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

Leonardo Ranaldi[1,2,3], Aria Nourbakhsh[2,3]
, Francesca Fallucchi[1], and Fabio Massimo Zanzotto[2]

[1] Department of Innovation and Information Engineering,
Gugliemlo Marconi University, Roma, Italy
(l.ranaldi,f.fallucchi)@unimarconi.it
[2] Department of Enterprise Engineering,
University of Rome Tor Vergata, Roma, Italy
(nrbrai,fabio.massimo.zanzotto)@uniroma2.it
[3] Equal Contribution

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 corporate 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.

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 [37], Moderna [26], and Sputnik [9] 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 [51, 18, 20, 47, 30]. Monitoring DWMs is therefore critical to enable police and public health agencies to be prepared and effectively counter these threats.

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 [19, 21]. 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 [23]. 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.

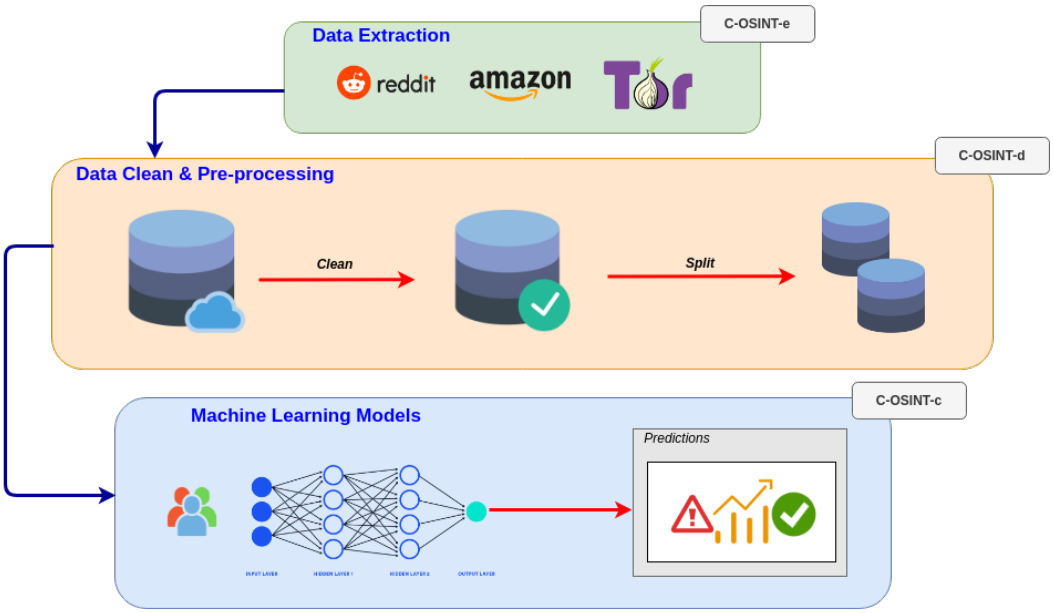

Figure 1: C-OSINT Framework.

DWFs are a great place to get into illicit online activities, and DWMs can be easily accessed through specialized browsers, such as Tor [16], I2P [41] and FreeNet [27]. 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 [14]. 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 [5, 2, 36, 57, 8] and DWFs [61, 25, 55]. Among the most credited are studies that propose automatic recognition and classification of activities and analysis of lexicon used in DWMs and DWFs [1, 11, 7]. According to numerous reports, law enforcement has successfully closed several illegal DWMs [22, 24]. 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 [17, 53, 32]. More recently, some attempts to classify the non-indexed part of the web, called the Deep Web [52, 60], and then with the ancestor of today's dark web, [42, 38] have been published.

With the growth in popularity, the dark web has become a research subject in many studies. Barratt et al. [5] and Aldridge et al. [2] have done extensive investigations of customers of DWMs taking the *'Silk Road'* phenomenon as a use case. Yang et al. [61] and Pete et al.[45], on the 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 dark part of the web have remained highly contextualized. One of the first works that shifted the focus to automatic content classification was done by Biryukov et al. [7]. Biryukov et al. [7] classified the content of the dark web, restricting the study only to Tor's hidden services resulting in 18 topical categories. By limiting the topic to drug trades, Graczyk et al.[28] combined unsupervised feature selection and an SVM classifier to classify drug selling services. This work started by classifying Tor's texts, but they did not directly address the legal/illegal distinction by placing all activities in the dark.

The first real distinction between activities, as selling services rather than forums, was proposed in [1, 3]. They presented DUTA (Darknet Usage Text Addresses), the first publicly available Darknet dataset, with a classification into topical categories and subcategories. Avarikioti et al.[4] 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.[11], following [4] and using the updated version of the publicly available DUTA [39], studied the style and structure of hidden illegal and legal services. Choshen et al.[11] 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 dark web 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 [19, 21, 23, 51] 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 B, C), to create a framework capable of: a) collecting information from '.oinion' 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 1) a crawler and scraper to collect the data (C-OSINT-e) described in the subsection 3.1.1; 2) a pre-processor of the extracted text and

a labeling step (C-OSINT-d) described in the subsection 3.1.2; 3) 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 [11].

One of the first public datasets obtained from dark web was the first version of "Darknet Usage Text Addresses" (DUTA) [1] that was built thanks to the contribution of a group of experts who provided the authors with access to the keyword lists and pre-classified samples of each category. However, this dataset suffers from dead links that could be particularly taken down by police forces or shut down by their hosts. Although an updated version, DUTA-10K [39], 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 [1], C-OSINT-e is based on an extraction step, cleaning phase, and finally, labeling of the extracted samples. From the [47] 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 [40]. 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[1] and Amazon[2]. 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.

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 [7]. 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.[39]. 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[3]. The time period of data collection for corpus construction

[1] https://www.reddit.com/
[2] https://www.amazon.com/
[3] https://github.com/ART-Group-it/C-OSINT

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.

is from November 2020 through March 2021. Appendix B and Appendix C 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.[11]. In all experiments, we apply a cleaning to the text of the corpora web pages. HTML markups are removed from the original dataset; the same is done for non-linguistic contents such as buttons, encryption keys, metadata, URLs, common words like "Show more result" etc.. Despite applying these pre-processing steps, the remaining textual elements are uncanonical, 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 were defined in [1] which consist of the following points: 1) an author tags a domain-based only on the textual content visible to the user, 2) A domain must receive only one tag based on its activity. 3) In case there are any uncertainties if any author hesitates on the tag, 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. Moreover, repetitions were avoided, and contents tagged as [*deleted*] were removed, typically deleted comments and posts. 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.

After extracting and cleaning the data by C-OSINT-e and C-OSINT-d, 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 [58] 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:

- $BERT_{base}$ [15], that is Bidirectional Encoder Representations from Transformers, and is trained on the BooksCorpus [65] 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 [62] 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 on 32.89 billion tokens, taken from datasets gathered from the surface web or publicly available datasets, such as Wikipedia, Bookcorpus, Giga5, Clueweb, and Common Crawl.
- ERNIE [54] 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 [12] 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.

Corpus	*Size*
BooksCorpus [65]	800M words
2010-and-2014-English Wikipedia dump	2,500M words
Giga5 [43]	16GB
Common Crawl [13]	110GB
ClueWeb [10]	19GB
Penn Treebank [35]	1M words

Table 2: Corpora used in training pre-trained Transformers and word embeddings. All corpora are derived from the surface web.

- 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 [33] 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 [59].

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 [56] is a model proposed to capture the style of writing at the word level. This model has initially proposed for cross-lingual authors' gender prediction to abstract away from the surface word representation. 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) [6] 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 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 [44] 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) [63]. 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.

The KERMIT model exploits the parse trees produced by a traditional parser. As advised by Zanzotto el al. [63], we used the English constitution-based parser available in CoreNLP library [64].

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

Table 3: Experiments with neural networks are obtained over 5 runs with different seeds.

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' depending on the source. In the experiments, the datasets were merged, building four balanced comparisons. The merged corpus pairs were split into training and test sets with a 70/30 ratio in each experiment. The evaluation was done by extracting the accuracy

of classification outputs, defined as the number of correct predictions divided by the number of total predictions.

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 Table 3 with the configurations described in section 4.

These results show the unexpected behavior of the applied models. Surprisingly, holistic Transformers have poor performance on these uncovered domains. $BERT_{base}$, $BERT_{multi}$, $Electra$, $XLNet$ and $Ernie$, $RoBERTa$, $DistilBERT$ have worse performances with respect to all the other models. Although the *BoE(GloVe)* lexical model scores better than the Transformer model, 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 holistic Transformers. Stylistic models perform on par with syntactic models. The tasks where stylistic models are performing better are the ones that are about classifying surface web services against their dark web counterparts. These results are two folds: 1- the data from the dark web has a different nature of writing style that can be captured through distinguishing features such as all capital letters, abbreviations, and punctuation. 2- A bag of words representation of POS tags can be a distinguishing factor between dark and surface web services and forums (Reddit) and markets (Amazon). However, the distinction is less evident for DFMs and DWMs.

Neural network models based on syntax have engaging performances on this dataset. Here, KERMIT [63] works better than holistic Transformers, showing that these tasks are sensitive concerning syntactic information that the Transformers cannot transfer to another unseen domain. Moreover, although KERMIT uses a parser that is trained on the surface web to parse sentences [34], syntactic rules are more restricted than semantic and discourse-level information captured by the Transformers, and 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. Although the results are still high and consistent among the four settings, possibly this combination adds to the dimensionality of the representation and suppresses better individual syntactic or POS information.

In conclusion, monitoring the activities on the dark web and comparing them with their similar real-world services is an ongoing challenge. Using pre-trained language models, such as Transformers, to solve text classification tasks is not consistently successful. 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 [31, 29] they also can overfit to their training data. In other words, they cannot transfer these types of knowledge to a new unseen domain.

To further explore this theory in future work, we propose a comprehensive analysis that takes advantage of a training data extraction and validation mechanism used in pre-trained models such as Transformers. In parallel, we would like to investigate the control mechanisms of neural networks as initiated in [48, 49]. 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 [51] and international security agencies [18, 19, 20] 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 Dark Web 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 models in many downstream tasks in NLP, 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 artificial 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.

7 Acknowledgments

References

[1] Mhd Wesam Al Nabki, Eduardo Fidalgo, Enrique Alegre, and Ivan 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*, pages 35–43, Valencia, Spain, April 2017. Association for Computational Linguistics.

[2] Judith Aldridge and David DDcary-HHtu. Not an 'ebay for drugs': The cryptomarket 'silk road' as a paradigm shifting criminal innovation. *SSRN Electron. J.*, 2014.

[3] Georgia Avarikioti, Roman Brunner, Aggelos Kiayias, Roger Wattenhofer, and Dionysis Zindros. Structure and content of the visible darknet. *CoRR*, abs/1811.01348, 2018.

[4] Georgia Avarikioti, Roman Brunner, Aggelos Kiayias, Roger Wattenhofer, and Dionysis Zindros. Structure and content of the visible darknet, 2018.

[5] Monica J Barratt, Jason A Ferris, and Adam R Winstock. Use of silk road, the online drug marketplace, in the united kingdom, australia and the united states. *Addiction*, 109(5):774–783, May 2014.

[6] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[7] Alex Biryukov, Ivan Pustogarov, Fabrice Thill, and Ralf-Philipp Weinmann. Content and popularity analysis of tor hidden services. *2014 IEEE 34th International Conference on Distributed Computing Systems Workshops (ICDCSW)*, pages 188–193, 2014.

[8] Alberto Bracci, Matthieu Nadini, Maxwell Aliapoulios, Damon McCoy, Ian Gray, Alexander Teytelboym, Angela Gallo, and Andrea Baronchelli. Dark web marketplaces and covid-19: After the vaccines, 2021.

[9] Burki and Talha Khan. The russian vaccine for covid-19. *The Lancet. Respiratory medicine*, 8, 2020.

[10] Jamie Callan, Mark Hoy, Changkuk Yoo, and Le Zhao. Clueweb09 data set, 2009.

[11] Leshem Choshen, Dan J. Eldad, Daniel Hershcovich, Elior Sulem, and Omri Abend. The language of legal and illegal activity on the darknet. In *ACL*, 2019.

[12] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*, 2020.

[13] Common Crawl. Common crawl. URL: http://commoncrawl.org, 2019.

[14] Tim de Boer and Vincent Breider. Invisible internet project(report). Master's thesis, University of Amsterdam, February 2019.

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

[16] Roger Dingledine, Nick Mathewson, and Paul Syverson. Tor: The second-generation onion router. In *Proceedings of the 13th Conference on USENIX Security Symposium - Volume 13*, SSYM'04, page 21, USA, 2004. USENIX Association.

[17] Susan Dumais and Hao 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, page 256–263, New York, NY, USA, 2000. Association for Computing Machinery.

[18] EUROPOL. Covid-19 sparks upward trend in cybercrime., 2020.

[19] EUROPOL. Eu drug markets: Impact of covid-19., 2020.

[20] EUROPOL. Europol predictions correct for fake covid-19 vaccines., 2020.

[21] EUROPOL. How covid-19-related crime infected europe during 2020, 2020.

[22] EUROPOL. Eu terrorism situation trend report (te-sat)., 2021.

[23] EUROPOL. Europol warning on the illicit sale of false negative covid-19 test certificates., 2021.

[24] FBI. Darknet takedown., 2017.

[25] Tianjun Fu, A. Abbasi, and Hsinchun Chen. A focused crawler for dark web forums. *J. Assoc. Inf. Sci. Technol.*, 61:1213–1231, 2010.

[26] James Gallagher. Moderna: Covid vaccine shows nearly 95% protection., 2020.

[27] Robert W. Gehl. Weaving the dark web: a trial of legitimacy on freenet, tor, and i2p. The MIT Press, 2018. ISBN: 9780262038263.

[28] Michał Graczyk and Kevin Kinningham. Automatic product categorization for anonymous marketplaces. 2015.

[29] Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, 2020.

[30] Intelligence and Security Committee of Parliament. Annual report, 2022.

[31] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does bert learn about the structure of language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, 2019.

[32] Min-Yen Kan. Web page classification without the web page. In *Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers amp; Posters*, WWW Alt. '04,

page 262–263, New York, NY, USA, 2004. Association for Computing Machinery.

[33] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019.

[34] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[35] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

[36] James Martin. Lost on the silk road: Online drug distribution and the 'cryptomarket'. *Criminology & Criminal Justice*, 14(3):351–367, 2014.

[37] Roberts Michelle. Covid: Pfizer-biontech vaccine approved for eu states., 2020.

[38] Steven 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, page 27–36, New York, NY, USA, 2006. Association for Computing Machinery.

[39] Mhd Wesam Al Nabki, Eduardo FIDALGO, Enrique Alegre, and Laura Fernández-Robles. Torank: Identifying the most influential suspicious domains in the tor network. *Expert Syst. Appl.*, 123:212–226, 2019.

[40] Saiba Nazah, Shamsul Huda, Jemal H. Abawajy, and Mohammad Mehedi Hassan. An unsupervised model for identifying and characterizing dark web forums. *IEEE Access*, 9:112871–112892, 2021.

[41] Phong Hoang Nguyen, Panagiotis Kintis, Manos Antonakakis, and Michalis Polychronakis. An empirical study of the i2p anonymity network and its censorship resistance. In *Proceedings of 2018 Internet Measurement Conference (IMC '18)*, October 2018.

[42] L. Overlier and P. Syverson. Locating hidden servers. In *2006 IEEE Symposium on Security and Privacy (S P'06)*, pages 15 pp.–114, 2006.

[43] R Parker, D Graff, J Kong, K Chen, and K Maeda. English gigaword fifth edition ldc2011t07 (tech. rep.). Technical report, Technical Report. Linguistic Data Consortium, Philadelphia, 2011.

[44] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[45] Ildiko Pete, Jack Hughes, Yi Ting Chua, and Maria Bada. A social network analysis and comparison of six dark web forums. In *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS PW)*, pages 484–493, 2020.

[46] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual bert?, 2019.

[47] Broadhurst R, Ball M, and Jiang. Availability of covid-19 related products on tor darknet markets. *Statistical Bulletin. Canberra: Australian Institute of Criminology.*, 2020.

[48] Leonardo Ranaldi, Francesca Fallucchi, Andrea Santilli, and Fabio Massimo Zanzotto. Kermitviz: Visualizing neural network activations on syntactic trees. In Emmanouel Garoufallou, María-Antonia Ovalle-Perandones, and Andreas Vlachidis, editors, *Metadata and Semantic Research*, pages 139–147, Cham, 2022. Springer International Publishing.

[49] Leonardo Ranaldi, Francesca Fallucchi, and Fabio Massimo Zanzotto. Dis-cover ai minds to preserve human knowledge. *Future Internet*, 14(1), 2022.

[50] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

[51] Sicurezzanazionale. Relazione al parlamento 2021., 2022.

[52] Weifeng Su, Jiying Wang, and Frederick Lochovsky. Automatic hierarchical classification of struc-

tured deep web databases. In *Proceedings of the 7th International Conference on Web Information Systems*, WISE'06, page 210–221, Berlin, Heidelberg, 2006. Springer-Verlag.

[53] Aixin Sun, Ee-Peng Lim, and Wee-Keong Ng. Web classification using support vector machine. In *Proceedings of the 4th International Workshop on Web Information and Data Management*, WIDM '02, page 96–99, New York, NY, USA, 2002. Association for Computing Machinery.

[54] Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *ArXiv*, abs/2107.02137, 2021.

[55] Nazgol Tavabi, Nathan Bartley, Andres Abeliuk, Sandeep Soni, Emilio Ferrara, and Kristina Lerman. Characterizing activity on the deep and dark web. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 206–213, New York, NY, USA, 2019. Association for Computing Machinery.

[56] Rob van der Goot, Nikola Ljubešić, Ian Matroos, Malvina Nissim, and Barbara 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)*, pages 383–389, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[57] Marie Claire Van Hout and Tim Bingham. 'silk road', the virtual drug marketplace: a single case study of user experiences. *Int. J. Drug Policy*, 24(5):385–391, September 2013.

[58] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[59] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.0, 2019.

[60] He-Xiang Xu, Xiu-Lan Hao, Shu-Yun Wang, and Yun-Fa Hu. A method of deep web classification. In *2007 International Conference on Machine Learning and Cybernetics*, volume 7, pages 4009–4014, 2007.

[61] Ying Yang, Lina Yang, Meihong Yang, Huanhuan Yu, Guichun Zhu, Zhenya Chen, and Lijuan Chen. Dark web forum correlation analysis research. In *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, pages 1216–1220, 2019.

[62] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*, 2019.

[63] Fabio Massimo Zanzotto, Andrea Santilli, Leonardo Ranaldi, Dario Onorati, Pierfrancesco Tommasino, and Francesca 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)*, pages 256–267, Online, November 2020. Association for Computational Linguistics.

[64] Muhua Zhu, Yue Zhang, Wenliang Chen, Min Zhang, and Jingbo 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)*, pages 434–443, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[65] Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books, 2015.

A Example of Listing

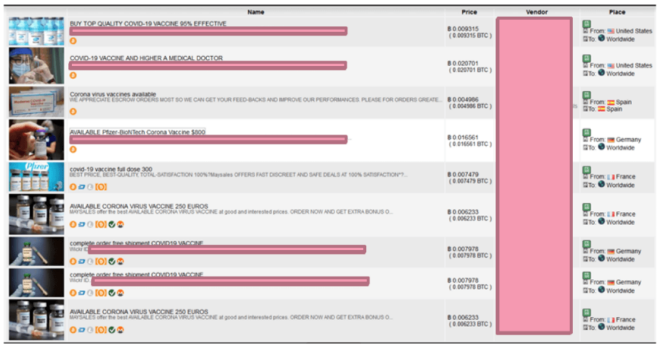

Figure 2: Screenshot of an ad in the vaccines category offering Pfizer/BioNTech vaccine and other vaccines. We have removed the seller's contact information, which invites the potential customer to have direct contact. The site screenshot was taken in April 2021.

B Dark Web Forums

Topic	*DWF*
COVID-19	RAID, dread, Nulled, 4chan, The Stock Insiders, Hidden Answers, Acropolis Forum, torBBS Teddit forum, SuprBay, DeaChan

Table 4: List of Dark Web Forums analyzed.

C Dark Web Marketplaces

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

Table 5: List of Dark Web Marketplaces analyzed.