

# C-OSINT: COVID-19 Open Source artificial INTelligence framework

Leonardo Ranaldi<sup>1,2,\*</sup>, Aria Nourbakhsh<sup>2</sup>, Francesca Fallucchi<sup>1</sup> and Fabio Massimo Zanzotto<sup>2</sup>

<sup>1</sup>*Department of Innovation and Information Engineering,  
Guglielmo Marconi University, Roma, Italy*

<sup>2</sup>*Department of Enterprise Engineering,  
University of Rome Tor Vergata, Roma, Italy*

## Abstract

With the emergence of COVID-19 disease worldwide, a market of the products related to this disease formed across the Internet. By the time these goods were in short supply, many uncontrolled Dark Web Marketplaces (DWM) were active in selling these products. At the same time, Dark Web Forums (DWF) became proxies for spreading false ideas, fake news about COVID-19, and advertising products sold in DWMs. This study investigates the activities entertained in the DWMs and DWFs to propose a learning-based model to distinguish them from their related counterparts on the surface web. To this end, we propose a COVID-19 Open Source artificial INTelligence framework (C-OSINT) to automatically collect and classify the activities done in DWMs and DWFs. Moreover, we incorporate linguistic and stylistic solutions to leverage the classification performance between the content found in DWMs and DWFs and two surface web sources. Our results show that using syntactic and stylistic representation outperforms the Transformer based results over these domains.

## Keywords

Machine Learning, Natural Language Processing, COVID-19, Dark Web, Cyberspace

## 1. Introduction

By the end of 2019, COVID-19, a respiratory disease, emerged that caused financial and health crises around the world. Consequently, many countries and health organizations started to respond to the pandemic. To stop and slow down the mortality rate of the disease, many vaccines were proposed, and the first batch of them in late 2020 was officially approved. Vaccines from Pfizer/BioNTech [1], Moderna [2], and Sputnik [3] were among the most famous and utilized brands. The unbalanced distribution of vaccine doses and the race to access the first dose soon generated concerns about illegal trades of the vaccine. Europol and other national security agencies reported the sale of fake COVID-19 vaccines on Dark Web Marketplaces (DWMs) on December 2020 [4, 5, 6, 7, 8]. Monitoring DWMs is therefore critical to enable police and public health agencies to be prepared and effectively counter these threats.


---


*ITASEC'22: Italian Conference on Cybersecurity, June 20–23, 2022, Rome, Italy*

\*Corresponding author.

✉ l.ranaldi@unimarconi.it (L. Ranaldi); nrbrai@uniroma2.it (A. Nourbakhsh); f.fallucchi@unimarconi.it (F. Fallucchi); fabio.massimo.zanzotto@uniroma2.it (F. M. Zanzotto)

ORCID 0000-0001-8488-4146 (L. Ranaldi)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

Interpol and Europol said that DWMs had become proxies for online trafficking of masks, COVID-19 tests, and alleged drugs constantly advertised on these platforms. A similar issue happened with the use of vaccines and the start of vaccination campaigns [9, 10]. The matter got exacerbated by the birth of the green pass as a document that would enable people to have public activities such as using public transportation and visiting public spaces [11]. At the same time, Dark Web Forums (DWFs) have been the subject of the proliferation of arguments and the spreading of fake information related to COVID-19. Linking these activities is not an easy task [12].

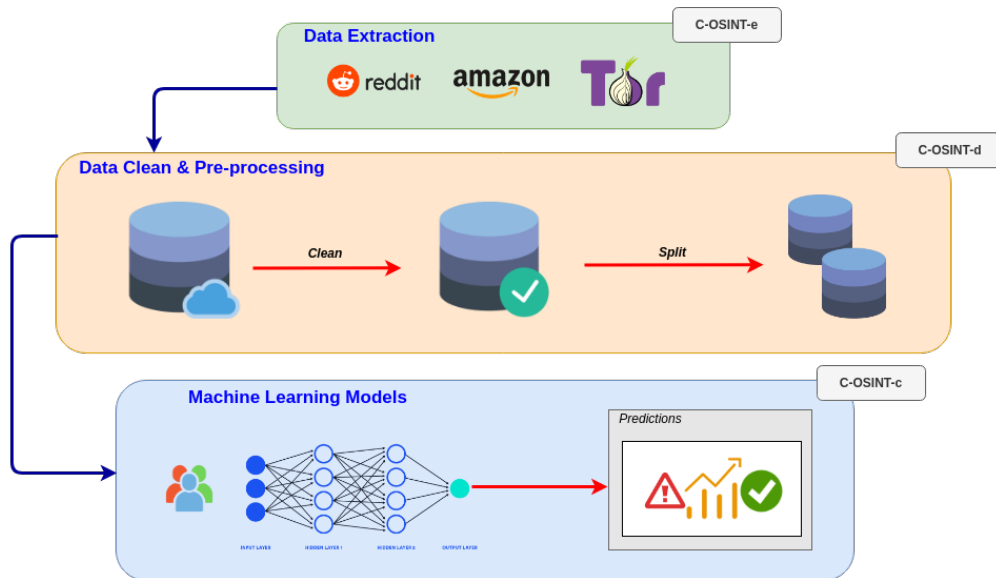
DWFs are a great place to get into illicit online activities, and DWMs can be easily accessed through specialized browsers, such as Tor [13], I2P [14] and FreeNet [15]. These browsers guarantee users' anonymity, and in turn, trades of many illegal goods such as drugs, firearms, credit cards, and fake documents are being conducted in them [16]. The growing popularity of Dark Web activities has attracted the interest of the scientific community, and security researchers to provide comparative analyses of the different DWMs [17, 18, 19, 20, 21] and DWFs [22, 23, 24]. Among the most credited are studies that propose automatic recognition and classification of activities and analysis of lexicon used in DWMs and DWFs [25, 26, 27]. According to numerous reports, law enforcement has successfully closed several illegal DWMs [28, 29]. Still, DWMs are inherently resilient to these interventions, and in 2020 COVID-19 disease provided another reason to analyze and classify the content produced in this particular domain.

In this study, we investigate the activities entertained in the DWMs and DWFs to propose a learning-based model to recognize their contents compared to the data from the surface web. To this end, we propose a COVID-19 Open Source artificial INTElligence framework (C-OSINT) to automatically collect and classify the textual content created in DWMs and DWFs compared to Reddit and Amazon. Our C-OSINT model consists of three parts: (1) the corpus extraction system C-OSINT-e, which is used to extract data from DWMs and DWFs; (2) the cleaning system C-OSINT-d, which cleans and does some pre-processing to build the final corpus; (3) the classification system C-OSINT-c, which classifies the '*onion*' service to determine whether the service is from a marketplace or forum (from Dark Web and surface web) by using the HTML text of the pages and applying Natural Language Processing algorithms. The rest of the paper is organized as follows. Section 2 describes state-of-the-art studies on Dark Web activities and how to identify them with automatically generated heuristics. Section 3 describes our C-OSINT-e, C-OSINT-d and Section 4 describes our C-OSINT-c. Finally, in section 5 we present the result of C-OSINT-c classifications and provide a discussion of the obtained results.

## 2. Background and Related Work

Since the 2000s, many have researched methods of classifying surface web content [30, 31, 32]. More recently, some attempts to classify the non-indexed part of the web, called the Deep Web [33, 34], and then with the ancestor of today's Dark Web (DW), [35, 36] have been published.

With the growth in popularity, the DW has become a research subject in many studies. Barratt et al. [17] and Aldridge et al. [18] have done extensive investigations of customers of DWMs taking the '*Silk Road*' phenomenon as a use case. Yang et al. [22] and Pete et al. [37], on the



**Figure 1:** C-OSINT Framework.

other hand, addressed the social relationships undertaken by users of DWFs. The two analytical activities, while fundamental to understanding the dynamics of the social networks that are created around the DW have remained highly contextualized. One of the first works that shifted the focus to automatic content classification was done by Biryukov et al. [27]. They classified the content of the DW, restricting the study only to Tor's hidden services resulting in 18 topical categories. By limiting the topic to drug trades, Graczyk et al.[38] combined unsupervised feature selection and an SVM classifier to classify drug selling services.

The first real distinction between activities, as selling services rather than forums, was proposed in [25, 39]. They presented DUTA (Darknet Usage Text Addresses), the first publicly available Darknet dataset, with a classification into topical categories and subcategories. Avarikioti et al.[40] on the other hand, were the first to focus only on the classification of illegal and legal activities, so they built a new dataset and used an SVM classifier in an active learning setting with a bag-of-words feature representation and got very good results. Recently, Choshen et al.[26], following [40] and using the updated version of the publicly available DUTA [41], studied the style and structure of hidden illegal and legal services. Choshen et al.[26] proposed some excellent classifiers that were based on shallow heuristics and converted the input text into part of speech (POS) tags. Their obtained results were satisfactory but at the same time evaded much important information such as sentence structure and basic semantics, and they converted some different symbols into a single symbol, ignoring many typical symbols peculiar to the DW domain.

In this paper, we propose an Open Source artificial INTelligence (OSINT) framework to automatically collect and classify activities entertained in DWMs and DWFs on a new emerging topic: COVID-19, from the same type of services on the surface web, namely Amazon and Reddit. Since the start of the production of vaccines and the obligation of the green pass certificate,

some stores began to sell them [9, 10, 11, 4] which may pose a significant risk to public health. Consequently, we propose a comparative analysis using two surface web platforms to show that our framework can differentiate the domain where the activity is taking place.

### 3. Data

In this article, we aim to analyze the COVID-19 topic in the most popular Dark Web Marketplaces (DWMs) and Dark Web Forums (DWFs) between 2020-2021 (see Appendix A.2, A.3), to create a framework capable of: a) collecting information from '.onion' services, b) recognizing activities in DWMs and DWFs for monitoring and warning of abuse. To solve this need, we analyzed the current methods to extract and classify the activities in subsection 3.1. To obtain data, we propose our framework, which consists of: a crawler and scraper to collect the data (C-OSINT-e) described in the subsection 3.1.1; a pre-processor of the extracted text and a labeling step (C-OSINT-d) described in the subsection 3.1.2; a set of classifiers based on machine learning models (C-OSINT-c).

#### 3.1. DarkNet Dataset

Obtaining and investigating data from the Dark Web is very complex due to the nature of the service and obstacles such as text and image-based CAPTCHAs or the absence of public DNS.

Current monitoring pipelines have the first objective of isolating suspicious domains from normal ones and classifying them into categories. These components are based on keyword heuristics, which are difficult to keep up to date and prone to false positives given the high rate of polysemy. There are other heuristics based on automatic learning, but they are highly dependent on datasets [26]. One of the first public datasets obtained from Dark Web was the first version of "Darknet Usage Text Addresses" (DUTA) [25]. Although an updated version, DUTA-10K [41], has been released, the dataset is obsolete because many of its links are currently down.

In this research, our first contribution is the system C-OSINT-e, which extracts text from '.onion' services from DWMs and DWFs. Similar to the strategy proposed in [25], C-OSINT-e is based on an extraction step, cleaning phase, and finally, labeling of the extracted samples. From the [7] report, it is possible to identify several DWMs that have COVID-19 related products available. Similarly, it is possible to analyze some DWFs, as proposed in [42]. Furthermore, to perform a comparative analysis and have corpora from DWMs and DMFs at our disposal, we did the same process on two very famous surface web services: Reddit<sup>1</sup> and Amazon<sup>2</sup>. These two surface web services were chosen because Reddit is very similar to the structure of DWFs, and Amazon is the largest online store. A screenshot of DWFs is shown in Figure 2 in the appendix, and some examples of the cleaned corpus can be seen in table 1.

---

<sup>1</sup><https://www.reddit.com/>

<sup>2</sup><https://www.amazon.com/>

| Sentence  | Corpus               |
|---|----------------------|
| We provide COVID-19 vaccine, Green Pass, Fake Tests   | Dark Web Marketplace |
| We ship Green Pass and QR code valid throughout Europe payment in BTC and immediate delivery.   | Dark Web Marketplace |
| Fake pandemic and vaccine speculation   | Dark Web Forum       |
| The fake pandemic is caused by the Jews who are ready to speculate on human as in Israel all lined up to vaccinate                          | Dark Web Forum       |
| Polonord Adeste 5 Nasal Rapid Test Kit for SARS-CoV-2 Antigen (Nasal Swab) for Self-Diagnosis, 5 Units (1 pack of 5 rapid tests)            | Amazon               |
| CLINTEST Rapid Covid-19 Antigen Self-Test   | Amazon               |
| To be extra cautious, rotate such masks every three days  | Reddit               |
| Safety was fine, not able to show efficacy. Since they didn't release the data we don't know how ineffective but that is what was reported. | Reddit               |

**Table 1**  
Examples taken from the DWMs, DWFs, Reddit, and Amazon corpora.

### 3.1.1. Extraction

Extracting domains using Tor is a complex task as there is no public DNS server where all hidden service addresses (HS) are registered. In Tor, there is a Hidden Service Directory (HSDir), which Tor relies on it, and it functions as an intermediate point between an HS, as it publishes its descriptors and clients, which communicate with it to learn the address of the HS introduction points [27]. However, a Tor needs a specific flag to be assigned by Tor to authorities to function as an HSDir.

Our C-OSINT-e works similar to the method proposed by Al Nabki et al.[41]. Instead of querying the flag, we use a custom crawler that uses a Tor socket to retrieve onion web pages and new addresses through the 9050 port using: online notepad services on the Surface Web, Tor network search engines, and hyperlinks from the DUTA dataset. Each service is being visited and then recursively extracts '.onion' links which are then cleaned, and duplicate and inactive links are being removed. Finally, the HTML code gets downloaded using the functions implemented in the selenium library.

The code used to perform scraping of both the Dark Web and surface web corpora is available at the following GitHub repository<sup>3</sup>. The time period of data collection for corpus construction is from November 2020 through March 2021. Appendix A.2 and Appendix A.3 show the list of '.onion' services analyzed.

### 3.1.2. Pre-processing & Labeling

The division into paragraphs and the cleaning of the dataset are done by the C-OSINT-d module, following the methodology proposed by Choshen et al.[26]. In all experiments, we apply a cleaning to the text of the corpora web pages. HTML markups are removed from the original

<sup>3</sup><https://github.com/ART-Group-it/C-OSINT>

dataset; the same is done for non-linguistic contents such as buttons, encryption keys, metadata and URLs. Despite applying these pre-processing steps, the remaining textual elements are unclear, and in some cases, unintelligible as domain-specific slang and abbreviations are widely used on the Dark Web.

The labeling process of new samples is carried out in two steps: 1) text classifier proposed previously; 2) sharing of manually assigned tags. The main rules defined in [25] consist of labeling a domain based only on the textual content visible to the user; a domain should receive only one tag based on its activity. In case of uncertainties, an open discussion is established with the rest of the authors.

### **3.2. Surface Web**

For the other two additional datasets from legal sources, we compiled a corpus of Amazon and Reddit pages of similar sizes and characteristics. Amazon is the largest hosting site for sellers of various goods. The corpus from Amazon contains 630 item descriptions, each consisting of more than one sentence. The item descriptions vary by price, item sold, and seller. The descriptions were selected by searching Amazon for terms related to COVID-19 and selecting search patterns to avoid excessive repetition. The search queries also included filtering by price so that each query would result in different items. Due to sellers' advertising strategies or geographic dispersion, the Amazon corpus contains both formal and informal language, and some item descriptions contain abbreviations and domain-specific words. Reddit is a social news, entertainment, and forum website where registered users can publish content in textual posts or hyperlinks. The corpus from Reddit contains 630 discussions on topics related to COVID-19. The source codes to reconstruct the two datasets can be found in the GitHub repository.

## **4. Methods**

In this section, we experiment with our C-OSINT framework to investigate which Natural Language Processing (NLP) algorithm achieves the best result in classifying the activities entertained in DWMs and DWFs.

The C-OSINT-c module is where our text classification experiments are done to find the essential linguistic features that distinguish the activities entertained in different services. Another goal of the classification task is to observe whether these tasks can be solved by holistic Transformers, lexical models, syntactic models, stylistic models, and models derived from the union of the previous ones.

### **4.1. Methods: Classification Models**

The models proposed in this section aim to cover all linguistic needs in the study of style, lexicon, and semantics.

**Holistic Transformers** These classifiers are based on Transformers-models [43] and seem to achieve state-of-the-art results in many text classification tasks.

We tested the following Transformer models to cover the majority of cases of pre-training size (see Table 2) and models:

- $BERT_{base}$  [44], that is Bidirectional Encoder Representations from Transformers, and is trained on the BooksCorpus [45] and English Wikipedia.
- $BERT_{multi}$ , that is the Multi-Language version of BERT [46] and is trained on a Wikipedia dump of 100 languages.
- XLNet [47] is based on a generalized autoregressive pre-training technique that allows the learning of bidirectional contexts by maximizing the expected likelihood over all permutations of the factorization order. This architecture is trained from datasets gathered from the surface web such as Wikipedia, Bookcorpus, Giga5, Clueweb, and Common Crawl.
- ERNIE [48] to improve some of BERT’s problems, introduced a language model representation that uses an external knowledge graph for named entities. ERNIE is pre-trained on Wikipedia corpus and Wikidata knowledge base.
- ELECTRA [49] proposes a mechanism of “corrupting” the input token that is replaced with a token that potentially fits the place. The training procedure is a classification of each token, whether it is a corrupted input or not. This model is trained on the same dataset as BERT.
- DistilBERT [50] proposes a method for pre-training a smaller, general-purpose language representation model, much like BERT, that can then be tuned with good performance on a wide range of tasks like its larger counterparts.
- RoBERTa [51] appears to be a replication study of the pre-training BERT, with the major difference being the focus on the impact of many key hyperparameters and the size of the training data. Indeed, it appears that BERT is under-trained in some respects, and changing the choice of hyperparameters may make a difference on some tasks.

All models described above were implemented using the official implementations coming from the Huggingface Transformers library [52].

**Stylistic Classifier** This classifier is used to determine if the proposed tasks are sensitive to syntactic and lexical information, and thus there is a stylistic difference between the source domains. We would expect texts associated with selling merchandise to be written more formally with pre-defined structures. In contrast, users utilize different styles to express their ideas in texts from forums with no strict rules. Among other differences, we can point to the use of capital letters, possible emoticons, and interjections in forum texts. For this purpose, we apply two models, one purely based on word-level features and one based on shallow syntactic structure.

*Bleaching text* [53] is a model proposed to capture the style of writing at the word level. A linear SVM classifier is applied over the final representation, which concatenates all the ‘bleached’ strings treated as a binary bag-of-word model.

*Part-of-speech tags (POS)* [54] are unique labels assigned to each token (word) to indicate the grammatical categories and other information such as tense and number (plural/singular) of the words. A vanilla feed-forward neural networks (FFNN) classifier is applied to the final



| <i>Corpus</i>                        | <i>Size</i>  |
|--------------------------------------|--------------|
| BooksCorpus [45]                     | 800M words   |
| 2010-and-2014-English Wikipedia dump | 2,500M words |
| Giga5 [55]                           | 16GB         |
| Common Crawl [56]                    | 110GB        |
| ClueWeb [57]                         | 19GB         |
| Penn Treebank [58]                   | 1M words     |

**Table 2**

Corpora used in training pre-trained Transformers and word embeddings. All corpora are derived from the surface web.

representation, the concatenation of all converted strings treated as a binary bag-of-word model. This model is trained with 300 dimensions for five epochs. The FFNN consists of an input layer of dimension 300 and 2 hidden layers of 150 and 50 dimensions with the *ReLU* activation function.

**Lexical-based Neural Networks** We used a classifier based on a vanilla feed-forward neural networks (FFNN) over a bag-of-word-embedding (BoE) representation of sentences to answer this question. This classifier is used to determine whether the proposed tasks can be described and classified through pre-trained word embeddings. In BoE, sentence representations are computed as the summation of the embedding of each constituent word of samples in our dataset. For this classification method, we used *GloVe* word embeddings [59] trained on 2014 Wikipedia dumps and Giga5. The FFNN used with Glove representation consists of an input layer of 300 dimensions and two hidden layers of 150 and 50 dimensions with the *ReLU* activation function, and it was trained for five epochs as well.

**Syntactic-based Neural Networks** Finally, to evaluate the role of “pre-trained” universal syntactic models, we used the Kernel-inspired Encoder with Recursive Mechanism for Interpretable Trees (KERMIT) [60]. This model positively exploits parse trees in neural networks as it increases pre-trained Transformers’ performance when used in combined models. The version used in the experiments encodes parse-trees in vectors of 4,000 dimensions. The rest of the FFNN comprises two hidden layers of 4,000 and 2,000 dimensions. Finally, the output layer consists of 2 dimensions for classification. Between each layer, the *ReLU* activation function and a dropout of 0.1 was used to avoid overfitting on the train data.

KERMIT model exploits the parse trees produced by a traditional parser. As advised by Zanzotto et al. [60], we used the English constitution-based parser, CoreNLP library [61].

## 4.2. Experimental set-up

Using C-OSINT-e and C-OSINT-d, we created four datasets: two from Dark Web Marketplaces and Forums and two from Surface Web. Each corpus contains 630 examples labeled either ‘forum’ or ‘market’. In the experiments, the datasets were merged, building four balanced comparisons, they were split into training and test sets with a 70/30 ratio. The evaluation was done by extracting the accuracy of classification outputs.



|                                     | <b>Dark Forums</b><br>vs<br><b>Dark Market</b> | <b>Reddit</b><br>vs<br><b>Amazon</b> | <b>Dark Market</b><br>vs<br><b>Amazon</b> | <b>Dark Forum</b><br>vs<br><b>Reddit</b> |
|-------------------------------------|--|--------------------------------------|---|--|
| <b>Holistic Transformers</b>        |  |                                      |   |  |
| <i>BERT<sub>base</sub></i>          | 66.83(±3.8)                                    | 75.56(±3.7)                          | 66.17(±4.4)                               | 71.11(±3.2)                              |
| <i>BERT<sub>multi</sub></i>         | 59.98(±2.8)                                    | 62.49(±1.9)                          | 54.66(±4.9)                               | 61.08(±4.5)                              |
| <i>Electra</i>                      | 62.54(±1.9)                                    | 73.86(±3.6)                          | 63.49(±4.3)                               | 72.22(±3.4)                              |
| <i>XLNet</i>                        | 54.29(±2.3)                                    | 67.72(±2.1)                          | 52.49(±4.9)                               | 64.6(±4.2)                               |
| <i>Ernie</i>                        | 65.08(±1.8)                                    | 76.59(±1.7)                          | 67.67(±3.8)                               | 75.56(±2.7)                              |
| <i>RoBerta</i>                      | 51.7(±1.8)                                     | 51.3(±3.2)                           | 53.49(±2.9)                               | 50.9(±1.9)                               |
| <i>DistilBERT</i>                   | 68.02(±5.1)                                    | 67.72(±4.2)                          | 66.83(±5.1)                               | 67.28(±2.7)                              |
| <b>Lexical Models</b>               |  |                                      |   |  |
| <i>BoE(GloVe)</i>                   | 84.38(±0.6)                                    | 87.3(±0.9)                           | 82.54(±0.8)                               | 73.54(±1.8)                              |
| <b>Syntactic Models:</b>            |  |                                      |   |  |
| <i>KERMIT</i>                       | <b>91.21</b> (±1.1)                            | <b>97.86</b> (±1.4)                  | 88.89(±1.2)                               | 94.37(±1.3)                              |
| <b>Stylistic models:</b>            |  |                                      |   |  |
| <i>Bleaching text</i>               | 89.79(±0.5)                                    | 94.66(±0.8)                          | 96.39(±0.6)                               | 92.92(±0.7)                              |
| <i>FFNN (POS)</i>                   | 90.07(±1.3)                                    | 97.22(±2.1)                          | <b>97.63</b> (±0.9)                       | <b>95.8</b> (±0.8)                       |
| <b>Lexical and Syntactic Models</b> |  |                                      |   |  |
| <i>BoE(GloVe) + KERMIT</i>          | 90.21(±1.3)                                    | 96.56(±1.6)                          | 96.03(±1.5)                               | 94.71(±1.8)                              |

**Table 3**

Accuracy of the different models. Experiments with neural networks are obtained over 5 runs with different seeds.

## 5. Results and Discussion

Looking at the performance of different approaches in the same dataset setting helps us compare their ability to tackle the problem of classifying markets and forums from dark and surface net. Results of the experiments are reported in Tab. 3 with the configurations described in Sec. 4.

These results show the unexpected behavior of the applied models because Transformers have poor performance on these uncovered domains. Although the *BoE(GloVe)* lexical scores better than the Transformer, it still lags behind the other syntactic and stylistic approaches. This poor performance can be attributed to the data that these models were trained on: all these representations were trained on surface web datasets which cannot generalize to the data from the Dark Web.

The other results for the proposed tasks are mixed, but the trend is that all work better than the Transformers. Stylistic models perform on par with syntactic models. The tasks where stylistic models perform better are those that classify surface web services against their Dark Web counterparts. These results are of two kinds: (1) data from the Dark Web have a writing style that can be captured through distinctive features such as all capital letters, abbreviations, punctuation. (2) A bag-of-words representation of POS-tags can be a distinguishing factor between Dark and surface Web services Reddit and Amazon. The distinction is less evident for DFMs and DWMs.

Neural network models based on syntax have engaging performances on this dataset. Here, KERMIT [60] works better than Transformers, showing that these tasks are sensitive concern-

ing syntactic information that the Transformers cannot transfer to another unseen domain. Moreover, although KERMIT uses a parser trained on the surface web to parse sentences [62], syntactic rules are more restricted than semantic and discourse-level information captured by the Transformers. Yet, it can find the variations among these different domains. However, the combined “pre-trained” lexical and syntactic model, *BoE(GloVe) + KERMIT*, do not outperform the two models separately.

In conclusion, monitoring the activities on the Dark Web and comparing them with their similar surface services is an ongoing challenge. Using models, such as Transformers, to solve text classification tasks is not consistently successful [63]. Possibly, activities on the Dark Web domain are written with a different style and grammar and require a different representation than what pre-trained embeddings offer. Taking into account that these models can handle lexical and syntactic information [64, 65] they also can overfit to their training data. In other words, they cannot transfer these types of knowledge to a new unseen domain.

In future work, we propose mechanisms for extracting and validating training data [66] and investigating the control mechanisms of neural networks [67], as initiated in [68]. Although these avenues of research are exciting and compelling, they still cannot be developed easily because of the lack of data from obscure and hard-to-find domains.

## 6. Conclusion

In this research, we investigated a new type of activity on the web that emerged due to the global pandemic. The products and discussions around COVID-19 on two parts of the web, namely surface and Dark Web, allowed us to investigate the performance of classification methods over these two domains. Although national security agencies [4] and international security agencies [5, 9, 6] continuously monitor these activities, they are not easily found, and automatic analysis could produce false truths.

For this matter, we proposed the C-OSINT framework to detect the activity related to the COVID-19 issue in Dark Web Marketplaces and Forums. COSINT-e and COSINT-d are used to extract and process data from heterogeneous sources such as '.onion' services and surface web pages. COSINT-c proposes a set of learning-based classifiers to classify the extracted corpora using COSINT-e and COSINT-d.

With the success of Transformer in many downstream tasks, we were expecting the same results on our extracted dataset. However, the results show that they cannot transfer their knowledge to an unseen domain. Finally, we observed that other subtle features such as style and syntactic information could be better clues in finding and distinguishing the activities between dark and surface web.

In summary, our contribution is two folds: (1) We build an Open Source Intelligence frameworks for activity recognition in the far reaches of the web around COVID-19 topic; (2) Reaching to the conclusion that adding external knowledge to the classification task in the form of syntactic and stylistic information would be more helpful than solely relying on pre-trained and automatic Transformer based classification.

## References

- [1] R. Michelle, Covid: Pfizer-biontech vaccine approved for eu states., 2020. URL: <https://www.bbc.co.uk/news/world-europe-55401136>.
- [2] J. Gallagher, Moderna: Covid vaccine shows nearly 95% protection., 2020. URL: <https://www.bbc.com/news/health-54902908>.
- [3] Burki, T. Khan, The russian vaccine for covid-19, *The Lancet. Respiratory medicine* 8 (2020). doi:10.1016/S2213-2600(20)30402-1.
- [4] Sicurezza nazionale, Relazione al parlamento 2021., 2022. URL: <https://www.sicurezzanazionale.gov.it/sisr.nsf/relazione-annuale/relazione-al-parlamento-2021.html>.
- [5] EUROPOL, Covid-19 sparks upward trend in cybercrime., 2020. URL: <https://www.europol.europa.eu/media-press/newsroom/news/covid-19-sparks-upward-trend-in-cybercrime>.
- [6] EUROPOL, Europol predictions correct for fake covid-19 vaccines., 2020. URL: <https://www.europol.europa.eu/media-press/newsroom/news/europol-predictions-correct-for-fake-covid-19-vaccines>.
- [7] B. R, B. M, Jiang, Availability of covid-19 related products on tor darknet markets, *Statistical Bulletin*. Canberra: Australian Institute of Criminology. (2020). doi:<https://doi.org/10.52922/sb04534>.
- [8] Intelligence, S. C. of Parliament, Annual report, 2022. URL: <https://isc.independent.gov.uk/>.
- [9] EUROPOL, Eu drug markets: Impact of covid-19., 2020. URL: <https://www.europol.europa.eu/media-press/newsroom/news/eu-drug-markets-impact-of-covid-19>.
- [10] EUROPOL, How covid-19-related crime infected europe during 2020, 2020. URL: [https://www.europol.europa.eu/sites/default/files/documents/how\\_covid-19-related\\_crime\\_infected\\_europe\\_during\\_2020.pdf](https://www.europol.europa.eu/sites/default/files/documents/how_covid-19-related_crime_infected_europe_during_2020.pdf).
- [11] EUROPOL, Europol warning on the illicit sale of false negative covid-19 test certificates., 2021. URL: <https://www.europol.europa.eu/media-press/newsroom/news/europol-warning-illicit-sale-of-false-negative-covid-19-test-certificates>.
- [12] L. Ranaldi, F. M. Zanzotto, Hiding your face is not enough: user identity linkage with image recognition, *Social Network Analysis and Mining* 10 (2020). URL: <https://doi.org/10.1007/s13278-020-00673-4>.
- [13] R. Dingledine, N. Mathewson, P. Syverson, Tor: The second-generation onion router, in: *Proceedings of the 13th Conference on USENIX Security Symposium - Volume 13, SSYM'04*, USENIX Association, USA, 2004, p. 21.
- [14] P. H. Nguyen, P. Kintis, M. Antonakakis, M. Polychronakis, An empirical study of the i2p anonymity network and its censorship resistance, in: *Proceedings of 2018 Internet Measurement Conference (IMC '18)*, 2018.
- [15] R. W. Gehl, *Weaving the dark web: a trial of legitimacy on freenet, tor, and i2p*, The MIT Press, 2018. ISBN: 9780262038263.
- [16] T. de Boer, V. Breider, *Invisible Internet Project(Report)*, Master's thesis, University of Amsterdam, 2019.
- [17] M. J. Barratt, J. A. Ferris, A. R. Winstock, Use of silk road, the online drug marketplace, in the united kingdom, australia and the united states, *Addiction* 109 (2014) 774–783.
- [18] J. Aldridge, D. DDcary-HHtu, Not an 'ebay for drugs': The cryptomarket 'silk road' as a paradigm shifting criminal innovation, *SSRN Electron. J.* (2014).

- [19] J. Martin, Lost on the silk road: Online drug distribution and the ‘cryptomarket’, *Criminology & Criminal Justice* 14 (2014) 351–367. URL: <https://doi.org/10.1177/1748895813505234>. doi:10.1177/1748895813505234. arXiv:<https://doi.org/10.1177/1748895813505234>.
- [20] M. C. Van Hout, T. Bingham, ‘silk road’, the virtual drug marketplace: a single case study of user experiences, *Int. J. Drug Policy* 24 (2013) 385–391.
- [21] A. Bracci, M. Nadini, M. Aliapoulos, D. McCoy, I. Gray, A. Teytelboym, A. Gallo, A. Baronchelli, Dark web marketplaces and covid-19: After the vaccines, 2021. URL: <https://arxiv.org/abs/2102.05470>. doi:10.48550/ARXIV.2102.05470.
- [22] Y. Yang, L. Yang, M. Yang, H. Yu, G. Zhu, Z. Chen, L. Chen, Dark web forum correlation analysis research, in: 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), 2019, pp. 1216–1220. doi:10.1109/ITAIC.2019.8785760.
- [23] T. Fu, A. Abbasi, H. Chen, A focused crawler for dark web forums, *J. Assoc. Inf. Sci. Technol.* 61 (2010) 1213–1231.
- [24] N. Tavabi, N. Bartley, A. Abeliuk, S. Soni, E. Ferrara, K. Lerman, Characterizing activity on the deep and dark web, in: Companion Proceedings of The 2019 World Wide Web Conference, WWW ’19, Association for Computing Machinery, New York, NY, USA, 2019, p. 206–213. URL: <https://doi.org/10.1145/3308560.3316502>. doi:10.1145/3308560.3316502.
- [25] M. W. Al Nabki, E. Fidalgo, E. Alegre, I. de Paz, 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, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 35–43. URL: <https://aclanthology.org/E17-1004>.
- [26] L. Choshen, D. J. Eldad, D. Hershovich, E. Sulem, O. Abend, The language of legal and illegal activity on the darknet, in: ACL, 2019.
- [27] A. Biryukov, I. Pustogarov, F. Thill, R.-P. Weinmann, Content and popularity analysis of tor hidden services, 2014 IEEE 34th International Conference on Distributed Computing Systems Workshops (ICDCSW) (2014) 188–193.
- [28] EUROPOL, Eu terrorism situation trend report (te-sat), 2021. URL: <https://www.europol.europa.eu/publications-events/main-reports/tesat-report>.
- [29] FBI, Darknet takedown., 2017. URL: <https://www.fbi.gov/news/stories/alphabay-takedown>.
- [30] S. Dumais, H. Chen, Hierarchical classification of web content, in: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’00, Association for Computing Machinery, New York, NY, USA, 2000, p. 256–263. URL: <https://doi.org/10.1145/345508.345593>. doi:10.1145/345508.345593.
- [31] A. Sun, E.-P. Lim, W.-K. Ng, Web classification using support vector machine, in: Proceedings of the 4th International Workshop on Web Information and Data Management, WIDM ’02, Association for Computing Machinery, New York, NY, USA, 2002, p. 96–99. URL: <https://doi.org/10.1145/584931.584952>. doi:10.1145/584931.584952.
- [32] M.-Y. Kan, Web page classification without the web page, in: Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers and Posters, WWW Alt. ’04, Association for Computing Machinery, New York, NY, USA, 2004, p. 262–263. URL:

<https://doi.org/10.1145/1013367.1013426>. doi:10.1145/1013367.1013426.

- [33] W. Su, J. Wang, F. Lochovsky, Automatic hierarchical classification of structured deep web databases, in: Proceedings of the 7th International Conference on Web Information Systems, WISE'06, Springer-Verlag, Berlin, Heidelberg, 2006, p. 210–221. URL: [https://doi.org/10.1007/11912873\\_23](https://doi.org/10.1007/11912873_23). doi:10.1007/11912873\_23.
- [34] H.-X. Xu, X.-L. Hao, S.-Y. Wang, Y.-F. Hu, A method of deep web classification, in: 2007 International Conference on Machine Learning and Cybernetics, volume 7, 2007, pp. 4009–4014. doi:10.1109/ICMLC.2007.4370847.
- [35] L. Overlier, P. Syverson, Locating hidden servers, in: 2006 IEEE Symposium on Security and Privacy (S P'06), 2006, pp. 15 pp.–114. doi:10.1109/SP.2006.24.
- [36] S. J. Murdoch, Hot or not: Revealing hidden services by their clock skew, in: Proceedings of the 13th ACM Conference on Computer and Communications Security, CCS '06, Association for Computing Machinery, New York, NY, USA, 2006, p. 27–36. URL: <https://doi.org/10.1145/1180405.1180410>. doi:10.1145/1180405.1180410.
- [37] I. Pete, J. Hughes, Y. T. Chua, M. Bada, A social network analysis and comparison of six dark web forums, in: 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS PW), 2020, pp. 484–493. doi:10.1109/EuroSPW51379.2020.00071.
- [38] M. Graczyk, K. Kinningham, Automatic product categorization for anonymous marketplaces, 2015.
- [39] G. Avarikioti, R. Brunner, A. Kiayias, R. Wattenhofer, D. Zindros, Structure and content of the visible darknet, CoRR abs/1811.01348 (2018). URL: <http://arxiv.org/abs/1811.01348>. arXiv:1811.01348.
- [40] G. Avarikioti, R. Brunner, A. Kiayias, R. Wattenhofer, D. Zindros, Structure and content of the visible darknet, 2018. arXiv:1811.01348.
- [41] M. W. A. Nabki, E. FIDALGO, E. Alegre, L. Fernández-Robles, Torank: Identifying the most influential suspicious domains in the tor network, Expert Syst. Appl. 123 (2019) 212–226.
- [42] S. Nazah, S. Huda, J. H. Abawajy, M. M. Hassan, An unsupervised model for identifying and characterizing dark web forums, IEEE Access 9 (2021) 112871–112892. doi:10.1109/ACCESS.2021.3103319.
- [43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.
- [44] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.
- [45] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, S. Fidler, Aligning books and movies: Towards story-like visual explanations by watching movies and reading books, 2015. arXiv:1506.06724.
- [46] T. Pires, E. Schlinger, D. Garrette, How multilingual is multilingual bert?, 2019. arXiv:1906.01502.
- [47] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, in: NeurIPS, 2019.
- [48] Y. Sun, S. Wang, S. Feng, S. Ding, C. Pang, J. Shang, J. Liu, X. Chen, Y. Zhao, Y. Lu, W. Liu, Z. Wu, W. Gong, J. Liang, Z. Shang, P. Sun, W. Liu, X. Ouyang, D. Yu, H. Tian, H. Wu, H. Wang, Ernie 3.0: Large-scale knowledge enhanced pre-training for language

- understanding and generation, ArXiv abs/2107.02137 (2021).
- [49] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, ELECTRA: Pre-training text encoders as discriminators rather than generators, in: ICLR, 2020. URL: <https://openreview.net/pdf?id=r1xMH1BtvB>.
  - [50] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBert, a distilled version of bert: smaller, faster, cheaper and lighter, ArXiv abs/1910.01108 (2019).
  - [51] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, ArXiv abs/1907.11692 (2019).
  - [52] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Brew, HuggingFace's Transformers: State-of-the-art Natural Language Processing, ArXiv abs/1910.0 (2019).
  - [53] R. van der Goot, N. Ljubešić, I. Matroos, M. Nissim, B. Plank, Bleaching text: Abstract features for cross-lingual gender prediction, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 383–389. URL: <https://aclanthology.org/P18-2061>. doi:10.18653/v1/P18-2061.
  - [54] S. Bird, E. Klein, E. Loper, Natural language processing with Python: analyzing text with the natural language toolkit, "O'Reilly Media, Inc.", 2009.
  - [55] R. Parker, D. Graff, J. Kong, K. Chen, K. Maeda, English gigaword fifth edition ldc2011t07 (tech. rep.), Technical Report, Technical Report. Linguistic Data Consortium, Philadelphia, 2011.
  - [56] C. Crawl, Common crawl, URL: <http://commoncrawl.org>, 2019.
  - [57] J. Callan, M. Hoy, C. Yoo, L. Zhao, Clueweb09 data set, 2009.
  - [58] M. P. Marcus, B. Santorini, M. A. Marcinkiewicz, Building a large annotated corpus of English: The Penn Treebank, Computational Linguistics 19 (1993) 313–330. URL: <https://aclanthology.org/J93-2004>.
  - [59] J. Pennington, R. Socher, C. Manning, GloVe: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1532–1543. URL: <https://aclanthology.org/D14-1162>. doi:10.3115/v1/D14-1162.
  - [60] F. M. Zanzotto, A. Santilli, L. Ranaldi, D. Onorati, P. Tommasino, F. Fallucchi, KERMIT: Complementing transformer architectures with encoders of explicit syntactic interpretations, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 256–267. URL: <https://aclanthology.org/2020.emnlp-main.18>. doi:10.18653/v1/2020.emnlp-main.18.
  - [61] M. Zhu, Y. Zhang, W. Chen, M. Zhang, J. Zhu, Fast and accurate shift-reduce constituent parsing, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 434–443. URL: <https://aclanthology.org/P13-1043>.
  - [62] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, D. McClosky, The Stanford CoreNLP natural language processing toolkit, in: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Baltimore, Maryland, 2014, pp. 55–60. URL:



<https://aclanthology.org/P14-5010>. doi:10.3115/v1/P14-5010.

- [63] L. Ranaldi, F. Ranaldi, F. Fallucchi, F. M. Zanzotto, Shedding light on the dark web: Authorship attribution in radical forums, *Information* 13 (2022). URL: <https://www.mdpi.com/2078-2489/13/9/435>.
- [64] G. Jawahar, B. Sagot, D. Seddah, What does bert learn about the structure of language, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3651–3657.
- [65] J. Hu, J. Gauthier, P. Qian, E. Wilcox, R. Levy, A systematic assessment of syntactic generalization in neural language models, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 1725–1744.
- [66] L. Ranaldi, F. Fallucchi, A. Santilli, F. M. Zanzotto, Kermitviz: Visualizing neural network activations on syntactic trees, in: E. Garoufallou, M.-A. Ovalle-Perandones, A. Vlachidis (Eds.), *Metadata and Semantic Research*, Springer International Publishing, Cham, 2022, pp. 139–147.
- [67] D. Onorati, P. Tommasino, L. Ranaldi, F. Fallucchi, F. M. Zanzotto, Pat-in-the-loop: Declarative knowledge for controlling neural networks, *Future Internet* 12 (2020). URL: <https://www.mdpi.com/1999-5903/12/12/218>. doi:10.3390/fi12120218.
- [68] L. Ranaldi, F. Fallucchi, F. M. Zanzotto, Dis-cover ai minds to preserve human knowledge, *Future Internet* 14 (2022). URL: <https://www.mdpi.com/1999-5903/14/1/10>. doi:10.3390/fi14010010.



# A. Appendix

## A.1. Example of Listing

| Name   | Price                         | Vendor     | Place                                 |
|--|-------------------------------|------------|---------------------------------------|
| BEST-QUALITY COVID-19 VACCINE B1612EFFECTIVE   | € 208.015<br>(€ 208.015 ETC.) | [REDACTED] | From: United States<br>2%   Worldwide |
| COVID-19 VACCINE AND ADVISE A MEDICAL DOCTOR   | € 208.015<br>(€ 208.015 ETC.) | [REDACTED] | From: United States<br>2%   Worldwide |
| COVID-19 VACCINE AVAILABLE                     | € 208.015<br>(€ 208.015 ETC.) | [REDACTED] | From: Spain<br>2%   Worldwide         |
| AVAILABLE Pfizer/BioNTech Corona Vaccine B1612 | € 208.015<br>(€ 208.015 ETC.) | [REDACTED] | From: Germany<br>2%   Worldwide       |
| COVID-19 VACCINE FOR SALE 300                  | € 208.015<br>(€ 208.015 ETC.) | [REDACTED] | From: France<br>2%   Worldwide        |
| AVAILABLE COVID-19 VACCINE 250 EUROS           | € 208.015<br>(€ 208.015 ETC.) | [REDACTED] | From: France<br>2%   Worldwide        |
| AVAILABLE Pfizer/BioNTech Corona Vaccine B1612 | € 208.015<br>(€ 208.015 ETC.) | [REDACTED] | From: Germany<br>2%   Worldwide       |
| AVAILABLE COVID-19 VACCINE 250 EUROS           | € 208.015<br>(€ 208.015 ETC.) | [REDACTED] | From: France<br>2%   Worldwide        |

**Figure 2:** Screenshot of an ad in the vaccines category offering Pfizer/BioNTech vaccine and other vaccines. We have removed the seller’s contact information, which invites the potential customer to have direct contact. The site screenshot was taken in April 2021.

## A.2. Dark Web Forums

| <i>Topic</i> | <i>DWF</i>  |
|--------------|---|
| COVID-19     | RAID, dread, Nulled,<br>4chan, The Stock Insiders, Hidden Answers,<br>Acropolis Forum, torBBS<br>Tedit forum, SuprBay,<br>DeaChan |

**Table 4**

List of Dark Web Forums analyzed.

## A.3. Dark Web Marketplaces

| <i>Product</i>          | <i>DWM</i>  |
|-------------------------|---|
| Vaccines                | Royal, Cypher, Asap, Bigblue,<br>Dark fox, Hydra, Invictus, Kilos,<br>Liberty, Yakuza, Recon,<br>Televend, The Canadian Headquarters,<br>Agartha, World market, Yukon |
| fake Tests<br>&<br>more | Magbo, Recon, Televend,<br>The Canadian Headquarters,<br>Torrez, Versus, White house, Yakuza<br>MagBO, 24HoursPPC, ASAP,<br>Dark Fox, Dark Leak Market,               |

**Table 5**

List of Dark Web Marketplaces analyzed.