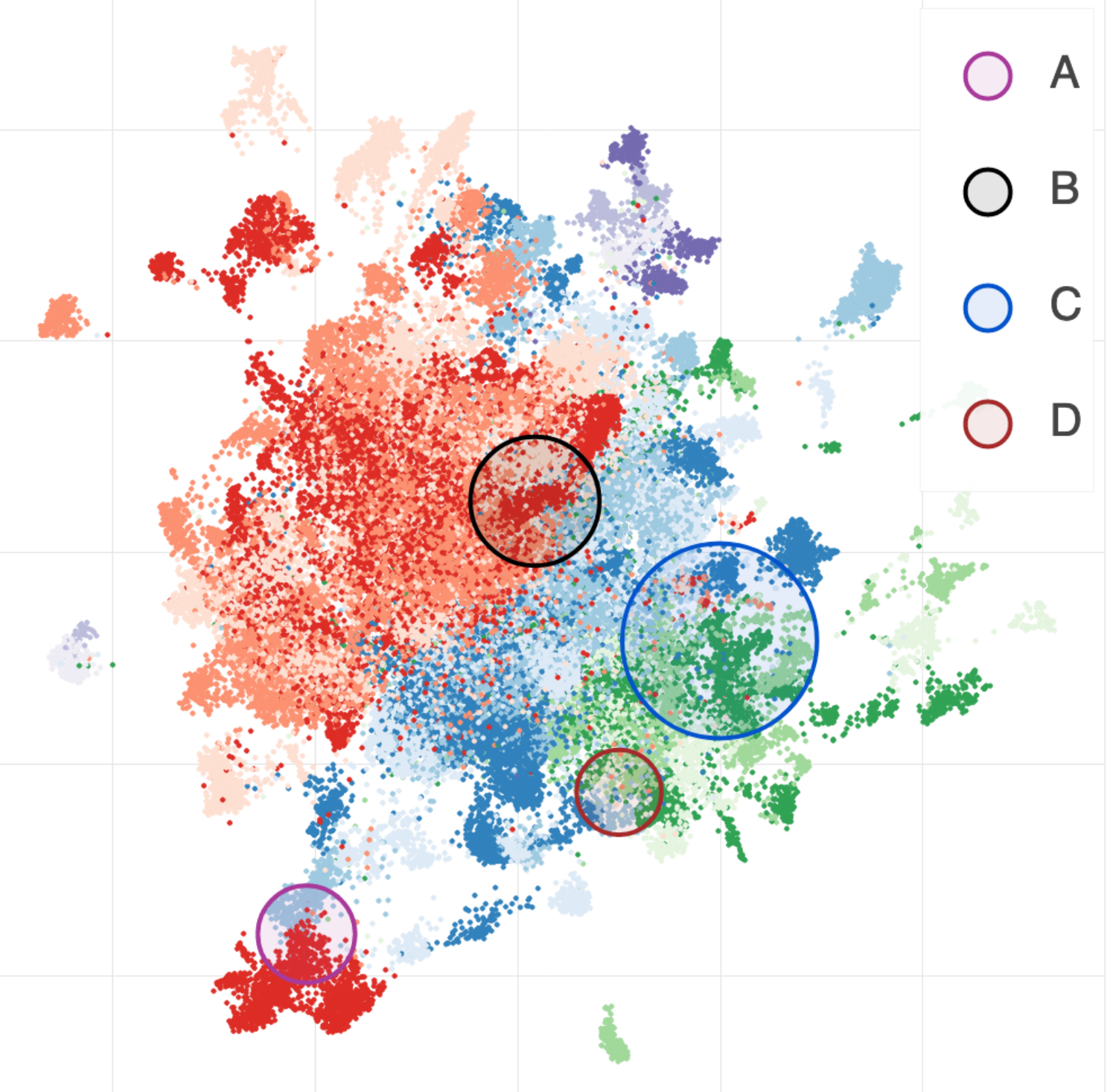

Figure: UMAP visualization of cross dataset embeddings for the top 200 authors, one hue per market. Circles denote the same user in two different markets. The circles in the figure highlight the episodes of the top four pairs of cross-market matched users (top candidate sybils).

References

Nicholas Andrews and Marcus Bishop. 2019. Learning invariant representations of social media users. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1684–1695.

Rasmus Munksgaard and Jakob Demant. 2016. Mixing politics and crime–the prevalence and decline of political discourse on the cryptomarket. *International Journal of Drug Policy*, 35:77–83.

Prasha Shrestha, Sebastian Sierra, Fabio A González, Manuel Montes, Paolo Rosso, and Thamar Solorio. 2017. Convolutional neural networks for authorship attribution of short texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 669–674.

Work funded by NSF grants SES-1949037, CCF-2028944, and OAC-2018627. All content represents the opinion of the authors, and is not necessarily endorsed by their sponsors. Author contact information: maneriker.1@osu.edu