MINING THE HIDDEN: ETHICAL DARK WEB DATA COLLECTION STRATEGIES FOR DEEPFAKE THREAT ANALYSIS

Shaikh Junaid Ahmad
Research Scholar, Institute of Technology & Management (SSBES' ITM), Nanded
shaikh.junaid24@gmail.com

Dr. P.A. Kadam
Assistant Professor, Institute of Technology & Management (SSBES' ITM), Nanded
puru.kadam@gmail.com

Abstract
One of the main features of the article is the deepfake technologies that authors introduced and described their tremendous influence on the areas of cybersecurity, privacy, and society as a whole. Besides, the advent of deepfake technologies in the dark web is also outlined in the paper. Authors propose a concept that is both ethical and technical to unearth dark web data for the sake of deepfake research. The research describes the character of the dark web in terms of the kinds of domains that it consists of, such as forums, marketplaces, and hidden services, and it also specifies potential multimodal data types for dark web research, like text, metadata, and media references. The paper illustrates the design of a data gathering system that comprises encrypted storage, Tor-based access, crawler tools, and thorough Preprocessing. Fundamental ethical concerns like institutional review, compliance with the law, and the prohibition of the distribution of illicit media have been included in the research. The article also highlights the difficulties of anonymity, lack of data, and changing dark web structures, as well as recommends the methods of conducting research that are safe, can be repeated, and are legally compliant. An instance is used here to demonstrate the ethical analysis of the metadata of deepfake software sales. Overall, this piece moves the needle on the adoption of the ethics-first and the hands-on approach of exploring nefarious deepfake activities in hidden networks, thereby, paving the way for academic and law enforcement-led inquiries to further study this area.
***Keywords**: Deepfake Technologies, Cyber Security, Dark web mining, Cybersecurity research methods, Web crawling (Tor, Scrapy, Selenium).*

1. Introduction

1.1 Background

Deepfakes are synthetic media created with deep learning that now pose a major cyberthreat. Initially developed for creative purposes, they are increasingly used in harmful ways such as disinformation, identity theft, blackmail, and political manipulation (Kumari & Baggam, 2025). The dark web, with its hidden markets, forums, and onion sites, plays a key role by enabling anonymous trade in deepfake tools and services (Zhao et al., 2023; Vidanage et al., 2024).

Accessing reliable data from these spaces is very challenging. Tor's anonymity, encrypted communication, and intentional secrecy make it difficult for researchers to gather information (Laliberte et al., 2022). In addition, legal and ethical issues make data collection risky, as scraping dark web platforms can expose researchers to dangerous material or potential legal problems (Scanlon, 2016).

Unlike earlier research focused on deepfake detection (Rossler et al., 2019), this paper emphasizes safe methodology for collecting and analysing dark web data in a reproducible and legally compliant way.

1.2 Objectives

The purposes of this study are the following:

1. To specify the moral boundaries for collecting data from the dark web for deepfake research.
2. To develop the research design for data crawling, storage, and preparation.
3. To achieve the goal of ethical and legal compliance, as well as non-harmful downloads, simultaneously.
4. To demonstrate the approach with a metadata extraction sample.
5. To evaluate it against current dark web mining methods.

1.3 Literature Review

1.3.1 Dark Web Ecosystems and Data Collection Challenges

The dark web is a haven for illegal activities such as drug markets, cybercrime services, and the distribution of deepfakes. Researchers have found that Tor-based onion services are the hubs of marketplaces and communities where synthetic media tools are promoted and shared (Bradbury, 2020; Everitt, 2021). Collecting data in such places involves dealing with the challenges posed by the anonymity networks, encryption layers, the changing nature of marketplaces, and the danger of researchers coming into contact with harmful

contents (Laliberte et al., 2022; Dupont et al., 2022).

To a great extent, these studies have come up with the darknet specialized crawlers for monitoring. One of the earliest frameworks of an onion crawler for Tor hidden services was developed by Scanlon (2016), while the more recent work has been concentrating on the scalability of crawlers that can be integrated with Scrapy and Selenium for the automation of data acquisition (Goode et al., 2023). Rather than focusing on illegal markets, most of these methods and tools are directed towards deepfake-related research leaving a gap in the methodologies.

1.3.2 Technical Approaches to Dark Web Mining.

The predominant way to gather information from the deep web is through the use of crawling methods. In their work, *Laliberte et al. (2022)* describe a typical approach to automation of data scraping that takes into consideration the allowable speed of operation and the need for stealth to avoid being detected by site administrators. The authors *(Zhao et al., 2023)* state that hybrid methods are one such example, where the rigorous nature of a focused crawl is combined with machine learning algorithms trained to pick the most suitable forums to proceed with.

Besides this, storage and pre-processing are also significant. To securely and safely manage sensitive data, fully encrypted databases such as MongoDB or SQL with disk encryption have been suggested *(Weimann, 2016)*. Data pre-processing is usually done through anonymization as a way to prevent the risk of exposure and to be fair to the ethical concerns, while the application of natural language processing (NLP) frameworks like SpaCy and NLTK for the extraction of significant features from the raw text (Zhao et al., 2023).

1.3.3 AI-Powered Threat Detection and Data Scarcity

Over time, the role of AI in the detection of cyber threats – malware, phishing, and deepfakes – has grown tremendously, yet its progress is limited by the lack of datasets. *Vishwamitra et al*. (2023) contend that if threat detection models are trained on limited or surface web datasets, they lose their ecological validity when applied to hidden networks. The well-known deepfake datasets FaceForensics++ (*Rossler et al., 2019*) and Celeb-DF (*Li et al., 2020*) are helpful for benchmarking but are far from being the right sources if one is to track the underground distribution of malicious media.

This point has been underlined very clearly in the cybersecurity literature, where the researchers have repeatedly expressed the need for datasets that are specific to the domain and which capture the dynamics of the threat actors' behaviours rather than those that are collected in a laboratory setting (*Everitt, 2021; Goode et al., 2023*). They further warn that without such resources, the detection models run the risk of being out of sync with the real-world adversarial environments.

1.3.4 Ethical Considerations in Cybercrime and Dark Web Research.

The question of ethics and legality are still major issues in dark web research. *Parsons et al.* (2019) and *Scanlon* (2016) both emphasize that researchers should comply with the requirements of the institutional review board; they must also reduce the risks of coming across or storing illegal materials, and use methods that protect the health of the researcher. Some of the best methods to implement are gathering only metadata instead of the complete illicit media, anonymizing the sensitive identifiers, and using a combination of strict access controls (*Décary-Hétu & Giommoni, 2017; Weimann, 2016*).

The professional principles found in the ACM Code of Ethics and IEEE focus on the three main ideas – **openness, prevention of harm,** and informed consent. In the case of cybercrime, however, it is almost impossible to get the consent, so anonymization and minimization of data exposure have been used as forms of protection against each other (*Parsons et al., 2019*). Moreover, recent works are also referring to "ethics by design" in the domains of cyber threat research, where every stage of data handling activities has embedded compliance verification and ethical safeguards (*Dupont et al., 2022*).

1.3.5 Gaps and Emerging Needs

While the dark web data mining methods and ethical cybersecurity research have gone a long way, the application of the same methods in deepfake threat analysis is still an open-field with many gaps. Most of the development of the dark web, existing crawling methods have been towards the drug market or deceitful financial activities; on the other hand, synthetic media has been barely touched despite its increased significance. Moreover, the presently used deepfake datasets are not only disconnected from the dark web, but also they downgrade the detection studies' fidelity. Although there are ethical frameworks, those few responsible dark web collection guidelines that are reproducible and provide step-by-step instructions are not available for this field of study.

The present work closes these gaps by presenting a methodology-first framework for an ethical dark web data collection on deepfake-related activities.

Contrary to the alert-total focus of previous works on the development of detection models, this study is about supporting the accomplishment of safe, reproducible, and lawfully defensible data acquisition that can be the basis for future AI-based deepfake threat analyses.

2. Methodology

2.1 Framework Overview

The research presents a methodology composed of five layers aiming at the collection of data from the dark web in an ethical and technically sound manner. The study focuses on the collection of data related to deepfake technology. The framework is intended to perform a balance between methodological rigor, legal compliance, and researcher safety.

1. Access Layer

One is gained to the dark web via Tor, in addition to VPN tunneling and onion routing being used to keep anonymity and protect the identity of the researcher. This layer is the one that certifies the safety of the following crawling operations.

2. Crawling Layer:

The data is obtained through a combined crawling method that involves the use of Scrapy, Selenium, and Python requests. These instruments allow an automated browsing of forums and marketplaces, without the focus on only text-based posts but also advertisements and metadata that refer to deepfake tools and services.

3. Data Storage Layer

Information gathered from the public is kept in a secure and encrypted database such as MongoDB or SQL with full-disk encryption. The strict rules on who can have access to the database and the audit trail of the change made to the data are the two ways to guarantee that there is accountability and the data is kept in its integrity. Only metadata and text snippets are retained, explicitly excluding any illicit multimedia files.

4. Preprocessing Layer

Preprocessing is one that changes the raw, unstructured content into a format that can be analyzed. NLTK and SpaCy are required to handle the processes of text cleaning, tokenization, and topic modeling. OpenCV is used for feature extraction in images or videos that are the subject of metadata referencing (e.g., file size, timestamps) without transferring the actual content. A strictly enforced protocol for the removal of any information that can identify a person is being implemented.

5. Ethical Guardrails

The framework is equipped with the mechanisms for ethical compliance that include consent from the Institutional Review Board (IRB), strict observance of GDPR regarding privacy of data, and keeping with national cyber laws, for example, the Information Technology (IT) Act in India. The idea of "no-illegal-download" forms another layer of the guard which is there to make sure that the research is not exposed to any kind of unlawful media scenarios.

2.2 Detailed Reproducibility

The experimental method is repeatable through a detailed stepwise approach:

1. **Secure Research Environment**
   - A sandboxed virtual machine (VM) is established, isolated from the host system.
   - Network access is routed through a VPN followed by the Tor browser for layered anonymity.
   - Logging mechanisms document researcher interactions while preventing storage of harmful content.
2. **Crawling Protocols**
   - Crawlers are configured to search for keywords such as *"deepfake," "AI video," "synthetic media,"* and *"face swap."*
   - Rate limiting and randomized delays are applied to prevent detection and reduce ethical concerns related to overloading services.
   - Target domains include onion-based forums and marketplaces where synthetic media tools are advertised.
3. **Data Storage**
   - Captured data (primarily text and metadata) is encrypted at rest and transmitted via secure channels.
   - Metadata, such as post timestamps, categories, and anonymized author IDs, is preserved.
   - No raw images or videos are retained; only derived metadata is stored to avoid illegal content handling.
4. **Preprocessing and Analysis Tools**
   - **Textual data:** Preprocessed using **NLTK** for tokenization and stop-word removal, and **SpaCy** for named entity recognition and dependency parsing.
   - **Image/Video references:** Metadata extracted using **OpenCV** (e.g., resolution tags, file formats) without media retrieval.

- **Anonymization:** Hashing of identifiers ensures privacy and compliance with ethical research norms.

5. **Data Handling Protocols**
   - Strict prohibition of downloading or storing explicit or harmful deepfake content.
   - Automatic filtering scripts are applied to detect and discard suspicious payloads.
   - Safe disposal procedures, including secure file deletion, are implemented for any temporary data fragments.

2.3 Flow Diagram

The process is easily imagined as a step-by-step, but still, sequential flow that retains the possibility of going back to a previous stage:

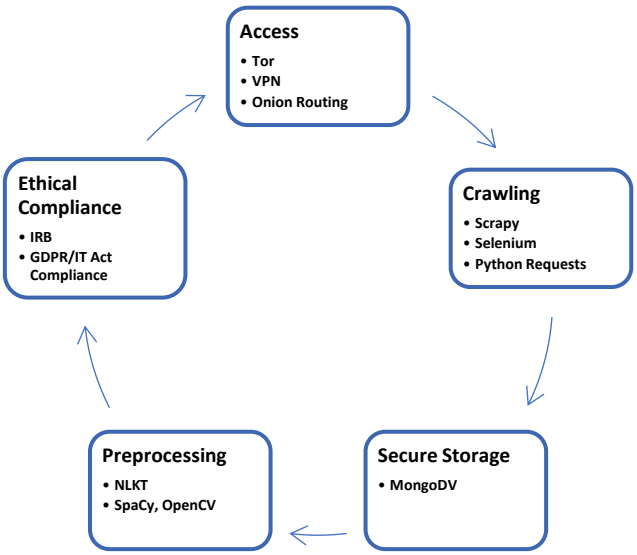

Figure 1 the flow Diagram of the new Ethical Dark Web Data Collection Framework for Deepfake Threat Analysis.

3. Results and Discussion

3.1 Findings

To illustrate the proposed methodology, we conducted a **case study on metadata crawling** of dark web forums referencing deepfake technologies. Using keyword-driven crawlers (*"deepfake," "AI video," "face swap,"* and *"synthetic media"*), we collected anonymized text snippets and post metadata without downloading any illicit media.

- **Sample Dataset:** The dataset included post timestamps, anonymized author identifiers, and text fragments extracted from discussion threads. No raw media was stored.
- **Visual Analysis:**
  - A **word cloud** generated from frequent terms highlighted keywords such as *"deepfake software," "tutorials,"* and *"AI face swap."*
  - A **timeline graph** demonstrated fluctuations in the frequency of deepfake-related discussions, with spikes corresponding to major news events involving synthetic media.
- **Thematic Insights:** Content analysis revealed three recurring themes:

1. Advertisements for deepfake software packages.
2. Service offerings where actors sell customized synthetic media.
3. Tutorials and guides on producing or distributing manipulated videos.

This case illustration shows that metadata alone can reveal meaningful insights into the evolving underground ecosystem of deepfakes.

3.2 Interpretation

The case study exemplifies the effectiveness of metadata-driven approaches:

• Without the need to directly access illicit multimedia files, researchers can still discover the typical behaviour of threats and the sudden increase of activities of that community in the underground.

• The gathering of metadata is the alternative that is considered safe compared to the direct handling of illegal deepfake content, in this case, the legal and ethical risks are minimized.

• The fact that the architecture is versatile hints at the possibility of the continuous observation of the situation where metadata flows can be followed and thus the monitoring of the spread of synthetic media tools may be facilitated.

Therefore, the methodology effectively combines the analyst's intellectual potential with the legal, ethical, and research productivity-matching protective measures.

3.3 Comparison with Previous Work.

Most of the studies that try to detect deepfakes have quite a heavy dependency on the public datasets of the surface web, like FaceForensics++ (*Rossler et al., 2019*), and Celeb-DF (*Li et al., 2020*). These datasets represent good benchmarks but have a limitation in that they do not reveal the dark web dynamics of the distribution of deepfakes. There are a few research works that have gone deep into the dark web for mining purposes, but many of them are missing the embedded safety provisions that can protect the researchers from possible risks (**Scanlon, 2016; Laliberte et al., 2022**).

The following highlights are features that differentiate the proposed framework from all alternate ones:

1. Reproducibility of a Technical Nature: Methodologically clear use of Tor, Scrapy, Selenium, MongoDB, and NLP tools.

2. Regulation Observe: Inclusion of safety measures that are congruent with IRB, GDPR, and IT Act guidelines.

3. Multi-modal Metadata Capability: The feature that enables the framework to gather and process textual, image as well as video metadata without the need to expose the user to the harmful content of the media.

Within the context of deepfake research, the framework stands out as an ethical, methodological, and safety partner that addresses the concerns raised by previous studies.

4. Conclusion

4.1 Summary

This paper presented a **methodology-focused framework** for ethically collecting and analyzing dark web data relevant to deepfake threats. Unlike traditional detection-centered research, the contribution here lies in proposing a structured, reproducible pipeline that integrates **technical feasibility with robust ethical safeguards**. The five-layered framework—spanning secure access, crawling, encrypted storage, preprocessing, and ethical guardrails—offers a comprehensive approach for studying underground ecosystems without engaging in harmful or illegal practices. Its effectiveness was illustrated through a case study on **metadata crawling**, which demonstrated how anonymized text and temporal data can provide meaningful insights into the circulation of deepfake tools and services.

4.2 Implications and Future Work

The proposed methodology has significant implications for both academic research and practice. It provides a **safe template for researchers and law enforcement agencies** to monitor deepfake-related activity in hidden networks without handling illicit media. By focusing on metadata and anonymized content, the framework also opens the possibility of creating **benchmark datasets** that preserve analytical value while remaining compliant with ethical and legal requirements.

Future directions include extending this approach to **real-time monitoring systems** capable of detecting emerging trends across dark web platforms, integrating **AI-powered analytics** to automate theme detection, and fostering **collaboration with cybersecurity authorities** for proactive threat intelligence. By embedding methodological rigor with ethical responsibility, this work contributes a sustainable pathway for advancing deepfake research while safeguarding both researchers and society.

References

1. Bradbury D. Unveiling the dark web. *Network Security*. 2014;2014(4):14-17. doi:https://doi.org/10.1016/s1353-4858(14)70042-x
2. Lusthaus J, Kleemans E, Leukfeldt R, Levi M, Holt T. Cybercriminal networks in the UK and Beyond: Network structure, criminal cooperation and external interactions. *Trends in Organized Crime*. Published online February 3, 2023. doi:https://doi.org/10.1007/s12117-022-09476-9
3. Raman R, Kumar Nair V, Nedungadi P, Ray I, Achuthan K. Darkweb research: Past, present, and future trends and mapping to sustainable development goals. *Heliyon*. 2023;9(11). doi:https://doi.org/10.1016/j.heliyon.2023.e22269
4. Bugajewska M. A Survey of Challenges in Dark Web Crawling: Technical, Security, and Ethical Perspective. *Communications in Computer and Information Science*. Published online 2025:349-355. doi:https://doi.org/10.1007/978-3-031-79086-7_27
5. Brilingaitė A, Bukauskas L, Juozapavičius A, Kutka E. Overcoming information-sharing challenges in cyber defence exercises. *Journal of Cybersecurity*. 2022;8(1). doi:https://doi.org/10.1093/cybsec/tyac001
6. Zhang K, Luc Van Gool, Radu Timofte. Deep Unfolding Network for Image Super-Resolution. *Journal of Cybersecurity*. Published online June 1, 2020. doi:https://doi.org/10.1109/cvpr42600.2020.00328
7. Li Y, Yang X, Sun P, Qi H, Lyu S. Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. *arxivorg*. Published online September 27, 2019. https://arxiv.org/abs/1909.12962
8. Álvarez González, Mailyn Moreno Espino, Moreno C, Yahima Hadfeg Fernández, Nayma Cepero Pérez. Ethics in Artificial Intelligence: an Approach to Cybersecurity. *Inteligencia artificial*. 2024;27(73):38-54. doi:https://doi.org/10.4114/intartif.vol27iss73pp38-54
9. Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M. FaceForensics++: Learning to Detect Manipulated Facial Images. *arXiv:190108971 [cs]*. Published online August 26, 2019. https://arxiv.org/abs/1901.08971
10. Leimich P, Harrison J, Buchanan WJ. A RAM triage methodology for Hadoop HDFS forensics. *Digital Investigation*. 2016;18:96-109.

doi:https://doi.org/10.1016/j.diin.2016.07.003
11. Pascale DD. CRATOR: a Dark Web Crawler. Arxiv.org. Published 2017. https://arxiv.org/html/2405.06356v1
12. European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Europa.eu. Published April 27, 2016. https://eur-lex.europa.eu/eli/reg/2016/679/oj