IDTraffickers: An Authorship Attribution Dataset to link and connect Potential Human-Trafficking Operations on Text Escort Advertisements

Vageesh Saxena
v.saxena@maastrichtuniversity.nl

Benjamin Bashpole
bashpole@idtraffickers.com

Gijs Van Dijck
gijs.vandijck@maastrichtuniversity.nl

Jerry Spanakis
jerry.spanakis@maastrichtuniversity.nl

EMNLP 2023

Problem Statement: Can Authorship Attribution approaches be used to link and connect potential Human Trafficking (HT) advertisements?

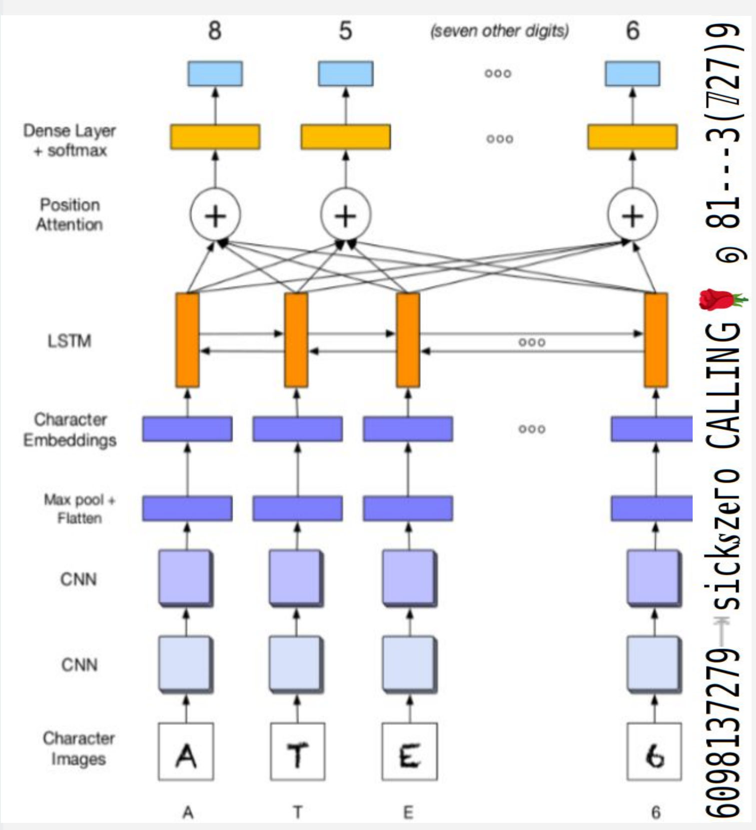
- HT indicators studied in ads linked to individuals/organizations.
- Law Enforcement (LEA) connect ads using phone numbers, images, and emails.
- Studies show HT involvement in escort market.
- Only 37% of Backpage escort ads had these features.

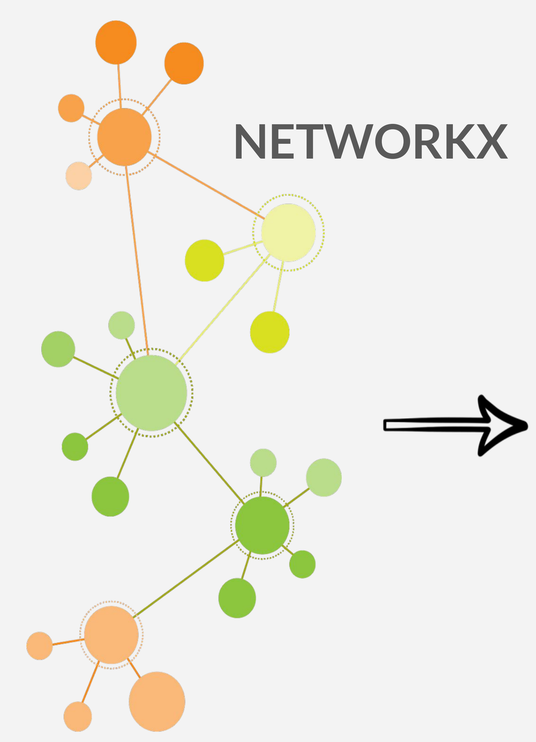
Our Contributions:

- Novel authorship attribution dataset with potential HT instances to analyze unique writing styles.
- Establishing authorship identification benchmark as a closed-set classification task.
- Utilizing trained representations for identifying potential aliases in open-set ranking task.

(i) IDTraffickers: An Authorship Attribution dataset with advertisements from Backpage Escort Market

- **Input**: Text Advertisement
- **Labels:** phone numbers
- **Output:** Phone number extraction (Classification) + Network Analysis (NetworkX) = Vendor Labels
- **Dataset:** 100k human annotated advertisements from DARPA dataset
- **Evaluation:** Lev Accuracy, Perfect Accuracy, Digit Accuracy, and Consistency





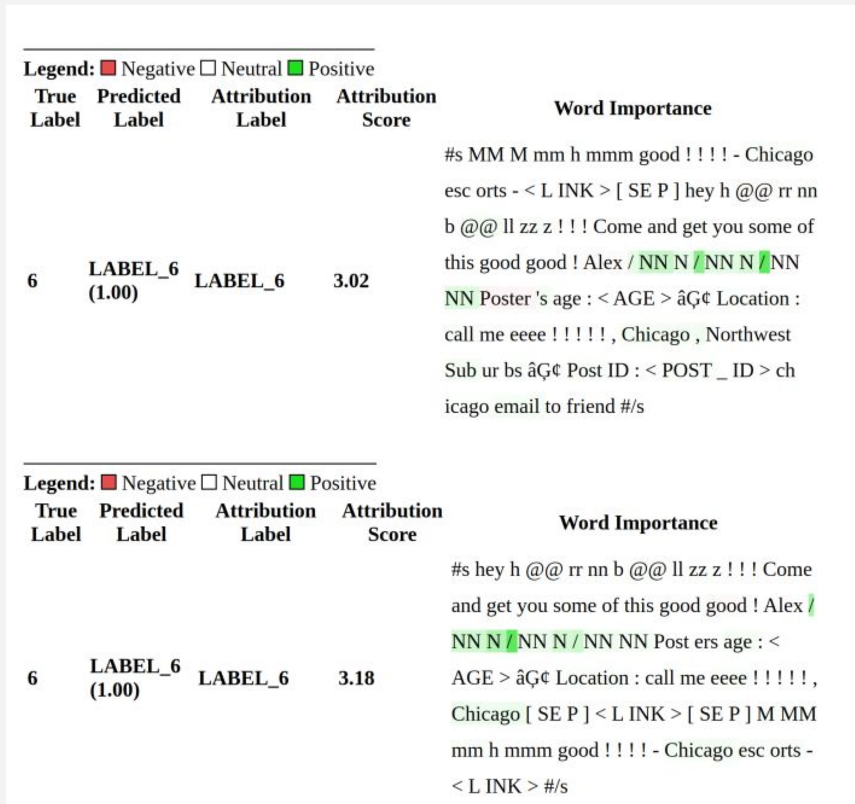
Geography	Advertisements	Vendors
East	24,000	5,029
West	22,556	2,576
North	3,124	254
South	27,871	2,291
Central	21,124	2,928
Overall	87,595	5,244

(ii) Authorship Identification Task: Identifying HT vendors through a closed-set classification task

- **Input:** Text Sequence (Title + Description)
- **Labels:** Vendor IDs
- **Dataset:** IDTraffickers
- **Evaluation:** Balance Accuracy, Micro-F1, Weighted-F1, and Macro-F1

Models	Acc.	Micro-F1	Weighted-F1	Macro-F1
Distilled Models				
BERT	0.9110	0.9147	0.9143	0.8467
RoBERTa	0.9199	0.9230	0.9229	0.8603
GPT2	0.9132	0.9172	0.9166	0.8500
Smaller Models				
ALBERT	0.7832	0.7891	0.7925	0.6596
DeBERTa-v3	0.8703	0.8757	0.8756	0.7825
T5	0.9157	0.9192	0.9190	0.8535
Contrastive Learning Models				
miniLM	0.8888	0.8934	0.8935	0.8101
DeCLUTR	**0.9230**	**0.9261**	**0.9259**	**0.8656**
Style-Emb	0.8887	0.8936	0.8932	0.8112
HT Language Model				
LM-Classifier	0.9294	0.9317	0.9316	0.8726

Table 3: Balanced Accuracy, Micro-F1, Weighted-F1, and Macro-F1 performances of the transformers-based classifiers on the author identification task.



(iii) Authorship Verification Task: Verifying potential aliases using open-set ranking task

- **Input:** Pre-trained representations of text advertisements from the trained classifier
- **Model:** DeCLUTR-small and Style-Embedding classifiers
- **Similarity-Search:** FAISS (clustering of dense vectors)
- **Red color:** performance before training
- **Green color:** performance after training
- **Evaluation:** Precision@K, Recall@K, MAP@K, and R-Precision

K	@1	@3	@5	@10	@20	@25	@50	@100	@X
Precision@K									
Style	0.0442 ± 0.20	0.0410 ± 0.16	0.0391 ± 0.15	0.0366 ± 0.13	0.0329 ± 0.11	0.0319 ± 0.10	0.0270 ± 0.08	0.0227 ± 0.07	-
DeCLUTR	0.3198 ± 0.46	0.2883 ± 0.39	0.2671 ± 0.36	0.2278 ± 0.32	0.1837 ± 0.27	0.1693 ± 0.26	0.1277 ± 0.21	0.0893 ± 0.15	-
Style	0.9616 ± 0.19	0.9437 ± 0.19	0.9124 ± 0.21	0.8175 ± 0.27	0.6818 ± 0.33	0.6328 ± 0.35	0.4815 ± 0.36	0.3551 ± 0.36	-
DeCLUTR	0.9672 ± 0.17	0.9532 ± 0.17	0.9221 ± 0.19	0.8253 ± 0.26	0.6868 ± 0.33	0.6367 ± 0.34	0.4835 ± 0.36	0.3561 ± 0.36	-
Recall@K									
Style	0.0023 ± 0.01	0.0063 ± 0.04	0.0091 ± 0.05	0.0146 ± 0.07	0.0233 ± 0.09	0.0269 ± 0.10	0.0394 ± 0.12	0.0580 ± 0.15	-
DeCLUTR	0.0242 ± 0.06	0.0567 ± 0.12	0.0792 ± 0.16	0.1136 ± 0.20	0.1539 ± 0.24	0.1676 ± 0.25	0.2122 ± 0.29	0.2590 ± 0.31	-
Style	0.0828 ± 0.09	0.2348 ± 0.24	0.3485 ± 0.32	0.5092 ± 0.37	0.6552 ± 0.37	0.6945 ± 0.36	0.7909 ± 0.32	0.8600 ± 0.27	-
DeCLUTR	0.0836 ± 0.09	0.2397 ± 0.25	0.3563 ± 0.32	0.5192 ± 0.37	0.6653 ± 0.37	0.7041 ± 0.36	0.7988 ± 0.32	0.8664 ± 0.27	-
MAP@K									
Style	0.0442 ± 0.20	0.0562 ± 0.21	0.0598 ± 0.21	0.0640 ± 0.21	0.0673 ± 0.21	0.0681 ± 0.21	0.0700 ± 0.21	0.0712 ± 0.21	-
DeCLUTR	0.3198 ± 0.46	0.3587 ± 0.45	0.3681 ± 0.45	0.3750 ± 0.44	0.3794 ± 0.44	0.3803 ± 0.44	0.3823 ± 0.44	0.3833 ± 0.44	-
Style	0.9616 ± 0.19	0.9687 ± 0.16	0.9698 ± 0.15	0.9706 ± 0.15	0.9709 ± 0.14	0.9710 ± 0.14	0.9710 ± 0.14	0.9710 ± 0.14	-
DeCLUTR	0.9672 ± 0.17	0.9735 ± 0.14	0.9746 ± 0.14	0.9752 ± 0.13	0.9755 ± 0.13	0.9755 ± 0.13	0.9756 ± 0.13	0.9756 ± 0.13	-
R-Precision@X									
Style	-	-	-	-	-	-	-	-	0.0199 ± 0.07
DeCLUTR	-	-	-	-	-	-	-	-	0.1641 ± 0.23
Style	-	-	-	-	-	-	-	-	0.8601 ± 0.22
DeCLUTR	-	-	-	-	-	-	-	-	0.8850 ± 0.20

Table 2: Precision@K, Recall@K, MAP@K, and R-Precision@X scores for the DeCLUTR and Style-Embedding models before and after being trained on the IDTraffickers dataset

Data Insights

- % of Punctuations, emojis, white spaces, and random characters:
  - 47% in IDTraffickers.
  - 10.6% in PAN2023
  - 12.4% in Reddit dataset

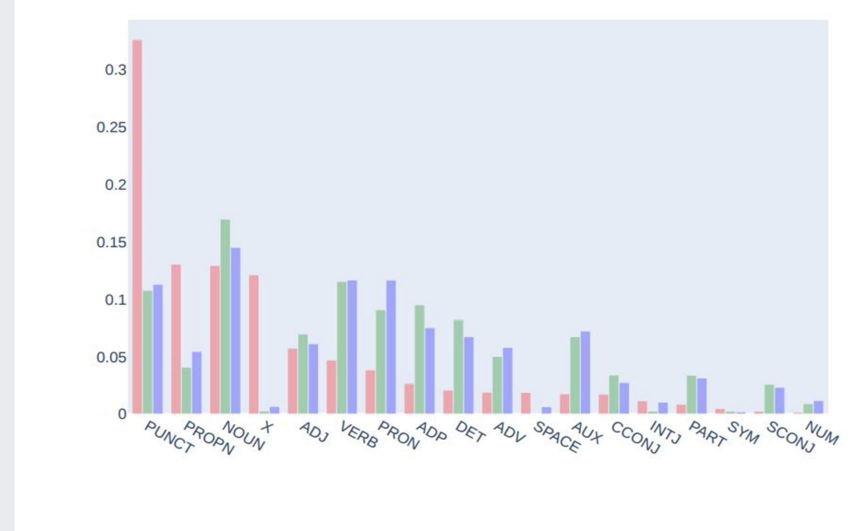


Figure 3: **POS-distribution:** Normalized POS-distribution for IDtraffickers, PAN2023, and Reddit-Conversations datasets.

- Wikifiability refers to the presence of entities with corresponding wikipedia mentions.
- IDTraffickers has higher wikifibility than PAN2023 and Reddit-Conv datasets.

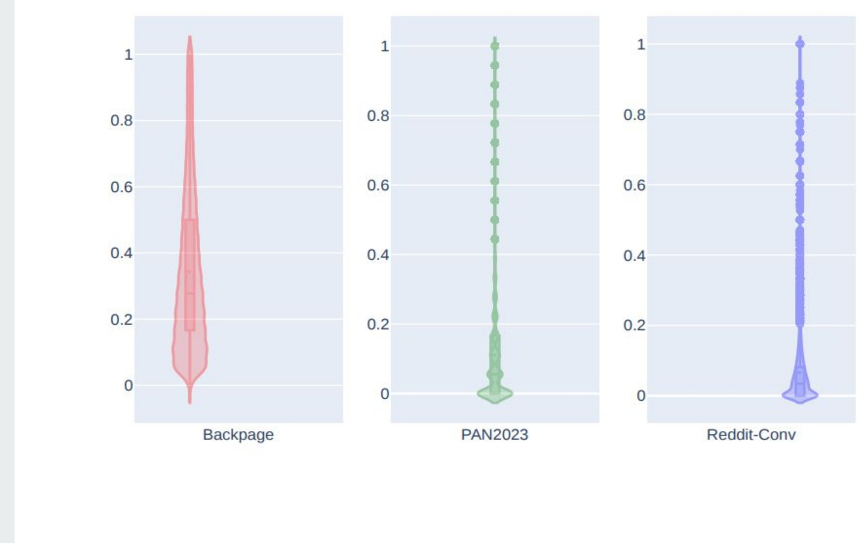


Figure 4: **Wikifiability:** No. of entities per advertisement with Wikipedia mentions in the IDtraffickers, PAN2023, and Reddit-Conversations datasets.

- Majority of recognized entities are primarily related to locations, names, and organizations.
- This finding aligns with the nature of ads, as they often include posting locations, escort names, and nearby landmarks.

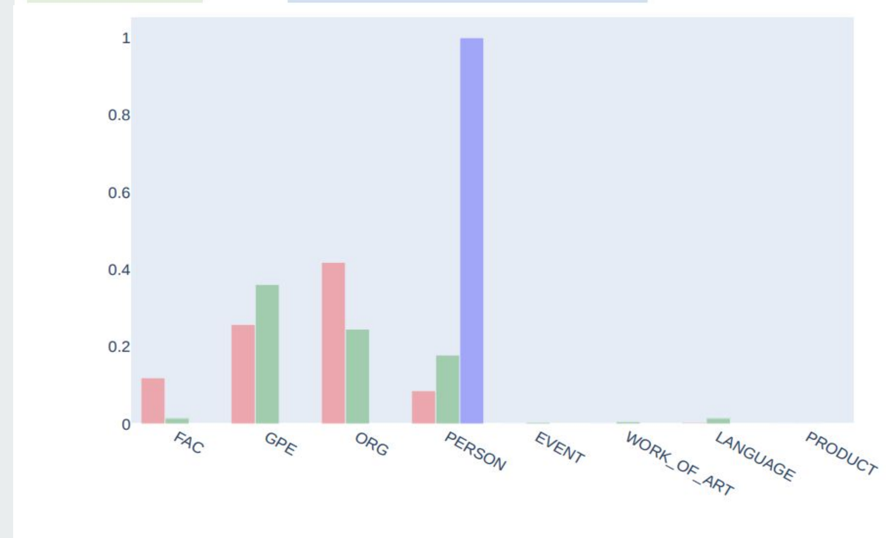


Figure 5: **Wiki-entities-distribution:** Extracted entities from the wikification of IDtraffickers, PAN2023, and Reddit-Conversations datasets.

Summary

Key Findings :

- Trained CNN-LSTM-CRF classifier effectively generates ground truth.
- The DeCLUTR classifier identifies unique writing styles with high accuracy.
- Trained classifiers can be used to identify potential aliases through ranking task.

Results:

- CNN-CRF-CRF classifier
  - Lev Accuracy: 0.9986
  - Perfect Accuracy: 0.9892
  - Digit Accuracy: 0.9950
  - Consistency: 0.9899
- Author Verification / Classification task
  - DeCLUTR-small model with Macro-F1 of 0.8656
- Author Identification / Ranking Task
  - Supervised pre-training helps
  - R-Precision of 0.8850 with a std. of 0.20
  - Outperforms the existing SOTA

Limitations :

- Vendors may not indicate all operable phone numbers.
- Lack of ground truth (Human Trafficking instances)
- Larger Architectures may yield better performance
- Lack of similar datasets to evaluate zero-shot performance
- Some advertisements don't have text description
- LLMs can be used to automatically generate advertisements
- Explainability is required amongst LEA to establish trust
- Misuse of such approaches can harm individuals

arXiv GitHub

SCAN ME SCAN ME