information

*Article*

# Towards Safe Cyber Practices: Developing a Proactive Cyber-Threat Intelligence System for Dark Web Forum Content by Identifying Cybercrimes

**Kanti Singh Sangher [1], Archana Singh [2], Hari Mohan Pandey [3] and Vivek Kumar [4,*]**

[1] School of IT, Centre for Development of Advanced Computing, Noida 201307, India; kantisingh@cdac.in
[2] Amity School of Engineering and Technology, Amity University, Noida 201313, India; asingh27@amity.edu
[3] Department of Computing and Informatics, Bournemouth University, Fern Barrow, Poole BH12 5BB, UK; profharimohanpandey@gmail.com
[4] Department of Mathematics and Computer Science, University of Cagliari, 09124 Cagliari, Italy
* Correspondence: vivek.kumar@unica.it

**Abstract:** The untraceable part of the Deep Web, also known as the Dark Web, is one of the most used "secretive spaces" to execute all sorts of illegal and criminal activities by terrorists, cybercriminals, spies, and offenders. Identifying actions, products, and offenders on the Dark Web is challenging due to its size, intractability, and anonymity. Therefore, it is crucial to intelligently enforce tools and techniques capable of identifying the activities of the Dark Web to assist law enforcement agencies as a support system. Therefore, this study proposes four deep learning architectures (RNN, CNN, LSTM, and Transformer)-based classification models using the pre-trained word embedding representations to identify illicit activities related to cybercrimes on Dark Web forums. We used the *Agora* dataset derived from the DarkNet market archive, which lists 109 activities by category. The listings in the dataset are vaguely described, and several data points are untagged, which rules out the automatic labeling of category items as target classes. Hence, to overcome this constraint, we applied a meticulously designed human annotation scheme to annotate the data, taking into account all the attributes to infer the context. In this research, we conducted comprehensive evaluations to assess the performance of our proposed approach. Our proposed BERT-based classification model achieved an accuracy score of 96%. Given the unbalancedness of the experimental data, our results indicate the advantage of our tailored data preprocessing strategies and validate our annotation scheme. Thus, in real-world scenarios, our work can be used to analyze Dark Web forums and identify cybercrimes by law enforcement agencies and can pave the path to develop sophisticated systems as per the requirements.

**Keywords:** dark web forum; cyber security; cybercrimes; deep learning; natural language processing; *Agora* marketplace; BERT; law enforcement agencies



**Citation:** Sangher, K.S.; Singh, A.; Pandey, H.M.; Kumar, V. Towards Safe Cyber Practices: Developing a Proactive Cyber-Threat Intelligence System for Dark Web Forum Content by Identifying Cybercrimes. *Information* **2023**, *14*, 349. https://doi.org/10.3390/info14060349

Academic Editors: Eftim Zdravevski, Petre Lameski and Ivan Miguel Pires

Received: 2 May 2023
Revised: 11 June 2023
Accepted: 12 June 2023
Published: 18 June 2023



## 1. Introduction

There is a general perception that the *Internet* and *World Wide Web* (WWW) are similar, but apparently, they are related but not synonymous. While the Internet is a network of networks, the WWW provides a uniform and user-friendly interface to access the information available on the Internet [1]. The WWW can be divided into three parts: *Surface Web*, *Deep Web*, and *Dark Web*. The *Surface Web* is the visible and accessible part of the WWW, and its contents are indexed by search engines such as Google (https://www.google.com/, accessed on 27 March 2023), Yahoo (https://www.yahoo.com/, accessed on 27 March 2023), Bing (https://www.bing.com/, accessed on 27 March 2023), etc. The *Surface Web* only comprises a small percentage (∼4%) of the WWW. The remainder belongs to the *Deep Web*, which is the unindexed part inaccessible by search engines [2,3]. Unindexed does not always mean illegal, but it is protected as it is intended for specific users and purposes.

Some common examples in our day-to-day usage of the *Deep Web* are Internet banking, email mailboxes, government databases, medical records, etc. The *Dark Web* (also called the DarkNet) is a subset and the deepest layer of the *Deep Web* and is accessible only illegally. The *Dark Web* is accessed by special software that ensures high encryption anonymity, such as The Onion Router (TOR) browser [4,5]. The *Dark Web* is manifested in several ways [6–12], such as having DarkNet marketplaces (DNMs) for illegal contraband, terrorism, spreading propaganda and hatred, human trafficking, hiring and recruiting anti-social elements, leaking government data, untraceable financial transactions, weapons, etc. [13]. A few examples include Agora (https://en.wikipedia.org/wiki/Agora_(online_marketplace), accessed on 27 March 2023), Silkroad [14] 2.0 (https://it.wikipedia.org/wiki/Silk_Road, accessed on 27 March 2023), and Alphabay (https://en.wikipedia.org/wiki/AlphaBay, accessed on 27 March 2023); these are among the most popular and well-known DNMs belonging to the *Dark Web* [15].

According to the recent "Digital Defense Report" (https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RWMFIi#page=7, accessed on 27 March 2023) by *Microsoft*, the *Dark Web* has become a hub to sell and purchase cybercrime-related services, and an amateur with no technical knowledge or prior experience to conduct a cybercrime attack can also buy a range of services with just one click. As shown in Figure 1, which shows average prices, activities include hiring attackers, hiring for spearphishing, stealing user credentials, performing denial-of-service attacks, and other services. As evident from the figure, these services in the dark markets are inexpensive, making attacks cheap and easy to execute, and as a consequence, attack numbers increase sharply.

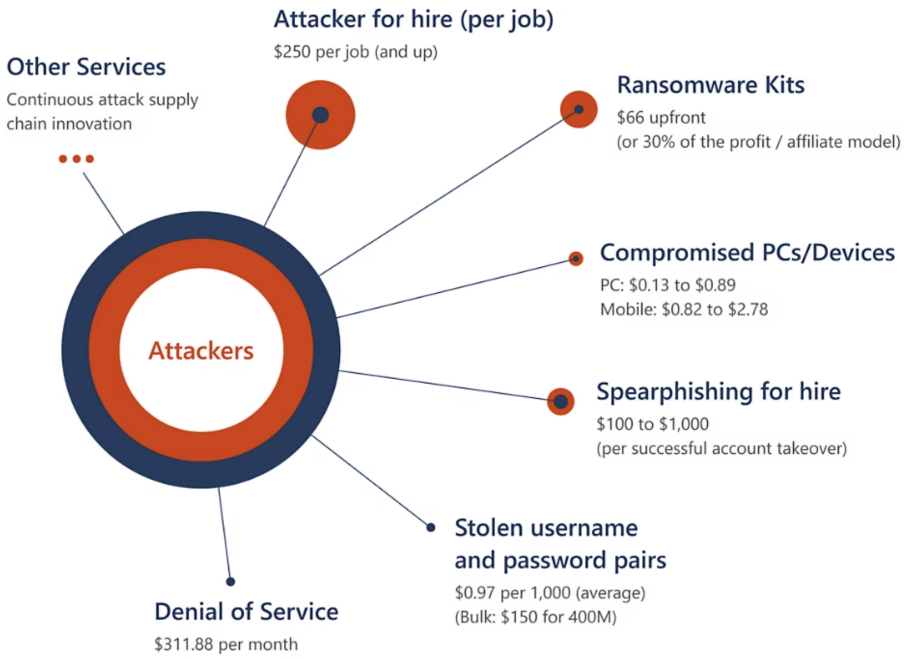


**Figure 1.** Average prices of cybercrime services for sale (source: "Microsoft Digital Defense Report").

Anonymous services such as Tor, Freenet, I2P, and JonDonym are frequently used to access the materials and services offered by the DarkNet [16]. The *Dark Web* provides a venue for collaboration, communication, and diverse acts through its forums. Identifying the activities happening on the *Dark Web* in order to prevent them is very challenging due to the anonymity of this platform and the intractability provided to users [17], and manually analyzing this content is resource intensive and unproductive. A viable approach in this direction is to extract the content of the *Dark Web* for analysis, followed by identifying research gaps and implementing scalable state-of-the-art machine learning (ML)-based approaches that can help with analysis, identification, and categorization of the content as

activity, items, tools, etc. This would enable law enforcement agencies to take preventive measures to enforce the law. The proposition can provide answers to questions such as:

- Are the activities on the DarkNet a potential indicator of possible cybercrimes? If so, what are the target domains and people at risk?
- What stolen/breached information is there, and what is the aftermath?
- Which preventive measures can be taken to minimize the losses?

While there are several DarkNet datasets available from 2011 to 2019 [18], including Silk Road 1.0, Silk Road 2.0, Sheep, Black Rank, Pandora, *Agora*, Blue Sky, Dream, Evolution, Middle Earth, Wall Street, Hydra, Silk Road 3.1, Olympus, Appolon, and Alphabet Marketplace, *Agora* and SilkRoad 1.0 are the most extensive datasets to date in terms of the number of data points and vast categories. The attributes available in the *Agora* marketplace dataset are helpful for information retrieval regarding cybercrime-based category identification, which aligns with our research goals. Therefore, in this work, we take into account the *Agora* DarkNet Market Archives (2013–2015) (https://gwern.net/dnm-archive#works-using-this-dataset, accessed on 27 March 2023) as a source to access the content of the *Dark Web*. The archive can help with the following suggested uses:

- Providing information on vendors across markets, such as Pretty Good Service (PGP (https://en.wikipedia.org/wiki/Pretty_Good_Privacy, accessed on 27 March 2023)) key and feedback ratings [19,20].
- Identifying the popularity of individual and categories of drugs [21–24].
- Identifying arrested/convicted vendors.
- Identifying vendor activities (products sold and ratings).
- Generating information to help operations, such as *Operation Onymous* (https://www.europol.europa.eu/operations-servi,ces-and-innovation/operations/operation-onymous, accessed on 27 March 2023).
- Topic modeling of forums.

Several existing works in the literature have uncovered the uses of this archive, such as *drug trafficking* [19,22,25–28], *author verification* [29], *cryptocurrency* and *Bitcoin transaction-related analysis* [30–33], *malware analysis* [34], *vendor identification* [19,20,35–37], *social media analysis* [38–40], and identifying services provided by DarkNet markets [41–45]. However, very little to no work has been done to determine prospective cybercrime by classifying the contents of *Dark Web* forums. Thus, to bridge this gap, our research focuses on identifying cybercrime-related activities based on various inputs, such as account stealing, data theft, financial fraud, hacking, software piracy, etc., of the *Agora* dataset. Thus, we draw research motivation from these pertaining factors to investigate the information on *Dark Web* forums to gather intelligence. This research provides an extensively intelligent analysis of the DNM dataset by combining the attributes' context, developing a set of keywords, and using natural language processing (NLP) approaches. The main contributions of this paper are as follows:

- We performed the annotation to label the data for creating the ground-truth labels for the large *Agora Dark Web* dataset.
- We applied heuristic approaches to finally select the preprocessing strategies suitable and useful for our experimental dataset.
- Based on our annotation, we modeled a novel multiclass classification problem to identify activities (cybercrime) on *Dark Web* forums.
- We implemented several deep learning approaches (baseline and state-of-the-art) using pre-trained word-embedding representations.
- Finally, we provide an in-depth discussion of the experimental outcome and present our key findings of this research work.

The rest of the paper is organized as follows. Section 2 presents the literature background of the works related to *Dark Web* forums. Section 3 mentions the problem statement and provides details of the dataset and preprocessing strategies used for the experiments.

Section 4 presents the architecture of the employed pipeline and sums up the total experiments performed. Section 5 presents the experimental outcomes and provides a discussion of the results. Finally, Section 6 presents the concluding remarks and the future research direction.

## 2. Literature Review

The criminal activities on DNMs caught the general public's attention in October 2013 when the Federal Bureau of Investigation (FBI), USA, closed down the marketplace Silk Road [46]. However, Silk Route 2.0 emerged a month later and closed again in 2014. The problem is not only the *Dark Web* but the social platforms that provide opportunities to hype the *Dark Web*, such as *Reddit* and discussion forums containing links to *Dark Web* websites [47,48]. Some notable works have thoroughly analyzed *Dark Web* forums to analyze the most frequent activities and products in marketplaces [49,50]. Since *Agora* DNM is a comprehensive dataset, several crucial works have grounded their research on this dataset to analyze *Dark Web* forums.

For instance, Ref. [50] presents an unsupervised model to monitor and categorize *Dark Web* forums using decision trees and clustering algorithms, making it adaptable to new and evolving forums without needing labeled training data. This two-step approach model applies topic modeling algorithms. The model classifies the forums into different categories by analyzing the extracted topics after web crawling. Finally, based on extracted topics, the proposed model classifies the forums into different categories, enabling researchers and law enforcement agencies to gain insights into the nature of the forums and the activities taking place within them.

The research in [41] highlights the inability of traditional user representation methods to capture the temporal content. The work also shows how recent works, mainly using CNN-based models, fail to handle the context and text length effectively. prices of cybercrime services for sale (source: "Microsoft Digital Defense Report"). To address these problems, the work proposed a model named *URM4DMU* that uses self-attention with an adaptive gate mechanism to improve post representation using temporal content.

The work [27] presents the DreamDrug dataset, a comprehensive and crowdsourced resource for training and evaluating Named Entity Recognition models to detect drugs in DarkNet markets. The dataset enables researchers and developers to develop more effective tools and techniques for monitoring illegal drug activities on online platforms, contributing to the broader goal of combating drug trafficking and ensuring public safety.

The work [29] highlights the limitations of works using literary texts and authorship analysis tools for cybercrime prevention. This work released VeriDark, a benchmark comprised of one authorship-identification dataset and three large-scale authorship-verification datasets to address these issues and provide competitive NLP baselines on these datasets.

In [28], a system named dSytle-GAN is introduced that considers both style-aware and content-based information to automate drug identification in DNMs. The work is focused on distinguishing the similarity between given pairs of drugs based on an attributed heterogeneous information network (AHIN) and a generative adversarial network (GAN). The authors claim that, unlike existing approaches, their proposed GAN-based model jointly considers the heterogeneity of the network and relatedness of drugs formulated by domain-specific meta-paths for robust node (i.e., drug) representation learning.

The research in DNM analysis is constantly evolving, and in general, researchers use several NLP tools and techniques. For instance, topic modeling has been very effective for such analysis [51–53] to infer information about drugs [12,21–24,27], extremists, terror activities, and resources. The work in [54] focuses on modeling topics related to homeland security threats and proposes combining traditional network analysis methods with topic-model-based text-mining techniques. The experiments of this work are performed using an English-language-based *Dark Web* portal (IslamicAwakening (http://forums.islamicawakening.com/, accessed on 27 March 2023)). The work in [55] emphasized the importance of *Dark Web* analysis in counter-terrorism (CT) to identify the

various websites used as sources for spreading propaganda and ideologies and recruiting new members. The work proposed a *Dark Web* analysis model to anticipate possible terror threats/activities to prevent terrorist attacks through analyzing *Dark Web* forums for CT. Another work [56] implemented a dynamic-systems approach for unsupervised anomaly detection to identify evolving threats in unlabeled and time-dependent datasets. The proposed method used finite-time Lyapunov exponents to characterize the time evolution of the distribution of text attributes in the forum content and the directed network structure. The works in [51,53] proposed a Latent Dirichlet Allocation (LDA) algorithm-based approach to analyze documents/corpus and discover the latent topics from websites of terrorists/extremists.

This situation has drawn interest in *Dark Web*-related research in monitoring and extracting information for cyber-threat intelligence [57,58] (CTI) through *Dark Web* forum analysis [59]. The current works take into account the relationship between suppliers and users and transaction statistic discoveries [60], developing automated approaches for discovering evidence of potential threats within hacker forums to aid in cyber-threat detection [61] and data-driven security game frameworks to model attackers and provide policy recommendations to the defender [62]. The study [63,64] discussed various strategies for monitoring the hidden areas of the Internet and suggested monitoring the DarkNet to find possible dangerous threats and activities [65,66]. Furthermore, tracking the hackers based on their illicit activities can lead to much crucial information [67,68]. Machine learning (ML) is extensively used to analyze *Dark Web* forums, and different approaches are used to extract the retrieved information. For instance, Refs. [69–71] are focused on malware detection in DNMs traffic.

Identifying Jihadist community groups and decoding their messages and communication is also prevalent among researchers. The article [72] presents a general framework (web-based knowledge that incorporates data collected from different international Jihadist forums). The work provides several analysis functions, such as forum browsing and searching, multilingual translation of forums, statistical analysis, and visualization of social networks. The work [73] considers extremist social media websites to introduce methods for identifying recruitment activities in violent groups. The work used data from the Western jihadist website *Ansar al-Jihad Network* that had been compiled by the University of Arizona's *Dark Web* project. Manual annotation was carried out on a sample of these data, engaging multiple judges who marked 192 randomly sampled posts as recruiting (Yes) or non-recruiting (No). The authors claim it to be the first result reported on such a task. Ref. [74] presents an automated method for sentiment and affects analysis incorporating ML and rich textual feature representation techniques to identify and measure the sentiment polarities and affect intensities expressed on Al-Firdaws (www.montada.com, accessed on 27 March 2023) and Montada (www.alfirdaws.org/vb, accessed on 27 March 2023 ) *Dark Web* forums. The work in [75] applied semiautomated methodologies to capture and organize domestic extremist website data of the USA to track and gather information from US radical online forums using human experts. The work used a three-step approach: forum identification, collection and parsing, and analysis.

This section shows that the *Agora* DNM is used for several information extractions works, but no work has been done specifically to identify cybercrimes. Therefore, the novelty of our work lies in the tailored annotation of cybercrime identification to facilitate law enforcement agencies to better interpret the context of conversations in DarkNet markets and cybercrime.

## 3. Dataset, Preprocessing, and Problem Formulation

This section explains the experimental dataset used, the preprocessing strategies applied to prepare the input data, and the problem statement modeled for this research work.

### *3.1. Dataset Description*

We used the Agora (https://www.kaggle.com/datasets/philipjames11/dark-net-marketplace-drug-data-agora-20142015, accessed on 27 March 2023) DarkNet dataset for

this research work. The *Agora* DarkNet raw dataset is a data parse of marketplaces extracted from the *DarkNet Market Archives* [76], (a dark/deep web) marketplace from 2014 to 2015. The raw dataset contains items such as weapons, drugs, services, etc. A sample of the dataset is shown in Figure 2. The raw dataset has 109,692 data points and 9 attributes, which are briefly mentioned below:

- **Vendor:** The items of this attribute are related to vendors, types of vendors, etc. There are 3192 distinct items listed in this attribute, and the distribution of the top 40 items is shown in Figure 3.
- **Category:** This attribute contains where the marketplace items are listed. There are a total of 109 specific items listed in the raw dataset.
- **Item:** This attribute contains the title of the listed items.
- **Description:** This attribute contains the description of the items.
- **Price:** This attribute contains the cost of the items. The cost is averaged for duplicate listings between 2014 and 2015)
- **Origin:** This attribute contains the place of origin from where the item is shipped. Several data points in this attribute are empty or missing.
- **Destination:** This attribute contains the place where the item is to be shipped (blank means no information was provided, but most likely worldwide.) Several data points in this attribute are also empty or missing.
- **Rating:** This attribute contains the seller's rating, typically on a scale of 5. A rating of "[0 deals]" or anything else indicates that the number of deals is too small for a rating to be displayed.
- **Remarks:** This attribute contains remarks such as "[0 deals]" or "Average price may be skewed outlier > 0.5 BTC found". In this attribute also, several data points are empty or missing.

ATM Skimmer Cashing/Installing Safety tutorial I will try to describe here the basic moments which are important in the work about which it is not necessary to forget and I will also write s
...
ATM Skimmer Cashing/Installing Safety
UNLIMITED STEALTH PAYPAL ACCOUNTS
Mac Address Spoofer Software (Must Have)
HOW TO HACK ANY FACEBOOK ACCOUNT EASY
MAC ADRESS CHANGER FOR MAC OS SOFTWARE
Anonymity Table of Contents: 1.What is a proxy? 2.Where can I get a proxy? 3.How do I put the proxy into effect? 4.How do I chain proxies? 5.My proxy*f*??s not working what*f*??s wrong? What should I do? 6.Wh ...
Transfer Balance from Hacked PayPals
How to make all Trojan-Virus-Keylogger 100% undetectable
GUIDE FOR ANONYMOUS INTERNET USAGE
Create a Realistic Credit Card in Photoshop
DarkNet Dictionary For those of you that are just beginning with the DarknetMarkets we have compiled a list of the terms you might come across while browsing around including links to important resources
BE SECURE AND ANONYMOUS ON ANDROID
Anonymous Instant Messenger Guide [ICQ Jabber Torchat]
How To Set Up A PGP Key Easily [Images Included]
Free VPN For Windows Mac & Linux
Anonymous SIM card
[ANDROID] Phone vip72 SOCKS + TOR Tutorial
Cell Phone Tracking
[SOCKS5 Finder] Proxy Scanner Proxy For Windows
Your Own .onion Domain + 1 month Tor hosting (no CP)
Hacker for Hire/ Hacking Service / Blackhat Help Agent
Totally Anonymous Bank Account - Euro/USD 15k Daily Withdrawl Limit - Limited Amount
All Agora's and Silk Road's Money Making Methods & Guides
How To Profit Illegally From Bitcoin Cybercrime And Much More

**Figure 2.** A sample text of the items and their description.

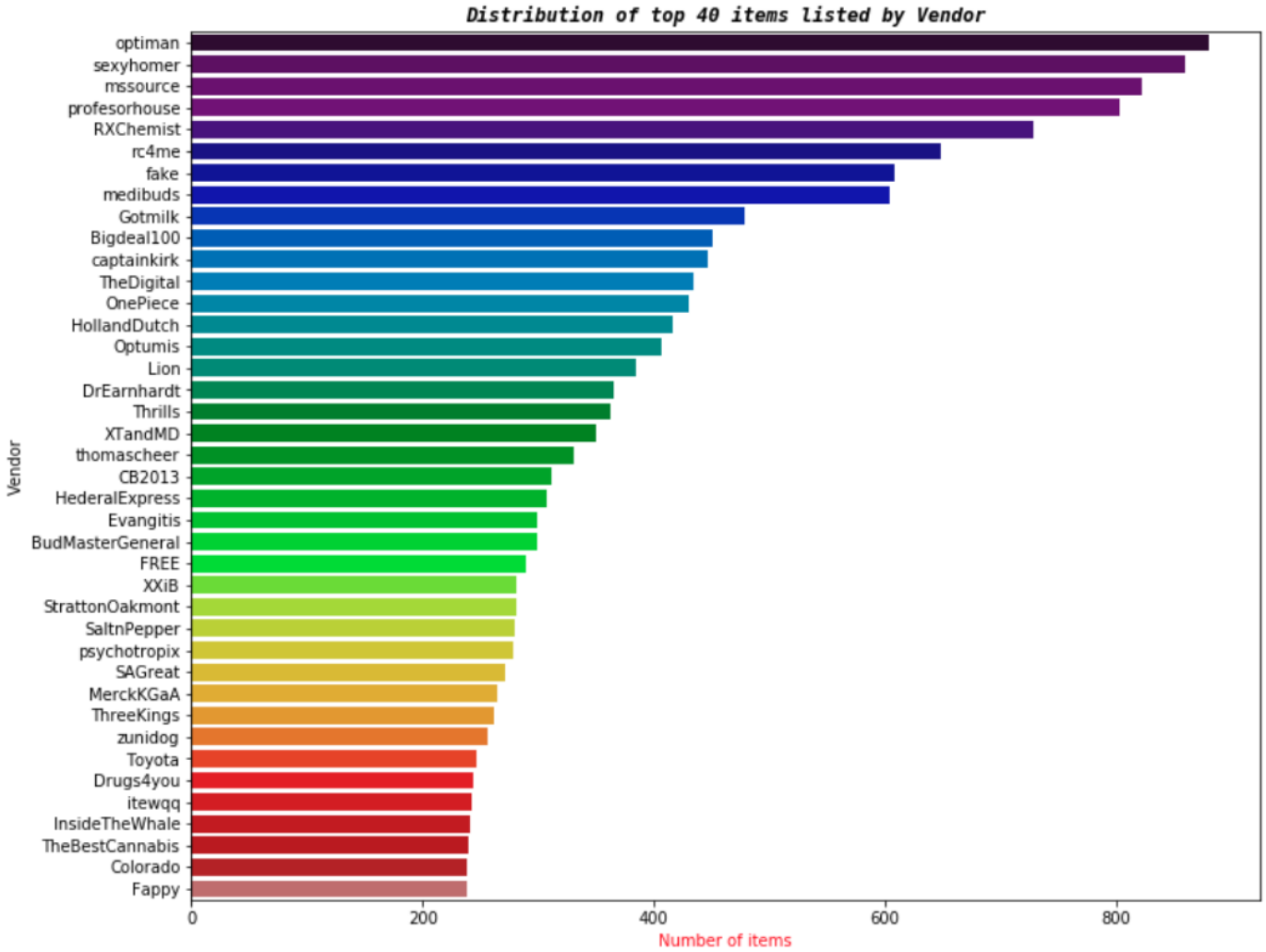


**Figure 3.** The distribution of top 40 items under Vendor of the *Agora* dataset.

The most frequent words of the textual data are shown in Figure 4 as a word cloud created by using the Python library (https://pypi.org/project/wordcloud/, accessed on 27 March 2023) to show the kind of activity engagements and related materials.

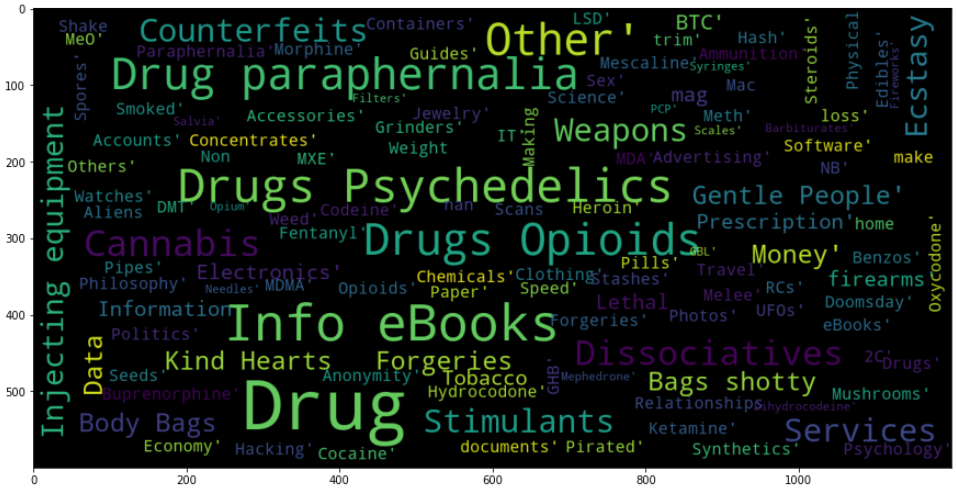


**Figure 4.** Word cloud presenting the most frequent items of the dataset.

### *3.2. Dataset Preprocessing*

As mentioned in the dataset description, several attributes of the raw dataset have missing values and contain outliers, redundancies, duplicate values, special characters, and symbols. In addition, the raw dataset is highly imbalanced; therefore, we paid particular attention to implementing optimal pre-processing strategies to interpret special characters/symbols better and to meticulously remove outliers and redundant data without losing the context of the item description [77–82]. The key pre-processing techniques implemented to pre-process the dataset are as below:

- The entire text is converted to lowercase to make the dataset uniform in terms of representation (e.g., "Category" and "category" are represented by a common token: "category").
- Punctuation is removed since it does not add valuable semantic information to the text.
- We removed stopwords for the above-mentioned reason.
- We removed newlines, whitespaces, and extra spaces from the text.
- We removed the special characters, symbols, and elements that are not part of standard English language.
- We performed stemming and lemmatization alternatively to observe the impact of classification models.
- Finally, we tokenized the text data to get words/tokens.
- We removed the blank and outlier values of the attribute "category".

After applying the above-mentioned pre-processing strategies, the processed dataset contained a total of 109,684 data points and 104 *Category* items. The distribution of the *Category* items is mentioned in Table 1.

**Table 1.** The overall distribution of category items across the *Agora* dataset.

| Sr. No. | Category (Items) | Count | Sr. No. | Category (Items) | Counts |
|---|---|---|---|---|---|
| **1** | Drugs/Cannabis/Weed | 21,272 | **53** | Drugs/Opioids/Buprenorphine | 284 |
| **2** | Drugs/Ecstasy/Pills | 7534 | **54** | Drugs/Psychedelics/Other | 272 |
| **3** | Drugs/Ecstasy/MDMA | 6116 | **55** | Drugs/Weight loss | 252 |
| **4** | Drugs/Stimulants/Cocaine | 6007 | **56** | Counterfeits/Accessories | 250 |
| **5** | Drugs/Prescription | 5561 | **57** | Drugs/Opioids/Morphine | 248 |
| **6** | Drugs/Benzos | 5393 | **58** | Drugs/Dissociatives/GHB | 226 |
| **7** | Drugs/Cannabis/Concentrates | 4257 | **59** | Drugs/Psychedelics/5-MeO | 216 |
| **8** | Drugs/Psychedelics/LSD | 3775 | **60** | Info/eBooks/Anonymity | 204 |
| **9** | Drugs/Cannabis/Hash | 3241 | **61** | Drug paraphernalia/Pipes | 195 |
| **10** | Drugs/Steroids | 2779 | **62** | Drugs/Opioids/Hydrocodone | 191 |
| **11** | Drugs/Stimulants/Meth | 2467 | **63** | Drug paraphernalia/Containers | 186 |
| **12** | Drugs/Stimulants/Speed | 2401 | **64** | Info/eBooks/Science | 163 |
| **13** | Drugs/RCs | 2182 | **65** | Drug paraphernalia/Stashes | 149 |
| **14** | Drugs/Stimulants/Prescription | 1956 | **66** | Info/eBooks/Relationships/Sex | 145 |
| **15** | Drugs/Opioids/Heroin | 1799 | **67** | Info/eBooks/IT | 144 |
| **16** | Services/Money | 1481 | **68** | Weapons/Ammunition | 138 |
| **17** | Other | 1425 | **69** | Services/Advertising | 132 |
| **18** | Drugs/Opioids/Oxycodone | 1360 | **70** | Drugs/Cannabis/Shake/trim | 121 |
| **19** | Counterfeits/Watches | 1309 | **71** | Drugs/Psychedelics/Others | 106 |
| **20** | Drugs/Opioids | 1236 | **72** | Drug paraphernalia/Grinders | 106 |
| **21** | Data/Accounts | 1233 | **73** | Weapons/Melee | 103 |
| **22** | Drugs/Psychedelics/Mushrooms | 1140 | **74** | Forgeries/Other | 100 |
| **23** | Drugs/Cannabis/Edibles | 1109 | **75** | Drugs/Opioids/Codeine | 92 |
| **24** | Drugs/Ecstasy/Other | 1004 | **76** | Services/Travel | 90 |
| **25** | Drugs/Dissociatives/Ketamine | 992 | **77** | Chemicals | 90 |
| **26** | Drugs/Psychedelics/NB | 974 | **78** | Drugs/Opioids/Opium | 87 |
| **27** | Drugs/Psychedelics/2C | 932 | **79** | Drugs/Psychedelics/Mescaline | 86 |

**Table 1.** *Cont.*

| Sr. No. | Category (Items) | Count | Sr. No. | Category (Items) | Counts |
|---|---|---|---|---|---|
| **28** | Information/Guides | 927 | **80** | Drugs/Psychedelics/Spores | 80 |
| **29** | Information/eBooks | 918 | **81** | Drugs/Dissociatives/GBL | 76 |
| **30** | Drugs/Other | 872 | **82** | Info/eBooks/Economy | 76 |
| **31** | Drugs/Opioids/Fentanyl | 848 | **83** | Drugs/Dissociatives/Other | 63 |
| **32** | Drugs/Psychedelics/DMT | 723 | **84** | Drug paraphernalia/Paper | 61 |
| **33** | Info/eBooks/Other | 691 | **85** | Counterfeits/Electronics | 59 |
| **34** | Drugs/Opioids/Other | 643 | **86** | Weapons/Non-lethal firearms | 57 |
| **35** | Drugs/Cannabis/Synthetics | 637 | **87** | Drugs/Opioids/Dihydrocodeine | 54 |
| **36** | Forgeries/Physical documents | 616 | **88** | Drug paraphernalia/Scales | 47 |
| **37** | Electronics | 599 | **89** | Drug paraphernalia/Injecting equipment/Syringes | 45 |
| **38** | Data/Pirated | 529 | **90** | Info/eBooks/Doomsday | 43 |
| **39** | Drugs/Cannabis/Seeds | 529 | **91** | Drugs/Stimulants/Mephedrone | 40 |
| **40** | Services/Other | 487 | **92** | Info/eBooks/Psychology | 40 |
| **41** | Services/Hacking | 453 | **93** | Drugs/Psychedelics/Salvia | 37 |
| **42** | Jewelry | 418 | **94** | Drugs/Barbiturates | 30 |
| **43** | Drugs/Dissociatives/MXE | 408 | **95** | Drug paraphernalia/Injecting equipment/Other | 30 |
| **44** | Tobacco/Smoked | 393 | **96** | Tobacco/Paraphernalia | 27 |
| **45** | Counterfeits/Money | 387 | **97** | Info/eBooks/Politics | 26 |
| **46** | Counterfeits/Clothing | 364 | **98** | Info/eBooks/Philosophy | 25 |
| **47** | Data/Software | 356 | **99** | Drug paraphernalia/Injecting equipment/Needles | 15 |
| **48** | Weapons/Lethal firearms | 344 | **100** | Weapons/Fireworks | 14 |
| **49** | Drugs/Ecstasy/MDA | 329 | **101** | Info/eBooks/Aliens/UFOs | 10 |
| **50** | Forgeries/Scans/Photos | 327 | **102** | Forgeries | 8 |
| **51** | Info/eBooks/Making money | 313 | **103** | Drug paraphernalia/Injecting equipment/Filters | 6 |
| **52** | Info/eBooks/Drugs | 289 | **104** | Drugs/Dissociatives/PCP | 4 |

### *3.3. Problem Formulation*

The primary objective of this work is to identify cybercrimes based on the context and item description of criminal actives of the *Agora* dataset. To do so, we analyzed the *Agora* dataset, and we modeled a novel problem statement of identifying the *Category* of the items by taking into account the description of the category items and their title. Since our objective is to identify the activities of the *Dark Web*, they essentially fall into three categories:

- The activities are clearly indicative of "**Cybercrime** ".
- The activities are clearly indicative of "**Not Cybercrime**".
- It is difficult to say if the activity is explicitly Cybercrime "**Can't say if cybercrime**".

Categorizing the dataset is a daunting task for the given form and data type. The reason is that most of the data points' attributes (item and item description) are not explicitly clear or indicative of the categories mentioned above. Therefore, first, we analyzed the data at a very fine-grained level, considering all the attributes to infer information that could help identify the related classes of the data points. After that, from the types of existing cybercrimes, we finalized the set of keywords representative of the *Agora* dataset. Finally, we executed the annotation using this handcrafted list of cybercrimes to categorize the input dataset, taking into account attributes, items, and item descriptions. We took references from large-scale annotations for generating the Anno-MI dataset [83–86] for designing our semi-automatic annotation scheme. The list was used to annotate the attribute *Category* into three labels: "Not Cybercrime" (0), "Cybercrime" (1), and "Can't say if Cybercrime" (2). The post-annotation human evaluation involved meticulous verification and validation of all the labels. Due to the inherent complexity of accurately inferring crime-related information using NLP techniques, the large size of the experimental dataset, and the cost of annotation,

we opted for a combination of manual and automated annotation to ensure the accuracy and reliability of the labeling process. The final distribution of the post-annotation input dataset is shown in Figure 5, and the 104 categories are shown in Table 2.

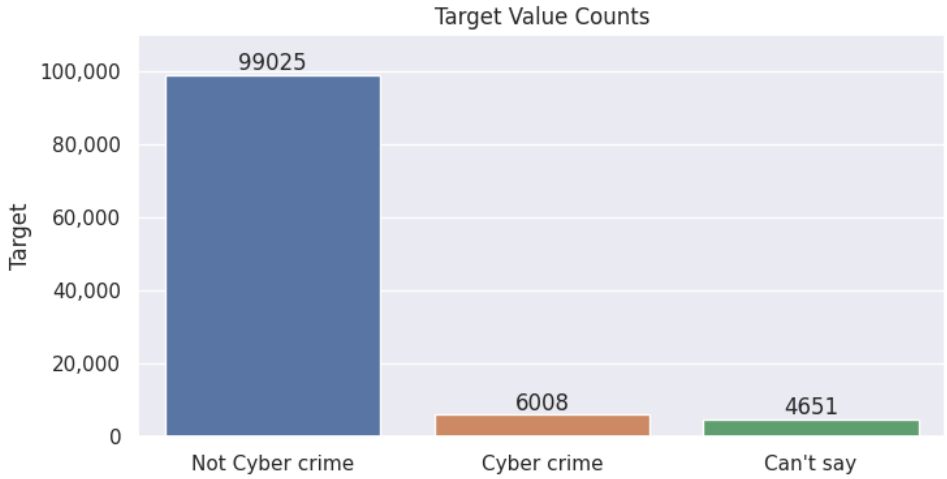


**Figure 5.** Distribution of the annotated dataset representing each target class.

**Table 2.** Targets assigned to different categories.

| Target | Meaning | Category |
|---|---|---|
| **1** | **Cybercrime** | Services/Hacking<br>Services/Money<br>Electronics<br>Data/Accounts<br>Data/Software<br>Data/Pirated |
| **2** | **Can't say if cybercrime** | Services/Other<br>Forgeries/Physical documents<br>Forgeries/Other<br>Forgeries/Scans/Photos<br>Information/Guides<br>Information/eBooks<br>Info/eBooks/Making money<br>Info/eBooks/Other<br>Info/eBooks/Anonymity<br>Other |
| **0** | **Not cybercrime** | Rest of the categories |

Posts assigning the target classes (annotation) to the attribute "**Category** " and the attributes "**Item**" and "**Item description**" are concatenated to generate the final text, which is further used to generate feature vectors. The overall distribution of target class **Cybercrime** listed crimes is shown in Figure 6. In this target class, the major stack is dedicated to Services/money, i.e., the use of the *Dark Web* platform to hire different cybercrime-related services in exchange for money. Other important categories are related to accounts, electronics, piracy, hacking, and software (crack).

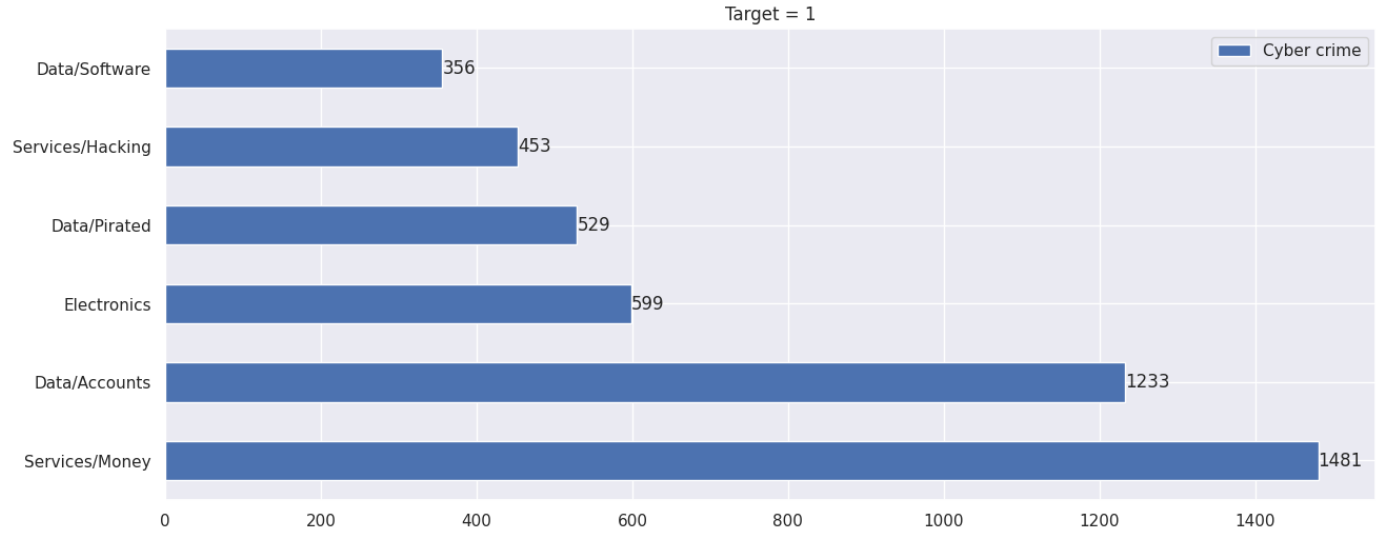


**Figure 6.** Distribution of target class "Cybercrime" items.

The overall distribution of target class **Can't say if cybercrime** is shown in Figure 7.

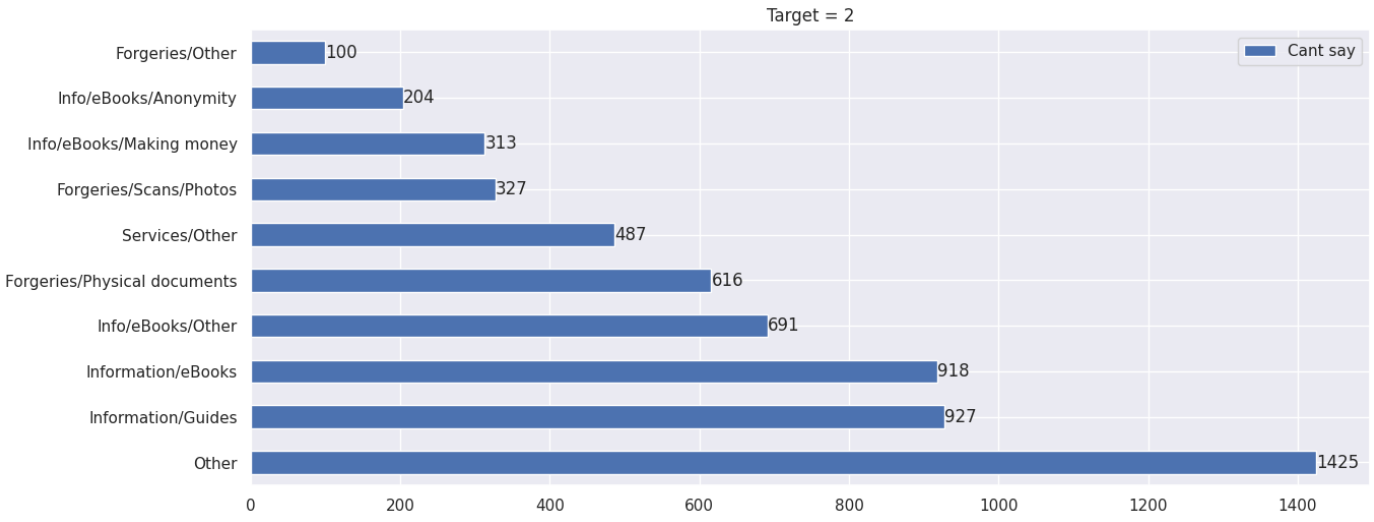


**Figure 7.** Distribution of target class "Can't say if cybercrime" items.

Hence, to sum up, in this research work, we tackle the problem statement of identifying *Dark Web* activities as a multiclass classification problem where the target labels are *Not Cybercrime* (0), *Cybercrime* (1), and *Can't say if cybercrime* (2).

## 4. Materials and Methods

In this section, we provide the details of computational resources used, the architecture of the employed classification model, and experiments performed.

### *4.1. Resource Description*

The computational resource used for this work is mentioned in Table 3.

**Table 3.** Resource specification.

| Item | Specification |
|---|---|
| CPU | AMD Radeon (TM) Graphics |
| GPU | NVIDIA GeForce RTX 3060 |
| RAM | 16 GB |
| CUDA | CUDA 11.7 + CuDNN8.4.1.50 |
| OS | Windows 11 |
| Python | Version 3.10 |
| TensorFlow | Version 2.10.1 |

### *4.2. Architecture of Classification Models*

For our research, we employed four DL-based classification models to perform the multiclass classification problem for analyzing *Dark Web* forum data. The end-to-end pipeline of the employed approaches is presented in Figure 8.

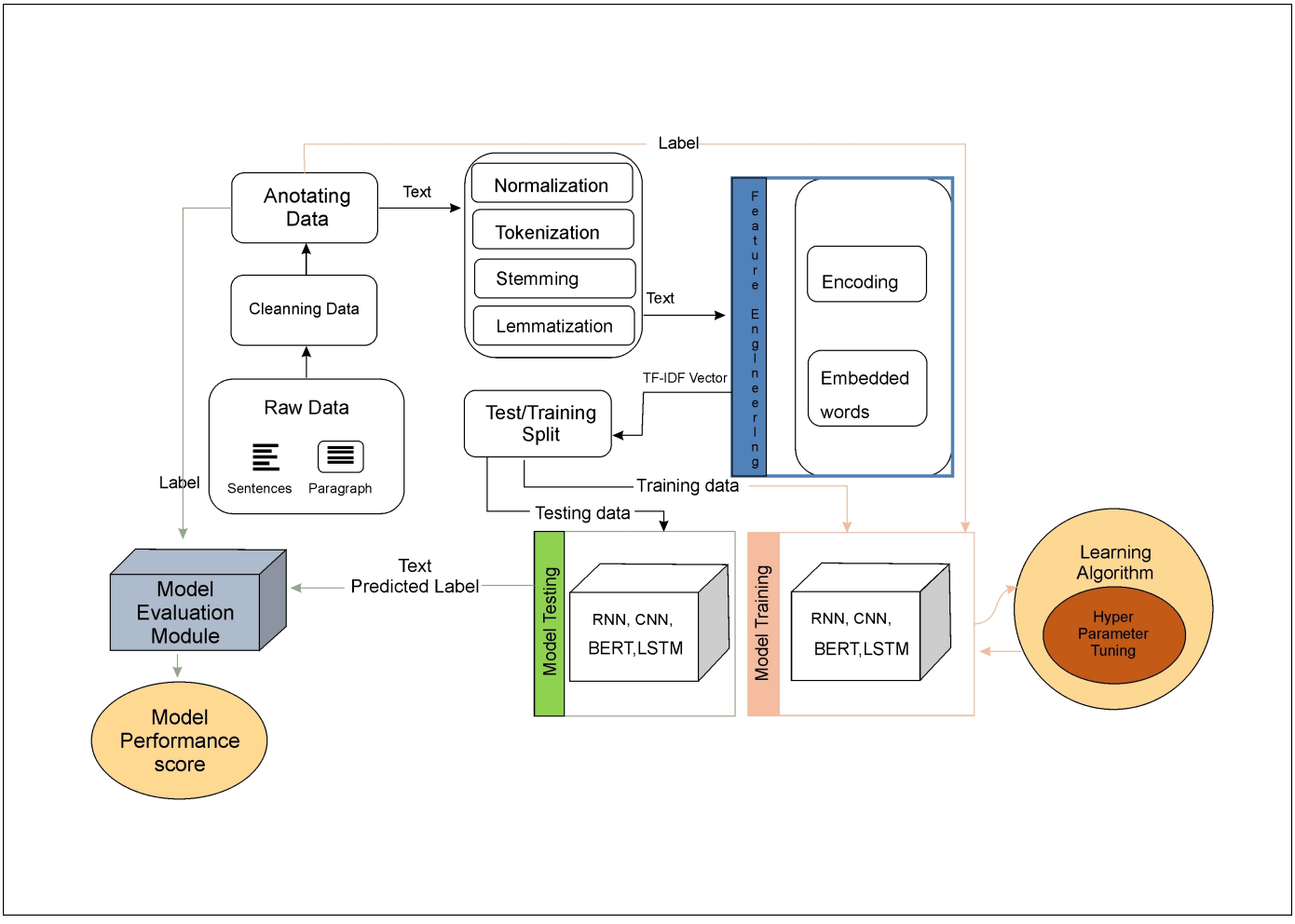


**Figure 8.** The pipeline employed to perform the classification incorporates the stages (a) data preprocessing, (b) feature engineering, (c) model training, and (d) model inference.

The architecture comprises several blocks that perform the different stages, namely data preprocessing, data normalization, feature engineering, classification model training, and model inference. These steps are presented in a simplified manner below:

1. First, the raw dataset is cleaned.
2. Then, annotation of the dataset is executed.

3. Post-annotation, the dataset is normalized, which essentially executes the preprocessing steps.
4. Then, the tokenized dataset is integer encoded.
5. Post encoding, the embedding matrix is generated.
6. The embedding matrix and encoded text are used to train the classification models.
7. As a last step, the prediction is done on the test dataset.

DL-based approaches are the current state-of-the-art and are very effective in capturing the context and subtle nuances of a domain if they are provided with sufficient training data. Further, benchmarking results obtained by recent research work [27,29,50] for DNM analysis proves the efficiency of DL approaches. Therefore, we took motivation from existing crucial work and used four DL classification models based on Recurrent Neural Networks (RNNs) [87,88], Convolutional Neural Networks (CNNs) [89,90], Long Short-Term Memory (LSTM) [91,92], and the Transformers architecture [93]. The RNN, CNN, and LSTM models used GloVe [94] pre-trained embeddings to generate the embedding matrix. The Bidirectional Encoder Representations from Transformers (BERT) (https://huggingface.co/docs/transformers/model_doc/bert, accessed on 27 March 2023) model used BERT embeddings (https://pypi.org/project/bert-embedding/, accessed on 27 March 2023) generated by the BERT Tokenizer (https://huggingface.co/docs/transformers/main_classes/tokenizer, accessed on 27 March 2023). For all our experiments, we used early stopping, training, and test sets in a ratio of 4:1 (80% and 20%), and the validation split percentage was set to 20.

## 5. Results and Discussion

We measured the performance of employed DL classification models by the metrics accuracy, precision, recall, and F-1 score. The formulas to calculate accuracy, precision, recall, and F-1 score are as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

where $TP$ , $FP$, $TN$, and $FN$ represent true positive, false positive, true negative, and false negative, respectively. The results obtained from the experiments with the classification models are summed up in Table 4. CNN has proven its efficacy in several ML downstream tasks, especially in the computer vision domain. Given the nature of our experimental dataset, the lowest performance of CNN, with 73% overall accuracy, is somewhat expected. RNN scored an accuracy of 88%, which is justified by the fact that LSTM and BERT are more advanced networks. LSTM and BERT significantly outperformed the previous two models and attained an accuracy of 96%. Further, given that the input data are heavily unbalanced, the performance of LSTM and BERT with each class is consistent, which is not always the case. To further analyze the results of the employed DL models with each target class, we present the confusion matrices [95] of each of the four models in Figure 9. The matrices clearly show the lower misclassification rate of LSTM and BERT compared to RNN and CNN-based DL models. The receiver operating characteristic (ROC) [96,97] curves for the four models are shown in Figure 10. ROC curves indicate the true positive rate (TPR) against the false positive rate (FPR) correlation; the higher the area under the curve (AUC), the better the classifier is. As evident from the plot, the Transformer model showed the best results with 0.99 AUC for each of the three classes, followed by LSTM. Thus, our results

show that the sophisticated DL models can better understand the peculiarities of domain and context, which leads to more reliable prediction for unstructured data such as ours.

**Table 4.** Classification results of employed DL models.

| Model | *Target Class* | *Precision* | *Recall* | *F1-Score* | *Accuracy* |
|---|---|---|---|---|---|
| ***CNN*** | Can't Say | 0.17 | 0.85 | 0.28 | 0.73 |
| | Cybercrime | 0.6 | 0.53 | 0.56 | |
| | Not Cybercrime | 0.99 | 0.73 | 0.84 | |
| ***RNN*** | Can't Say | 0.37 | 0.7 | 0.48 | 0.88 |
| | Cybercrime | 0.43 | 0.78 | 0.55 | |
| | Not Cybercrime | 0.99 | 0.89 | 0.94 | |
| ***LSTM*** | Can't Say | 0.64 | 0.79 | 0.71 | 0.96 |
| | Cybercrime | 0.72 | 0.83 | 0.77 | |
| | Not Cybercrime | 1 | 0.97 | 0.98 | |
| ***BERT*** | Can't Say | 0.65 | 0.86 | 0.74 | 0.96 |
| | Cybercrime | 0.81 | 0.77 | 0.79 | |
| | Not Cybercrime | 0.99 | 0.98 | 0.99 | |

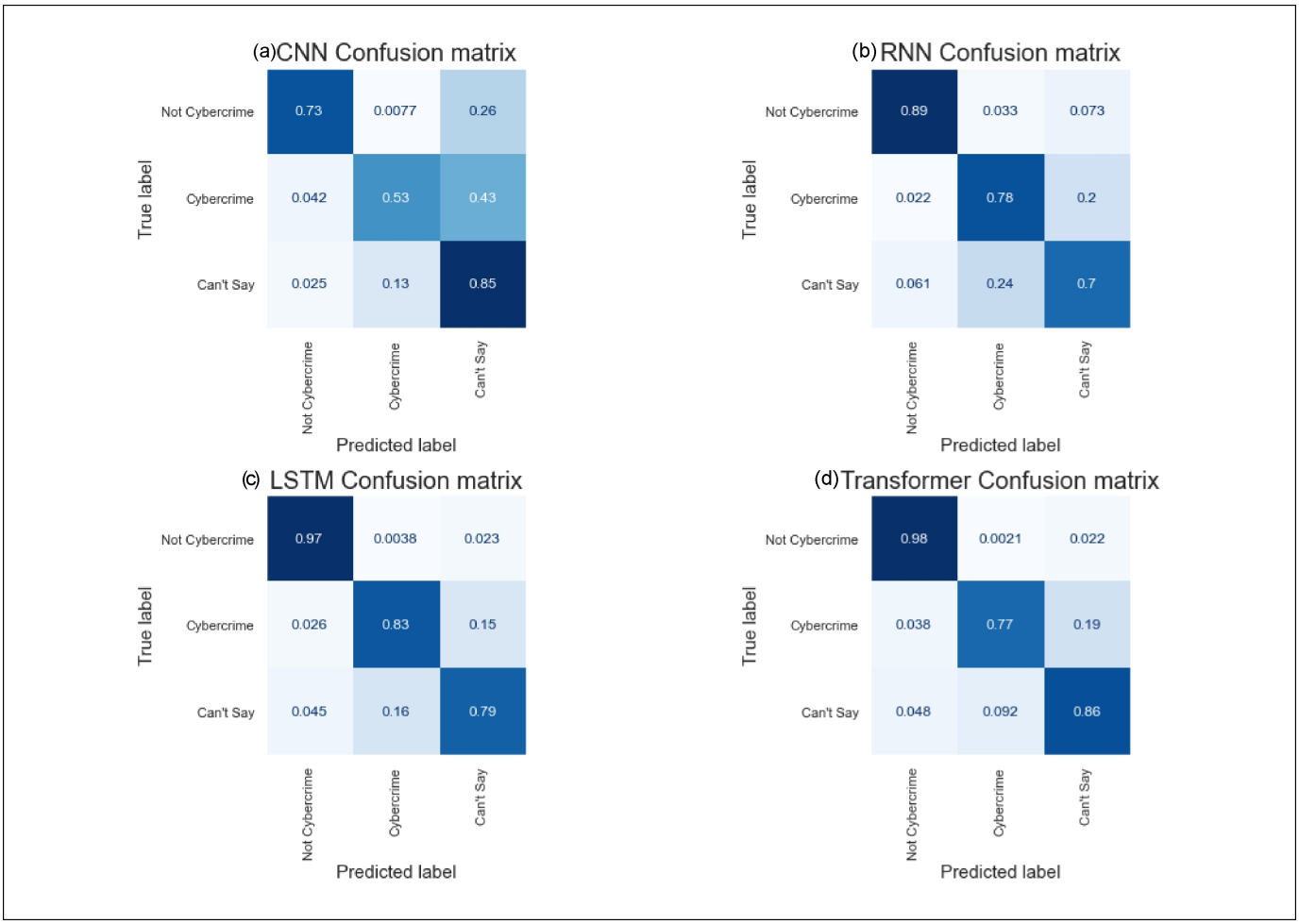


**Figure 9.** The confusion matrices of (**a**) CNN, (**b**) RNN, (**c**) LSTM, and (**d**) BERT for multiclass classification.

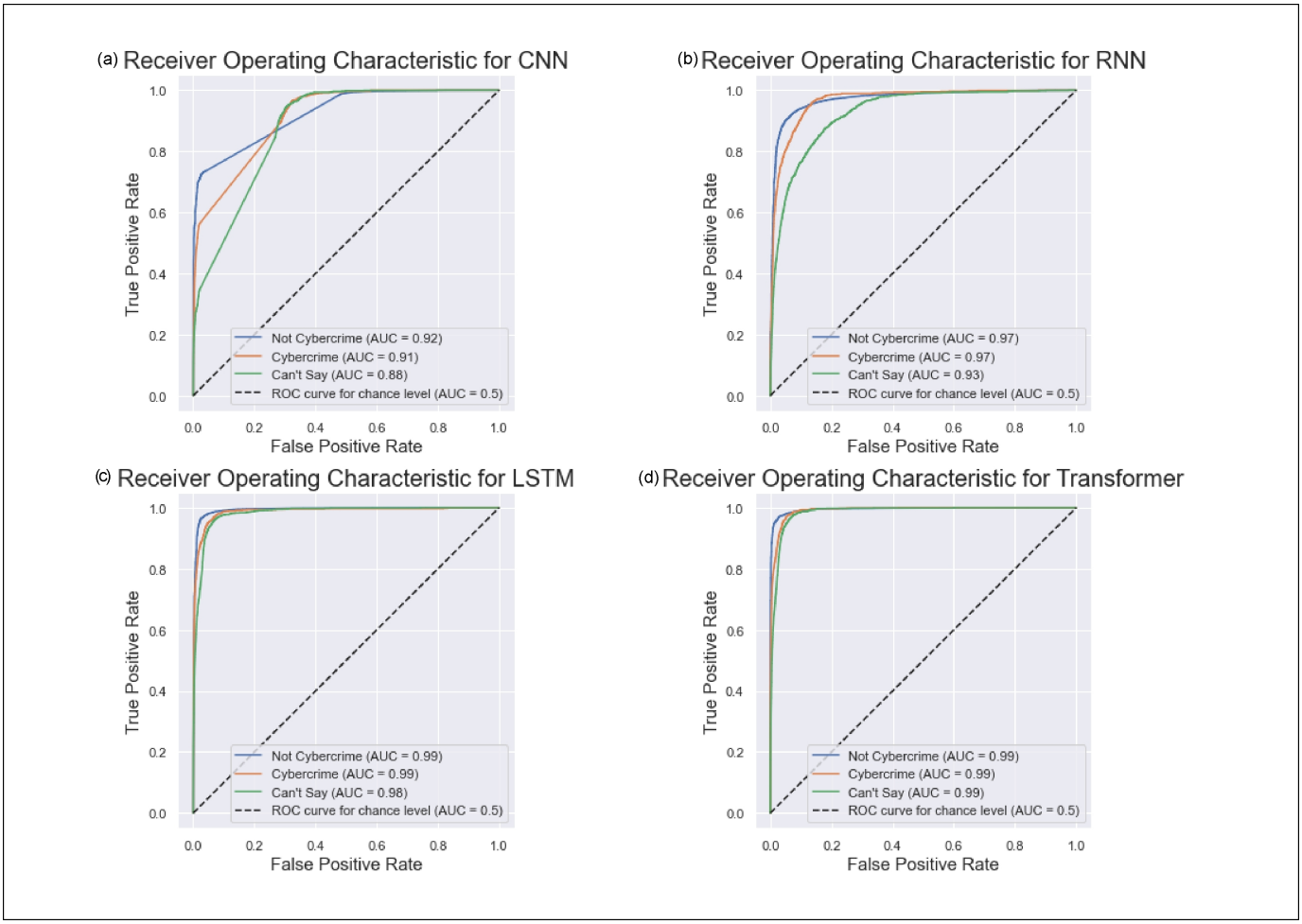


**Figure 10.** The AUC ROC curves of (**a**) CNN, (**b**) RNN, (**c**) LSTM, and (**d**) BERT for multiclass classification.

## 6. Conclusions and Future Work

The *Dark Web* is a platform that is a budding ground for criminals and criminally motivated people as it provides an untraceable and convenient way to carry out a wide range of illegal activities. The anonymity of this platform also boosts users' confidence to indulge more in such criminal practices, because in the back of their minds, the criminals have an idea of being safe from law enforcement agencies. Therefore, to prevent any threat to a person/country, it is imperative to thoroughly inspect all aspects of information gathering, exchange, and interactions over different sources of the *Dark Web*. This will help law enforcement agencies to monitor and track suspicious persons and activities constantly.

To this end, we apply a semi-automated annotation scheme that enables us to identify if activities are directly or remotely related to cybercrimes and to provide contextual cues for classification models to determine the *Category* items for analysis of DNM data. Our observation based on experiments is that context is essential, and the annotation has to consider all the attributes to infer the peculiar and subtle information indicative of the target. Our tailored and hand-picked preprocessing strategies have been beneficial in modeling the raw dataset that ultimately contributed to obtaining optimal performances of employed DL models. The highest accuracy of 96% validates our opted annotation method, which is cost-effective and less resource-demanding. Therefore, the significant contribution of our work is that our results can pave the path for further research and serve as baselines. Other researchers can benefit from the semi-automated annotation method to analyze other DNM datasets for identifying cybercrime.

A minor limitation of such a dataset is that unsupervised approaches (such as clustering) cannot be seen as an alternative to human annotation since the dataset has vague and incomplete item information, resulting in large clusters. Therefore, in future work, we aim to adopt a more human-expert-based annotation scheme to develop a comprehensive *Dark Web* dataset using expert annotators and crowdsourcing (Amazon M-Turk (https://www.mturk.com/, accessed on 27 March 2023 )). The mentioned annotation is proposed to reduce the items belonging to the target class *Can't say if cybercrime* and to also provide subcategories for cybercrime. We did not use this human-expert-based annotation scheme because the *Agora* dataset is large, which would demand resources, time, and expense. Another future goal is to analyze other existing DNMs for extracting useful data related to cybercrimes to create a larger dataset and knowledge graphs to address the domain adaptation challenges and improve DL model performance. Finally, we aim to inspect the fairness of employed models to field test them for real-world scenarios.

**Author Contributions:** Conceptualization, K.S.S., A.S., H.M.P. and V.K.; methodology, K.S.S., A.S., H.M.P. and V.K.; software, K.S.S., A.S., H.M.P. and V.K.; formal analysis, K.S.S., A.S., H.M.P. and V.K.; investigation, K.S.S., A.S. and V.K.; resources, K.S.S., A.S., H.M.P. and V.K.; data curation, K.S.S., A.S. and V.K.; writing—original draft preparation, K.S.S., A.S., H.M.P. and V.K.; writing—review and editing, V.K.; visualization, K.S.S., A.S. and V.K.; supervision, A.S., H.M.P. and V.K.; project administration, A.S. and V.K. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not Applicable.

**Informed Consent Statement:** Not Applicable.

**Data Availability Statement:** The raw *Agora dataset* is available for download at (https://www.kaggle.com/datasets/philipjames11/dark-net-marketplace-drug-data-agora-20142015, accessed on 27 March 2023), and the potential original source of the data can be accessed at (https://gwern.net/dnm-archive#works-using-this-dataset, accessed on 27 March 2023).

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| AI | Artificial Intelligence |
| DL | Deep Learning |
| NLP | Natural Language Processing |
| CTI | Cyber-Threat Intelligence |
| TOR | The Onion Router |
| WWW | World Wide Web |
| CNN | Convolution Neural Network |
| LSTM | Long Short-Term Memory |
| BERT | Bidirectional Encoder Representations from Transformers |
| CT | Counter-Terrorism |
| GAN | Generative Adversarial Network |

## References

1. Pallen, M. Guide to the Internet: The world wide web. *BMJ* **1995**, *311*, 1552–1556.
2. Gehl, R.W. Archives for the dark web: A field guide for study. In *Research Methods for the Digital Humanities*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 31–51.
3. Mancini, S.; Tomei, L.A. The Dark Web: Defined, Discovered, Exploited. *Int. J. Cyber Res. Educ.* **2019**, *1*, 1–12.
4. Jardine, E. The Dark Web dilemma: Tor, anonymity and online policing. *Glob. Comm. Internet Gov. Pap. Ser.* **2015**, *21, 1–13*.
5. Chertoff, M.; Simon, T. The Impact of the Dark Web on Internet Governance and Cyber Security. 2015. Available online: https://policycommons.net/artifacts/1203086/the-impact-of-the-dark-web-on-internet-goverannce-and-cyber-security/1756195/ (accessed on 27 March 2023).
6. Weimann, G. Going dark: Terrorism on the dark web. *Stud. Confl. Terror.* **2016**, *39*, 195–206.

7. Ablon, L.; Libicki, M.C.; Golay, A.A. *Markets for Cybercrime Tools and Stolen Data: Hackers' Bazaar*; Rand Corporation: Santa Monica, CA, USA, 2014.
8. Weimann, G. Terrorist migration to the dark web. *Perspect. Terror.* **2016**, *10*, 40–44.
9. Gupta, A.; Maynard, S.B.; Ahmad, A. The Dark Web Phenomenon: A Review and Research Agenda. 2019. Available online: https://aisel.aisnet.org/acis2019/1/ (accessed on 27 March 2023).
10. Lacson, W.; Jones, B. The 21st century darknet market: Lessons from the fall of Silk Road. *Int. J. Cyber Criminol.* **2016**, *10*, 40.
11. Buxton, J.; Bingham, T. The rise and challenge of dark net drug markets. *Policy Brief* **2015**, *7*, 1–2.
12. Rhumorbarbe, D.; Staehli, L.; Broséus, J.; Rossy, Q.; Esseiva, P. Buying drugs on a Darknet market: A better deal? Studying the online illicit drug market through the analysis of digital, physical and chemical data. *Forensic Sci. Int.* **2016**, *267*, 173–182.
13. Lacey, D.; Salmon, P.M. It's dark in there: Using systems analysis to investigate trust and engagement in dark web forums. In Proceedings of the Engineering Psychology and Cognitive Ergonomics: 12th International Conference, EPCE 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, 2–7 August 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 117–128.
14. Van Hout, M.C.; Bingham, T. Responsible vendors, intelligent consumers: Silk Road, the online revolution in drug trading. *Int. J. Drug Policy* **2014**, *25*, 183–189.
15. Cherqi, O.; Mezzour, G.; Ghogho, M.; El Koutbi, M. Analysis of hacking related trade in the darkweb. In Proceedings of the 2018 IEEE International Conference on Intelligence and Security Informatics (ISI), Miami, FL, USA, 9–11 November 2018; pp. 79–84.
16. Ghosh, S.; Das, A.; Porras, P.; Yegneswaran, V.; Gehani, A. 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, Halifax, NS, Canada, 13–17 August 2017; pp. 1793–1802.
17. Montieri, A.; Ciuonzo, D.; Aceto, G.; Pescapé, A. Anonymity services tor, i2p, jondonym: Classifying in the dark (web). *IEEE Trans. Dependable Secur. Comput.* **2018**, *17*, 662–675.
18. ElBahrawy, A.; Alessandretti, L.; Rusnac, L.; Goldsmith, D.; Teytelboym, A.; Baronchelli, A. Collective dynamics of dark web marketplaces. *Sci. Rep.* **2020**, *10*, 1–8.
19. Broséus, J.; Rhumorbarbe, D.; Mireault, C.; Ouellette, V.; Crispino, F.; Décary-Hétu, D. Studying illicit drug trafficking on Darknet markets: Structure and organisation from a Canadian perspective. *Forensic Sci. Int.* **2016**, *264*, 7–14.
20. Dwyer, A.C.; Hallett, J.; Peersman, C.; Edwards, M.; Davidson, B.I.; Rashid, A. How darknet market users learned to worry more and love PGP: Analysis of security advice on darknet marketplaces. *arXiv* **2022**, arXiv:2203.08557.
21. Zaunseder, A.; Bancroft, A. Pricing of illicit drugs on darknet markets: A conceptual exploration. *Drugs Alcohol Today* **2021**, *21*, 135–145.
22. Zambiasi, D. Drugs on the web, crime in the streets. the impact of shutdowns of dark net marketplaces on street crime. *J. Econ. Behav. Organ.* **2022**, *202*, 274–306.
23. Armona, L. Measuring the Demand Effects of Formal and Informal Communication: Evidence from Online Markets for Illicit Drugs. *arXiv* **2018**, arXiv:1802.08778.
24. Miller, J.N. The war on drugs 2.0: Darknet fentanyl's rise and the effects of regulatory and law enforcement action. *Contemp. Econ. Policy* **2020**, *38*, 246–257.
25. Andrei, F.; Barrera, D.; Krakowski, K.; Sulis, E. Trust intermediary in a cryptomarket for illegal drugs. *Eur. Sociol. Rev.* **2023**, jcad020. https://doi.org/10.1093/esr/jcad020..
26. Hiramoto, N.; Tsuchiya, Y. Are Illicit Drugs a Driving Force for Cryptomarket Leadership? *J. Drug Issues* **2022**, *53*, 451–474
27. Bogensperger, J.; Schlarb, S.; Hanbury, A.; Recski, G. DreamDrug-A crowdsourced NER dataset for detecting drugs in darknet markets. In Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021), Gyeongju, Republic of Korea, 11 November 2021; pp. 137–157.
28. Zhang, Y.; Qian, Y.; Fan, Y.; Ye, Y.; Li, X.; Xiong, Q.; Shao, F. dstyle-gan: Generative adversarial network based on writing and photography styles for drug identification in darknet markets. In Proceedings of the Annual Computer Security Applications Conference, Austin, TX, USA, 7–11 December 2022; pp. 669–680.
29. Manolache, A.; Brad, F.; Barbalau, A.; Ionescu, R.T.; Popescu, M. VeriDark: A Large-Scale Benchmark for Authorship Verification on the Dark Web. In *Advances in Neural Information Processing Systems*; Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2022; Volume 35, pp. 15574–15588.
30. Dearden, T.E.; Tucker, S.E. Follow the Money: Analyzing Darknet Activity Using Cryptocurrency and the Bitcoin Blockchain. *J. Contemp. Crim. Justice* **2023**, *39*, 257–275.
31. Akcora, C.G.; Purusotham, S.; Gel, Y.R.; Krawiec-Thayer, M.; Kantarcioglu, M. How to not get caught when you launder money on blockchain? *arXiv* **2020**, arXiv:2010.15082.
32. Gomez, G.; Moreno-Sanchez, P.; Caballero, J. Watch Your Back: Identifying Cybercrime Financial Relationships in Bitcoin through Back-and-Forth Exploration. In Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, Los Angeles, CA, USA, 7–11 November 2022; pp. 1291–1305.
33. Demant, J.; Munksgaard, R.; Houborg, E. Personal use, social supply or redistribution? Cryptomarket demand on Silk Road 2 and Agora. *Trends Organ. Crime* **2018**, *21*, 42–61.
34. Chen, C.; Peersman, C.; Edwards, M.; Ursani, Z.; Rashid, A. Amoc: A multifaceted machine learning-based toolkit for analysing cybercriminal communities on the darknet. In Proceedings of the 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 15–18 December 2021; pp. 2516–2524.

35. Saxena, V.; Rethmeier, N.; Van Dijck, G.; Spanakis, G. VendorLink: An NLP approach for Identifying & Linking Vendor Migrants & Potential Aliases on Darknet Markets. *arXiv* **2023**, arXiv:2305.02763.
36. Maras, M.H.; Arsovska, J.; Wandt, A.S.; Logie, K. Keeping Pace With the Evolution of Illicit Darknet Fentanyl Markets: Using a Mixed Methods Approach to Identify Trust Signals and Develop a Vendor Trustworthiness Index. *J. Contemp. Crim. Justice* **2023**, *39*, 276–297.
37. Booij, T.M.; Verburgh, T.; Falconieri, F.; van Wegberg, R.S. Get Rich or Keep Tryin'Trajectories in dark net market vendor careers. In Proceedings of the 2021 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), Vienna, Austria, 6–10 September 2021; pp. 202–212.
38. Szigeti, Á.; Frank, R.; Kiss, T. Trust factors in the social figuration of online drug trafficking: A qualitative content analysis on a darknet market. *J. Contemp. Crim. Justice* **2023**, *39*, 167–184.
39. Lokala, U.; Phukan, O.C.; Dastidar, T.G.; Lamy, F.; Daniulaityte, R.; Sheth, A. "Can We Detect Substance Use Disorder?": Knowledge and Time Aware Classification on Social Media from Darkweb. *arXiv* **2023**, arXiv:2304.10512.
40. Cork, A.; Everson, R.; Levine, M.; Koschate, M. Using computational techniques to study social influence online. *Group Process. Intergroup Relations* **2020**, *23*, 808–826.
41. Liu, H.; Zhao, J.; Huo, Y.; Wang, Y.; Liao, C.; Shen, L.; Cui, S.; Shi, J. URM4DMU: An User Representation Model for Darknet Markets Users. In Proceedings of the ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
42. Luong, H.T. Preliminary Findings of the Trends and Patterns of Darknet-Related Criminals in the Last Decade. 2022. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4174766 (accessed on 27 March 2023).
43. Ogbanufe, O.; Baucum, F.; Benjamin, J. Network Analysis of a Darknet Marketplace: Identifying Themes and Key Users of Illicit Networks. 2022. Available online: https://aisel.aisnet.org/wisp2022/15/ (accessed on 27 March 2023).
44. Stoddart, K. Non and Sub-State Actors: Cybercrime, Terrorism, and Hackers. In *Cyberwarfare: Threats to Critical Infrastructure*; Springer International Publishing: Cham, Switzerland, 2022; pp. 351–399. https://doi.org/10.1007/978-3-030-97299-8_6.
45. Maneriker, P.; He, Y.; Parthasarathy, S. SYSML: StYlometry with Structure and Multitask Learning: Implications for Darknet Forum Migrant Analysis. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing; Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 7–11 November 2021; pp. 6844–6857. https://doi.org/10.18653/v1/2021.emnlp-main.548.
46. Baravalle, A.; Lopez, M.S.; Lee, S.W. Mining the dark web: Drugs and fake ids. In Proceedings of the 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), Barcelona, Spain, 12–15 December 2016; pp. 350–356.
47. Fu, T.; Abbasi, A.; Chen, H. A focused crawler for Dark Web forums. *J. Am. Soc. Inf. Sci. Technol.* **2010**, *61*, 1213–1231.
48. Raghavan, S.; Garcia-Molina, H. Crawling the hidden web. In Proceedings of the Vldb, Roma, Italy, 11-14 September 2001; Volume 1, pp. 129–138.
49. Zulkarnine, A.T.; Frank, R.; Monk, B.; Mitchell, J.; Davies, G. Surfacing collaborated networks in dark web to find illicit and criminal content. In Proceedings of the 2016 IEEE Conference on Intelligence and Security Informatics (ISI), Tucson, AZ, USA, 28–30 September 2016; pp. 109–114.
50. Nazah, S.; Huda, S.; Abawajy, J.H.; Hassan, M.M. An Unsupervised Model for Identifying and Characterizing Dark Web Forums. *IEEE Access* **2021**, *9*, 112871–112892. https://doi.org/10.1109/ACCESS.2021.3103319.
51. Yang, L.; Liu, F.; Kizza, J.M.; Ege, R.K. Discovering topics from dark websites. In Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Cyber Security, Nashville, TN, USA, 31 March–2 April 2009; pp. 175–179.
52. L'huillier, G.; Alvarez, H.; Ríos, S.A.; Aguilera, F. Topic-based social network analysis for virtual communities of interests in the dark web. *ACM Sigkdd Explor. Newsl.* **2011**, *12*, 66–73.
53. Porter, K. Analyzing the DarkNetMarkets subreddit for evolutions of tools and trends using LDA topic modeling. *Digit. Investig.* **2018**, *26*, S87–S97.
54. Ríos, S.A.; Muñoz, R. Dark web portal overlapping community detection based on topic models. In Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics, New York, NY, USA, 12 August 2012; pp. 1–7.
55. Sachan, A. Countering terrorism through dark web analysis. In Proceedings of the 2012 Third International Conference on Computing, Communication and Networking Technologies (ICCCNT'12), Coimbatore, India, 26–28 July 2012; pp. 1–5. https://doi.org/10.1109/ICCCNT.2012.6396055.
56. Kramer, S. Anomaly detection in extremist web forums using a dynamical systems approach. In Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics, Washington, DC, USA, 25 July 2010; pp. 1–10.
57. Arnold, N.; Ebrahimi, M.; Zhang, N.; Lazarine, B.; Patton, M.; Chen, H.; Samtani, S. Dark-net ecosystem cyber-threat intelligence (CTI) tool. In Proceedings of the 2019 IEEE International Conference on Intelligence and Security Informatics (ISI), Shenzhen, China, 1–3 July 2019; pp. 92–97.
58. Dalvi, A.; Patil, G.; Bhirud, S. Dark Web Marketplace Monitoring-The Emerging Business Trend of Cybersecurity. In Proceedings of the 2022 International Conference on Trends in Quantum Computing and Emerging Business Technologies (TQCEBT), Maharashtra, India, 13–15 October 2022; pp. 1–6.
59. Nazah, S.; Huda, S.; Abawajy, J.; Hassan, M.M. Evolution of dark web threat analysis and detection: A systematic approach. *IEEE Access* **2020**, *8*, 171796–171819.

60. Nunes, E.; Diab, A.; Gunn, A.; Marin, E.; Mishra, V.; Paliath, V.; Robertson, J.; Shakarian, J.; Thart, A.; Shakarian, P. Darknet and deepnet mining for proactive cybersecurity threat intelligence. In Proceedings of the 2016 IEEE Conference on Intelligence and Security Informatics (ISI), Tucson, AZ, USA, 28–30 September 2016; pp. 7–12.
61. Benjamin, V.; Li, W.; Holt, T.; Chen, H. Exploring threats and vulnerabilities in hacker web: Forums, IRC and carding shops. In Proceedings of the 2015 IEEE International Conference on Intelligence and Security Informatics (ISI), Baltimore, MD, USA, 27–29 May 2015; pp. 85–90.
62. Robertson, J.; Paliath, V.; Shakarian, J.; Thart, A.; Shakarian, P. Data driven game theoretic cyber threat mitigation. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; Volume 30, pp. 4041–4046.
63. Pineau, T.; Schopfer, A.; Grossrieder, L.; Broséus, J.; Esseiva, P.; Rossy, Q. The study of doping market: How to produce intelligence from Internet forums. *Forensic Sci. Int.* **2016**, *268*, 103–115.
64. Al Nabki, M.W.; Fidalgo, E.; Alegre, E.; De Paz, I. 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, Valencia, Spain, 3–7 April 2017; pp. 35–43.
65. Abbasi, A.; Chen, H. Affect intensity analysis of dark web forums. In Proceedings of the 2007 IEEE Intelligence and Security Informatics, New Brunswick, NJ, USA, 23–24 May 2007; pp. 282–288.
66. Glancy, F.H.; Yadav, S.B. A computational model for financial reporting fraud detection. *Decis. Support Syst.* **2011**, *50*, 595–601.
67. Holt, T.J.; Strumsky, D.; Smirnova, O.; Kilger, M. Examining the social networks of malware writers and hackers. *Int. J. Cyber Criminol.* **2012**, *6*, 891–903.
68. Jordan, T.; Taylor, P. A sociology of hackers. *Sociol. Rev.* **1998**, *46*, 757–780.
69. Habibi Lashkari, A.; Kaur, G.; Rahali, A. Didarknet: A contemporary approach to detect and characterize the darknet traffic using deep image learning. In Proceedings of the 2020 the 10th International Conference on Communication and Network Security, Tokyo, Japan, 27–29 November 2020; pp. 1–13.
70. Ebrahimi, M.; Nunamaker, J.F., Jr.; Chen, H. Semi-supervised cyber threat identification in dark net markets: A transductive and deep learning approach. *J. Manag. Inf. Syst.* **2020**, *37*, 694–722.
71. Iliadis, L.A.; Kaifas, T. Darknet traffic classification using machine learning techniques. In Proceedings of the 2021 10th International Conference on Modern Circuits and Systems Technologies (MOCAST), Thessaloniki, Greece, 5–7 July 2021; pp. 1–4.
72. Zhang, Y.; Zeng, S.; Fan, L.; Dang, Y.; Larson, C.A.; Chen, H. Dark web forums portal: Searching and analyzing jihadist forums. In Proceedings of the 2009 IEEE International Conference on Intelligence and Security Informatics, Dallas, TX, USA, 8–11 June 2009; pp. 71–76.
73. Scanlon, J.R.; Gerber, M.S. Automatic detection of cyber-recruitment by violent extremists. *Secur. Inform.* **2014**, *3*, 1–10.
74. Chen, H. Sentiment and affect analysis of dark web forums: Measuring radicalization on the internet. In Proceedings of the 2008 IEEE International Conference on Intelligence and Security Informatics, San Antonio, TX, USA, 2–3 November 2008; pp. 104–109.
75. Zhou, Y.; Reid, E.; Qin, J.; Chen, H.; Lai, G. US domestic extremist groups on the Web: Link and content analysis. *IEEE Intell. Syst.* **2005**, *20*, 44–51.
76. Branwen, G.; Christin, N.; Décary-Hétu, D.; Andersen, R.M.; Presidente, E.; Lau, D.; Sohhlz, D.K.; Cakic, V. Dark Net Market Archives, 2011–2015. Available online: https://gwern.net/dnm-archive (accessed on 27 March 2023 ).
77. Dessì, D.; Helaoui, R.; Kumar, V.; Recupero, D.R.; Riboni, D. TF-IDF vs. Word Embeddings for Morbidity Identification in Clinical Notes: An Initial Study. In Proceedings of the First Workshop on Smart Personal Health Interfaces Co-Located with 25th International Conference on Intelligent User Interfaces, SmartPhil@IUI 2020, Cagliari, Italy, 17 March 2020; pp. 1–12.
78. Kumar, V.; Verma, A.; Mittal, N.; Gromov, S.V. Anatomy of Preprocessing of Big Data for Monolingual Corpora Paraphrase Extraction: Source Language Sentence. *Emerg. Technol. Data Min. Inf. Secur.* **2019**, *3*, 495.
79. Kumar, V.; Recupero, D.R.; Riboni, D.; Helaoui, R. Ensembling Classical Machine Learning and Deep Learning Approaches for Morbidity Identification From Clinical Notes. *IEEE Access* **2021**, *9*, 7107–7126. https://doi.org/10.1109/ACCESS.2020.3043221.
80. Uysal, A.K.; Gunal, S. The impact of preprocessing on text classification. *Inf. Process. Manag.* **2014**, *50*, 104–112.
81. Bhandari, A.; Kumar, V.; Thien Huong, P.T.; Thanh, D.N. Sentiment analysis of COVID-19 tweets: Leveraging stacked word embedding representation for identifying distinct classes within a sentiment. In *Artificial Intelligence in Data and Big Data Processing: Proceedings of ICABDE 2021*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 341–352.
82. Kumar, V.; Mishra, B.K.; Mazzara, M.; Thanh, D.N.; Verma, A. Prediction of malignant and benign breast cancer: A data mining approach in healthcare applications. In *Advances in Data Science and Management*; Springer: Berlin/Heidelberg, Germany, 2020.
83. Wu, Z.; Balloccu, S.; Kumar, V.; Helaoui, R.; Reforgiato Recupero, D.; Riboni, D. Creation, Analysis and Evaluation of AnnoMI, a Dataset of Expert-Annotated Counselling Dialogues. *Future Internet* **2023**, *15*, 110. https://doi.org/10.3390/fi15030110.
84. Wu, Z.; Balloccu, S.; Kumar, V.; Helaoui, R.; Reiter, E.; Recupero, D.R.; Riboni, D. Anno-mi: A dataset of expert-annotated counselling dialogues. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 6177–6181.
85. Kumar, V.; Balloccu, S.; Wu, Z.; Reiter, E.; Helaoui, R.; Recupero, D.; Riboni, D. Data Augmentation for Reliability and Fairness in Counselling Quality Classification. In *1st Workshop on Scarce Data in Artificial Intelligence for Healthcare-SDAIH, INSTICC*; SciTePress: Setúbal, Portugal, 2023; pp. 23–28. https://doi.org/10.5220/0011531400003523.
86. Kumar, V.; Reforgiato Recupero, D.; Helaoui, R.; Riboni, D. K-LM: Knowledge Augmenting in Language Models Within the Scholarly Domain. *IEEE Access* **2022**, *10*, 91802–91815. https://doi.org/10.1109/ACCESS.2022.3201542.

87. Liu, P.; Qiu, X.; Huang, X. Recurrent neural network for text classification with multi-task learning. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, New York, NY, USA, 9–15 July 2016; pp. 2873–2879.
88. Medsker, L.; Jain, L.C. *Recurrent Neural Networks: Design and Applications*; CRC Press: Boca Raton, FL, USA, 1999.
89. LeCun, Y.; Kavukcuoglu, K.; Farabet, C. Convolutional networks and applications in vision. In Proceedings of the 2010 IEEE International Symposium on Circuits and Systems, Paris, France, 30 May–2 June 2010; pp. 253–256.
90. Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, G.; Cai, J.; et al. Recent advances in convolutional neural networks. *Pattern Recognit.* **2018**, *77*, 354–377.
91. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780.
92. Graves, A.; Graves, A. Long short-term memory. In *Supervised Sequence Labelling with Recurrent Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 37–45.
93. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–11.
94. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
95. Beauxis-Aussalet, E.; Hardman, L. Simplifying the visualization of confusion matrix. In Proceedings of the 26th Benelux Conference on Artificial Intelligence (BNAIC), Nijmegen, The Netherlands, 6–7 November 2014.
96. Mandrekar, J.N. Receiver operating characteristic curve in diagnostic test assessment. *J. Thorac. Oncol.* **2010**, *5*, 1315–1316.
97. Hajian-Tilaki, K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Casp. J. Intern. Med.* **2013**, *4*, 627.