eDarkFind: Unsupervised Multi-view Learning for Sybil Account Detection

Ramnath Kumar
Birla Institute of Technology and Science
Hyderabad, India
ramnathkumar181@gmail.com

Shweta Yadav
Wright State University
Dayton, Ohio, USA
shweta@knoesis.org

Raminta Daniulaityte
Wright State University
Dayton, Ohio, USA
raminta.daniulaityte@wright.edu

Francois Lamy
Mahidol University
Thailand
flamy1978@gmail.com

Krishnaprasad Thirunarayan
Wright State University
Dayton, Ohio, USA
t.k.prasad@wright.edu

Usha Lokala
Wright State University
Dayton, Ohio, USA
usha@knoesis.com

Amit Sheth
University of South Carolina
Columbia, South Carolina, USA
amit@sc.edu

ABSTRACT

Darknet crypto markets are online marketplaces using crypto currencies (e.g., Bitcoin, Monero) and advanced encryption techniques to offer anonymity to vendors and consumers trading for illegal goods or services. The exact volume of substances advertised and sold through these crypto markets is difficult to assess, at least partially, because vendors tend to maintain multiple accounts (or Sybil accounts) within and across different crypto markets. Linking these different accounts will allow us to accurately evaluate the volume of substances advertised across the different crypto markets by each vendor. In this paper, we present a multi-view unsupervised framework (eDarkFind) that helps modeling vendor characteristics and facilitates Sybil account detection. We employ a multi-view learning paradigm to generalize and improve the performance by exploiting the diverse views from multiple rich sources such as BERT, stylometric, and location representation. Our model is further tailored to take advantage of domain-specific knowledge such as the Drug Abuse Ontology to take into consideration the substance information. We performed extensive experiments and demonstrated that the multiple views obtained from diverse sources can be effective in linking Sybil accounts. Our proposed eDarkFind model achieves an accuracy of 98% on three real-world datasets which shows the generality of the approach.

CCS CONCEPTS

• **Security and privacy** → **Web application security**; • **Networks** → *Online social networks*; • **Computing methodologies** → Machine learning algorithms.

WWW '20, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7023-3/20/04.
https://doi.org/10.1145/3366423.3380263

KEYWORDS

Darknet Market, Drug Trafficker Identification, Multi-view Learning, Stylometry, Sybil Detection, Correlation Analysis

ACM Reference Format:
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), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3366423.3380263

1 INTRODUCTION

Over the last few years, "darknet markets" or "crypto markets" have been playing a substantial role in the distribution of illicit substances and goods [21]. These crypto markets are located in the "Darknet", a subsection of the Deep Web accessible only through specific browsers (e.g., The Onion Router (TOR) or Invisible Internet Project (I2P)). Crypto markets are online marketplaces hosting several vendors, using crypto currencies (e.g., Bitcoin, Monero, Litecoin) and advanced encryption techniques to offer anonymity to vendors and consumers trading in illegal goods or services [5]. Approximately 62% of these crypto market posts were dedicated to psychoactive substances (i.e., pharmaceuticals, drug-related chemicals, or illicit drugs) (Europol 2017). Our study focuses on synthetic opioid, heroin, and other opioid-related advertisements on the crypto market. Our focus on opioid-related advertisements have significant public health implications as opioid-related morbidity and mortality have increased significantly in the US and other countries [20]. Tracking and analyzing opioid-related ads on the crypto markets could help identify emerging substances and aid in detection of potential shifts in illicit opioid availability trends and preferences. However, to generate more accurate and reliable information about the trends and shifts in the availability and marketing of illicit drugs on crypto markets, it is important to have tools to understand vendor behaviors as some vendors may maintain multiple accounts (or Sybil accounts) within and across different crypto markets. It

is necessary to link these different accounts to accurately evaluate the volume of substances advertised across the different darknet markets. Unfortunately, it is extremely labor-intensive to manually explore and link various accounts owing to the increasing number of the darknet markets. An example of the darknet listing is shown in Fig-1. Hence, developing novel techniques that can automatically connect various accounts of the same vendors on darknet markets is extremely desirable.

To address this problem, existing methods [3, 16] utilize the stylometric analysis to link the Sybil accounts to the same vendor according to the writing styles. Some of the recent studies [35, 39] have also exploited the vendor's distinct photography styles by analyzing the product images. This concept is driven by the fact that vendors on Darknet often have to take photographs of their products (instead of using inventory pictures) to demonstrate ownership of illegal goods or stolen items. Such images may represent the personal photography style of a vendor which provide essential cues to identify the Sybil accounts.

The recent research [39] leverages the information from both of these modalities (product description and image) to link the vendors across the multiple accounts. Their study shows that the product images, in addition to the stylometric features, can help in boosting the performance of the model. Most of these sophisticated methods are, however, based on the Deep Learning framework. Unfortunately, impressive performance only come with the massive amounts of annotated data and high training time.

Our approach is unique in that we create an unsupervised approach that allows a cross-domain analysis of the market. Previous approaches have failed to work with cross-domain data and focus on splitting a given vendor into different parts. Although they can be certain of the ground-truth label, their model is highly influenced by the distribution of the data after the split, which may not represent the real-world examples [35][39]. Instead, we take real-world data from the same vendor across different sites to conduct our study. Our model does not depend on how the data is split. This concept allows the model to perform well on a larger scale and is tailored to work on real-world examples, unlike previous approaches which modified the dataset to create a test set. Furthermore, ours is the first approach to integrate domain-specific knowledge with vendor classification. Moreover, the simplicity of an unsupervised model is more than sufficient for the task on hand.

In this paper, we propose to achieve the author attribution task by employing an unsupervised multi-view learning framework (named *eDarkFind*) dedicated to the crypto market environment. Multi-view learning uses multiple views to improve learning performance. Multiple views can be obtained from heterogeneous sources, multiple feature extractors, and feature transformation. These multiple views describe the distinct, diverse, and complementary information of the data. Multi-view learning is an ideal learning mechanism for the data from real applications where examples are characterized by distinct (often orthogonal) feature sets (views). For example, the vendor profile on a social networking site can be described by their social network, comments, status, and other activities.

We characterize vendors using five *metaviews*. We start with the pre-trained language representation networks (Bidirectional Encoder Representations from Transformers (BERT) [12]) to generate the first view. These models are highly efficient in generating the task-agnostic input representation from the transformer architecture [34]. This enables even the low-resource tasks to benefit from deep bi-directional architectures [12]. For the second meta-view, we obtain domain-specific contextual information from all the listings by the vendor. This is achieved by exploiting the technique of the *ParagraphVector* proposed by [25]. Since the pre-trained language representation was trained on a generic corpus, it is dense and not ideal for modeling sentences in the drug domain. To capture domain-specific contextual information and cater to multi-lingual data, we try to capture domain-specific embedding.

We extract the stylometric features, content-based features, *Location information* and *Substance information*. Location information such as *shipping information* of a vendor can provide important clues to resolve the vendor's identity. The Drug Abuse Ontology (DAO) [24] can be used to extract the substance information. This information can be used to compute the similarity among vendors (authors). Finally, these views are fused using the Generalized canonical correlation analysis (GCCA) technique discussed in Section 4.4.

The major contributions of this work are as follows:

(1) We propose a robust unsupervised method for the Sybil account detection on the real-life Darknet market dataset. Our system exploits the capabilities of multi-view learning to capture diverse and complementary information of the data. To the best of our knowledge, this is the first attempt to study the effect of multi-view learning for Sybil account detection task. Furthermore, our unsupervised approach can predict the similarity of vendors with an accuracy of 98%.
(2) Unlike previous approaches, we performed cross-domain analysis and were able to test our approach on real-world datasets.
(3) We validate our proposed methodology on 3 different DarkNet Market datasets, namely *Dream Market*, *Tochka*, and *Wall Street* and our model consistently outperforms all other defined baselines.
(4) Our model is the first to classifying vendors using a domain-specific knowledge graph of drugs such as the DAO.

2 RELATED WORKS

In this section, we describe prior studies that are broadly related to the DarkNet Market Data Analysis and authorship-attribution task (Authorship attribution task is similar to the current Dark Web vendor classification task). Below, we divide prior research under two prominent subheadings: (i) DarkNet Market Data Analysis, and (ii) Author attribution.

2.1 DarkNet Market Data Analysis

In the recent past, several research efforts have tackled drug trafficking in darknet markets. Broséus et al. [7] conducted a preliminary study on the Canadian drug market. They collected the data from 8 crypto markets listing 4000 products and 200 vendor profiles. They further investigated the structure and organization of the online market through a combined analysis of vendor names and their corresponding public PGP keys. This research provides information on how vendors across crypto markets diversify and replicate, revealing vendor behavior. Dittus et al. [13] performed an extensive

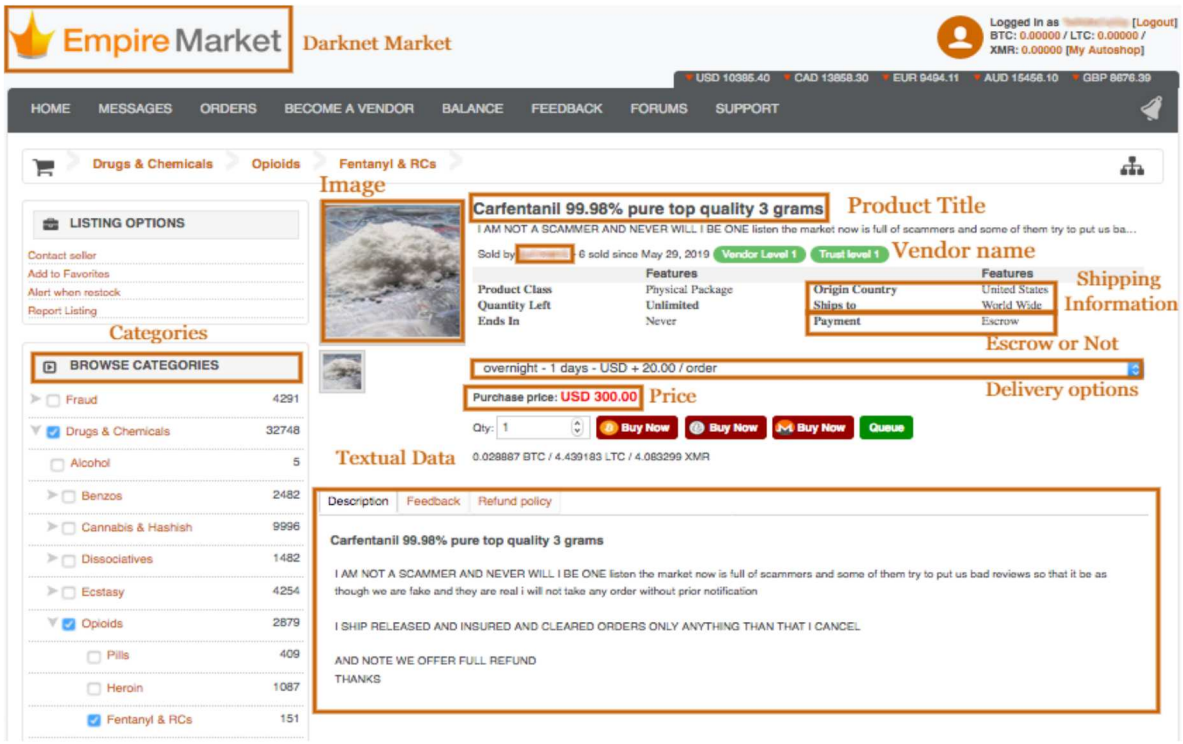

Figure 1: Snapshot of Darknet Market

empirical survey to analyze the economic geography of the darknet market, and provided strong evidence that vendors are mostly concentrated in consumer countries. Ho and Ng [17] investigated the role of stylometric features that work at the character level, word level, sentence level, and paragraph level. They uncovered that writing style could be used to attribute and correlate authorship between vendors on multiple Dark Web forums with high accuracy. Apart from capturing the stylometric information, some studies have also exploited the image embedded in the vendor sites. [35] utilized a deep neural network and transfer learning framework to extract distinct features from a vendor's photos automatically. They discovered that image-based approach outperforms existing stylometry based techniques in both accuracy and coverage. Another study conducted by [39] integrated both the stylometric and the photography styles using an attributed heterogeneous information network (AHIN).

2.2 Author attribution (AA)

Authorship attribution is the task of identifying an unknown text's author among a set of candidate authors. It has applications ranging from plagiarism detection to Forensic Linguistics. The critical insight behind statistical authorship attribution is that measuring textual features enables distinguishing between texts written by different authors. Several natural language processing techniques have been developed for AA, mainly using machine learning techniques in concert with hand-crafted features. These features vary from naive bag-of-words model-based features which provide content divergence indicator between authors to the stylometric features which capture the distinctive writing patterns of an author. Mosteller and Wallace [26] first established the role of the stylometric features by showing that the frequency of common words can be a pivotal clue to distinguish the authors. Researchers subsequently proposed new attributes to characterize writing styles [19, 31] after this seminal study. Popular features used for this task include statistical properties of words and characters (such as word/character frequency, the average length of the word, long term correlation and the richness of vocabulary) [15].

Recently, various deep learning methods have been proposed for the AA task based on the Convolutional Neural Network (CNN). Ruder et al. [27] proposed a novel CNN based AA technique that combines both word and character channels. This model has also been tested on social media such as Twitter and Reddit datasets. Ferracane et al. [14] integrated discourse features into state of the art character bi-gram CNN classifiers to boost their performance. Shrestha et al. [28] used the sequence of character n-grams as input to the CNN model for author attribution of short texts. Zhang et al. [38] encoded a syntax tree into a learnable distributed representation and fed these as inputs to the CNN model. Another study, conducted by [32], utilized SVM, and replaced the softmax as an activation layer to improve the performance. The best performing system in PAN 2015 [4] developed a multi-head Recurrent Neural Network character language model that provides the set of next character probabilities for each author at every step of the model.

3 DATASET

The dataset was obtained using eDarkTrends, a semi-automated platform, to extract posts about fentanyl, fentanyl analogs and other non-pharmaceutical synthetic opioids in the crypto market. Data sources include three different crypto markets:

- **Dream Market:** The market was established in late 2013. Before 2017, Dream Market was the second-largest darknet market in the world after AlphaBay. However, with the shutdown of AlphaBay in 2017, Dream Market soon became the biggest darknet market in the world [1]. The market had a total of 261 withdrawals between November 2014 and April 2019. The market had almost $197,589.12 worth of transactions during this period [36].
- **Tochka:** The market was established in 2015. It is a very small market that operates predominantly in Europe and North America. The site sells more than 3621 products, including drugs, malware, and others. The market rebranded itself as the Point market and operates under this name currently [29]. The market had a total of 2,990 withdrawals between November 2014 and April 2019. The market had almost $5,072.33 worth of transactions during this period [36].
- **Wall Street:** The market offered a platform for selling illegal drugs, weapons, hacking software, and stolen login credentials. However, the market has been suffering from the exit scam since April 2019 [2]. The admins switched the platform into maintenance and transferred the customers' funds reportedly stealing a total of $14 million to $30 million worth of XMR and bitcoins from vendor accounts [37]. The market was later taken down in May 2019. Wall Street was the second-largest darknet market in the world before being seized in May 2019 by German Federal Criminal Police, under the authority of the German Public Prosecutor office. The market had a total of 7,755 withdrawals between November 2014 and April 2019. The market had almost $18,729.40 worth of transactions during this period [36].

Extracted data included various features about vendors including product name, vendor screen name (vendor name), drug category, product description, price (Bitcoin or US$), country/region of origin and destination, and others. Extracted crypto market data were further processed to extract relevant drug mentions using Drug Abuse Ontology (DAO) [8], [22]. One of the key benefits of using an ontology-enhanced semantic approach is the ability to identify all variants of a concept in data (e.g., generic names, slang terms, and scientific names). DAO contains names of psychoactive substances such as heroin, fentanyl, and also synthetic substances such as U-47,700, MT-45. They also contain brand and generic names of pharmaceutical drugs such as Duragesic and fentanyl transdermal system and slang terms such as Roxy and Fent. It also contains information regarding the route of administration (e.g., oral, IV), unit of dosage (e.g., gr, gram, pint, tablets), physiological effects (e.g., dysphoria, vomiting) and substance form (e.g., powder, liquid, HCl). Further details regarding the pre-processing have been mentioned in [24]. We combine all vendors based on the vendor name and create vendor name level data. The result is a dictionary where the key "vendor name" leads to the characteristics of the vendor such as substance, ships from, ships to, and other textual data including the product description, terms and conditions, and rating data. The intuition behind including rating data lies in the fact that branding and ratings can enable detection of aliases meant to provide deceptive/fake good reviews.

Our dataset has 1992 unique vendors, of which 1448 vendors operate on Dream Market, 408 vendors operate on Tochka and 466 vendors on the Wall Street market. Using the above datasets, we form three annotated datasets that provide pairs of vendors that are the same but appear in multiple markets. Our intuition is that the vendor name in the crypto market plays the role of their brand. Hence the vendors on different sites having the same vendor name can be assumed to be the same vendor with high confidence. We created four different datasets as follows:

- **Dream_Tochka:** This dataset contains vendors who were present in both Dream Market and Tochka.
- **Dream_Wallst:** This dataset contains vendors who were present in both Dream Market and Wall Street.
- **Tochka_Wallst:** This dataset contains vendors who were present in both Wall Street and Tochka.
- **All:** This dataset is created using vendors who were present in at least two different markets.

A view of all datasets are the same and look as shown in Table-1. A brief overview of the sites which we include in our dataset are shown in Table-2 and Table-3.

Ethics of Data Analysis: We follow the standard ethical practices to analyze our datasets [11] [30]. The dataset contains only the publicly available information. We cannot personally identify the vendors. Furthermore, our dataset does not include any interaction with human subjects. Our dataset does not contain any images as per our data usage safety agreement. Our research provides a novel approach to detect similar accounts (Sybil accounts) by combining ideas from author attribution techniques, synthesizing multiple views and exploiting domain-specific knowledge. The benefits of our research severely outweigh the potential risks.

4 METHODOLOGY

4.1 Task Definition

The task involves the detection of similarity between two vendors on online forums, i.e., Darknet, Reddit, and Twitter. Let us denote the set of n vendors V as $\{v_1, ..., v_n\}$, where each vendor v_i can be associated with a subset of m sites $S = \{s_1, ..., s_m\}$. Given any two vendors v_a and v_b associated with the respective sites s_i and s_j, our goal is to develop a similarity measure $sim(v_a^{s_i}, v_b^{s_j})$ between the two vendors using various views.

4.2 Summary of Proposed Approach

Given a vendor v_i for the site s_j, our model leverages various features such as pre-trained representation, contextual, stylometric, substance-related, image-related and others features to produce a learnable distributed embedding U_{ij}. We use the textual information available from the given vendor v_i to extract the contextual and stylometric features of the vendor. We use the knowledge graph to extract the class of drugs the vendor sells and create embeddings for the same. We use the location from which the drug was shipped to

Table 1: Example case of similar vendors on different sites

vendorname	Text	Location	Substance
@vendor1	This listing is for CANADA & USA ONLY USA Customers....	Norway <END>U.S.A.	Oxycodone
@vendor2	This listing is for CANADA ONLY –Limited Time–....	Norvege <END>US	Tramadol

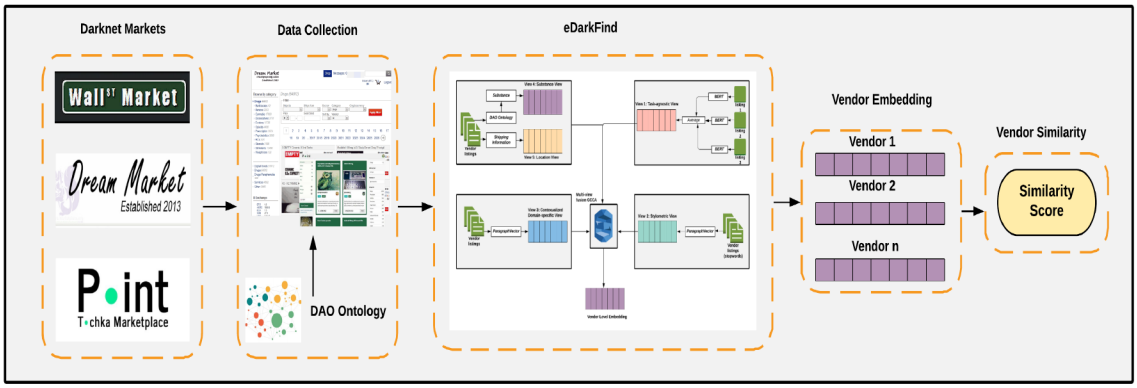

Figure 2: Overview of the proposed model architecture.

Table 2: Description of the Markets

Marketplace	Number of Withdrawal Transactions	Bitcoin	USD Equivalent
Dream	261	99.1503695	$197,589.12
Tochka	2,990	0.70483642	$5,072.33
Wall Street	7,755	2.572515	$18,729.40

Table 3: Summary of Dataset

Data	*Dream Market*	*Tochka*	*Wall street*	*All*
Unique # Vendors	1448	408	466	1992
Unique # Substance	852	313	290	1148
Unique # Location	356	44	29	389
Unique # Descriptions	16800	1829	1723	18472

understand the region of operation better. All these views (features) are merged using an unsupervised approach, such as the weighted generalized canonical correlation analysis [18]. The summary of creating vendor level embedding is shown in Fig-2. The eDarkFind model is represented in Fig-3.

4.3 Vendor Embedding

We aim to generate the vendor embedding by incorporating various views to model the association of the vendor with the corresponding site. These views are fused using CCA to obtain a single representation, which we call Vendor embedding. In the following subsection, we discuss the generation of each view in detail.

4.3.1 ***Task-agnostic View.*** We obtained the first view[1] from the pre-trained language representation networks. In this work, we use the Bidirectional Encoder Representations from Transformers (BERT) [12] to generate the representation of the sentences appearing on a site associated with the vendor. These pretrained models are highly efficient in generating the task-agnostic input representation from the transformer architecture [34]. This enables even the low-resource tasks to benefit from deep bi-directional architectures [12] and the unsupervised training framework to obtain the pre-trained network. The sentence representation is usually obtained by aggregating the last few layers of the BERT model. The number of layers you want to aggregate can vary from problem to problem. We noticed that the last four layers produce the best representation of the sentence. To obtain the representation for a vendor v_i, we compile all the textual description from vendor associated site s_j as a list $C_{ij} = \{c_1, ..., c_n\}$ of n unique textual descriptions. The representation e_k of each listing c_k is obtained from the pre-trained network. The final vendor level representation is obtained by computing the mean value of all the representations obtained from the pre-trained network. We generate the two task-agnostic views, one from each pre-trained network.

4.3.2 ***Contextualized Domain-Specific View.*** We generate another view of the vendor by focusing on their domain knowledge. Since the pre-trained language representation was trained on a generic corpus, it is dense and not ideal for modeling sentences in the drug domain. To capture domain-specific contextual information and cater to multi-lingual data, we capture domain-specific embedding as follows.

[1]In this paper view and embedding are interchangeably used.

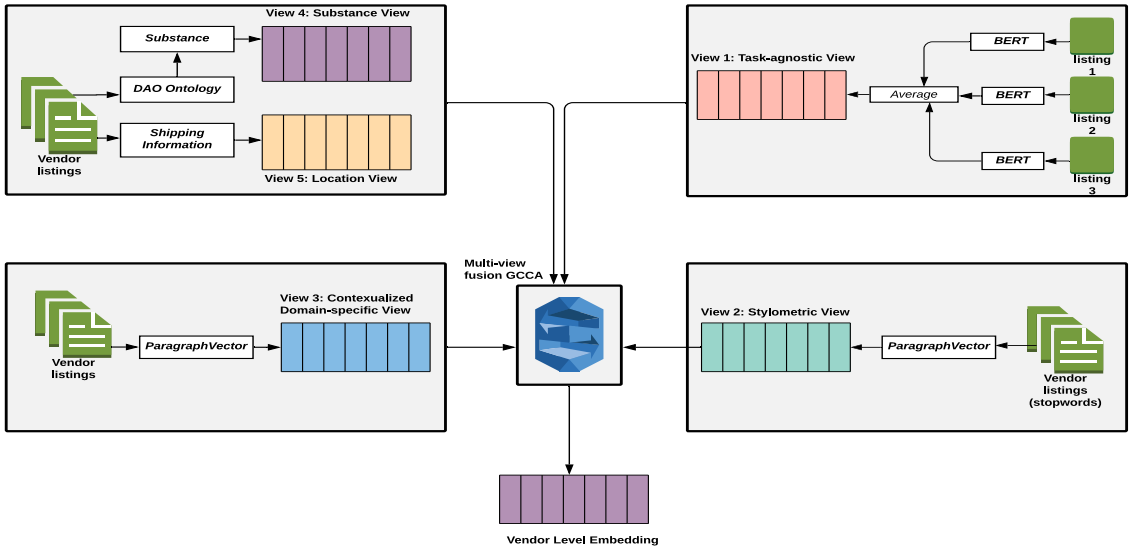

Figure 3: eDarkFind Model

We first collect all textual descriptions (listing) by the vendor and create a document by appending them with a special delimiter <END>. An unsupervised learning method *ParagraphVector* [23] is then applied on this document. This method generates a fixed-sized vector for each vendor by performing the auxiliary task of predicting the words within the documents. The use of *ParagraphVector* allows us to apply it on a document of variable length, which captures the domain-specific vendor level information. The algorithm to compute the domain-specific view is as follows:

Every vendor-document and all words within them are first mapped to unique vectors such that each vector is represented by a column in matrix $D \in \mathbb{R}^{d_s \times N_v}$ and $W_s \in \mathbb{R}^{d_s \times |V|}$, respectively. Here, d_s is the embedding size, N_v is the number of vendors and $|V|$ represents the size of the vocabulary. *Continuous-bag-of-words* approach [25] is then performed where a target word is predicted given the word vectors from its context-window. The key idea here is to use the document vector of the associated document as part of the context words. More formally, given a vendor-document d_i for vendor v_i comprising a sequence of n_i-words $w_1, w_2, ..., w_{n_i}$, we calculate the average log probability of predicting each word within a sliding context window of size k_s. This average log probability is:

$$\frac{1}{n_i} \sum_{t=k_s}^{n_i - k_s} \log p(w_t | d_i, w_{t-k_s}, ..., w_{t+k_s}) \tag{1}$$

To predict a word within a window, we take the average of all the neighboring context word vectors along with the document vector $\vec{d_i}$ and use a neural network with softmax prediction:

$$p(w_t | d_i, w_{t-k_s}, ..., w_{t+k_s}) = \frac{e^{\vec{y}_{w_t}}}{\sum_i e^{\vec{y}_i}} \tag{2}$$

Here, $\vec{y} = [y_1, ..., y_{|V|}]$ is the output of the neural network, i.e.,

$$\vec{y} = \mathrm{U}_d h(\vec{d_i}, \vec{w}_{t-k_s}, ..., \vec{w}_{t+k_s}; D, W_s) + \vec{b}_d \tag{3}$$

$\vec{b}_d \in \mathbb{R}^{|V|}, U_d \in \mathbb{R}^{|V| \times d_s}$ are parameters and $h(\cdot)$ represents the average of vectors $\vec{d_i}, \vec{w}_{t-k_s}, ..., \vec{w}_{t+k_s}$ taken from D and W_s. Finally, after training, D learns the vendors' document vectors which represent their domain-specific view.

4.3.3 ***Stylometric View.*** People possess their characteristic authorship style, which is reflected in their writings. These styles are generally affected by attributes such as gender and influences [10]. Based on a long history of authorship attribution approaches, one innate characteristic worth exploring is the usage of stopwords and special characters. Although the content and the context may vary between documents, their style of usage of the stopwords and special characters are roughly similar. To do so, we extracted only the stopwords from the listing's of the vendor and used the *ParagraghVector* model. The approach is precisely the same as the domain-specific embedding (c.f. Subsection-4.3.2).

4.3.4 ***Location View.*** The location from which the vendor operates is an important feature in our vendor embedding. However, the location information can be expressed in various formats. For example, the United States of America can also be expressed as the USA and Suomi is the Finnish way of saying Finland (an example of multilingual representation). To create a more uniform embedding, we abstracted this information to represent the alpha-2 code for the various countries. In our dataset, we observed 68 such locations from which the vendors were operating. These were then made into a vector L = $\{l_1, ..., l_n\}$ for n locations. To create location level embedding (view) for vendor i, we first create a binary vector B =

Table 4: Performance metric of our model on different datasets.

	Dream_Tochka		***Dream_Wallst***		***Tochka_Wallst***	
Model	*Accuracy*	*F-Score*	*Accuracy*	*F-Score*	*Accuracy*	*F-Score*
BERT (B)	0.12	0.11	1	1	0.03	0.03
Universal Sentence Encoder (U)	0.15	0.08	1	1	0.13	0.18
Domain-Specific (DS)	0.88	0.82	0	0	0.68	0.49
Stylometric(S)	0.88	0.82	0	0	0.97	0.95
Domain-Adapted (B+DS)	0.87	0.87	0.75	0.86	0.9	0.92
Domain-Adapted-Style (B+DS+S)	0.95	0.94	0.71	0.83	0.93	0.93
Substance (Sub)	0.93	0.92	0.52	0.68	0.97	0.95
Location (Loc)	0.95	0.94	0.88	0.94	**0.97**	**0.95**
Final_Model ((B+DS+S)+Sub+Loc)	**0.97**	**0.97**	**1**	**1**	0.93	0.93

$\{F_1^i, ..., F_n^i\}$ where,

$$F_j^i = \begin{cases} 1 & \text{if vendor } v_i \text{ operates from location } l_j \\ 0 & \text{otherwise} \end{cases}$$

This simple approach can only work under the assumption that all the features are of equal importance. However, it is more common that vendors operate from countries such as the US in comparison to Norway or Sweden. Hence, two vendors sharing the same location as Norway are more likely to be similar in comparison to them being from the US. To incorporate this idea into our model, we include the self-information weight into the model. The final vector L = $\{l_1, ..., l_n\}$ for n unique features can be written as

$$L = \{F_1^i * w_1, ..., F_n^i * w_n\} \tag{4}$$

where,

$$w_i = -log(Pr(F_i^{all} = 1)) \tag{5}$$

and *all* indicates the total vendor corpus.

4.3.5 ***Substance View.*** The substance/class of drugs the vendor sells is generally an important feature of the vendor. Like most economic markets, the sellers tend to specialize in a class of drugs and try to dominate the market. This feature was extracted using our Drug Abuse Ontology. In this section, we focus predominantly on creating substance embedding from this information. There are 362 unique substances in our dataset. To reduce the number of dimensions, we grouped them by the class of drugs. This method allows us to create a smaller set of features, specifically 16 classes. A methodology similar to the location embedding was used to obtain the substance view S = $\{s_1, ..., s_{16}\}$.

4.4 Fusion of the Views

We take a multi-view learning approach to combine the various views discussed above into a comprehensive embedding for each vendor. We use Canonical Correlation Analysis (CCA) [18] to perform this fusion. CCA captures maximal information between two views and creates a combined representation. Since we fuse more than two views, we employ an extension of CCA called the Generalized CCA (GCCA) [9]. GCCA finds G_i,U_i that maximizes:

$$arg \min_{G_i, U_i} \sum_i ||G - X_i U_i||_F^2 \tag{6}$$

such that $G^{'}G = I$.

In the above representation, $G \in \mathbb{R}^{n \times k}$ contains all vendor representations, $X_i \in \mathbb{R}^{n \times d_i}$ corresponds to the data matrix for the i^{th} view and $U_i \in \mathbb{R}^{d_i \times k}$ maps from latent space to observable view i [33]. However, since all the views are not equally important, we employ the weighted GCCA (wGCCA) [6]. In this representation, we add a weight term to the above equation as follows:

$$arg \min_{G_i, U_i} \sum_i w_i ||G - X_i U_i||_F^2 \tag{7}$$

such that $G^{'}G = I$ and $w_i \geq 0$ and represents the importance of the i^{th} view in the fusion process.
The columns of G are the eigenvectors of $\sum_i w_i X_i (X_i^{'} X_i)^{-1} X_i^{'}$, and the solution for $U_i = (X_i^{'} X_i)^{-1} X_i^{'} G$.

For obtaining the final feature representation, we used wGCCA to fuse different views as follows:
The **domain adapted embedding (DA)** was created by fusing the task-agnostic embedding (G) and the contextualized domain specific embedding (DS) as follows:

$$DA = G \otimes DS \tag{8}$$

The **domain adapted with style embedding (DAS)** was created by fusing the domain adapted embedding (DA) and the stylometric embedding (S) as follows:

$$DAS = DA \otimes S \tag{9}$$

The final **eDarkFind embedding (eDF)** was created by fusing the domain adapted embedding with style embedding (DAS), location embedding (L), and the substance embedding (Sub) as follows:

$$eDF = DAS \otimes L \otimes Sub \tag{10}$$

Table 5: Performance metric of various models on All sites compiled. '✓' indicates the usage of the operation to fuse the embedding and '-' indicates that the said operation is not used.

Model	*wGCCA*	*concat*	*Precision*	*Recall*	*F-score*	*Accuracy*
BERT (B)	-	-	0.25	0.5	0.33	0.5
Universal Sentence Encoder (U)	-	-	0.76	0.53	0.39	0.53
Domain-Specific (DS)	-	-	0.25	0.5	0.33	0.5
Stylometric (S)	-	-	0.25	0.5	0.33	0.5
Domain-Adapted (B+DS)	-	✓	0.25	0.5	0.33	0.5
Domain-Adapted (B+DS)	✓	-	0.84	0.82	0.82	0.82
Domain-Adapted-Style (B+DS+S)	-	✓	0.25	0.5	0.33	0.5
Domain-Adapted-Style (B+DS+S)	✓	-	0.87	0.84	0.83	0.84
Substance (Sub)	-	-	0.92	0.9	0.9	0.9
Location (Loc)	-	-	0.94	0.93	0.93	0.93
Final_Model ((B+DS+S)+Sub+Loc)	-	✓	0.76	0.54	0.42	0.54
Final_Model ((B+DS+S)+Sub+Loc)	✓	-	**0.98**	**0.98**	**0.98**	**0.98**

4.5 Final Prediction

Our main goal in this paper is to create promising vendor embedding and use simple cosine similarity to compute the vendor similarity as follows:

$$S(V_x, V_y) = cos(\boldsymbol{V}_x, \boldsymbol{V}_y) = \frac{\boldsymbol{V}_x \cdot \boldsymbol{V}_y}{||\boldsymbol{V}_x|| \cdot ||\boldsymbol{V}_y||} \tag{11}$$

where V_j refers to the vendor embedding of individual j.

5 EXPERIMENTAL RESULTS

5.1 Baseline Models

Here we describe the state-of-the-art baseline methods that we compare eDarkFind against. To justify multi-view embeddings, we take all single view embeddings as our baseline model.

- **BERT:** This model uses only the textual data and extracts pre-trained features from BERT. We add the last four layers of BERT to obtain sentence level embedding of the text. We then average these sentence-level embeddings to get the embedding of all sentences by the given vendor.
- **Universal Sentence Encoder:** This model also uses only the textual data and extracts another set of pre-trained features from Universal Sentence Encoder. We then average the sentence level embeddings to get the embedding of all sentences by the given vendor.
- **Domain Specific:** This model has been created by training a *ParagraphVector* model on the textual data available for each vendor. This allows us to create paragraph-level embeddings for the vendor.
- **Stylometric:** This model has been created by running the *ParagraphVector* model on the stylometric data extracted for each vendor.
- **Location:** This model has been created using the *shipping from* information available on most of these sites. We created embedding of the location as described in the previous section.
- **Substance:** This model has been created using the substance information available on most of these sites. We created the embedding of the substance as described in the previous section.

5.2 Evaluation Metrics

As mentioned earlier, our intuition is that the vendor name in the crypto market plays the role of their brand. Hence the vendors on different sites having the same vendor name can be assumed to be the same vendor with high confidence. Hence, we label these pairs as positive and the others as negative. Such a solution negates the need for splitting the user into two equal parts for training and evaluation as used by previous approaches [35, 39]. In the above setting, we evaluate the cosine similarity between the two vendors and label that pair as positive if the similarity is greater than 0.5 and negative otherwise. We have chosen the optimal threshold value by considering the system performance in terms of F-Score and Accuracy on the validation-dataset for a number of threshold-values and discovered 0.5 to be the sweet-spot. We compute the mean-average F-score/Accuracy-(all models) vs the threshold and notice that the range 0.4-0.6 is a plateau and is not biased by any means. Hence, we chose 0.5 as the threshold for our models.

On each model, we compute various metrics such as precision, recall, F-score and accuracy to compare different models.

5.3 Results

Table-4 presents the performance of our model on various metrics using our datasets. eDarkFind manages to outperform all other chosen baselines and shows significant improvement on all datasets.

The lowest performance is achieved by a simple single view embedding of textual data. This model includes task-agnostic, domain-specific, and stylometric embeddings. Furthermore, a single view embedding of location and substance performs even better. However, our approach of merging the substance, location and textual embedding brings a significant improvement in our model.
In Table-5, we present the performance on the dataset comprising of all three sites and by using the '*wGCCA*' (weight GCCA) and '*concat*' operation to fuse the embedding. The results shows that the fusion of the views using the wGCCA operation is better than the simple concatenate operation. We also evaluated our model with other metrics such as AUC and observed that the eDarkFind model outperformed other models and achieved a score of 0.99. In Fig-4, we present the ROC curve of various models for further comparison by readers.

5.4 Ablation Study

Now We discuss the performance only on the comprehensive dataset represented in Table-5. We experimented with multiple variants of this model to analyze the importance of various features present in our architecture. We noticed that single view embedding of textual data performs poorly. The sites in Dream_Wallast and Tochka_Wallast dataset have very different structure and content. Therefore, when we compare the features of one site related to "product-descriptions" with that of another site related to "rating" and obtain orthogonal features, we observe performance differences. Furthermore, pre-trained embeddings fail to distinguish between textual-data in our dataset well since preponderance of data is on the topic of drugs that forces the model to give high similarity if two texts describe drugs.

However, by merging the views of domain-specific embedding and pre-trained embedding, our domain-adapted embedding outperforms both the individuals' components by almost 32%. Furthermore, the infusion of stylometric features into our model further improves the performance of our model by 2%. Although substance and location alone were able to outperform our textual model, the infusion and integration of the three improve it by almost 6%. This improvement is very impressive since it gets progressively harder to improve a model once it reaches a high (saturating) performance. Our final model shows an improvement of almost 48% compared to our initial models.
We also experimented with the concatenation of multiple views and observed that the concatenation model performs poorly compared to the wGCCA. This can be shown from the fact that our final model outperforms it by almost 44%.

5.5 Domain Specific Error Analysis

A small sample of product ads/descriptions were examined and compared across two different crypto markets by substance use researchers. The following observation were made:

- It appears that vendors tend to re-use their advertisement texts across different platforms when selling the same type of substances.
- Some users tend to use the same slang terms across various listings (e.g., horse for heroin).
- Differences in the textual data among vendors stem from the inclusion of additional customer reviews. Some differences in product ads/descriptions across different platforms arose from the variation in product reviews allegedly written by clients and added to the product advertisements to bolster client confidence.
- Various sites have various fields, and there is no uniform feature set for all. Hence, there are cases where our model is forced to compare product descriptions with rating data. This adds unnecessary noise to our model. Such situations arise since our model tries to use all available relevant information.
- Most vendors tend to specialize in a subclass of drugs. This practice is a common economic strategy in the market and is widely known as a competitive advantage. Thus, allowing the vendor to dominate or compete in only a subclass of drugs. Hence, we see a high bias towards fields such as substance and the feature alone models the vendor very well.
- Some vendors may have the same textual data but may operate from different locations or may even be selling different drugs.

5.6 Case Studies

Most vendors tend to copy their textual data between sites, and a complex model is not required to capture the similarity between them. Upon closer inspection of our dataset by domain experts, we found the following insightful examples about the working of the crypto market.

- The two vendors share the same vendor name but operate on different sites. Both entities have not entered any product description. The only textual data available for both of them are their rating data. Both share the same rating data, even the timestamp. This is an obvious example of branding using bots. Such cases motivated us to include the rating data in our model.
- Due to differences in the structure of the sites, there may be some noticeable differences. Some vendors avoid certain fields on different sites. Such cases forces us to sometimes compare product descriptions of a vendor with the rating data of the vendor. Such cases are an anomaly and usually, lead to wrong predictions.
- Some vendors replicate their style of the text. This may include various emojis and special characters. This includes emojis such as ôö, ñÜ,Üí, âá and many more which are shown here in UTF-8.

The above cases have been illustrated in the Table-6.

6 CONCLUSION

In this paper, we presented an unsupervised model based on the concept of multi-view learning which can be used to fingerprint vendors and track them across different sites. Our model leverages and judiciously combines a variety of features such as stylometric features, domain-specific contextual features, location, and substance features to give a vendor level multi-view embedding. By evaluating the dark find on real-world datasets, we have shown that our model achieves an accuracy of 98%. The results demonstrated

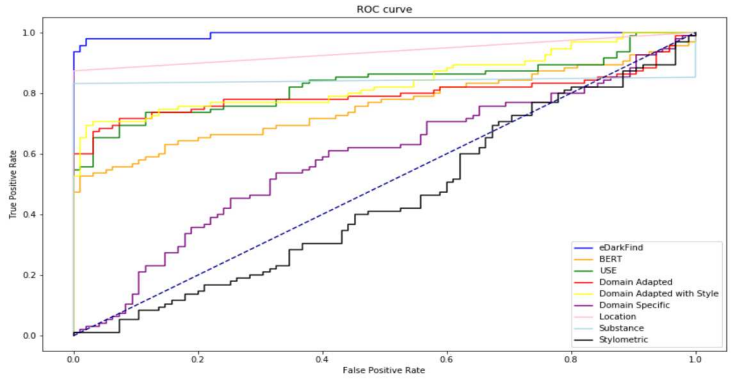

Figure 4: ROC curve comparison between true positive rate & false positive rate over the baselines and proposed models.

Table 6: Use Case examples in our dataset. The highlighted text shows that the vendors use the near-same text pattern or style on two different sites.

Case Study	@Vendor 1	@Vendor 2
Branding	5//02/14 09:49 am,5/Thanks alles schick/11/10 01:46 pm, <END>Tilidin 50MG/4MGOriginal Apothekenware	5//02/14 09:49 am,5/Thanks alles schick/11/10 01:46 pm, <END>Tilidin 50MG/4MGOriginal Apothekenware <END> 5/Thanks alles schick/11/10 01:46 pm,
Comparing product description and rating since the vendor did not enter product description in other site	Percocet Oxycodone 5/325mg 200 TabletsFinalize Early and get 20 Free bonus sent for a total of 220!US Made Mallinckrodt 5mg/325 (made in St. Louis, Miss. USA) ...	5//02/07 01:03 pm,5/Thanks Again. A++/01/21 11:49 pm,5/Trustworthy/01/16 12:22 pm,4.33//01/07 08:50 am,5/Great communication, trustworthy, and overdelivered./12/31 11:09 pm,5//11/29 03:25 pm,5/FAST A+++ Best Stealth I've seen yet. ...
Similar stylometric features captured by the use of special characters or emojis	——————————— ——————————— ****** NEWS 25.12.2018 NEWS ****** ——————————— ——————————— We ship all new ...	——————————— ——————————— PRODUCTS ——————————— ——————————— AFGHAN HEROIN A+++COCAINE #3 ...

by our model on different datasets demonstrate that our model outperforms other alternative approaches to this task.

REFERENCES

[1] 2017. Dark Web Users Suspect "Dream Market" Has Also Been Backdoored by Feds. https://thehackernews.com/2017/07/dream-market-darkweb.html

[2] 2019. Dark web marketplace Wall Street Market busted by international police. https://nakedsecurity.sophos.com/2019/05/07/dark-web-marketplace-wall-street-market-busted-by-international-police/

[3] Sadia Afroz, Aylin Caliskan Islam, Ariel Stolerman, Rachel Greenstadt, and Damon McCoy. 2014. Doppelgänger finder: Taking stylometry to the underground. In *2014 IEEE Symposium on Security and Privacy*. IEEE, 212–226.

[4] Douglas Bagnall. 2015. Author identification using multi-headed recurrent neural networks. *arXiv preprint arXiv:1506.04891* (2015).

[5] Monica J Barratt and Judith Aldridge. 2016. Everything you always wanted to know about drug cryptomarkets*(* but were afraid to ask). *The International journal on drug policy* 35 (2016), 1.

[6] Adrian Benton, Raman Arora, and Mark Dredze. 2016. Learning multiview embeddings of twitter users. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 14–19.

[7] Julian Broséus, Damien Rhumorbarbe, Caroline Mireault, Vincent Ouellette, Frank Crispino, and David Décary-Hétu. 2016. Studying illicit drug trafficking on Darknet markets: structure and organisation from a Canadian perspective. *Forensic science international* 264 (2016), 7–14.

[8] Delroy Cameron, Gary A Smith, Raminta Daniulaityte, Amit P Sheth, Drashti Dave, Lu Chen, Gaurish Anand, Robert Carlson, Kera Z Watkins, and Russel Falck. 2013. PREDOSE: a semantic web platform for drug abuse epidemiology using social media. *Journal of biomedical informatics* 46, 6 (2013), 985–997.

[9] J Douglas Carroll. 1968. Generalization of canonical correlation analysis to three or more sets of variables. In *Proceedings of the 76th annual convention of the American Psychological Association*, Vol. 3. 227–228.

[10] Na Cheng, Rajarathnam Chandramouli, and KP Subbalakshmi. 2011. Author gender identification from text. *Digital Investigation* 8, 1 (2011), 78–88.

[11] 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*. ACM, 213–224.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[13] Martin Dittus, Joss Wright, and Mark Graham. 2018. Platform Criminalism: The'last-mile'geography of the darknet market supply chain. In *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee, 277–286.

[14] Elisa Ferracane, Su Wang, and Raymond Mooney. 2017. Leveraging discourse information effectively for authorship attribution. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 584–593.

[15] Graeme Hirst and Olǵa Feiguina. 2007. Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing* 22, 4 (2007), 405–417.

[16] Thanh Nghia Ho and Wee Keong Ng. 2016. Application of stylometry to darkweb forum user identification. In *International Conference on Information and Communications Security*. Springer, 173–183.

[17] Thanh Nghia Ho and Wee Keong Ng. 2016. Application of Stylometry to DarkWeb Forum User Identification. In *Information and Communications Security*, Kwok-Yan Lam, Chi-Hung Chi, and Sihan Qing (Eds.). Springer International Publishing, Cham, 173–183.

[18] Harold Hotelling. 1992. Relations between two sets of variates. In *Breakthroughs in statistics*. Springer, 162–190.

[19] Patrick Juola et al. 2008. Authorship attribution. *Foundations and Trends® in Information Retrieval* 1, 3 (2008), 233–334.

[20] KD Kochanek, SL Murphy, JQ Xu, and E Arias. 2017. Mortality in the United States, 2016. NCHS Data Brief, no 293. *National Center for Health Statistics* (2017).

[21] Kristy Kruithof. 2016. *Internet-facilitated drugs trade: An analysis of the size, scope and the role of the Netherlands*. RAND.

[22] Francois R Lamy, Raminta Daniulaityte, Ramzi W Nahhas, Monica J Barratt, Alan G Smith, Amit Sheth, Silvia S Martins, Edward W Boyer, and Robert G Carlson. 2017. Increases in synthetic cannabinoids-related harms: Results from a longitudinal web-based content analysis. *International Journal of Drug Policy* 44 (2017), 121–129.

[23] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*. 1188–1196.

[24] Usha Lokala, Francois R Lamy, Raminta Daniulaityte, Amit Sheth, Ramzi W Nahhas, Jason I Roden, Shweta Yadav, and Robert G Carlson. 2019. Global trends, local harms: availability of fentanyl-type drugs on the dark web and accidental overdoses in Ohio. *Computational and Mathematical Organization Theory* 25, 1 (2019), 48–59.

[25] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[26] Frederick Mosteller and David L Wallace. 1963. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers. *J. Amer. Statist. Assoc.* 58, 302 (1963), 275–309.

[27] Sebastian Ruder, Parsa Ghaffari, and John G Breslin. 2016. Character-level and multi-channel convolutional neural networks for large-scale authorship attribution. *arXiv preprint arXiv:1609.06686* (2016).

[28] Prasha Shrestha, Sebastian Sierra, Fabio Gonzalez, 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*. 669–674.

[29] Kevin Smith. 2016. Tochka Market URL: Links - Buy Items from Tochka Darknet Marketplace. https://www.deepweb-sites.com/tochka-market-url-links-darknet-reddit-review/

[30] Kyle Soska and Nicolas Christin. 2015. Measuring the longitudinal evolution of the online anonymous marketplace ecosystem. In *24th {USENIX} Security Symposium ({USENIX} Security 15)*. 33–48.

[31] Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology* 60, 3 (2009), 538–556.

[32] Yichuan Tang. 2013. Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239* (2013).

[33] Michel Van De Velden and Tammo HA Bijmolt. 2006. Generalized canonical correlation analysis of matrices with missing rows: a simulation study. *Psychometrika* 71, 2 (2006), 323–331.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *NIPS*.

[35] 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*. ACM, 431–442.

[36] Zack Whittaker. [n. d.]. Deep Dot Web Indictment. https://www.documentcloud.org/documents/5993699-Deep-Dot-Web-Indictment.html

[37] Aaron van Wirdum. 2019. Major Darknet Marketplace Wall Street Market Shuttered by Law... https://bitcoinmagazine.com/articles/major-darknet-marketplace-wall-street-market-shuttered-law-enforcement

[38] Richong Zhang, Zhiyuan Hu, Hongyu Guo, and Yongyi Mao. 2018. Syntax Encoding with Application in Authorship Attribution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2742–2753.

[39] 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*. ACM, 3448–3454.