Name Entity Recognition (NER) Based Drug Related Page Classification on Dark Web

Ashwini Dalvi
Veermata Jijabai Technological Institute, India
aadalvi_p19@ce.vjti.ac.in

Vaishnavi Shah
Veermata Jijabai Technological Institute, India
vrshah_b19@ce.vjti.ac.in

Dhruvin Gandhi
Veermata Jijabai Technological Institute, India
drgandhi_b19@ce.vjti.ac.in

Siddharth Shah
Veermata Jijabai Technological Institute, India
ssshah_b19@ce.vjti.ac.in

S G Bhirud
Veermata Jijabai Technological Institute, India
sgbhirud@ce.vjti.ac.in

***Abstract*—While researching the dark web marketplaces, it was observed that the drugs' names varied on different marketplaces. Therefore, the same drug might refer to different names on different dark web marketplaces. For example, some marketplaces use the chemical name or medical name as their product name to ensure the exact product; meanwhile, some marketplaces use the street name as the product name to attract users and get more orders.**

The present work discussed a NER based method to find if a website on the dark web has mentioned drugs. First, the dark web crawler crawled data from the dark web. Then, the authors introduced the Named Entity Recognition (NER) drug dataset with two categories of drug-named entities: Street name and Chemical name. Further, to identify drug-related web pages comprising street and chemical names of drugs with the NER model employed on scraped data. The proposed NER model was tested with the Drug-NER dataset. The DrugcrossNER project contains a predefined Drug-NER dataset with over 3500 listings from the dark web markets. The proposed work also generates a DRUG entity for the NER model in spaCy, an open-source NLP library in python, as it does not have the in-built ability to classify objects into custom categories.

Keywords—Dark web, Drug marketplace, Street names, Chemical names, NER model, spaCy

I. INTRODUCTION

The emergence of darknet services revolutionized drug dealing through the dark web marketplace. As a result, drug dealers and users increasingly use dark web marketplaces because they are perceived as safe and anonymous. However, dark web monitoring to control marketplaces requires extensive data collection from different sources, including dark web marketplaces, hacker forums, and messaging apps [1],[2],[3].

Many studies have been conducted on darknet markets over the past few years. Although researchers have been using reliable data collection methodologies to collect darknet market data, the work on collecting and analyzing darknet market data is still evolving. Researchers and criminologists have collected and analyzed online traces of dark web marketplaces for years with manual or automatic methods. Researchers can access large, robust datasets through automated data collection to analyze darknet websites [4],[5],[6]. However, the collected data from the dark web marketplace is mostly in unstructured text format.

Identifying predetermined types of entities in unstructured data is possible using named entity recognition text. For example, collecting information from dark web drug marketplaces relies on extracting cybersecurity entities from unstructured texts. In addition, natural language processing methods like machine translation and information retrieval from textual data rely on named entity recognition. However, it is rare for Named Entity Recognition (NER) models to address cybersecurity-related entity extraction. Researchers offered reviews on NER techniques and approaches in the cyber security domain [7] and proposed future directions for evolving NER in cyber security.

NER methods for cyber security data have traditionally been based on linguistic characteristics. However, researchers have proposed an approach incorporating Conditional Random Fields (CRF) based on deep learning [8]. However, researchers proved that the LSTM-CRF improved NER extraction accuracy compared to the traditional pure statistical CRF method [9].

It is common for cyber security texts to contain a large number of long sentences. In addition, these sentences are often challenging to understand due to their structure. In order to extract these features accurately, it is not easy to rely solely on neural networks. Instead, researchers proposed a data-driven attention mechanism with a deep learning model to improve the recognition of complex entities, enrich the text's local features, and model the context better [10].

There is a great deal of diversity in the cyber security domain, which makes it difficult for named entity recognition (NER) to identify security entities. Furthermore, identifying named entities within a noisy text is a challenging task that is usually enhanced when an external source of information is incorporated. Thus creating a NER corpus to address domain-specific problems like recognition of drug-related pages on the dark web marketplace would be proven helpful in related research.

The proposed work implements three approaches to evaluate the performance of NER models for drug entity recognition.

- The NER corpus of drug-named entities: Street name and Chemical name
- Evaluating model performance with open source Drug-NER dataset

- Generating a DRUG entity for the NER model in spaCy, an open-source NLP python library

Following is a brief outline of the paper. Section II provides a review of recent research in this field. Section III presents models and results, and section IV concludes the discussion.

II. Related Work

Dark web drug marketplace investigation intrigues researchers for different reasons and purposes. Researchers offered an overview of dark web marketplaces to comprehend their functions and processes [11]. In addition, some researchers focus on investigating the dark web for vendor profiling [12]. Using dark web marketplace data, researchers analyzed the types of cyberattacks available for sale and which are more valuable [13].
Following the proposed design by researchers, law enforcement authorities are required to evaluate a collection of dark web pages [14].
Based on free tools available on the World Wide Web, the proposed study provides an analytical framework for automating the dark web scraping and analyzing the data [15]. A Web crawler extracted marketplace listings and seller information from a case study marketplace. According to a manual investigation, the Dark Web marketplace researchers chose 12 top-level categories, including "Drugs and Chemicals, and 60 subcategories. Furthermore, there were over 100 third-level categories. Researchers examined drug users' lives using a Finnish website [16]. Nine thousand three hundred posts were analyzed to understand the socioeconomic conditions of the users. This analysis used usernames and forum posts to represent a user's way of life and drug use. The language in which user names themselves can reveal their identity. Researchers observed that user names were based on various topics, such as personal names, fictitious characters, place names, and invented words. Applying NER based approach could result in labeling such content retrieved from the dark web drug marketplaces.

There are few attempts in the literature to use the NER approach for dark web investigation.

Researchers described DarkNER, an application of Named Entity Recognition (NER) based on neural networks to recognize six types of named entities in onion domains on the Tor network: Location, Person, Products, Corporation, Group, and Creative-work [17]. The proposed NER model was trained on The W-NUT-2017 dataset and tested using samples of manually tagged Tor hidden services. Researchers emphasized that the NER model can still recognize relevant elements connected to suspicious actions even if it is trained with a dataset unrelated to the final application.
Research presented DreamDrug, a crowdsourced dataset to detect drug mentions in darknet market listings [18]. Manually annotated drug entities were created for the DreamDrug dataset, which can be used to train and evaluate named entity recognition systems (NER). Approximately 15,000 drug entities have been carefully annotated in the dataset from over 3,500 item listings initially collected from the darknet market platform "DreamMarket" in 2017. Besides providing an in-depth explanation of the creation and annotation of datasets, the authors optimize and evaluate traditional NER models based on the new dataset.
In the field of illegal drugs, researchers focused on identifying Chinese public hazard entities [19]. However, it is challenging to recognize illegal drug entities using traditional entity recognition methods because disguised forms, such as homophones and multi-entities, were used. Therefore, researchers proposed a deep learning-based multi-information fusion approach to identify Chinese public hazards on the Darknet.

Researchers proposed a novel multi-tasking approach for social media data by combining Named Entity (NE) segmentation and fine-grained NE categorization [20]. Further, the mentioned approach was improved for darknet data by introducing the Local Distance Neighbour feature [21], and researchers tested the proposed model on the WNUT-2017 dataset.

The related work overview concludes that the dark web drug marketplace is one of the research interests of researchers. However, limited attempts have been noted to apply NERdriven drug-related page classification. Thus proposed work is the first attempt to classify drug-related pages with the NER model.

III. Implemented Approaches

The dark web crawler crawled pages from the dark web. The collected data consists of uncleaned text data from HTML pages. Data preprocessing for natural language processing (NLP) is performed on raw data to detect drug-related pages. The street and chemical drug names are created to prepare the NER dataset. The street names and chemical names of drugs are complied with by crawling surface websites hosting drug-related information. For example, the website of Newport academy, the Teen Rehab Center, offers brief information on street names for drugs [22]. The other rehab center website, Addiction Center, presents a glossary for different drug street names. The drug's atomic or molecular structure is described by its chemical name. The chemical name data set is created by authors by scraping information from PubChem, a database maintained by the National Institutes of Health (NIH), which provides public access to chemistry information [23].

The proposed work attempted drug-related page classification using two methods.

a) With Crawled dataset and Drug-NER dataset

Method - 1A
The information from the retrieved lists of drug names is transformed into the entities STREET and CHEMICAL, respectively. Customized dark web crawler collected pages from the dark web—the collected data comprised raw HTML and related scripting pages. The scraped data, which contains information from drug-related pages, comprises street and chemical names of drugs embedded into the TITLE and BODY content of HTML pages. First, the lists of street and chemical drug names are constructed with crawled data. Then, these names are transformed into distinct entities STREET and CHEMICAL, respectively. With the help of the newly generated entities, the drugs extracted from scraped TITLES were categorized into the respective

entities. Finally, the training titles were passed into the NER model, which is trained further. The model is attached with a well-defined preprocessing function to account for the sequential constraints in the input text. Since there is no training dataset for the drug-related Tor hidden services, the authors used the TITLES from the scraped data set for training.
In order to detect the presence of 'drug' entities in darknet markets, the researchers of the DRUG-CrossNER project created a dataset called Drug-NER.

The NER model classified the laypeople's names of the drugs into STREET and the CHEMICAL names into STREET and CHEMICAL entities, respectively. Table I depicts the performance measure of the NER model using crawled data as the training dataset and Drug-NER data as the testing dataset.

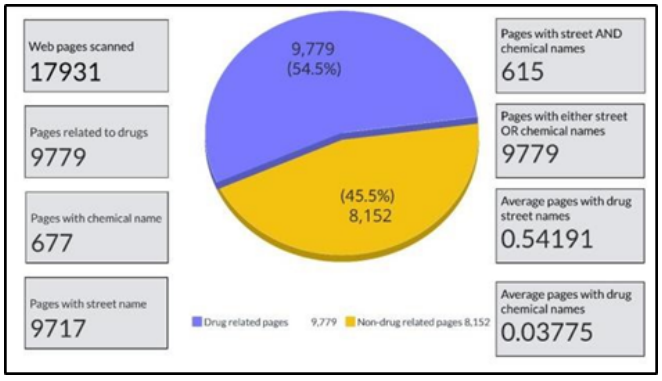
TABLE I. PERFORMANCE OF NER MODEL WITH METHOD 1A

Sr No	Model Performance Measures	Values
1	Precision	0.5217
2	Recall	1
3	Accuracy	0.9874
4	F1 score	0.3428

Fig1 Statistics and visualization of crawled data with method 1A

Figure 1 shows the related visualization of crawled data with method 1A.

Method -1B



Since there is no training dataset for the drug-related Tor hidden services, in other attempts, the authors split the crawled data into a training set of 70% and a testing set of 30% of the dataset. First, the authors used word matching to label the training and testing datasets, and the model was trained further. Later, the testing dataset was used to test the trained model's performance. Table II reports the model's performance in terms of Precision, Recall, and harmonic mean F1 score measures.

TABLE II. PERFORMANCE OF NER MODEL WITH METHOD 1B

Sr No	Model Performance Measures	Values
1	Precision	0.8851
2	Recall	0.9357
3	Accuracy	0.9850
4	F1 score	0.4548

Figure 2 shows the related visualization of crawled data with method 1B.

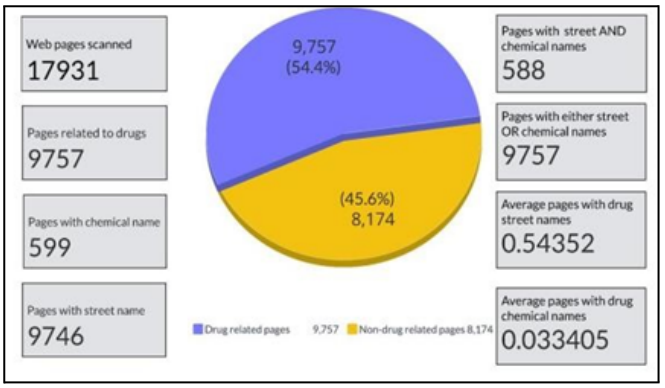


Fig 2 Statistics and visualization of crawled data with method 1B

b) With spaCy NER Model

By definition, spaCy includes a pipeline 'ner' for Named recognition. While it functions effectively, it is frequently inaccurate for most cybersecurity use cases. The context in which the word is referred to is crucial to understand. For example, a word can be classified as PERSON or ORG, depending on the context. spaCy does not have the in-built ability to classify objects into custom categories. For example, for the NER model to classify the drugs, authors need to generate a DRUG entity for the NER model. This approach shows the model's adaptability in detecting a completely new textual entity, such as drug names in hidden Tor services. The limited number of supervised training samples and the potential of numerous interpretations for a given word make building a NER system difficult. Furthermore, the quality of the input text significantly influences the system's performance.

TABLE III. PERFORMANCE OF NER MODEL WITH METHOD 2

Sr No	Model Performance Measures	Values
1	Precision	0.94631
2	Recall	0.68678
3	Accuracy	0.9874
4	F1 score	0.39767

Figure 3 shows the related visualization of crawled data with method 2.

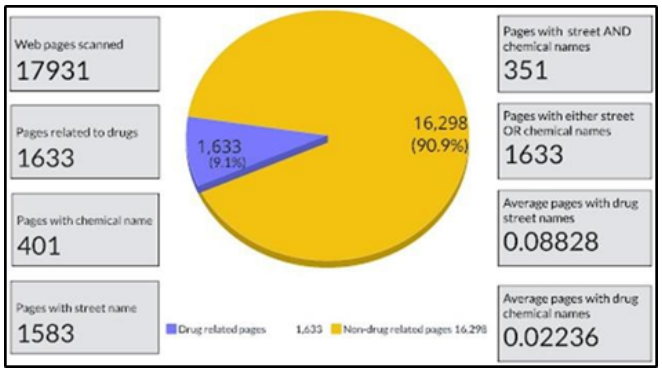


Fig 3 Statistics and visualization of crawled data with method 2

The appendix includes figures 4,5,6 depicting three bar graphs for visualization of web pages with mentioned Street names of drugs and bar graph visualization of web pages with mentioned Chemical names of drugs with methods 1A, 1B, and 2.

IV. Conclusion

In research on dark web investigation, works are cited on better learning exploits and hacker threats [24, 25]. However, drug-related investigation is majorly driven by manual supervision. The proposed work attempts to identify the dark web drug marketplace with an ML-based approach. The proposed work attempts NER model-based drug page classification on the dark web data. First, the authors introduced two categories of drug-named entities in their Named Entity Recognition (NER) dataset: street names and chemical names. Next, the work presents the classification of drug-related pages based on Street and Chemical drug name entities.

References

[1] Schäfer, M., Fuchs, M., Strohmeier, M., Engel, M., Liechti, M., & Lenders, V. (2019, May). BlackWidow: Monitoring the dark web for cyber security information. In 2019 11th International Conference on Cyber Conflict (CyCon) (Vol. 900, pp. 1-21). IEEE.

[2] Li, Z., Du, X., Liao, X., Jiang, X., & Champagne-Langabeer, T. (2021). Demystifying the dark web opioid trade: content analysis on anonymous market listings and forum posts. Journal of Medical Internet Research, 23(2), e24486.

[3] Samtani, S., Li, W., Benjamin, V., & Chen, H. (2021). Informing cyber threat intelligence through dark Web situational awareness: The AZSecure hacker assets portal. Digital Threats: Research and Practice (DTRAP), 2(4), 1-10.

[4] Alaidi, A. H. M., Al_airaji, R. A. M., ALRikabi, H. T., Aljazaery, I. A., & Abbood, S. H. (2022). Dark Web Illegal Activities Crawling and Classifying Using Data Mining Techniques. International Journal of Interactive Mobile Technologies, 16(10).

[5] Dalvi, A., Paranjpe, S., Amale, R., Kurumkar, S., Kazi, F., & Bhirud, S. G. (2021, May). SpyDark: Surface and Dark Web Crawler. In 2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC) (pp. 45-49). IEEE.

[6] Koloveas, P., Chantzios, T., Tryfonopoulos, C., & Skiadopoulos, S. (2019, July). A crawler architecture for harvesting the clear, social, and dark web for IoT-related cyber-threat intelligence. In 2019 IEEE World Congress on Services (SERVICES) (Vol. 2642, pp. 3-8). IEEE.

[7] Gao, C., Zhang, X., Han, M., & Liu, H. (2021). A review on cyber security named entity recognition. Frontiers of Information Technology & Electronic Engineering, 22(9), 1153-1168.

[8] Simran, K., Sriram, S., Vinayakumar, R., & Soman, K. P. (2019, December). Deep learning approach for intelligent named entity recognition of cyber security. In International Symposium on Signal Processing and Intelligent Recognition Systems (pp. 163-172). Springer, Singapore.

[9] Gasmi, H., Bouras, A., & Laval, J. (2018). LSTM recurrent neural networks for cybersecurity named entity recognition. ICSEA, 11, 2018.

[10] Gao, C., Zhang, X., & Liu, H. (2021). Data and knowledge-driven named entity recognition for cyber security. Cybersecurity, 4(1), 1-13.

[11] Liggett, R., Lee, J.R., Roddy, A.L., Wallin, M.A. (2020). The Dark Web as a Platform for Crime: An Exploration of Illicit Drug, Firearm, CSAM, and Cybercrime Markets. In: Holt, T., Bossler, A. (eds) The Palgrave Handbook of International Cybercrime and Cyberdeviance. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-319-78440-3_17

[12] Hämäläinen, L., Haasio, A., & Harviainen, J. T. (2021). Usernames on a Finnish Online Marketplace for Illegal Drugs. Names, 69(3).

[13] Meland, P. H., & Sindre, G. (2019, December). Cyber attacks for sale. In 2019 International Conference on Computational Science and Computational Intelligence (CSCI) (pp. 54-59). IEEE.

[14] Godawatte, K., Raza, M., Murtaza, M., & Saeed, A. (2019, December). Dark web along with the dark web marketing and surveillance. In 2019 20th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT) (pp. 483-485). IEEE.

[15] Hayes, D. R., Cappa, F., & Cardon, J. (2018). A framework for more effective dark web marketplace investigations. Information, 9(8), 186.

[16] Haasio, A., Harviainen, J. T., & Savolainen, R. (2020). Information needs of drug users on a local dark Web marketplace. Information Processing & Management, 57(2), 102080.

[17] https://github.com/jbogensperger/DRUG_CROSSNER Accessed on 10 September 2022

[18] Bogensperger, J., Schlarb, S., Hanbury, A., & Recski, G. (2021, November). DreamDrug-A crowdsourced NER dataset for detecting drugs in darknet markets. In Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021) (pp. 137-157).

[19] Zhang, P., Wang, X., Ya, J., Zhao, J., Liu, T., & Shi, J. (2021, December). Darknet Public Hazard Entity Recognition Based on Deep Learning. In Proceedings of the 2021 ACM International Conference on Intelligent Computing and its Emerging Applications (pp. 94-100).

[20] Aguilar, G., Maharjan, S., López-Monroy, A. P., & Solorio, T. (2019). A multi-task approach for named entity recognition in social media data. arXiv preprint arXiv:1906.04135.

[21] Al-Nabki, M. W., Janez-Martino, F., Vasco-Carofilis, R. A., Fidalgo, E., & Velasco-Mata, J. (2020). Improving Named Entity Recognition in Tor Darknet with Local Distance Neighbor Feature. arXiv preprint arXiv:2005.08746.

[22] https://www.newportacademy.com/ Accessed on 10 September 2022

[23] https://pubchem.ncbi.nlm.nih.gov/ Accessed on 10 September 2022

[24] Samtani, S., Chai, Y., & Chen, H. (2022). Linking exploits from the dark web to known vulnerabilities for proactive cyber threat intelligence: An attention-based deep structured semantic model. MIS Quarterly, 46(2), 911-946.

[25] Samtani, S., Zhu, H., & Chen, H. (2020). Proactively identifying emerging hacker threats from the dark web: A diachronic graph embedding framework (d-gef). ACM Transactions on Privacy and Security (TOPS), 23(4), 1-33.

APPENDIX

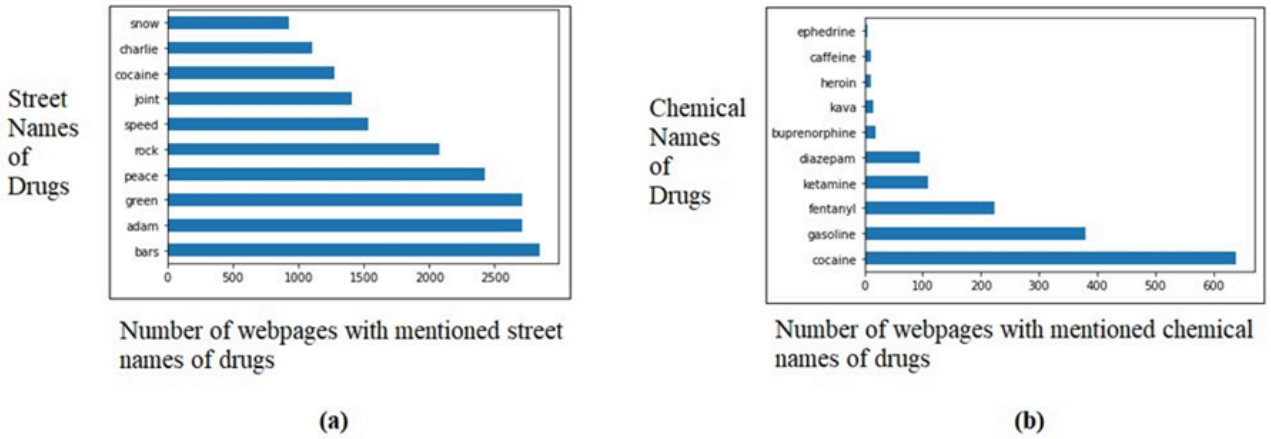


Fig. 4 (a) Bar graph visualization of web pages with mentions of Street names of drugs and (b) Bar graph visualization of web pages with mentions of Chemical names of drugs with method 1A

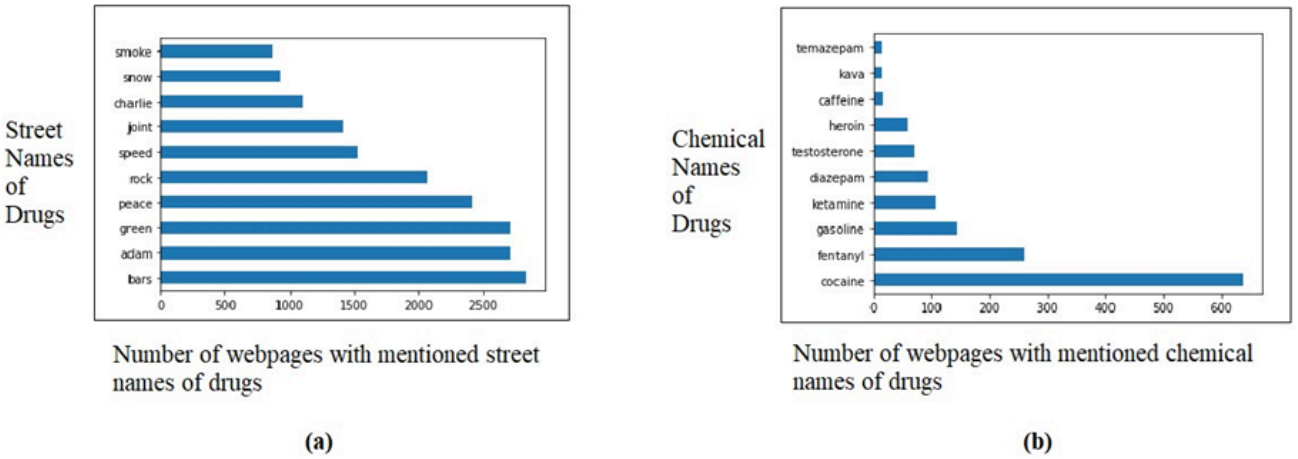


Fig. 5 (a) Bar graph visualization of web pages with mentions of Street names of drugs and (b) Bar graph visualization of web pages with mentions of Chemical names of drugs with method 1B

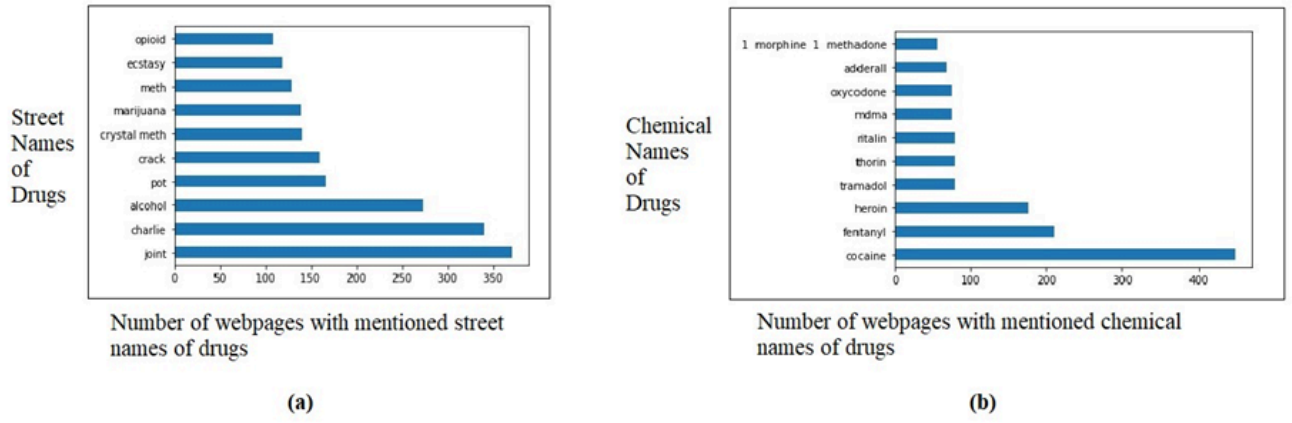


Fig. 6 (a) Bar graph visualization of web pages with mentions of Street names of drugs and (b) Bar graph visualization of web pages with mentions of Chemical names of drugs with method 2