VendorLink: An NLP approach for Identifying & Linking Vendor Migrants & Potential Aliases on Darknet Markets

Vageesh Saxena
Law & Tech Lab
Maastricht University
v.saxena@maastrichtuniversity.nl

Nils Rethmeier
Speech and Language Technology Lab
DFKI, Berlin
nils.rethmeier@dfki.de

Gijs Van Dijck
Law & Tech Lab
Maastricht University
gijs.vandijck@maastrichtuniversity.nl

Gerasimos Spanakis
Law & Tech Lab
Maastricht University
jerry.spanakis@maastrichtuniversity.nl

Abstract

The anonymity on the Darknet allows vendors to stay undetected by using multiple vendor aliases or frequently migrating between markets. Consequently, illegal markets and their connections are challenging to uncover on the Darknet. To identify relationships between illegal markets and their vendors, we propose VendorLink, an NLP-based approach that examines writing patterns to verify, identify, and link unique vendor accounts across text advertisements (ads) on seven public Darknet markets. In contrast to existing literature, VendorLink utilizes the strength of supervised pre-training to perform closed-set vendor verification, open-set vendor identification, and low-resource market adaption tasks. Through VendorLink, we uncover (i) 15 migrants and 71 potential aliases in the Alphabay-Dreams-Silk dataset, (ii) 17 migrants and 3 potential aliases in the Valhalla-Berlusconi dataset, and (iii) 75 migrants and 10 potential aliases in the Traderoute-Agora dataset. Altogether, our approach can help Law Enforcement Agencies (LEA) make more informed decisions by verifying and identifying migrating vendors and their potential aliases on existing and Low-Resource (LR) emerging Darknet markets. [1]

1 Introduction

Conventional search engines index surface-web websites that only constitute 4% of the entire internet (Georgiev, 2021). The remaining comprises 90% Deep Web (not indexed) and 6% Darknet, which uses advanced anonymity enhancing protocols (Georgiev, 2021). While the former serves legitimate purposes requiring anonymity, the latter is also used for illegal activities such as financial fraud (ENISA, 2018), child exploitation (Bruggen and Blokland, 2021), and trading of illicit weapons (Weimann, 2016; Persi Paoli et al., 2017), prohibited drugs, and chemicals (Kruithof et al., 2016).

Given the Darknet's scope, size, and anonymity, it is difficult for LEA to uncover connections between illegal marketplaces (Vogt, 2017). While manual detection of such connections is a time-consuming and resource-extensive process, the recent success of online scrapers (Fu et al., 2010; Hayes et al., 2018) and monitoring systems (Schäfer et al., 2019; Godawatte et al., 2019) has enabled researchers and LEA to analyze (Easttom, 2018; Faizan and Khan, 2019; Goodison et al., 2019; Davies, 2020) and automatically identify (Al Nabki et al., 2017; Ghosh et al., 2017; Ubbink et al., 2019; He et al., 2019) Darknet contents. This research proposes a vendor verification and identification approach to help LEA make better decisions by linking vendors, offloading manual labor, and generating similarity-based analyses. In contrast to the existing Darknet literature (He et al., 2015; Ekambaranathan, 2018; Tai et al., 2019; Kumar et al., 2020; Manolache et al., 2022), VendorLink, as illustrated in Figure 1, emphasizes the following contributions to the problem of verifying and identifying vendors on Darknet markets:

(i) Closed-Set Vendor Verification Task: Due to limited resources, LEA prioritizes investigating Darknet vendors based on the size and nature of their trade. Thus, Darknet vendors often distribute their business across multiple markets to stay undetected. Likewise, some vendors relocate and resume their business in other markets after a market seizes (Booij et al., 2021). We refer to these

[1] Our code implementation is publicly available at https://github.com/maastrichtlawtech/VendorLink.git

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 8619–8639

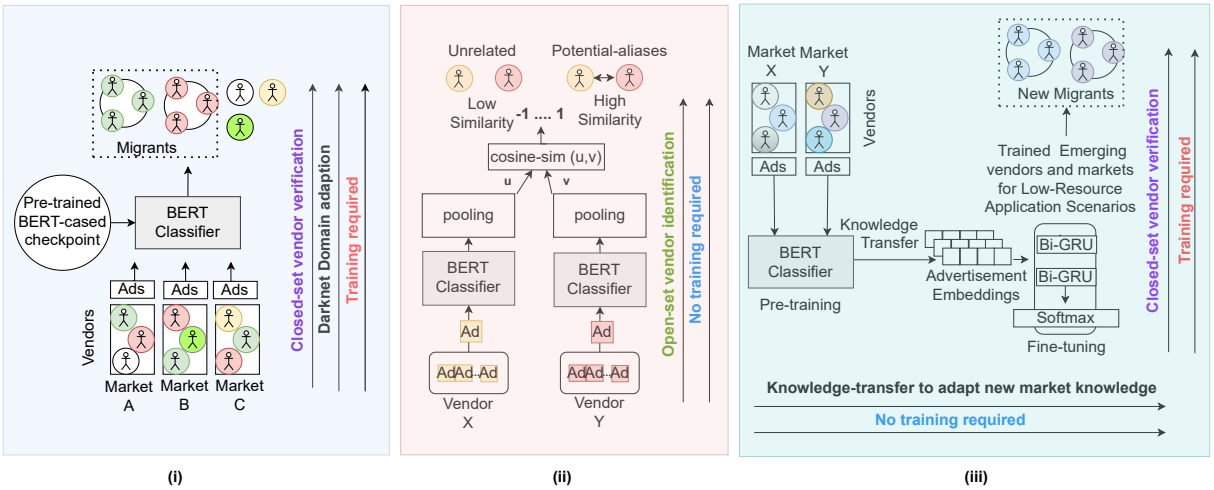


Figure 1: **(i) Closed-Set Vendor Verification Task:** A supervised pre-training task that performs classification using a BERT-cased classifier in a closed-set environment to verify unique vendor migrants across existing markets, **(ii) Open-Set Vendor Identification Task:** A text-similarity task in an open-set environment that utilizes style representations from the established BERT-cased classifier to verify known vendors and identify potential-aliases, **(iii) Low-Resource Market Adaptation Task:** A knowledge-transfer task in a closed-set environment to adapt new market knowledge and verify migrants across Low-Resource (LR) emerging markets.

migrating vendors as *migrants* for brevity. Unfortunately, this movement prevents LEA from correctly estimating the size of a vendor's operations. To aid LEA, we perform supervised pre-training by conducting multiclass classification in a closed-set environment (Zhou et al., 2021a) to analyze different writing styles in text ads and classify vendor migrants to unique vendor accounts across three Darknet markets. Moreover, researchers have observed a significant difference in language structure between Darknet and Surface net websites (Choshen et al., 2019; Jin et al., 2022). Since most contextualized models are trained on surface web data, the supervised pre-training step allows our model to adapt to the Darknet market domain knowledge.

(ii) Open-set Vendor Identification Task: Darknet vendors often create aliases and work in groups to distribute their products across multiple markets, allowing them to expand their business without being detected by LEA. Moreover, given the scope and anonymity of the Darknet, manually linking these profiles is infeasible. Hundreds of new markets and vendors emerge daily on the Darknet. While the existing literature has established impressive performance on the vendor verification task, any trained classifier will fail during inference to encounter unknown vendors from emerging markets in real-to-close-world scenarios. Therefore, in this research, we use the style representations from the pre-trained classifier to compute the cosine similarity between the text ads to verify existing vendors and identify potential aliases and unknown vendors in an open-set environment (Zhou et al., 2021a).

(iii) Low-Resource Market Adaptation task: While research has demonstrated impressive performance for the Darknet's vendor verification task (Kumar et al., 2020; Manolache et al., 2022), high computational and storage requirements pose a significant challenge to LEA. Furthermore, with the exponential growth of Darknet markets and vendors with new content every year, there is a dire need for systems that can verify existing vendors from a known database and simultaneously adapt to new market knowledge from emerging vendors and markets. After all, not all LEA have the resources to train computationally expensive models from scratch. Therefore, this experiment investigates our classifier's capability to benefit transfer learning in a low-resource setting (Ruder et al., 2019) for adapting new market knowledge and performing closed-set vendor verification on emerging (upcoming) vendors and markets. Finally, we evaluate the influence of knowledge transfer on our trained low-resource model against the zero-shot (Srivastava et al., 2018) and transformers-based baselines.

2 Related Research

Vendor Verification - a supervised Authorship Attribution (AA) task: Researchers previously have utilized various NLP (Ekambaranathan, 2018;

Tai et al., 2019; Manolache et al., 2022) and computer vision (Wang et al., 2018; He et al., 2015) techniques to identify and link vendors across Darknet markets. For example, in their research, Zhang et al. (2019) proposed uStyle-uID to leverage writing and photography styles to identify vendors in drug trafficking markets. Similarly, Kumar et al. (2020) proposed exploiting the multi-view learning paradigm and domain-specific knowledge to improve the cross-domain performance with both stylometric and location representation.

The Darknet ads consist of a product title and description, vendor name, price of the product, and occasionally some meta-data and images. While most of these details were enclosed in the ad's description, manual extraction of these features requires considerable labeling efforts. Therefore, we emphasize our research towards an end-to-end approach that only expects the advertisement's title and description to analyze the writing patterns for vendor verification and identification. Furthermore, since we perform multi-class classification over the text sequences of Darknet ads, we consider our approach similar to the AA task in NLP.

With the advances in NLP, there has been considerable research into the field of AA that has demonstrated the success of TFIDF-based clustering and classification techniques (Agarwal et al., 2019; İzzet Bozkurt et al., 2007), CNNs (Rhodes, 2015; Shrestha et al., 2017), RNNs (Zhao et al., 2018; Jafariakinabad et al., 2019; Gupta et al., 2019), and transformers architectures (Fabien et al., 2020; Ordoñez et al., 2020; Uchendu et al., 2020a). Moreover, researchers have also observed a significant difference in language structure between Darknet and Surface net websites (Choshen et al., 2019; Jin et al., 2022). Therefore, exploring the application of authorship tasks on the Darknet language is crucial.

Vendor Identification; A Text Similarity task: Text-similarity techniques are not new to the researchers in the field of AA (Sapkota et al., 2013; Castro Castro et al., 2015; Rexha et al., 2018; Boenninghoff et al., 2019). However, with the recent success of transformers (Reimers and Gurevych, 2019a; Yang et al., 2019b; Jiang et al., 2022), researchers are now investigating the application of semantically meaningful representations for paraphrasing detection (Timmer et al., 2021; Olney, 2021; Ko and Choi, 2020), text summarization (Miller, 2019; Cai et al., 2022), semantic parsing (Ge et al., 2019; Ferraro and Suominen, 2020), question answering (Yang et al., 2019a; Vold and Conrad, 2021; Louis and Spanakis, 2021), and AA (Fabien et al., 2020; Li et al., 2020; Custódio and Paraboni, 2021; Uchendu et al., 2020b).

The recent developments in style representations (Hay et al., 2020; Zhu and Jurgens, 2021) have revealed a promising avenue to explore for the authorship verification task. In their research, Wegmann et al. (2022) discovered that the success of these representations comes from their ability to represent style by latching on to spurious content correlations. Moreover, the authors suggest using content control in a contrastive setup to represent style better in a way that is more independent from content. In this research, we utilize a similar approach to extract the style representations from the advertisements of darknet vendors by passing it through a Transformer-based classifier pre-trained for a closed-set vendor verification task. Next, we use these representations to compute text similarity (cosine similarity) in the advertisements of different vendors. Despite our acknowledgment of the promises of using content control on style representations, this research focuses on establishing a baseline on Darknet markets. That being said, we intend to experiment with content control in our future experiments.

Knowledge Adaption; A Transfer Learning task: In their research, Ruder (2019) introduced transfer learning to extract knowledge from a source setting and transfer it to a target setting. Since then, many researchers have investigated the successful application of transfer learning on the cross-domain and topic AA task (Sapkota et al., 2014; Barlas and Stamatatos, 2021). Similar to the experiments in (Devlin et al., 2019; Horne et al., 2020), this work proposes utilizing knowledge transfer to adapt new market knowledge from the emerging Darknet vendors and markets. The transfer is applied using pre-trained style representations to train a computationally efficient BiGRU classifier for the closed-set vendor verification task.

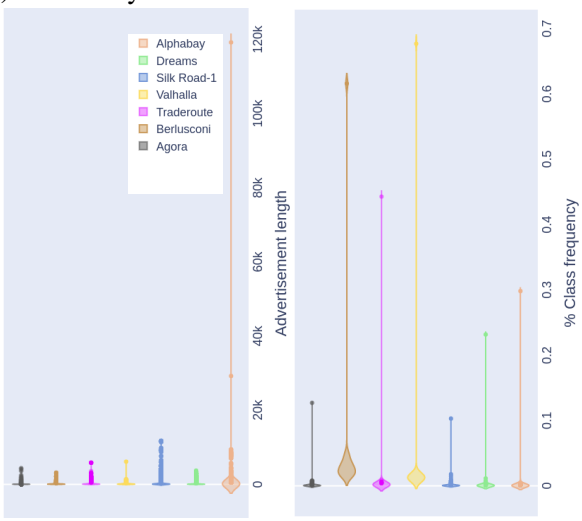
3 Datasets

Many researchers have conducted similar experiments on scraped data from active Darknet markets. However, since law enforcement has seized and shut down these markets now, we could not reproduce the results nor get access to their data. Therefore, for reproducibility and future research pur-

Task	Dataset	Ads.	Vendors
Baseline / Supervised Pre-Training	Alphabay	100,429	1,457
	Dreams	93,586	1,422
	Silk Road-1	78,681	1,392
	Alphabay-Dreams-Silk	**272,696**	**4,271**
Low-Resource Supervised Market Adaption	Valhalla	2,175	110
	Berlusconi	1,437	84
	Valhalla-Berlusconi	**3,612**	**194**
High-Resource Supervised Market Adaption	Traderoute	19,952	612
	Agora	109,644	3,187
	Traderoute-Agora	**129,586**	**3,799**

Table 1: Number of unique ads and vendor accounts per market.

poses, we conduct our analyses on public datasets from Alphabay (Van Wegberg et al., 2018; Baravalle and Lee, 2018; CMU, 2017-18a), Dreams, Traderoute, Valhalla, and Berlusconi (Carr et al., 2019; CMU, 2017-18b), Agora (Branwen et al., 2015), and Silk Road (Christin, 2013; CMU, 2012-13) non-anonymous markets.[2]



Dataset	Min	Median	Max
Alphabay	3	97	119.229k
Dreams	2	67	3,713
Silk Road-1	3	61	11.742k
Berlusconi	5	95	3,161
Traderoute	3	100	5,835
Valhalla	3	88	6,119
Agora	3	40	4,367

(a)

Dataset	Min	Median	Max
Alphabay	199.678u	348.480u	0.2981
Dreams	213.678u	384.623u	0.2316
Silk Road-1	167.515u	385.285u	0.1029
Berlusconi	0.0160	0.0271	0.6158
Traderoute	0.0001	0.0002	0.4427
Valhalla	0.0009	0.0124	0.6767
Agora	182.333u	392.017u	0.1270

(b)

Figure 2: Distribution of (a) Token length per advertisement (b) Number of ads per vendor.

Preprocessing: Figure 2(a) demonstrates the distribution of the number of tokens for all the input ads in our datasets. In a violin plot, the probability distribution is maximum around the median, and Table 2(a) shows that the median for our chosen datasets is between 40 and 100. Therefore, to compare other baseline classifiers and transformers-based models fairly, we truncate our ads to the first 512 tokens. On the other hand, figure 2(b) demonstrates a class imbalance in the number of ads per vendor account in our datasets. As can be seen, some markets are more imbalanced than others. Therefore, in contrast to earlier research emphasising the performance of the trained models on accuracy and micro-F1, we also evaluate our trained models on macro-F1, which weighs all classes equally.

Table 1 illustrates the number of unique ads (input sequences) and vendor accounts per market.[3] First, we merge the title and description of the ads using the "[SEP]" token to form the input sequences. Then, we drop all the duplicate ads for every vendor in our dataset. Most ads are in English, with a few exceptions where the vendors use multiple languages. We reason that the noise in the data roughly represents the unique writing style of individual vendors. For example, we found that the vendor "CaliforniaDreams420" refers to medicines as "medi...", "SAPIOWAX" uses multiple "-" for newline, and "QualityKing" only uses uppercase letters in its ads. Therefore, any cleaning and processing will only be counter-productive. However, since we consider the vendor accounts as the gold labels for our classification task, we lower-cased all the vendor names to minimize the number of vendors in our datasets. In other words, we assume the vendors "agentq" and "AgentQ" to be the same entity. The table illustrates how we divide our datasets for supervised pre-training, Low-Resource, and High-Resource fine-tuning steps. Finally, we assign all the vendors with less than 20 ads to a new class label, "others," allowing our classifier to operate in a zero-shot setting.

While we do not perform classification for vendors with less than 20 ads, we capture similarities in the ads for these vendors through our open-set vendor identification task. That said, the number 20 is not arbitrary and is established through experiments. We also experimented with the same setup by removing vendors with less than 5, 10, 20, 50, and 100 ads. The results demonstrate that our model requires at-least 20 ads from each vendor to

[2]Hosted by IMPACT cyber trust portal

[3]In this research, market data refers to the ads and vendor accounts from a single Darknet market. On the other hand, a dataset refers to the combined data from two or more markets.

perform the classification optimally.

4 Experiments

Before running our experiments, we conduct a sanity check to evaluate the need for ML algorithms by examining the similarity in Darknet ads using textdistance-based traditional stylometric approaches (orsinium, 2022) (refer appendix A.1.1). Our analyses show that these traditional methods fail to identify vendors with dissimilar ads, indicating the need for sophisticated feature-extraction techniques. Furthermore, these approaches help us discard identical ads from further analysis.

4.1 Closed-Set Vendor Verification Task

Architectural Baselines: To verify the vendor migrants existing across multiple markets, we first train multiple classifiers to examine different writing styles in Darknet ads and establish a benchmark amongst various ML and neural network-based algorithms. Given the resources at our disposal, training models on the combined Alphabay, Dreams, and Silk Road datasets would be computationally expensive and time-consuming. Therefore, we first establish an architectural baseline by training (i) TFIDF-based statistical (Multinomial Naive Bayes, Logistic Regressor, Random Forest, SVMs, and MLP network), (ii) Bi-directional GRU with Fasttext embeddings (Gupta et al., 2019), CNNs over character n-grams (Shrestha et al., 2017), (iii) Pre-trained BERT-base-cased (Devlin et al., 2019), RoBERTa-base (Liu et al., 2019), and a DistilBERT-base-cased (Sanh et al., 2019) sequence classifiers to identify 1,422 unique vendor accounts from 93,586 ads on the Dreams market.

Methodological Baselines: We further establish a methodological baseline to investigate the influence of different training approaches on the combined Alphabay, Dreams, and Silk Road 1 dataset with 272,696 ads and 3,896 unique vendors. First, we train BERT-base-cased and uncased classifiers to investigate the influence of uppercase and lowercase patterns in ads on the model's performance. Second, we investigate if applying knowledge transfer from a BERT-cased model, trained on the Darknet ads for the language task, improves the classification performance. In this research, we refer to the trained language model as *DarkBERT-LM* and the classifier as *DarkBERT-classifier*. In another study, Houlsby et al. (2019) suggests that rather than updating the weights of the pre-trained model, it is much more efficient to stitch adapter layers and update them while keeping the pre-trained model frozen. Therefore, we finally train a BERT-cased classifier with adapter layers (aka *Adapter-BERT*) and compute its performance.

4.2 Open-Set Vendor Identification Task

In their research, (Kornblith et al., 2019; Phang et al., 2021) proposed *Centered Kernel Alignment (CKA)* as a similarity metric to reliably compute correspondences between representations in networks trained from different initializations. In this research, we compute CKA similarity between the representational layers of our trained classifier and an available pre-trained checkpoint (not trained on Darknet data). Finally, we examine the least similar layers, i.e., the layers that changed most during training and have a low CKA similarity, to extract semantically-meaningful style representations from the ads of Darknet markets. [4]

Similar to Reimers and Gurevych (2019b), we compute the similarity between two vendors by computing cosine-similarity between the extracted style representations in their ads. Then, assigning one of the vendors as the parent vendor, we repeat the process for all the other vendors in our dataset. However, cosine distance represents a linear space with all dimensions weighted equally. Therefore, Xiao (2018) suggests that the emphasis be on the rank and not the absolute value representing the similarity between the two vendors. Besides, vendors on Darknet advertise their products across various categories. For two vendors, A, and B, selling their products under multiple categories, the cosine similarity between their ads would be low by default. Therefore, instead of comparing ads across similar trade categories (which requires labeling efforts and is counterproductive to our research), we propose normalized similarity (sim_{norm}) as a measure of cosine similarity (sim) in ads between two vendors, w.r.t. to the self-similarity (sim_{self}) in their ads through the equation below:

$$sim_{norm} = 2 * \frac{sim(A,B)}{sim_{self}(A,A) + sim_{self}(B,B)}$$

4.3 Low-Resource Market Adaption Task

To verify the vendor migrants from emerging markets, we conduct experiments on an LR dataset,

[4] Algorithm-2 in Appendix A.3 demonstrates the pseudo-code for computing CKA similarity across layers of our trained classifier and an available pre-trained checkpoint.

i.e., Valhalla-Berlusconi, with 3,612 ads and 194 vendors. First, we extract the style representations from the "[CLS]" token of the pre-trained classifier (Section 4.1) for all the ads in our LR dataset. Then, following (Devlin et al., 2019), we apply knowledge transfer from the pre-trained classifier to a two-layer bidirectional GRU classifier by initializing it with the extracted style representations. The Bi-GRU classifier is then fine-tuned to adapt new market knowledge and verify the migrants across the LR dataset. Our research refers to this as the *transfer-BiGRU* model. Performing knowledge transfer helps our existing classifier to evolve with emerging vendor and Darknet market data. During the evaluation, we compare the performance of our transfer-BiGRU against BERT-base-cased and two-layer BiGRU (with fasttext embeddings) classifiers (aka end-to-end baselines) when trained from scratch on the LR dataset. Finally, we also evaluate the zero-shot performance of our architectural and methodological classifiers (aka zero-shot baselines) against the transfer-BiGRU for the closed-set vendor verification task.

5 Results

5.1 Open-Set Vendor Verification Task

Architectural Baselines: Table 2 presents the performance of our architectural baselines evaluated on the Dreams market. Amongst all the statistical models, we found a Multilayer Perceptron (MLP) with bigram TF-IDF features to perform the best. While conventional neural networks such as character-based CNN and Bidirectional GRU with fasttext embeddings performed better than the statistical models, we noted a considerable increase in performance with the transformers-based architecture on our datasets. To our surprise, the RoBERTa-base model underperformed compared to the BERT-base-cased architecture. Although we propose to leverage writing styles to identify various vendors, the Darknet markets are intentionally designed with random noise to foil any automated system. Furthermore, since RoBERTa-tokenizer works on "byte-level BPE," we believe the trained model did not have enough data to learn these features. Consequently, we establish the trained BERT-cased classifier on the Dreams market as the benchmark classifier of our architectural baselines.

Methodological Baselines: Table 3 illustrates the performance of our methodological baselines evaluated on the combined Alphabay-Dreams-Silk Road-1 test dataset. Our first experiment investigates the influence of writing style, i.e., lowercase and uppercase patterns, on the classification task. As can be seen, the BERT-cased classifier outperforms the uncased classifier by a reasonable margin (Approx. 3% on 3,896 class labels). We believe that the increment in performance comes from adding upper and lowercase patterns during training. Next, we experiment with continued pre-training of the DarkBERT-LM on the ads for the language task [5] to achieve a test perplexity of 2.07. In comparison to the BERT-cased classifier, we observe a minor increase in the performance of the finetuned DarkBERT-Classifier. However, we reason that such a minor increase is not worth all the training. Furthermore, the low performance of the DarkBERT-LM depicts the unpredictable and noisy lingo used by Darknet vendors in their ads. We also suspect that further pre-training our models on an extensive dataset can help the baseline improve its performance. Finally, the Adapter BERT also underperforms compared to the vanilla BERT-cased classifier. Consequently, we establish the BERT-cased architecture trained on the closed-set vendor

Data	Models	Accuracy	Micro-F1	Macro-F1
Dreams Market	Statistical Models			
	Multinomial Naive Bayes	0.0183	0.0144	0.0059
	Random Forest	0.0102	0.1093	0.0449
	Logistic Regression	0.0045	0.0090	0.0037
	SVM	0.2480	0.3974	0.3703
	Neural Networks			
	MLP	0.6614	0.6603	0.6594
	Character-CNN	0.7266	0.7256	0.7248
	BiGRU-Fasttext	0.7374	0.7415	0.7360
	Transformers Networks			
	BERT-cased	**0.8978**	**0.8978**	**0.9002**
	DistilBERT-cased	0.8886	0.8885	0.8889
	RoBERTa-base	0.8776	0.8797	0.8736

Table 2: Performance of architectural baselines on the Dreams market.

Data	Models	Accuracy	Micro-F1	Macro-F1
Alphabay-Dreams-Silk Dataset	BERT-uncased	0.8947	0.8939	0.8768
	BERT-cased	**0.9046**	**0.9066**	**0.9013**
	DarkBERT-Classifier	0.9000	0.9090	0.9073
	Adapter BERT	0.8398	0.8330	0.8188

Table 3: Performance of methodological baselines on the combined Alphabay-Dreams-Silk dataset.

[5] Pre-training BERT for a language task is highly resource-intensive. Unfortunately, we did not have the resources to continue the pre-training until the convergence and only trained our model for 20 epochs.

verification task as the benchmark classifier for the Alphabay-Dreams-Silk Road Darknet dataset.

5.2 Open-Set Vendor Identification Task

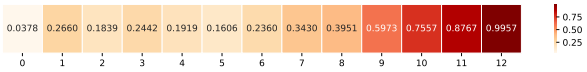


Figure 3: CKA distance between layers of the BERT-cased methodological classifier, compared before and after being trained on the Alphabay-Dreams-Silk dataset.

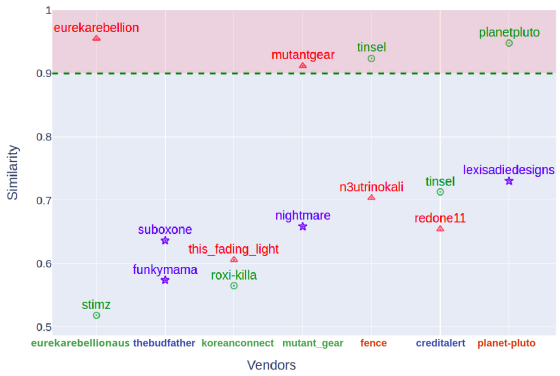


Figure 4: Scatter plot between parent-vendors (on the x-axis) and their potential aliases (scatter points on y-axis) from Alphabay, Dreams, and Silk Road-1 markets.

Figure 3 reveals a high CKA distance, i.e., low CKA similarity, between the representations for the last four layers of our BERT-cased classifier. Therefore, extracting information from the weighted sum of the final four layers provides the most meaningful style representations for our ads in the Alphabay-Dreams-Silk dataset. As described in section 4.2, we use these style representations to compute the cosine similarity between vendor ads. Figure 4 displays some randomly selected parent vendors on the x-axis and their two potential aliases (scatter points) with a similarity score in their ads on the y-axis. [6] Our analysis indicates "eurekarebellionaus" and "eurekarebellion," "mutant_gear" and "mutantgear", "fence" and "tinsel," and "planet-pluto" and "planetpluto" have very high similarity in their ads and can be from the same vendor. The higher the similarity, the more likely it is for two vendors to be the same entity. [7] For a better visibility, these vendors are highlighted inside the red box of our scatter plot.

	Parent Vendor	Alias / Copycat	Similarity
High (potential aliases)	houseofdank	houseofdank2.0	0.9844
	incorporated	incorporatedv2	0.9769
	castro6969	castro69696	0.9541
	thewizard	thewizzardnl	0.9480
	europills	europills2	0.9467
Low (potential Copycats)	topgear	topgear69	0.0367
	dutchpirates	dutchpiratesshop	-0.1015
	whitey	whiteyford	-0.1410
	g3cko	gecko	-0.2292
	aussieimportpills	aussieimportpillsv2	-0.2560

Table 4: Normalized similarity between parent vendors and their potential aliases/copycats aligned in decreasing order.

Often, vendor aliases have similar-looking vendor handles to have recognition and a monopoly over their business. While most similar-looking accounts can be detected using string-based matching techniques like string_grouper (Chris van den Berg, 2021), our experiments reveal the existence of copycats with very different writing styles and low similarity in their ads. For example, our experiments uncovered that only about 24% of similar-looking vendor-alias pairs in the Alphabay-Dreams-Silk dataset have a similarity score of 0.7 or above in their ads. Table 4 illustrates the similarity in ads between 10 such parent vendors and their likely aliases or copycats. Finally, we believe our experiments can also help law enforcement uncover potential vendor-alias pairs with completely unrelated vendor names, ex: "fence" and "tinsel" (see figure 4), but a high similarity between their ads.

5.3 Low Resource Market Adaption Task

To set the Zero-Shot baselines, we first use the established BERT-cased architectural and methodological classifiers to perform zero-shot vendor verification on the LR dataset, Valhalla-Berlusconi. Since the emerging LR dataset has new vendors, we assign all these new vendor accounts to the class label "others." However, since the macro-F1 score is computed for the unweighted arithmetic mean of F1 for all class labels, the absence of previously

[6] We generate the scatter plot using Plotly, which allows us to zoom infinitely for any vendor. However, we only show the chosen vendors with two potential aliases for better clarity and visibility.

[7] As mentioned, Darknet vendors often create aliases to hide from Law Enforcement Agencies (LEA). However, since numerous vendors appear on Darknet markets yearly, it becomes difficult for law enforcement to manually link these aliases to a parent vendor. The unavailability of ground truth poses a challenge in evaluating the existence of these aliases in our datasets. Therefore, we cannot confidently comment upon the accuracy of our similarity-based analyses without the qualitative case study. We encourage LEA not solely to rely on these similarities but use them as a starting point for their manual investigations. Furthermore, we strongly discourage LEA from abiding by these analyses as evidence for investigation or prosecution. The sole purpose of this research is to help LEA bring meaning to the online Darknet market data.

existing vendors in the LR emerging market leads us to unreliable macro-F1 results. Consequently, we emphasize the performance of our Zero-Shot baselines on the micro-F1 score. The baselines exhibit promising performance with a micro-F1 of 0.7702 and 0.7388 despite not being trained on LR data. The decrease in macro-F1 performance from architectural to methodological baseline is due to an increase in vendor accounts from 1,442 in the Dreams market to 3,896 in the Alphabay-Dreams-Silk Road dataset.

Models	Layer	Micro-F1	Macro-F1
Zero-Shot Baselines			
Architectural	-	0.7702	0.2927
Methodological	-	0.7388	0.2401
End-to-End Baselines			
BERT-cased	-	**0.8987**	**0.8148**
BiGRU-Fasttext	-	0.7797	0.6957
Transfer Baselines			
Transfer-BiGRU	Embedding	0.7653	0.6408
	Last	0.8590	0.7809
	Second-to-Last	0.8951	0.7884
	Weighted Sum All 12	0.8928	0.7837
	Weighted Sum Last 4	**0.8946**	**0.8132**

Table 5: Performance of Zero-Shot, End-to-End, and Transfer baselines on the Valhalla-Berlusconi dataset.

GPU	Models	Trainable parameters	Training Time (Hrs:Mins)
Tesla-V100 (32 GB)	BERT-cased	110M	0:54
	BiGRU-Fasttext	13M	0:12
	Transfer-BiGRU	24M	0:32
GE-MX110 (2 GB)	Transfer-BiGRU	24M	2:40

Table 6: Computational details of trained classifiers on the LR, Valhalla-Berlusconi, dataset.

Furthermore, we also train a BERT-cased and a BiGRU classifier with fasttext embeddings from scratch to adapt new market knowledge and vendors from the emerging LR dataset. As illustrated in table 5, compared to the Zero-Shot baselines, the End-to-End baselines show a significant increase in performance in both micro-F1 and macro-F1 scores. Finally, following (Devlin et al., 2019), we perform knowledge transfer by extracting the style representations from multiple layers of the BERT-cased methodological classifier and using them to initialize the BiGRU before the classification layer. Table 5 demonstrates that when initialized with the weighted sum of the last four layers, the transfer-BiGRU classifier benefits most from the knowledge transfer and performs comparably to the End-to-End BERT-cased classifier on the emerging LR dataset. Consequently, we establish the transfer-BiGRU architecture trained on the closed-set vendor verification task as the benchmark classifier for the LR, Valhalla-Berlusconi dataset. [8]

Finally, Table 6 reflects upon the computational aspects of the trained models by comparing the number of trainable parameters and training time for classifiers trained on the LR dataset. As can be seen, compared to the BERT-cased, our transfer-BiGRU classifier is carbon-efficient (refer to appendix 10), has 78% less trainable parameters, and takes approximately half the training time. Furthermore, we also show the training feasibility of our transfer-BiGRU on a low-end graphic card, GeForce-MX110, with 2 GB of GPU memory. Thus, our low-compute transfer-BiGRU classifier can significantly help law enforcement scale our approach to emerging markets without significant performance loss.

6 Error Analysis

Vendor	Pred	Text A	Text B
house ofdank	TP	** 1 Lb of Sour [DRUG1] (Greenhouse) **	** 1 oz of Greenhouse [drug1] greenhouse grown **
house ofdank2.0	TP	** 1 OZ of [drug2] Greehouse grown **	** 1 Lb of [drug2] Greehouse grown **
appleinc	FP	10 x €50 euro COUNTERFEIT notes (Very Good Quality)	5 x $100 DOLLAR COUNTERFEIT STRIP high quality bills
canadian pharmacy	FP	[usa to usa] [drug3] 80mg just 19.99 bucks per pill only	[usa to usa] 30 pills [drug3] 100mg 19.99 usd ultram

Table 7: Qualitative analysis of BERT-cased classifier (trained on Alphabay-Dreams-Silk Road Dataset) for True Positives (TP) and False Positives (FP) predictions.

To better understand the strengths and weaknesses of our trained models, we perform qualitative analysis on the predictions of the BERT-cased classifier (trained on the Alphabay-Dreams-Silk Road Dataset) in Table 7. Note that we only display the title of these advertisements due to space constraints and visibility reasons. As can be seen in the first two examples, our trained classifier can recognize many patterns in the ads, such as "**," "[DRUG1]", "[drug1]", and "greenhouse gown,." The first two examples also show how similar the

[8] We also test the performance of our baselines on an emerging High-Resource (HR) dataset, Traderoute-Agora. Results in the appendix table 10 show that the transfer-BiGRU model underperforms compared to the End-to-End BERT-cased classifier. In other words, applying knowledge transfer on emerging HR markets does not yield the best performance. Please refer to section A.1.3 in the appendix for more details.

advertisements are between the vendors "houseofdank" and "houseofdank2.0". This is also indicated by the high similarity in the advertisements of the two vendors (refer 4). Finally, the next two examples in the Table below indicate the cases of false positives. As can be seen, here, the network is confusing between vocabulary such as "COUNTERFEIT," "quality," "supernotes," source and destination locations, [drug3], and the price of the product.

Furthermore, we inspect cases where our trained BERT-cased classifier fails, but the transfer-GRU classifier succeeds after knowledge transfer. Table 8 demonstrates vendor advertisements where the writing style between advertisements changed drastically between the Alphabay-Dreams-Silk Road and Valhalla-Berlusconi datasets. Consequently, our BERT-cased classifier fails to verify vendors from the Valhalla-Berlusconi dataset in the zero-shot setting. Finally, after applying knowledge transfer and fine-tuning our transfer-BiGRU model, the model quickly adapts to the new writing styles from these vendor advertisements.

Vendor	**Alphabay-Dreams-Silk Road**	**Valhalla-Berlusconi**
cannacorner	[drug1] 3.5g —|MERCEDES|	7g [drug1] —|lambo|
medicalznl	5 GRAMS COLOMBIAN [DRUGX] 93% + FREE SHIPPING	2.5 grams - colombian [drugx] 90+% pure uncut
color	Credit Cards Can Be Without Security Code	lasted update credit cards in this file.

Table 8: Qualitative analysis of transfer-BiGRU classifier (trained on Valhalla-Berlusconi Dataset) for True Positives (TP) and False Positives (FP) predictions.

7 Discussion and Future Work

We discuss details about additional experiments and the training setup in appendix sections A.1 and A.2, respectively. In addition, the pseudo-code for the CKA algorithm is provided in appendix A.3.

In the future, we plan to work on the assumptions in section 9 by investigating content-control contrastive learning approaches (Wegmann et al., 2022) to perform vendor verification and identification on existing and emerging Darknet datasets.

8 Conclusion

This research presents an NLP-based vendor verification and identification approach, VendorLink, for law enforcement to verify, identify, and link vendor migrants and potential aliases on the existing and emerging Darknet markets. In this work, we first perform supervised pre-training to adapt Darknet market knowledge and establish a BERT-cased classifier to verify existing vendor migrants between markets in a closed-set environment. Then, we extract the style representations from the trained BERT-cased classifier to compute the text similarity in vendor ads in an open-set environment and link vendors to their potential aliases. Finally, we adapt new market knowledge by employing knowledge transfer from the trained BERT-cased classifier to a low-compute-resource BiGRU classifier and perform closed-set vendor verification on the emerging LR markets. Through our experiments, we uncover (i) 15 migrants and 71 potential aliases in the Alphabay-Dreams-Silk dataset, (ii) 17 migrants and 3 potential aliases in the Valhalla-Berlusconi dataset, and (iii) 75 migrants and 10 potential aliases in the Traderoute-Agora dataset with a cosine similarity of 0.8 and above, between the ads of vendors and their potential aliases.

9 Limitations

Assumptions: This work applies a lower-case transformation to the vendor names during the pre-processing step and assumes vendor accounts "agentq" and "AgentQ" to be from the same entity. However, in reality, these entities can refer to two different vendors. Additionally, we train our classifier in a multi-class classification setting, assuming that ads correspond to only one individual vendor account. However, our experiments uncover the existence of copycats on Darknet markets. In reality, it is always possible for multiple vendors to co-exist with similar vendor names; hence, any supervised approach will only generate skew results. In the future, we plan to look toward contrastive learning approaches (Pan et al., 2021; Zhou et al., 2021b; Wegmann et al., 2022) to avoid these assumptions.

Architectural limitations: This research establishes a BERT-base-cased classifier to verify migrating vendors across existing and emerging Darknet markets. While we acknowledge that using a bigger BERT model with a sliding window may improve our classification's performance, given the resources at our disposal, we decided against it. Moreover, as mentioned earlier, most of the ads used in this research are in English, with a few exceptions where the vendors use multiple languages. Therefore, applying a multilingual transformer-based model to the classification task (Wang and

Banko, 2021) can improve our approach's performance.

Unsupervised and HR settings: As described in the assumptions, the core of our approach lies in the availability of gold labels. VendorLink utilizes the supervised pre-training step to perform knowledge transfer and text-similarity tasks. Therefore, our approach suffers a significant limitation in the absence of these ground labels / unsupervised settings. Furthermore, as described in A.1.3, our approach could not scale well to verify vendor migrants in HR emerging datasets. In the future, we plan to expose VendorLink to contrastive learning approaches to learn universal representations and overcome the problem.

Diverse Advertisements: In the semi-supervised task, we compute the likelihood of two vendor accounts being from the same entity by calculating the similarity between the advertisements of the two vendors. Since one of the novelties of this research lies in the direction of End-to-End training, we have avoided using handcrafted labels for applying content control to generate content-independent style representation. However, as explained in section 4.2, an advertisement from the drug category can be very different from that of the weapon category. Therefore, in the future, we plan to train another classifier to classify Darknet advertisements into different trade categories before performing the vendor-verification task.

XAI limitations: eXplainaible Artificial Intelligence (XAI) is integral in promoting trust and understanding amongst the end-users. From LEA's perspective, its absence can be viewed as arguably negligent and unreliable. While we acknowledge that our approach currently lacks an XAI feature, in the future, we plan to build upon our experiments in A.1.5 and establish a reliable approach for understanding and explaining our model's decision.

10 Broader Impact

This section discusses mandatory data collection protocols, ethical considerations, potential risks, and legal, societal, and environmental impacts.

Data Collection Protocol: Ethical concerns associated with web scraping do not apply to our research as the online darknet data used is requested through a signed Memorandum of Agreement (MoA) with IMPACT Cyber Trust portal (ICC). As a result, the data is freely available, legally collected, and distributed for large-scale cybersecurity analytics, allowing researchers to advance the state-of-the-art cyber-risk R&D and decision support.

Legal Impact: This research emphasizes bringing structure and meaning to the massively available online data on Darknet markets for LEA. While we can not predict whether our research will impact the LEA process, the intent is to identify potential connections between vendors of illegal goods and present LEA with a broader information base for their internal processes. Please note that at no point do we claim to provide pieces of evidence necessary for prosecuting any criminal.

Ethical and Privacy Considerations: We acknowledge that using vendor names in our study could potentially be exploited and identified as a privacy concern. However, after going through a Data Privacy Impact Assessment (DPIA) at our institution, the committee concluded that the vendor names used in this study are pseudonyms and do not reflect any individual's identity. Furthermore, research suggests that the lifespan of Darknet vendors and marketplaces is between a few months and a couple of years. (Booij et al., 2021; Broadhurst et al., 2021; UNDOC, 2020). Since the market ads in our datasets span between 2011-2018, the likelihood of any vendor's existence with the same user name is very low. Finally, under article 6, Lawfulness of processing, the GDPR clause suggests that the processing of personal data is lawful as long as the task is carried out in the public interest. Given the nature of illegal activities on the Darknet and despite all its potential risks, we believe that our research can potentially benefit LEA and save human lives. That said, while using vendor names in our analyses promotes transparency and reproducibility amongst the readers, we encourage these vendors to reach out to us in case of any concerns. In such circumstances, we take complete responsibility for taking immediate action and removing their information from our research.

Societal Impact and Potential Risk: In their research, Juola (2020) described the dark side of authorship studies and social media analytics for target-based recommendation systems and employee, political, medical, gender, demographic, and racial profiling. While our approach can lend itself to abuses, we find it unlikely for anyone to

exploit our research as it is given the extreme difference in the language between the Darknet and surface web websites (Choshen et al., 2019). That said, we acknowledge the possibility of privacy infringement outside criminal markets to match user activity across public platforms. For instance, ill-intentioned third parties and organizations could use our research to circumvent an individual's identity on public social media platforms. Therefore, we encourage our readers to be aware of the ethical duality while using our research to develop authorship technologies inside and outside cybersecurity scenarios.

Environmental Impact: Keeping in mind that not all LEA have the resources to train computationally expensive architectures, we investigate utilizing knowledge transfer to train low-compute-resource models in this research. As a result, our transfer-BiGRU classifier has a carbon efficiency of 0.07 $kgCO_2eq/kWh$ and 2.25 $kgCO_2eq/kWh$ as opposed to the BERT-cased classifier with a carbon efficiency of 0.12 $kgCO_2eq/kWh$ and 4.21 $kgCO_2eq/kWh$ on the Vallhalla-Berlusconi and Traderoute-Agora datasets, respectively. These estimations were conducted on Tesla V100-SXM2-32GB (TDP of 300W) using the MachineLearning Impact calculator presented in (Lacoste et al., 2019). In other words, this research demonstrates that applying knowledge transfer from existing to emerging markets can help law enforcement train low-compute-resource models with high performance, faster training time, and lesser carbon footprint.

11 Acknowledgement

This research is supported by the Sector Plan Digital Legal Studies of the Dutch Ministry of Education, Culture, and Science and Cora4NLP project, funded by the German Federal Ministry of Education and Research (BMBF) under funding code 01IW20010. Finally, the experiments were made possible using the Data Science Research Infrastructure (DSRI) hosted at Maastricht University.

References

Lucky Agarwal, Kartik Thakral, Gaurav Bhatt, and Ankush Mittal. 2019. Authorship clustering using tf-idf weighted word-embeddings. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, FIRE '19, page 24–29, New York, NY, USA. Association for Computing Machinery.

Mhd Wesam Al Nabki, Eduardo Fidalgo, Enrique Alegre, and Ivan de Paz. 2017. Classifying illegal activities on tor network based on web textual contents. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 35–43, Valencia, Spain. Association for Computational Linguistics.

Andres Baravalle and Sin Lee. 2018. *Dark Web Markets: Turning the Lights on AlphaBay: 19th International Conference, Dubai, United Arab Emirates, November 12-15, 2018, Proceedings, Part II*, pages 502–514.

Georgios Barlas and Efstathios Stamatatos. 2021. A transfer learning approach to cross-domain authorship attribution. *Evol. Syst.*, 12(3):625–643.

Benedikt Boenninghoff, Steffen Hessler, Dorothea Kolossa, and Robert M. Nickel. 2019. Explainable authorship verification in social media via attention-based similarity learning.

Tim M. Booij, Thijmen Verburgh, Federico Falconieri, and Rolf S. van Wegberg. 2021. Get rich or keep tryin' trajectories in dark net market vendor careers. In *2021 IEEE European Symposium on Security and Privacy Workshops (EuroS PW)*, pages 202–212.

Gwern Branwen, Nicolas Christin, David Décary-Hétu, Rasmus Munksgaard Andersen, StExo, El Presidente, Anonymous, Daryl Lau, Delyan Kratunov Sohhlz, Vince Cakic, Van Buskirk, Whom, Michael McKenna, and Sigi Goode. 2015. Dark net market archives, 2011-2015. `https://www.gwern.net/DNM-archives`. Accessed: DATE.

Roderic Broadhurst, Matthew Ball, Chuxuan Jiang, Joy Wang, and Harry Trivedi. 2021. *Impact of darknet market seizures on opioid availability*.

Madeleine Bruggen and Arjan Blokland. 2021. *Child Sexual Exploitation Communities on the Darkweb: How Organized Are They?*, pages 259–280.

Xiaoyan Cai, Sen Liu, Libin Yang, Yan Lu, Jintao Zhao, Dinggang Shen, and Tianming Liu. 2022. Covidsum: A linguistically enriched scibert-based summarization model for covid-19 scientific papers. *Journal of Biomedical Informatics*, 127:103999.

Theo Carr, Jun Zhuang, Dwight Sablan, Emma LaRue, Yubao Wu, Mohammad Al Hasan, and George Mohler. 2019. Into the reverie: Exploration of the dream market. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 1432–1441.

Daniel Castro Castro, Yaritza Adame Arcia, María Pelaez Brioso, and Rafael Muñoz Guillena. 2015. Authorship verification, average similarity analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 84–90, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Leshem Choshen, Dan Eldad, Daniel Hershcovich, Elior Sulem, and Omri Abend. 2019. The language of legal and illegal activity on the Darknet. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4271–4279, Florence, Italy. Association for Computational Linguistics.

Chris van den Berg. 2021. string_grouper. [Online; accessed 2022-09-01].

Nicolas Christin. 2013. Traveling the silk road: A measurement analysis of a large anonymous online marketplace. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, page 213–224, New York, NY, USA. Association for Computing Machinery.

Carnegie Mellon University CMU. 2012-13. Traveling the silk road: Non-anonymized datasets.

Carnegie Mellon University CMU. 2017-18a. Alphabay marketplace: Non-anonymized dataset, 2017-18.

Carnegie Mellon University CMU. 2017-18b. Dream, traderoute, berlusconi and valhalla marketplaces, 2017-2018: Non-anonymized datasets.

José Eleandro Custódio and Ivandré Paraboni. 2021. Stacked authorship attribution of digital texts. *Expert Systems with Applications*, 176:114866.

Gemma Davies. 2020. Shining a light on policing of the dark web: An analysis of uk investigatory powers. *The Journal of Criminal Law*, 84(5):407–426.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

C Easttom. 2018. Conducting investigations on the dark web. *Journal of Information Warfare*, 17(4):26–37.

Anirudh Ekambaranathan. 2018. Using stylometry to track cybercriminals in darknet forums.

ENISA. 2018. Financial fraud in the digital space.

Maël Fabien, Esau Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. 2020. BertAA : BERT fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 127–137, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

Mohd Faizan and Raees Ahmad Khan. 2019. Exploring and analyzing the dark web: A new alchemy.

Gabriela Ferraro and Hanna Suominen. 2020. Transformer semantic parsing. In *Proceedings of the The 18th Annual Workshop of the Australasian Language Technology Association*, pages 121–126, Virtual Workshop. Australasian Language Technology Association.

Tianjun Fu, Ahmed Abbasi, and Hsinchun Chen. 2010. A focused crawler for dark web forums. *J. Am. Soc. Inf. Sci. Technol.*, 61(6):1213–1231.

Donglai Ge, Junhui Li, and Muhua Zhu. 2019. A transformer-based semantic parser for nlpcc-2019 shared task 2. In *Natural Language Processing and Chinese Computing*, pages 772–781, Cham. Springer International Publishing.

Deyan Georgiev. 2021. How much of the internet is the dark web in 2021? : Alarming dark web statistics.

Shalini Ghosh, Ariyam Das, Phil Porras, Vinod Yegneswaran, and Ashish Gehani. 2017. Automated categorization of onion sites for analyzing the darkweb ecosystem. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 1793–1802, New York, NY, USA. Association for Computing Machinery.

K. Godawatte, M. Raza, M. Murtaza, and A. Saeed. 2019. Dark web along with the dark web marketing and surveillance. In *2019 20th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT)*, pages 483–485.

Sean E. Goodison, Dulani Woods, Jeremy D. Barnum, Adam R. Kemerer, and Brian A. Jackson. 2019. *Identifying Law Enforcement Needs for Conducting Criminal Investigations Involving Evidence on the Dark Web*. RAND Corporation, Santa Monica, CA.

Shriya TP Gupta, Jajati Keshari Sahoo, and Rajendra Kumar Roul. 2019. Authorship identification using recurrent neural networks. In *Proceedings of the 2019 3rd International Conference on Information System and Data Mining*, ICISDM 2019, page 133–137, New York, NY, USA. Association for Computing Machinery.

Julien Hay, Bich-Lien Doan, Fabrice Popineau, and Ouassim Ait Elhara. 2020. Representation learning of writing style. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 232–243, Online. Association for Computational Linguistics.

Darren R Hayes, Francesco Cappa, and James Cardon. 2018. A framework for more effective dark web marketplace investigations. *Information*, 9(8):186.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Siyu He, Yongzhong He, and Mingzhe Li. 2019. Classification of illegal activities on the dark web. In *Proceedings of the 2019 2nd International Conference on Information Science and Systems*, ICISS 2019, page 73–78, New York, NY, USA. Association for Computing Machinery.

Leo Horne, Matthias Matti, Pouya Pourjafar, and Zuowen Wang. 2020. GRUBERT: A GRU-based method to fuse BERT hidden layers for Twitter sentiment analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 130–138, Suzhou, China. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp.

Fereshteh Jafariakinabad, Sansiri Tarnpradab, and Kien A. Hua. 2019. Syntactic recurrent neural network for authorship attribution.

Ting Jiang, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Liangjie Zhang, and Qi Zhang. 2022. Promptbert: Improving bert sentence embeddings with prompts. *arXiv preprint arXiv:2201.04337*.

Youngjin Jin, Eugene Jang, Yongjae Lee, Seungwon Shin, and Jin-Woo Chung. 2022. Shedding new light on the language of the dark web.

Patrick Juola. 2020. Authorship studies and the dark side of social media analytics. *Journal of Universal Computer Science*, 26:156–170.

Bowon Ko and Ho-Jin Choi. 2020. Paraphrase bidirectional transformer with multi-task learning. In *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 217–220.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for pytorch.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited.

Kristy Kruithof, Judith Aldridge, David Décary Hétu, Megan Sim, Elma Dujso, and Stijn Hoorens. 2016. *The role of the 'dark web' in the trade of illicit drugs*. RAND Corporation, Santa Monica, CA.

Ramnath Kumar, Shweta Yadav, Raminta Daniulaityte, Francois Lamy, Krishnaprasad Thirunarayan, Usha Lokala, and Amit Sheth. 2020. Edarkfind: Unsupervised multi-view learning for sybil account detection. In *Proceedings of The Web Conference 2020*, WWW '20, page 1955–1965, New York, NY, USA. Association for Computing Machinery.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *CoRR*, abs/1910.09700.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Antoine Louis and Gerasimos Spanakis. 2021. A statutory article retrieval dataset in french.

Andrei Manolache, Florin Brad, Antonio Barbalau, Radu Tudor Ionescu, and Marius Popescu. 2022. Veridark: A large-scale benchmark for authorship verification on the dark web.

Derek Miller. 2019. Leveraging bert for extractive text summarization on lectures.

Andrew M. Olney. 2021. Paraphrasing academic text: A study ofnbsp;back-translating anatomy andnbsp;physiology with transformers. In *Artificial Intelligence in Education: 22nd International Conference, AIED 2021, Utrecht, The Netherlands, June 14–18, 2021, Proceedings, Part II*, page 279–284, Berlin, Heidelberg. Springer-Verlag.

Juanita Ordoñez, Rafael Rivera Soto, and Barry Y. Chen. 2020. Will longformers pan out for authorship verification? notebook for pan at clef 2020. In *CLEF*.

orsinium. 2022. textdistance. [Online; accessed 2022-09-01].

Lin Pan, Chung-Wei Hang, Avirup Sil, and Saloni Potdar. 2021. Improved text classification via contrastive adversarial training.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Giacomo Persi Paoli, Judith Aldridge, Nathan Ryan, and Richard Warnes. 2017. *Behind the curtain: The illicit trade of firearms, explosives and ammunition on the dark web*. RAND Corporation, Santa Monica, CA.

Jason Phang, Haokun Liu, and Samuel R. Bowman. 2021. Fine-tuned transformers show clusters of similar representations across layers. *CoRR*, abs/2109.08406.

Charles Pierse. 2021. Transformers Interpret.

Nils Reimers and Iryna Gurevych. 2019a. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019b. Sentence-bert: Sentence embeddings using siamese bert-networks.

Andi Rexha, Mark Kröll, Hermann Ziak, and Roman Kern. 2018. Authorship identification of documents with high content similarity. *Scientometrics*, 115(1):223–237.

Dylan Rhodes. 2015. Author attribution with cnn's.

Sebastian Ruder. 2019. *Neural transfer learning for natural language processing*. Ph.D. thesis, NUI Galway.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Upendra Sapkota, Thamar Solorio, Manuel Montes, Steven Bethard, and Paolo Rosso. 2014. Cross-topic authorship attribution: Will out-of-topic data help? In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1228–1237, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Upendra Sapkota, Thamar Solorio, Manuel Montes-y Gómez, and Paolo Rosso. 2013. The use of orthogonal similarity relations in the prediction of authorship. In *Computational Linguistics and Intelligent Text Processing*, pages 463–475, Berlin, Heidelberg. Springer Berlin Heidelberg.

Matthias Schäfer, Markus Fuchs, Martin Strohmeier, Markus Engel, Marc Liechti, and Vincent Lenders. 2019. Blackwidow: Monitoring the dark web for cyber security information. In *11th International Conference on Cyber Conflict (CyCon)*, volume 900, pages 1–21.

Prasha Shrestha, Sebastian Sierra, Fabio 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, Valencia, Spain. Association for Computational Linguistics.

Shashank Srivastava, Igor Labutov, and Tom Mitchell. 2018. Zero-shot learning of classifiers from natural language quantification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 306–316, Melbourne, Australia. Association for Computational Linguistics.

Xiao Hui Tai, Kyle Soska, and Nicolas Christin. 2019. Adversarial matching of dark net market vendor accounts. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*, KDD '19, page 1871–1880, New York, NY, USA. Association for Computing Machinery.

Roelien C. Timmer, David Liebowitz, Surya Nepal, and Salil S. Kanhere. 2021. Can pre-trained transformers be used in detecting complex sensitive sentences? - a monsanto case study. In *2021 Third IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, pages 90–97.

Jeroen Ubbink, Dr. Luca Allodia, Dr. Alexander Serebrenik, and Dr. Decebal Mocanu. 2019. Characterization of illegal dark web arms markets.

Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020a. Authorship attribution for neural text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8384–8395, Online. Association for Computational Linguistics.

Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020b. Authorship attribution for neural text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8384–8395, Online. Association for Computational Linguistics.

UNDOC. 2020. In focus trafficking over the darknet 4 - united nations office on drugs and crime.

Guido Van Rossum and Fred L Drake Jr. 1995. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.

Rolf Van Wegberg, Samaneh Tajalizadehkhoob, Kyle Soska, Ugur Akyazi, Carlos Gañán, Bram Klievink, Nicolas Christin, and Michel Van Eeten. 2018. Plug and prey? measuring the commoditization of cybercrime via online anonymous markets. In *Proceedings of the 27th USENIX Conference on Security Symposium*, SEC'18, page 1009–1026, USA. USENIX Association.

Sophia Dastagir Vogt. 2017. The digital underworld: Combating crime on the dark web in the modern era.

Andrew Vold and Jack G. Conrad. 2021. *Using Transformers to Improve Answer Retrieval for Legal Questions*, page 245–249. Association for Computing Machinery, New York, NY, USA.

Cindy Wang and Michele Banko. 2021. Practical transformer-based multilingual text classification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 121–129, Online. Association for Computational Linguistics.

Xiangwen Wang, Peng Peng, Chun Wang, and Gang Wang. 2018. You are your photographs: Detecting multiple identities of vendors in the darknet marketplaces. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, ASIACCS '18, page 431–442, New York, NY, USA. Association for Computing Machinery.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. 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, Dublin, Ireland. Association for Computational Linguistics.

Gabriel Weimann. 2016. Terrorist migration to the dark web. *Perspectives on Terrorism*, 10(3):40–44.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Han Xiao. 2018. bert-as-service.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019a. End-to-end open-domain question answering with. In *Proceedings of the 2019 Conference of the North*. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019b. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2016. Character-level convolutional networks for text classification.

Yiming Zhang, Yujie Fan, Wei Song, Shifu Hou, Yanfang Ye, Xin Li, Liang Zhao, Chuan Shi, Jiabin Wang, and Qi Xiong. 2019. Your style your identity: Leveraging writing and photography styles for drug trafficker identification in darknet markets over attributed heterogeneous information network. In *The World Wide Web Conference*, WWW '19, page 3448–3454, New York, NY, USA. Association for Computing Machinery.

Chen Zhao, Wei Song, Xianjun Liu, Lizhen Liu, and Xinlei Zhao. 2018. Research on authorship attribution of article fragments via rnns. In *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, pages 156–159.

Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. 2021a. Learning placeholders for open-set recognition.

Wenxuan Zhou, Fangyu Liu, and Muhao Chen. 2021b. Contrastive out-of-distribution detection for pre-trained transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1100–1111, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jian Zhu and David Jurgens. 2021. Idiosyncratic but not arbitrary: Learning idiolects in online registers reveals distinctive yet consistent individual styles. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 279–297, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

İzzet Bozkurt, O. Baghoglu, and Erkan Uyar. 2007. Authorship attribution. *2007 22nd international symposium on computer and information sciences*, pages 1–5.

A Appendix

A.1 Additional Experiments

A.1.1 Sanity Check: stylometric approaches

As a sanity check, we investigate the need for ML algorithms by examining if traditional stylometric approaches can identify writing patterns in Darknet ads. Since languages are represented by characters, tokens, and sentence-level elements, we compute string, token, and sequence-based similarities between ads using the Damerau-Levenshtein distance, Jaccard Index, and Ratcliff-Obershelp pattern recognition technique from textdistance. We define the similarity between two vendor ads as the average of the above three metrics. For a vendor with multiple ads, say vendor A, we compute average similarity as the mean of similarities between all their ads. Similarly, for vendor B, existing across multiple markets, we take all the ads from market X and compute their similarity with ads

of market Y (one at a time). Finally, we compute the average similarity as the mean of similarities between the ads for vendor B across all markets. Algorithm 1 explains the pseudo-code for computing similarity between the ads within and across the Darknet markets.

Figure 5 demonstrates the performance of traditional stylometric approaches on a box plot. The plot represents the average similarity distribution and its skewness within the ads of Alphabay-Alphabay, Dreams-Dreams, Silk Road-Silk Road and across Alphabay-Dreams, Dreams-Silk Road, and Alphabay-Silk Road markets. As can be seen, most ads have an average similarity below 0.20. While there are outliers with higher similarities, only one vendor, "cyanspore", has a similarity score of 1.0 for the Alphabay-Dreams and Dreams-Silk datasets. Since the ads from this vendor are exactly similar, we remove them from all our further analyses.

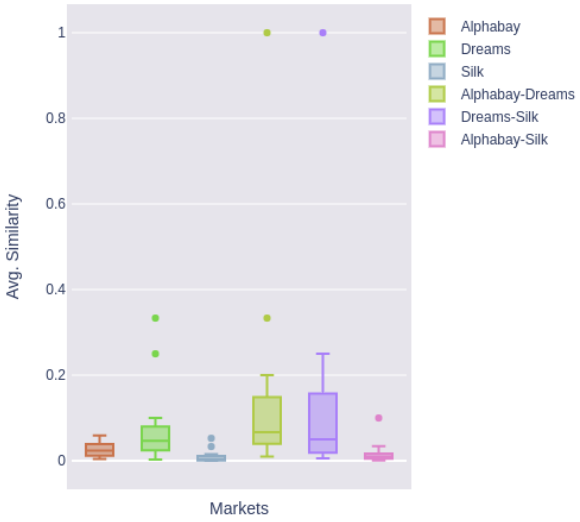


Figure 5: Performance of traditional stylometric techniques average similarity in ads for vendors within and across Darknet datasets.

The low similarity scores within and across datasets indicate the limited capabilities of traditional stylometric frameworks and suggest the need for mathematical models that can abstract features on higher levels. The low scores also serve as a sanity check indicating that vendors on Darknet use different vocabulary and styles in their ads within and across different markets, indicating the need for more profound feature-abstraction techniques.

Algorithm 1: TextDistance-based algorithm for computing stylometric similarity

Data: Alphabay (A), Dreams (D), and Silk Road-1 (S)
Input: len(A), len(D), len(S) > 1, and operation(Op)
$\forall Op \in [within, across]$
Output: Average similarity

```
/* For computing similarity within
   w and across a markets          */
list_w, list_a = [], []
Def Similarity(text_A, text_B):
    return normalized-mean(
     Levenshtein(text_A, text_B),
     jaccard(text_A, text_B),
     obershelp(text_A, text_B) )
if Op == within then
    /* Computing average similarity
       for a vendor within a Darknet
       market (say A)               */
    allVendors = uniqueVendors(A)
    for vendor in allVendors do
        for ad_A1 in A[vendor] do
            for ad_A2 in A[vendor] do
                list_w.append(Similarity(ad_A1,
                 ad_A2))
    averageSimilarity = MEAN(list_w)
else
    /* Computing average similarity
       for a vendor across multiple
       markets (say A and D)        */
    allVendors = commonVendors(A, D)
    for vendor in allVendors do
        for ad_A in A[vendor] do
            for ad_D in D[vendor] do
                list_a.append(Similarity(ad_A,
                 ad_D))
    averageSimilarity = MEAN(list_across)
```

A.1.2 Vendor Verification Task: Influence of advertisement frequency and trade categories on classifier's performance

The Alphabay-Dreams-Silk Road dataset consists of 272,696 unique ads and 3,896 vendors with 322 distinct categories. Table 9 illustrates the performance of our established BERT-cased classifier for

five vendors selling trades across the most distinct categories and five vendors selling trades across only one category. As can be seen, the classifier's performance remains unaffected (more or less) with the number of trade categories and advertisement frequencies. The consistent performance suggests that despite the trade being conducted amongst different categories, Darknet vendors tend to advertise their products similarly, allowing our classifier to distinguish between unique writing styles from different vendors.

Vendor	Ad Frequencies	Categories	F1-Score
googleyed	349	63	0.9340
etizolam	186	59	0.9462
gotmilk	842	48	0.9893
rinran	437	47	0.9816
uhrwerk	135	41	0.9925
citizen5	35	1	1.0000
corktech	35	1	0.9714
sabinas	26	1	0.9615
mrsupermario	24	1	0.9615
emperium	22	1	1.000

Table 9: F1-score w.r.t vendor advertisement frequency and trade categories.

A.1.3 Applying Knowledge Transfer: adapting to verify vendors from High Resource (HR) emerging markets

Models	Layer	Micro-F1	Macro-F1
Zero-Shot Baselines			
Architectural	-	0.7305	0.2173
Methodological	-	0.6498	0.1563
End-to-End Baselines			
BERT-cased	-	**0.8750**	**0.8700**
BiGRU-Fasttext	-	0.6577	0.6539
Transfer Baselines			
Transfer-BiGRU	Embedding	0.6707	0.6698
	Last	0.7061	0.7153
	Second-to-Last	0.6992	0.6911
	Weighted Sum All 12	0.6698	0.6703
	Weighted Sum Last 4	**0.8065**	**0.8177**

Table 10: Performance of Zero-Shot, End-to-End, and Transfer baselines on the Traderoute-Agora dataset.

GPU	Models	Trainable parameters	Training Time (Hrs:Mins)
Tesla-V100 (32 GB)	BERT-cased	112M	32:30
	BiGRU-Fasttext	31M	2:25
	Transfer-BiGRU	42M	17:23

Table 11: Computational details of trained classifiers on the Traderoute-Agora dataset.

In this research, we demonstrate the ability of our approach to adapt and verify migrating vendors from emerging LR markets using a compute-efficient network (transfer-BiGRU). Similar to the results presented in Section 5.3, tables 10 and 11 demonstrate the performance and computational details of a transfer-BiGRU classifier on an HR emerging, Traderoute-Agora, dataset. As can be seen, despite the lesser trainable parameters and training time, our transfer-BiGRU underperforms compared to the end-to-end BERT-cased baseline. Therefore, we do not claim that our knowledge transfer approach scales to emerging vendors in HR Darknet markets.

A.1.4 Seed Runs

Due to limited resource constraints, we only analyze the effects of different initializations on our model's performance for the established benchmarks. As seen in Table 12, the standard deviation, variance, and average performance suggest around 1% influence of initialization on the model's performance. We report all the performance in this work based on our experiments conducted with a seed value of 1111.

Seed Value	BERT-cased Alpha-Dreams-Silk Dataset	BERT-cased Valhalla-Berlusconi Dataset	transfer-BiGRU
40	0.8969	0.8039	0.7798
100	0.8824	0.8278	0.8005
500	0.8813	0.7837	17:23
1100	0.8861	0.8089	0.8019
1111	**0.9013**	**0.8290**	**0.8132**
Var.	6.46 x 10^{-5}	0.0002	0.0002
Std.	0.0080	0.0167	0.0143
Avg.	**0.8896**	**0.8106**	**0.8035**

Table 12: Influence of different initialization on macro-F1 performance.

A.1.5 Model Explanations

We also conduct various word attributions-based explainability experiments on our BERT-cased methodological classifier to understand our model's decisions. Figure 6 illustrates the word attributions of the same advertisement from a vendor, "pckabml", generated through the captum (Kokhlikyan et al., 2020) and transformers-interpret (Pierse, 2021) frameworks. As can be seen, despite the ads being the same, different explainability frameworks generates different word attributions causing inconsistency in our explanations.

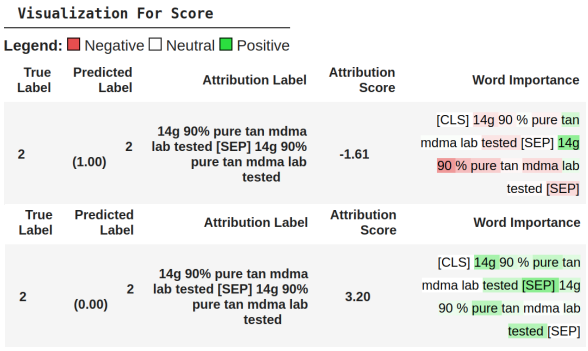


Figure 6: Inconsistency in model explanations within different explainability frameworks.

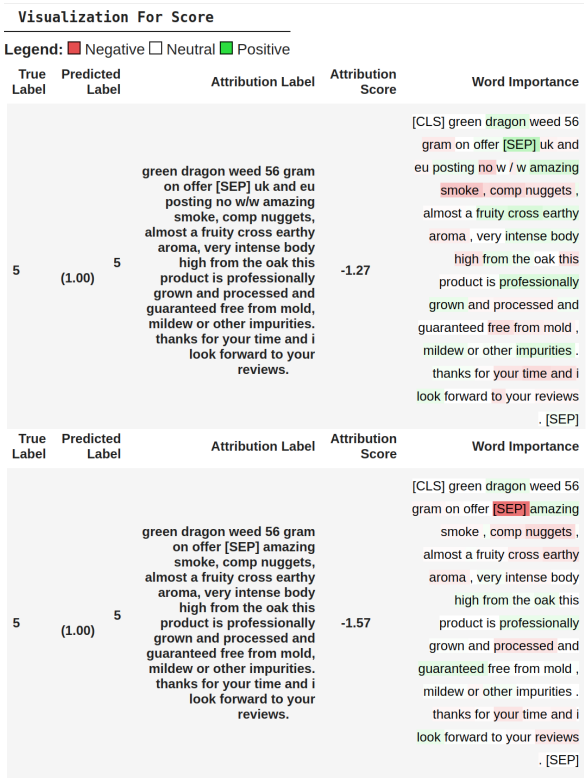


Figure 7: Inconsistency in model explanations for similar ads from the same vendor.

On the other hand, figure 7 illustrates the captum-based word attributions for similar ads from a vendor, "uridol". As can be seen, despite the similarity in ads and generating explanations from the same framework, we get different word attributions causing inconsistency in our explanations. We believe that computing the word attributions through the [CLS] token instead of the entire advertisement could be one of the reasons for these inconsistencies. While we do not clearly understand the reasoning behind the discrepancy in our explanations, we plan to investigate it in the future.

A.2 Infrastructure & Schedule

Data: We perform our experiments using the standard splitting ratio of 0.75:0.05:0.20 ratio for the train, validation, and test dataset.

Training: We perform the training and evaluation of our Neural Networks on a single Tesla V100 GPU with 32 GBs of memory. The training and evaluation of statistical classifiers are performed on a server with one Intel Xeon Processor E5-2698 v4 and 512 GBs of RAM. Finally, we train our distilled transfer-BiGRU model for the Low-Resource setting on a GeForce-MX110 graphic card with 2 GBs of memory.
We use Adam optimizer with $\beta 1 = 0.9$, $\beta 2 = 0.999$, L2 weight decay of 0.01, and a learning rate of 0.001 with warm-up over the first 500 steps, and a linear decay.

Architectures & Hyperparameters [9]: We train all our statistical models using unigrams and bi-grams features and balanced class weights. We experiment SVMs with both linear and Radial basis function (RBF) kernels, Random Forest with n_estimators of 100 and 1000, max_depth of 5, 10, and 20, and MLP with 100 layers and 100 neurons each. Finally, we evaluate our statistical models on the test dataset using a 5-fold nested cross-validation technique.

Our CNN architecture operates on sequences of n-grams characters extracted from the Darknet ads. We then pass the extracted embeddings through six convolutional with max-pooling and three fully connected layers. Inspired by (Zhang et al., 2016), we kept the input length to 1,014, dropout to 0.5 for the fully connected layers with 768 neurons each, a kernel size of 7 in the first two convolutional layers and 3 for the remaining layers. Finally, we set the filter size to 32 and train our models with a batch size of 32 until convergence.

The RNN architecture contains a two-layer Bidirectional-GRU model with two fully connected layers and fasttext embeddings. We first pack and pad the input sequence with variable length through a PyTorch function and then pass it to the embedding layer. After generating the text representation from the Bi-GRU layers, we finally pass the output through a softmax layer and perform classification over it. After some experimentations, we set the

[9] All the models are implemented in python (Van Rossum and Drake Jr, 1995) using Sklearn (Pedregosa et al., 2011), PyTorch (Paszke et al., 2019), and Hugging-face (Wolf et al., 2020) frameworks.

number of hidden units to 768, dropout to 0.65, batch size to 32, and trained the model until convergence.

Finally, we train several transformers models (BERT-base-cased, BERT-base-uncased, RoBERTa-base, and DistilBERT-base-cased) with a sequence classification head on top at a batch size of 32 [10] for 40 epochs (due to computational reasons) for the architectural baselines and till convergence for the methodological baselines. We also train a BERT-base-uncased model on the language task for 20 epochs. All the transformer-based architectures are initialized from a pre-trained model checkpoint.

Computational Details: Tables 13 and 14 presents details about the number of trainable parameters and execution time for all the trained models in the architectural and methodological baselines.

Models (trained on Dreams data)	Trainable parameters	Training time in hrs.
Multinomial Naive Bayes	-	53:56
Random Forest	-	68:27
Logistic Regression	-	79:42
SVM	-	81:08
MLP	-	94:18
Character-CNN	16M	0:54
GRU-Fasttext	39M	1:12
BERT	110M	25:14
RoBERTa	125M	23:40
DistilBERT	68M	17:57

Table 13: Number of trainable parameters and training time for architectural baselines.

Models (trained on Alphabay-Dreams -Silk Road dataset)	Trainable parameters	Training time in hrs.
BERT-uncased	111M	67:02
BERT-cased	112M	66:58
DarkBERT-LM	108M	156:14
DarkBERT Classifier	112M	49:39
Adapter BERT	4M	51:00

Table 14: Number of trainable parameters and training time for methodological baselines.

[10] The maximum batch size allowed by our resources without running into memory issues.

Evaluation Metrics: We evaluate our trained classifiers against accuracy, micro-average F1, and macro-average F1 (commonly known as macro-F1 and micro-F1) using the classification report from scikit-learn. We argue that macro-F1 computes the score independently for each class and then takes the average (treating majority and minority classes equally). Given the class imbalance we have in our dataset, we heavily emphasize our trained models' performance on macro-F1 scores. Furthermore, we evaluate the BERT-base language model on loss and perplexity. Finally, we use Centered Kernel Alignment (CKA) to evaluate and compute correspondences between our methodological baseline representations before and after finetuning.

A.3 CKA Algorithm

Algorithm 2: Computing CKA similarity between layers of BERT classifier

Data: Alphabay (A), Dreams (D), and Silk Road-1 (S)
Input: len(A), len(D), len(S) > 1
Output: CKA similarity

similarity = []
$X \leftarrow A + D + S$
$N \leftarrow len(X)$
Def CKA (Emb_A, Emb_B) :
    /* Embedding shape :- (N, 13, 512, 768) */
    /* Extracting embeddings from the CLS token */
    $\alpha \leftarrow CLS(Emb_A)$
    $\beta \leftarrow CLS(Emb_B)$
    $CKA_{RBF}(\alpha\beta) \leftarrow \frac{\langle K_\alpha, K_\beta \rangle_\mathcal{F}}{||K_\alpha||_\mathcal{F} ||K_\beta||_\mathcal{F}}$
    return $CKA_{RBF}(\alpha\beta)$
/* Extracting embeddings for the Darknet ads before and after training of BERT classifier */
$Emb_A \leftarrow BERTClassifier_{before}(X)$
$Emb_B \leftarrow BERTClassifier_{after}(X)$
/* Computing similarity between layers :- 13x13 matrix */
$CKA_{Layers} \leftarrow CKA(Emb_A, Emb_B)$

ACL 2023 Responsible NLP Checklist

A For every submission:

☐ A1. Did you describe the limitations of your work?
Left blank.

☐ A2. Did you discuss any potential risks of your work?
Left blank.

☐ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☐ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☐ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Left blank.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C ☐ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Left blank.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Left blank.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Left blank.

D ☐ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Left blank.