

An Analysis of Opioid Trafficking on the Surface Web

Kaleigh Powell
Louisiana Tech University
kaleigh.powell@outlook.com

Austin Adams
Louisiana Tech University
austinsadams2@gmail.com

William Bradley Glisson
Louisiana Tech University
glisson@latech.edu

Abstract

Turns out, your everyday search engine might be hiding more than comedic videos and cooking blogs. While opioid trafficking is often associated with the dark web, recent studies suggest that illegal drug markets are creeping into the more visible corners of the internet. This research set out to investigate whether opioid trafficking can be identified on the surface web by combining a custom-built web crawler with a Large Language Model (LLM) to analyze web page text. On an initial test crawl of 75 webpages, the system flagged 17 as high-confidence cases of trafficking, most involving semi-synthetic opioids like Hydrocodone and Oxycodone. Strikingly, every flagged instance was attributed to sellers, not buyers, highlighting the commercial nature of these listings. These findings suggest that the surface web is not only being used for illicit activity, but that automated tools can meaningfully detect and categorize this behavior.

Keywords: Surface Web, web crawler, text analysis, drug marketplaces, opioids.

1. Introduction

Opioid trafficking poses a highly profitable market for criminals. Per the United Nations Office on Drugs and Crime (UNODC) in 2019, a kilogram of heroin sells for as much as \$60,000 and a kilogram of pharmaceutical opioids sells for up to \$90,000 (United Nations Office on Drugs and Crime 2022). Since 2013, the number of deaths due to opioid abuse has skyrocketed. According to the Center for Disease Control, in 2013 the number of deaths per 100,000 people was 1.0, and in 2022 the number of deaths jumped to 22.7 per 100,000 people (Spencer et al., 2024). The Drug Enforcement Agency (DEA) classifies drugs by Schedules ranging from I to V (1 to 5), with more harmful and addictive drugs being closer to Schedule I and drugs with less potential for abuse being closer to Schedule V. Opioids generally fall within Schedule I and Schedule II, meaning they are some of the most harmful and addictive drugs with the most risk

for abuse (United States Drug Enforcement Administration 2018).

There are three categories of opioids: natural, synthetic, and semi-synthetic (United States Drug Enforcement Administration 2018). Natural opioids are also sometimes referred to as opiates. In general, opioids are used for pain management, because they can interact directly with pain receptors in the brain and body to reduce the perception of pain (American Psychiatric Association, 2025). Withdrawal symptoms play a large role in the tendency of opioid addiction.

Per information released by the American Psychiatric Association (APA) in 2025, “3% to 19% of people who take prescription pain medications develop an addiction to them”. The APA reports that dependence on opioids occurs in a relatively short time span, about 4-8 weeks. They describe that this means withdrawal symptoms can present very quickly if a patient is on opioids for chronic issues. The APA continues to explain that symptoms can appear as soon as 6 hours after the last dose and often peak at about 72 hours after. They can include nausea, vomiting, stomach cramps, diarrhea, goosebumps, depression, anxiety, insomnia, generalized pain, and extreme cravings. These severe symptoms encourage continued dosage to avoid withdrawal (American Psychiatric Association, 2025).

With the relatively high potential for opioids to cause addiction, there are likely marketplaces for those who do not have prescriptions to acquire these drugs. Specifically, as digital interconnectivity has increased over recent years, so has the presence and efficiency of Internet-based supply chain operations (Ben-Daya et al., 2017). This increase in digital marketplace efficiency has had a direct impact on the sale and distribution of illegal substances. Particularly, from 2017 to 2022, the United Nations Office on Drugs and Crime (2022) reported that the number of websites involved in transactions linked to illegal goods and services increased dramatically. From 2014 to 2022, the proportion of drug-using consumers who had purchased drugs on the dark web more than doubled from 4.7 to 10.8 percent (United Nations Office on Drugs and Crime 2023).

Accessing the dark web is most commonly done through an Anonymous Communication Network (ACN) called Tor. According to a research paper by Jesper Bergman and Oliver B. Popov, Tor is the most common ACN for citizens in non-democratic and semi-non-democratic countries, whistleblowers, journalists, and really anyone in need of end-to-end anonymity (Bergman & Popov, 2023). This anonymity provided by Tor allows users from all backgrounds to access the dark web with full anonymity, backed by Onion Routing protocols, encryption algorithms, and anonymous traffic routing. This high level of anonymity permits individuals, and organizations, to engage in various cybercrimes like the sale of illicit or illegal goods and the hosting of illegal servers and websites. Here, people are able to freely discuss topics that would normally be reported by the surface web.

While research often focuses on the anonymity of the dark web as a facilitator of opioid trafficking, a growing body of evidence suggests that similar transactions occur on the surface web (van der Sanden et al., 2021). Unlike the dark web, the surface web is readily accessible through traditional search engines and is frequented by a broader demographic of users (United Nations Office on Drugs and Crime 2023). This shift in accessibility raises a key hypothesis: Opioid trafficking networks on the surface web can be identified and differentiated by analyzing text artifacts. Furthermore, this raises three relevant research questions:

1. Is there a way to identify the specific opioid being trafficked?
2. Can the geographic scope or role of the actor (seller vs. distributor) be inferred from listings?
3. Is there a significant difference in the number of advertisements found for different categories of opioids (synthetic, natural, semi-synthetic)?

This research demonstrates that trafficking of opioids is not limited to the obscured corners of the Internet but is also present in more accessible spaces on the surface web. By deploying an automated web crawler and using a Large Language Model (LLM) for analysis, the experiment uncovered that a portion of public web pages contain clear indicators of drug sales, especially involving commonly abused semi-synthetic opioids.

This paper is structured as follows: Section II introduces relevant research for drug marketplaces, web crawling, and Artificial Intelligence (AI) tools to analyze text. Section III describes the methodology used to perform the experiment. Section IV is a collection and analysis of the data gathered. Section V includes the conclusion and suggestions for future work.

2. Literature review

Many researchers have explored topics related to various web crawlers and dangerous material that exists on the surface web (Da Molin et al., 2019; Leonard et al., 2019; Rawat & Rajavat, 2024; Stewart et al., 2025). This review specifically aims to provide insight into the current work done in the areas of drug marketplaces by Bichler et al. (2017), Pilarczyk (2011), and Moyle et al. (2019), surface web crawlers by Shrivastava (2018), Ahuja et al. (2014), and Iliou et al. (2017), LLM analysis tools by Törnberg (2024) and Alizadeh et al. (2024), and medical perspectives on opioids by Dallin et al. (2023) and Häuser et al. (2021).

2.1. Drug marketplaces

Bichler et al. (2017) explain that contrary to popular belief and social media, drug marketing networks follow a loose organizational structure like that of legitimate businesses. They discovered that there exists a hierarchy within this subset of organized crime. However, the hierarchy can vary based on the size of the operation, with local operations being of higher density and global operations being of lower density. Density, in this context, does not refer to the number of people but the number of connections each individual in the group has to another individual. Bichler et al. found that in local operations, there will be a higher density, meaning several individuals in the set of people will have connections with multiple other individuals in that same set. This kind of operational structure, however, is not secure, because if one subset is compromised then the rest of the set is also compromised. Larger operations use a lower density because of this, which creates a "chain of command" and better protects the organization as a whole by ensuring one subset of the group cannot compromise the whole.

Pilarczyk (2011) explains that instead of marketing being seen or used as a tool, it is simply seen as a process which creates a specific value for the customer. He explains that this process covers three essential layers in which a message is delivered to the consumer: message senders, message content, and the communication channels by which the message is delivered.

Moyle et al. (2019) investigates the role of social media and messaging applications such as Snapchat, Instagram, and WhatsApp in facilitating drug transactions on the surface web. Through global surveys and interviews with app-using drug consumers, they found that platforms like these serve as a middle ground between street-level drug dealing and dark web crypto markets. It was evaluated that though users have concerns about law enforcement surveillance and drug

legitimacy, social apps on the surface web are becoming increasingly viable tools for illicit drug distribution.

2.2. Surface Web crawlers

Shrivastava (2018) offers a detailed examination of web crawlers, emphasizing their essential role in indexing the surface web for search engines. The study begins by explaining the seed URLs a crawler takes to begin traversing hyperlinks to collect and index webpage content. Crawlers do not directly access each page, rather sending requests to servers and parsing responses to build searchable indexes. She highlights that a single search engine can run thousands of crawling instances in parallel across distributed servers. Web crawler policies including politeness (avoidance of overwhelming servers), revisit (data refresh), and filtering (to avoid duplicated data or blocked content from robots.txt) are also defined and explored.

Ahuja et al. (2014) provide a detailed taxonomy of web crawlers, categorizing them based on crawling strategies and operational contexts on the surface web. These include breadth-first crawlers for simple site-wide coverage, incremental crawlers for updating existing datasets, form-focused crawlers for extracting hidden content behind forms, and focused crawlers that prioritize topic relevance through classifiers and link analysis. Advanced architectures like parallel and distributed crawlers are noted for their ability to scale web indexing by retrieving hundreds of pages per second across multiple systems. The research also outlines critical operational challenges, such as managing crawler scale in response to the rapidly growing web, filtering duplicate and irrelevant content, handling server politeness via robots.txt, and mitigating legal concerns like copyright and privacy. Performance metrics include the baseline capacity of a simple crawler, approximately 86,400 pages per day, underscoring the need for optimization through algorithms like PageRank, Fish Search, and heuristic prioritization. These insights demonstrate the technical and ethical complexities involved in designing robust surface web crawlers.

Iliou et al. (2017) developed a crawler to search both the dark and surface web. Their web crawler focused on the discovery of resources related to the creation of homemade explosives. Their experiment further compared the effectiveness of the crawler, comparing the results from both the Surface and Dark Web.

2.3. LLM analysis tools

Törnberg (2024) presents a comprehensive guide to leveraging LLMs for text analysis, emphasizing their

potential ability to process unstructured surface web data with minimal preprocessing. His approach reflects a shift from traditional keyword-based or structured web crawling to more interpretive analysis enabled by neural networks. He used large, real-world datasets from the surface web and used LLMs to perform tasks like classification and discourse analysis, showcasing the use of LLMs to perform text analysis on webpage contents.

Alizadeh et al. (2024) evaluated proprietary and open-source LLMs to assess their performance and cost-effectiveness in text annotation tasks, including stance detection, topic modeling, and relevance classification. They compared GPT-3.5, GPT-4, FLAN-T5 (XL), LLaMA-1, LLaMA-2, and LLaMA-3 by using zero-shot, few-shot, and fine-tuning approaches across 11 annotation tasks and 9,777 text samples from tweets and news articles. They found that GPT-3.5 outperformed crowd-sourced MTurk workers by an average of 25 percentage points in zero-shot settings and fine-tuning the model with just 50 annotated examples led to an average increase in accuracy of 15.7%. Other open-source models like FLAN-T5 (XL) and LLaMA-2 also showed strong results. Overall, commercial models performed better, but open-source LLMs consistently outperformed in tasks with more than two classification labels. Their results reinforced the value of LLMs, especially fine-tuned LLMs, for efficient and scalable text analysis.

Wan et al. (2024) researched how effective it is to use LLMs for generating relevant taxonomies and labeling data. While testing, this team discovered that their framework could be used for efficient analysis for a variety of domains and applications that rely on large volumes of unstructured text. By allowing this framework to analyze large sets of unstructured data, data can be interpreted into natural language. By labeling the data using key specific algorithms, the LLM is able to read through a text, label what it is, and analyze it. This allows any individual to take large sets of unstructured data, or text, and pass it into an LLM. The individual would then be given an easy-to-read response which allowed for more efficient decisions to be made based on the data. According to this research, LLMs also outperformed humans in every tested field. While their research was an overall success, the LLMs were fairly slow and not cost-effective at all. They suggested finding faster and cheaper models because of the high cost of evaluations.

Drug marketplaces have been well documented and investigated, as well as their presence on the Dark Web. However, most research on surface web drug trafficking focuses on social media sites. Based on this, there is a gap in utilizing these tools and methods to analyze opioid trafficking networks on the surface web directly.

3. Methodology

To investigate the hypothesis that opioid trafficking networks on the surface web can be identified and differentiated by analyzing text artifacts, this experiment is approached in two parts. First the development of the crawler, then the analysis of text. This experiment uses the outline for a controlled experiment as defined by Shadish et al. (2001), with the controlled, confounding, independent, and dependent variables listed in Table 1.

Table 1. Experiment variables and their types.

Variable Type	Variables
Controlled	Scrapy version & configuration, Gemini model version, System prompt for Gemini, Web page types, Automation timing using Watchdog, Network conditions
Dependent	Presence of trafficking content, Classification score, Percentage of flagged websites
Independent	Seed URLs, Keywords, Crawling rules & filters
Confounding	Non-English languages, Website layout inconsistencies, Crawler IP blocking (robots.txt), or Rate limiting

The tools used to conduct this experiment are presented in Table 2. Each tool used to conduct the research is listed on the left with the tool's specifications listed on the right.

Table 2. Tools and their specifications.

Tool	Specification
Hewlett Packard Laptop	Spectre, Intel Core i7, 16 GB Memory, 1 TB Storage, Linux Mint v 21.3
Python	v3.31
Scrapy	v2.13.0
Watchdog	v6.0.0
BeautifulSoup	v4.13.4
Gemini	Model 1.0 Pro
ChatGPT	Model 4.1 Mini

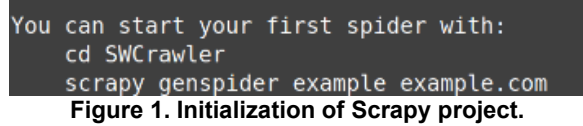
For this experiment, both the web crawler and Gemini client are created to run autonomously and independently of the other. This is done by creating separate subprocesses when the application is deployed. Meaning, the only steps the user must take to deploy the crawler is to download the crawler's code and run it.

By running the command `pip install -e` in the terminal, a file can be specified as main and therefore run as an executable. In this case, a file called `run_drugcrawl.py` was used. This file utilizes subprocesses and command line calls to run the crawler and Gemini client concurrently. Doing this allows for

the data to be analyzed while the crawler discovers additional webpages, allowing for fast data analysis.

3.1. Web crawler development

The crawler used in this experiment was constructed using Python and Scrapy, a free, open-source, web crawler framework (Scrapy Developers 2025). To initialize the Scrapy project and crawler, the `scrapy startproject SWCrawler` command is run. When this command is run, a guideline on how a spider can be started is printed to the terminal, shown in Figure 1. This output will also print the directory location where the spider and Scrapy project is stored.



After initializing the project, accessing the `swcrawler/spiders` directory allowed for further development and customization of the spiders.

First, the spider must have a set of URLs to begin crawling. For this, the directory and file `resources/seed_urls.txt` was created. Using ChatGPT, a list of seed URLs were generated using the following system prompt:

'I am developing a surface web crawler for my research project. My research involves finding websites that could be involved in possible drug trafficking and feeding them into my Scrapy crawler. Generate a list of seed urls to give my crawler.'

This list was then placed into the `seed_urls.txt` file, with the prompt being run several times after containing the previously generated list and an additional line stating:

'Do not reuse any URL in the list below.'

Next, the file was imported into the spider code and passed into the spider's `urls` variable as seen in Figure 2.

```
# Load seed URLs from file-
with open("./resources/seed_urls.txt") as f:-
    self.urls = [line.strip() for line in f if line.strip()]
```

Figure 2. Python code snippet showing the seed URLs being loaded into the spider.

After loading the crawler with the seed URLs, a filter is added so that only relevant websites' HTML is downloaded. Again, using ChatGPT, a list of keywords was created and placed into `swcrawler/keywords.py`. The system prompt used to generate this list of keywords is as follows:

'I am developing a surface web crawler for my research project. My research involves finding websites that could be involved in possible drug trafficking and

feeding them into my Scrapy crawler. Generate a list of keywords that may indicate drug trafficking of opioids to add into my spider so that only relevant HTML is pulled.'

To ensure that returned keywords are relevant to this research, each word in the list is reviewed. A report released by the Drug Enforcement Agency (DEA) was referenced to verify common street names and other drug slang terms and code words (DEA Houston Division, 2018). This adds a filter to the crawler so that only websites of potential relevance that contain the keywords can be downloaded.

Once a website's HTML is downloaded, it is then saved to a storage directory. An example directory is `scraped_data/DrugSpider/text/`. From here, the text will be analyzed by the AI client.

3.2. LLM text analysis

Using Watchdog, an observer is created to watch the `scraped_data/DrugSpider/text/` directory and wait for a file to be added. Once a file is added, the contents of the file are read, and the text is parsed out using BeautifulSoup. This text is then sent to the Gemini client through a free API key requested from the "Google AI for Developers" website (Google, n.d.).

Once the key is obtained, it is placed in an `.env` file. The directory and file `api/gemini.py` are created. This file contains a function that prompts Gemini with the carved text and feeds it a system prompt which outlines the parameters in which the agent should analyze data and how its response should be formatted as seen in Figure 3.

```
Return a JSON object with the following structure:-
{
  "confidence": "<low|medium|high>",-
  "drug": ["<Type of opioid present>"],-
  "method": "<selling|buying|making|transporting>"-
}
```

Figure 3. Gemini response specification prompt.

The prompt also adds restrictions that prevent hallucinations, making assumptions, and specifies how the LLM should classify specific results. These guidelines are shown in the following snippet of the overall system prompt:

'You are a language analysis model specialized in detecting indicators of drug trafficking. Your task is to analyze the provided text for any linguistic, semantic, or contextual evidence that may suggest involvement in or references to drug trafficking activity. Guidelines:

- Consider both direct and indirect references, including slang, euphemisms, or coded language.

- Assess the context and tone of the text to distinguish between innocuous and suspicious content.
- Do not infer beyond what is contextually supported in the text.
- Do not generate explanations or narrative responses.

Important:

- Do not list any drugs unless they are explicitly mentioned or unambiguously referenced in the text. Avoid assumptions or hallucinations.
- Do not generate, infer, or fabricate drug names beyond what the input text clearly supports.'

Results are classified by high, medium, and low confidence levels. This ranking is determined based on a snippet from the overall system prompt which is shown below:

'Drug Detection Confidence Guidelines:

- *High: Direct and explicit mention of drug trafficking or specific drug names commonly associated with illegal distribution, with clear context suggesting trafficking behavior (e.g., "selling fentanyl", "shipment of heroin").*
- *Medium: Indirect, coded, or euphemistic references that plausibly indicate drug trafficking (e.g., "moving powder", "white stuff delivery"), especially if accompanied by contextual clues such as logistics, pricing, or locations.*
- *Low: Vague, ambiguous, or speculative language that may suggest drug trafficking but lacks sufficient supporting context (e.g., "stuff", "package", "drop") and could be interpreted innocuously.'*

From here, the message is sent using the API. The response is formatted as a code block initially, meaning the JSON needs to be extracted before moving to the next request. This can be done as seen in Figure 4. This data is then immediately stored in a file called `processed_results.json` for further analysis.

```
# Clean the response text from Markdown code block
formatting-
cleaned_text = response.text.strip()-
-
# Remove starting ```json or ``` and ending ```-
if cleaned_text.startswith("```"):-
    cleaned_text =
cleaned_text.lstrip("```json").lstrip("```").strip()-
    if cleaned_text.endswith("```"):-
        cleaned_text =
cleaned_text.rstrip("```").strip()-
-
# Now parse JSON-
parsed_response = json.loads(cleaned_text)-
return parsed_response-
```

Figure 4. Python code snippet showing response parsing of Gemini response.

Due to time constraints, the crawler was manually stopped after a benchmark of 1,500 HTML pages were captured. The AI analysis code was then run manually so that data analysis could continue processing HTML. Three different API keys were generated and were changed frequently so that when an API ran out of usage tokens, another key could take its place until the tokens had enough time to regenerate. The command used to run this module alone was `python3 -m swcrawler.api.gemini`. Slight modifications were made to the module to allow it to be run by itself, without the dependence of Watchdog. These modifications can be seen in Figure 5.

```
if __name__ == "__main__":
    input_directory =
    "./swcrawler/scraped_data/DrugSpider/text/"
    output_directory =
    "./swcrawler/scraped_data/DrugSpider/processed/"
    process_directory(input_directory, output_directory)
```

Figure 5. Code snippet of modification to Gemini module, allowing it to run on its own.

Here, `process_directory` is a custom function built to take an input directory and an output directory as arguments. The first file stored in the input directory is analyzed using the AI agent, then placed into the output directory for storage. This allows for the sorting of processed and unprocessed files, which provides a faster way to examine downloaded or processed HTML files.

3.3. Scope and restrictions

This experiment focuses on the discovery of websites advertising opioid trafficking, the identification of the opioids being trafficked, and the method of distribution (i.e. selling, buying, making, and transporting) on the surface web.

Before performing this experiment, several restrictions were put in place. The first restriction was that robots.txt had to be obeyed. Robots.txt is a file that websites may configure that aims to prevent the site from being overloaded with requests, which blocks crawlers from being able to access the page. Ignoring robots.txt is illegal and a breach of privacy and terms of service, so the crawler is built to always respect websites with robots.txt configured. This experiment also does not attempt to bypass any authentication requirements or other built-in crawler blocking mechanisms.

The scope of this research is limited to the use of the free Gemini API for text analysis. Text analysis software, open-source tools, and other LLM APIs were not considered in this research. Additionally, to minimize the chance of hallucinations from the AI model, the system prompt is carefully constructed using

methods and procedures outlined by Barkley and van der Merwe (2024).

4. Results and analysis

The first run of the crawler simply aimed to verify the functionality of the code. This first run identified a total of 75 unique webpages containing potential opioid-related content. The second crawl identified 1,500 webpages. Each webpage is parsed and analyzed using the Gemini language model, returning classifications based on the confidence level of detected trafficking content and associated drug types.

An example JSON response is displayed in Figure 6. Here, each response from the agent has been saved to a file and can be easily referenced and analyzed. The confidence of a response is given at the beginning of each line, followed by a list of any identified drugs, and ending with the identified method of trafficking. In Figure 6, it is observed that 2 of the 7 responses returned a medium confidence, listed several drugs, and displayed 'selling' for the method of trafficking. All of the data presented in this section is analyzed by manually reviewing these JSON responses and tallying the data listed for each field (confidence, drugs, and method).

```
{"confidence": "low", "drug": [], "method": ""}
{"confidence": "low", "drug": [], "method": null}
{"confidence": "medium", "drug": ["Tramadol",
"Hydrocodone", "Temazepam", "Modafinil", "Diazepam",
"Klonopin", "Adderall", "Oxycontin", "Alprazolam",
"Zolpidem"], "method": "selling"}
{"confidence": "low", "drug": [], "method": null}
{"confidence": "medium", "drug": ["Xanax", "Alprazolam",
"Hydrocodone", "Temazepam", "Modafinil", "Diazepam",
"Klonopin", "Adderall", "Oxycontin"], "method":
"selling"}
{"confidence": "low", "drug": [], "method": ""}
{"confidence": "low", "drug": [], "method": null}
```

Figure 6. Responses saved to file after parsing.

To verify that responses labelled with medium or high confidence and indicated a specific method of trafficking were accurate, the HTML files associated with these entries were reviewed. One of these analyzed HTML files was 654 lines long. Two lines in this file mentioned purchase of an opioid with no prescription required or a specific payment method known to indicate potential drug trafficking (United Nations Office on Drugs and Crime 2023). Line 20 specifically says: `“description:’Buy Meds Online without Rx”`. Line 304 states: `“alt=’btc payment method”`.

Further inspection of this file showed mentions of specific drug names including Valium, Oxycontin, and Adderall. Line 20 indicates that an Rx, or prescription, is not required for the product to be purchased. Line 304 suggests that a known potentially illicit purchasing

method “btc”, also known as Bitcoin, can be used on this site. Based on this context, this site does have a high confidence level of drug trafficking and the method of trafficking was correctly identified as selling. This method of inspection is used to manually check the validity of the responses of all HTML files that were flagged with medium or high confidence.

Further analysis of the results is presented by confidence level, drug type, and chemical categorization to provide a thorough analysis of the collected data.

4.1. Confidence in classification

As shown in Figure 7, the majority of the classified responses for the first crawl (57 out of 75, or 76%) were labeled as low confidence by the model. A smaller subset (17 responses, or 22.7%) was classified as high confidence, while only 1 response (1.3%) was categorized as medium confidence. The one case categorized as medium confidence did not return a specific drug name in the response and was not considered further in the results analyzing the types and categories of drugs.

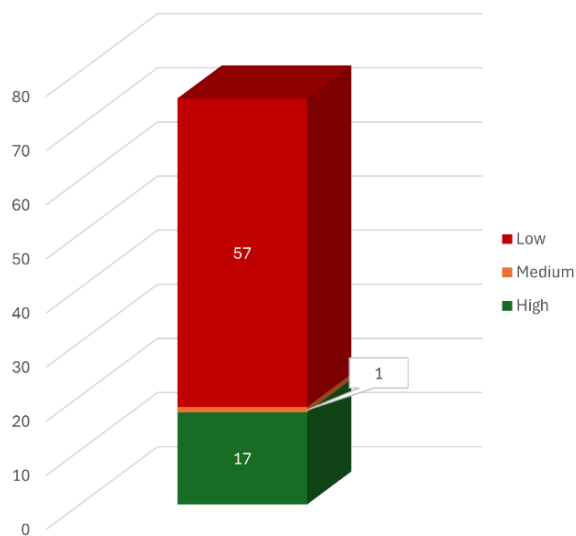


Figure 7. First crawl's distribution of confidence ratings.

For the second crawl, 1,453 out of 1,500, or 96.9% responses were labeled as low confidence. The responses returned with medium confidence were 19 out of 1,500 or 1.3%, and the responses returned with high confidence were 28 out of 1,500 or 1.9%. In total, about 3% of responses indicated there was potential drug trafficking. The number of medium and high responses is displayed in Figure 8.

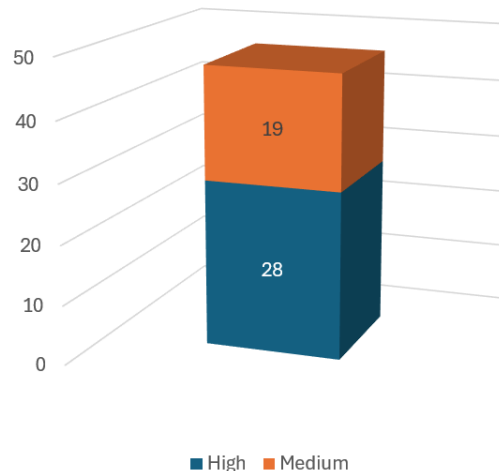


Figure 8. Second crawl's distribution of medium and high confidence ratings.

The high number of low-confidence classifications suggests that either the true absence of opioid-related content, keyword ambiguity, incomplete webpage data, or seller obfuscation strategies (e.g., code words, image-based communication) may be affecting automated detection capabilities.

4.2. Frequency and type of opioids identified

Among the 17 high-confidence classifications in the first crawl, two primary opioid categories emerged: Hydrocodone and Oxycodone, each appearing in 13 of the 17 cases. One additional opioid, Tramadol, appeared in 2 out of 17 responses, representing a smaller proportion of the flagged content. Table 3 shows the overall percentage for each category of opioid returned from the 17 high-confidence cases. Each response of high confidence may list more than one type of drug. If two opioid categories were listed in the same response, each was counted separately. Additionally, every high-confidence listing was categorized as a seller. These listings were categorized as selling because they all referenced a method of purchase or indicated products could be added to a cart and shipped.

Table 3. Percentage of identified opioid types in the first crawl.

Type of Opioid	Percentage of Responses
Hydrocodone	76.47%
Oxycodone	76.47%
Tramadol	11.76%

Hydrocodone and Oxycodone each account for 76.47% of the total identified high-confidence opioid

content, with Tramadol comprising 11.76%. Specific drug names and brands mentioned under Oxycodone are Oxycontin, Percocet, and Roxicodone. For Hydrocodone and Tramadol, the base opioid name was the only name mentioned.

A report from the Drug Enforcement Administration (2025) explains that semi-synthetic opioids, particularly those that are often combined with other medications are prescribed more frequently. The findings suggest that these medications are more common in surface web opioid mentions than opioids such as heroin, which has no accepted medical use in the United States.

In the second crawl, every medium confidence and high confidence response returned a drug type. The number of medium confidence and high confidence responses were added together for a total of 47 responses that indicated potential drug trafficking.

As in the first crawl, Hydrocodone and Oxycodone emerged as the primary opioid categories. Hydrocodone appeared in 38 out of 47 responses while Oxycodone appeared in 39 out of 47 responses. Tramadol once again lagged behind these with only 5 mentions of the 47 responses. Table 4 shows the overall percentage each of these categories represented out of the total 47 responses. Again, each response could return more than one drug, so each percentage is calculated independently of one another. All 47 positive responses were also labelled as sellers, meaning there was direct reference to purchasing methods, which further highlights the more distribution-oriented purpose of trafficking on the surface web.

Table 4. Percentage of identified opioid types in the second crawl.

Type of Opioid	Percentage of Responses
Hydrocodone	80.85%
Oxycodone	82.98%
Tramadol	10.64%

In the second crawl, Hydrocodone accounted for 80.85% of positive responses while Oxycodone accounted for 82.98%. Tramadol accounted for 10.64%. Comparing the percentages from the first and second crawls, it is observed that the overall percentages stayed relatively similar with Tramadol accounting for approximately 11% on average while Hydrocodone and Oxycodone represented approximately 79% on average.

When analyzing the specific brand names of the drugs identified, Oxycodone again returned Oxycontin, Percocet, and Roxicodone. Hydrocodone returned Vicodin, and Tramadol simply returned its generic name.

4.3. Synthetic vs. semi-synthetic opioid patterns

Each identified opioid was categorized based on its chemical origin: synthetic, semi-synthetic, or natural. Of the opioids mentioned in high-confidence contexts for the first crawl, 13 were semi-synthetic (Hydrocodone and Oxycodone), while 2 were fully synthetic (Tramadol). No natural opiates such as morphine or codeine were flagged. The distribution of these opioid categories is displayed in Table 5.

Table 5. Number of mentions of specific opioid chemical categories in the first crawl.

Category	Number of Mentions
Semi-Synthetic	13
Synthetic	2
Natural	0

In the second crawl, there were 39 mentions of semi-synthetic opioids (Hydrocodone and Oxycodone) and only 5 mentions of fully synthetic opioids (Tramadol). Once again, no natural opiates were flagged. The distribution of these opioid categories is displayed in Table 6.

Table 6. Number of mentions of specific opioid chemical categories in the second crawl.

Category	Number of Mentions
Semi-Synthetic	39
Synthetic	5
Natural	0

4.4. Non-opioid compounds

Interestingly, the crawler and language model also detected a considerable number of non-opioid substances mentioned in conjunction with trafficking-like content. In the first crawl, benzodiazepines appeared in 14 out of 17 high-confidence cases. These included specific drugs like Xanax, Alprazolam, Valium, Diazepam, Klonopin, Rivotril, Restoril, and Temazepam. Thienodiazepine-based drugs including Etizolam and Etilaam were also noted in 2 of 17 cases.

In the second crawl, benzodiazepines appeared in 41 of the 47 positive response cases and thienodiazepines appeared in only 2 of the 47 responses. Specific benzodiazepines included Xanax, Alprazolam, Diazepam, Klonopin, Temazepam, Restoril, Valium, Rivotril, and Lorazepam. For thienodiazepines, Etizolam and Etizolam were returned.

Another non-opioid substance was found in one response in the second crawl. This substance is Tianeptine, a drug that can act similarly to an opioid.

This drug is not approved for use by the U.S. Food and Drug Administration (FDA) for any real medical uses and has been marked as a drug of particular concern with a high potential risk for abuse (U.S. Food and Drug Administration, 2025).

This pattern of non-opioid drugs being returned alongside opioids aligns with known trends in poly-drug abuse where benzodiazepines are often co-used with opioids to enhance sedation or mitigate withdrawal (Jones et al., 2012). It also raises important questions regarding the broader scope of substance abuse detection and the utility of expanding crawler filters beyond opioid-related keywords.

5. Conclusion and future work

The hypothesis that opioid trafficking networks can be identified on the surface web through the analysis of text artifacts is supported by the findings of this study. The analysis of the initial 75 webpages revealed 17 high-confidence instances of trafficking content and the second returned 47 instances of trafficking from the total 1,500. This supports the idea that web crawling paired with language model analysis is a viable method for detecting illicit activity online. Furthermore, the language model was able to accurately identify the specific opioids being sold, with Hydrocodone, Oxycodone, and Tramadol being the specific types of opioids found by the crawler. In terms of the role of the actors, every single listing was classified as a seller, suggesting that most content flagged for trafficking on the surface web is oriented around direct sales rather than purchase, manufacture, or transport. Additionally, the study indicates that semi-synthetic opioids are represented in surface web content more than synthetic or natural alternatives, answering the final research question. The presence of seller behavior and repeated mention of specific substances reveal that even without anonymity tools like Tor, illicit actors are operating openly online. These findings highlight a growing need for proactive monitoring of surface web content and indicate that traditional Internet spaces are being leveraged for illegal pharmaceutical distribution.

Future research will explore the application of this methodology to the dark web, where access-controlled environments and greater anonymity may reveal a different landscape of trafficking behavior. Investigating seed URLs from dark web indexes, including those hosted behind authentication layers or chat-based forums, could yield richer datasets. Beyond opioid detection, future work could expand the scope of analysis to include other illicit substances or illegal items such as counterfeit documentation, unlicensed pharmaceuticals, and weapons. Moreover, the ability to extract and analyze image metadata, such as EXIF data

or steganographic content, would offer a deeper layer of insight into how information is being concealed and shared across platforms. Additionally, discovering methods to ethically and legally crawl sites on the surface web that have protections against web crawlers, such as a robots.txt file, should be considered to further the pool of URLs that may be investigated. Finally, this research will be extended to explore mobile applications or encrypted messaging services on the surface web, as these channels increasingly serve as hubs for drug-related transactions.

6. References

- Ahuja, M. S., Bal, J. S., & Varnica. (2014). Web Crawler: Extracting the Web Data. *International Journal of Computer Trends and Technology (IJCTT)*, 13(3), 132–137.
<https://doi.org/10.14445/22312803/IJCTT-V13P128>
- Alizadeh, M., Kubli, M., Samei, Z., Dehghani, S., Zahedivafa, M., Bermeo, J. D., Korobeynikova, M., & Gilardi, F. (2024). Open-source LLMs for text annotation: a practical guide for model setting and fine-tuning. *Journal of Computational Social Science*, 8.
<https://doi.org/https://doi.org/10.1007/s42001-024-00345-9>
- American Psychiatric Association. (2025). Opioid Use Disorder. <https://www.psychiatry.org/patients-families/opioid-use-disorder>
- Barkley, L., & van der Merwe, B. (2024). Investigating the Role of Prompting and External Tools in Hallucination Rates of Large Language Models. *arXiv*.
<https://doi.org/http://dx.doi.org/10.48550/arXiv.2410.19385>
- Ben-Daya, M., Hassini, E., & Bahroun, Z. (2017). Internet of things and supply chain management: a literature review. *International Journal of Production Research*, 57(15-16), 4719–4742.
<https://doi.org/https://doi.org/10.1080/00207543.2017.1402140>
- Bergman, J., & Popov, O. B. (2023). Exploring Dark Web Crawlers: A Systematic Literature Review of Dark Web Crawlers and Their Implementation. *IEEE Access*.
- Bichler, G., Malm, A., & Cooper, T. (2017). Drug supply networks: a systematic review of the organizational structure of illicit drug trade. *Crime Science*, 6(2), 1–23.
<https://doi.org/https://doi.org/10.1186/s40163-017-0063-3>
- Da Molin, G., Napoli, M. L., Sabella, E. A., & Veshi, A. (2019). The Youth and the Dangers of the Web: A Field Study. *RIEDS - Rivista Italiana di Economia, Demografia e Statistica - The Italian Journal of Economic, Demographic and Statistical Studies*, 73(1), 65–76.

- Dallin, J., King, C. R., & Galke, C. (2023). The Opioid Epidemic: A Review of the Contributing Factors, Negative Consequences, and Best Practices. *Cureus*, 15(7). <https://doi.org/10.7759/cureus.41621>
- DEA Houston Division. (2018). Slang Terms and Code Words: A Reference for Law Enforcement Personnel. <https://www.dea.gov/sites/default/files/2018-07/DIR-022-18.pdf>
- Drug Enforcement Administration. (2025). Hydrocodone. *Drug & Chemical Evaluation Section*.
- Google. (n.d.). *Gemini Developer API*.
- Häuser, W., Buchser, E., Finn, D. P., Dom, G., Fors, E., Heiskanen, T., Jarlbaek, L., Knaggs, R. D., Kosek, E., Krceviski-Škvarč, N., Pakkonen, K., Perrot, S., Trouvin, A.-P., & Morlion, B. (2021). Is Europe also facing an opioid crisis?--A survey of European Pain Federation chapters. *Eur J Pain*, 25, 1760–1769. <https://doi.org/https://doi.org/10.1002/ejp.1786>
- Iliou, C., Kalpakis, G., Tsikrika, T., Vrochidis, S., & Kompatsiaris, I. (2017). Hybrid focused crawling on the Surface and the Dark Web. *EURASIP Journal on Information*, 11. <https://doi.org/https://doi.org/10.1186/s13635-017-0064-5>
- Jones, J. D., Mogali, S., & Comer, S. D. (2012). Polydrug abuse: a review of opioid and benzodiazepine combination use. *Drug Alcohol Dependence*, 125(1-2), 8–18. <https://doi.org/10.1016/j.drugalcdep.2012.07.004>
- Leonard, J. B., Hines, E. Q., & Anderson, B. D. (2019). Prime Eligible Poisons: Identification of Extremely Hazardous Substances Available on Amazon.com. *Clinical Toxicology*, 58(1), 45–48. <https://doi.org/https://doi.org/10.1080/15563650.2019.1594870>
- Moyle, L., Childs, A., Coomber, R., & Barratt, M. (2019). #Drugsforsale: An exploration of the use of social media and encrypted messaging apps to supply and access drugs. *International Journal of Drug Policy*, 63, 101–110. <https://doi.org/10.1016/j.drugpo.2018.08.005>
- Pilarczyk, B. (2011). Marketing communications process on the pharmaceutical market. *International Marketing Trends Conference*.
- Rawat, R., & Rajavat, A. (2024). Perceptual Operating Systems for the Trade Associations of Cyber Criminals to Scrutinize Hazardous Content. *International Journal of Cyber Warfare and Terrorism*, 14(1). <https://doi.org/10.4018/IJCWT.343314>
- Scrapy Developers (2025). Scrapy Documentation. <https://docs.scrapy.org/en/latest/>
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and Quasi-Experimental Designs for Generalized Casual Inference* (2 ed.). Cengage Learning.
- Shrivastava, V. (2018). A Methodical Study of Web Crawler. *Journal of Engineering Research and Application*, 8(11), 1–8. <https://doi.org/10.9790/9622-0811010108>
- Spencer, M., Garnett, M., & Miniño, A. (2024). Drug Overdose Deaths in the United States. *NCHS Data Brief*, 491, 1–5. <https://www.cdc.gov/nchs/data/databriefs/db491.pdf>
- Stewart, B., Vessel, B., & Glisson, W. B. (2025). *Classifying Dark Web Executables Using Public Malware Tools* Proceedings of the 58th Hawaii International Conference on System Sciences,
- Törnberg, P. (2024). How to Use Large Language Models for Text Analysis.
- U.S. Food and Drug Administration. (2025). Tianeptine Products Linked to Serious Harm, Overdoses, Death. <https://www.fda.gov/consumers/consumer-updates/tianeptine-products-linked-serious-harm-overdoses-death>
- United Nations Office on Drugs and Crime (2022). Drug Trafficking & Cultivation: Drug prices. <https://dataunodc.un.org/dp-drug-prices>
- United Nations Office on Drugs and Crime (2023). World Drug Report 2023, Booklet 3: Drug Market Trends - Cannabis and Opioids, Chapter 7: The Role of the Dark Web. 12. https://www.unodc.org/res/WDR-2023/WDR23_B3_CH7_darkweb.pdf
- United States Drug Enforcement Administration (2018). Drug Scheduling.
- van der Sanden, R., Wilkins, C., Romeo, J., Rychert, M., & Barratt, M. (2021). Predictors of using social media to purchase drugs in New Zealand: Findings from a large-scale online survey. *International Journal of Drug Policy*, 98. <https://doi.org/10.1016/j.drugpo.2021.103430>
- Wan, M., Safavi, T., Jauhar, S. K., Kim, Y., Counts, S., Neville, J., Suri, S., Shah, C., White, R. W., Yand, L., Andersen, R., Buscher, G., Joshi, D., & Rangan, N. (2024). *TnT-LLM: Text Mining at Scale with Large Language Models* Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining,