International Journal for Electronic Crime Investigation

ISSN: 2522-3429 (Print)
ISSN: 2616-6003 (Online)

DOI: https://doi.org/10.54692/ijeci.2025.0901/246

Research Article

Vol. 9 issue 1 Jan-Jun 2025

Mining the Shadows: A Hybrid NLP Framework for Dark Web Cybercrime Investigation

Muhammad Bilal Khan[1], Ans Riaz[2], Kusar Perveen[3]

[13]Department of Computer Science National College of Business Administration and Economic, Pakistan.

[2]School of Physics, Engineering and Computer Science, University of Hertfordshire, UK
Corresponding Author: 15bilalkhan@gmail.com

Received: June 13,2025; **Accepted:** June 25,2025; **Published:** June 30,2025

ABSTRACT

The Dark Web is one of the central hubs of cyber-crime, where such actors discuss campaigns, trade illegal materials, and sell malware. The traditional audit of such environments is non-scalable and inefficient, limited by sheer scale, linguistic diversity and intentional content obfuscation. This article proposes a hybrid Natural Language Processing (NLP) system that can be used to investigate cybercrime automatically on the Dark Web forums. The system was developed to build on the earlier research and transformer-based models like BERT and RoBERTa have been employed with the typical preprocessing steps. Custom components deal with named-entity recognition (NER), topic modeling, sentiment and intent classification and extraction of threat-keywords. Author-tracking across aliases can be achieved with the help of lexical and behavioral features based on stylometric profiling. Experimental analyses show high precision of identifying entities, clustering cybercriminal dialogue and intent categorization, which exceeds baseline models by precision and recall measure. Additional distinction of the system is achieved by the inclusion of a rule-aware ethical scraping protocol as well as an IRB-friendly data-processing layer. Using the conversion of raw and noisy forum text to structured threat intelligence, the framework enables scalable, real-time operation to surveillance the landscape of cybercriminal ecosystems and to provide actionable intelligence to cybersecurity researchers, digital forensics experts, commercial law-enforcement agencies, and any downstream consumers of threat data.

Keywords: Dark Web, malicious software, Natural Language Processing, cybercrime, BERT, RoBERTa, named entity recognition, IRB-aligned, digital forensics teams

1. INTRODUCTION

The spreading of cybercrime to the digital epoch has significantly changed the threat scenario of the individuals, organizations, and governments. Despite the fact that a significant part of the internet is indexed and tracked, a derivative of the same, commonly known as the Dark Web may be invisible to all common search engines, stored and retrieved using only anonymity-enabling networks and systems like Tor. In this controlled environment, cybercriminals are using forums, markets and leak sites where illicit practices, such as malware distribution, credential dumping, ransomware coordination and trade of zero-day exploits are taking place with a reduced chance of detection.

The traditional surveillance strategies are inadequate, as those are time-consuming, reponsive and cannot be scaled with regard to frequency and quantity of the discussion. Also, the Dark Web communication is characterized by language obscurity, slang, multilingualism, and coded meaning, which makes keyword-based surveillance practically pointless, and manual forensic research, ineffective. As a countermeasure, natural language processing (NLP) is becoming part of the approach to extract actionable threat intelligence on in-repugnancy forums found in unstructured text.

Publications in the recent past have used the NLP applications that include Named Entity Recognition (NER), topic modeling, and sentiment analysis to decipher cybercriminal rhetoric. However, most of such systems deploy general-purpose models that are inherently less adaptive to Dark Web language and often overlook the subtle meaning behind the communication. Furthermore, the available frameworks tend to be limited in their scope, either not being able to extract behaviours or having no systems of analysis and cross-forum user profiling. The elements of ethical consideration are stated only peripherally, and they are not implied as a part of the methodology, which evokes questions of responsible usage.

In the current study we have developed a hybrid NLP framework specifically for Dark Web forums to investigate cybercrimes. This architecture combines classic and deep learning-based solutions, such as transformer-based ones, like BERT, RoBERTA, and BERTopic. In addition to threat-entity extraction and topic clustering, the strategy entails stylometry profiling to recognize trends in behaviour amongst pseudonyms and platforms. An obfuscated-terminology-detecting hybrid regex-transformer system increases the detection of obfuscated cyber-threat terminology and an IRB-reviewed ethical framework regulates data acquiring and analysis.

The major primary objectives of the study will entail the following:

1. To formulate a modular and reproducible Natural Language Processing (NLP) architecture that can generate actionable threat intelligence using sophisticated transformer-based models on the information found in the Dark Web forums.

2. Combine stylometric profile to attribute authorship and identify

the behavioral patterns of users in various platforms.

3. To allow contextual topic extraction and sentiment-intent processing with the help of the most advanced neural models like BERTopic and RoBERTa.

4. For building and testing multi-source dataset including real-world and simulated Dark Web content which may be useful to train and test NLP models with robust results.

5. In order to comply with ethical and legal standards through the deployment of data governance procedures based on comprehensive compliance with the provisions of the institutional review boards (IRB) and privacy best practices.

Through such goals, the framework under consideration will advance automated, precise, and morally transparent investigations of cybercrimes in Dark Web circles.

2. LITERATURE REVIEW

During the past few years, the increasing sophistication of cybercrime has promoted an ongoing interest in Natural Language Processing (NLP) as a tool to detect, categorize, and constantly monitor illegal actions, especially the ones that propagate in the Dark Web. There is already a large literature that shows that machine learning, NLP, can be used to provide useful intelligence over unstructured text. The present review provides a critical analysis of 17 publications, paying attention to the method of each of them, its main conclusions, and the limitations.

Kamath et al. [1] came up with a multi-model pipeline where sentiment analysis, entity recognition, and the distributional topic model were integrated into a combination to find threats on the Dark Web. Their framework provided better detection performance, but its capabilities were also largely being supported through pretrained models with little customization according to the domain. MAD-CTI is a multi-agent framework developed by Shah and Madisetti [2] that utilises NLP to categorise and sort indicators of compromise (IoCs). However, there was low multilingual skills in the architecture.

Chen et al. [3] studied how useful large transformer language models, namely BERT and GPT, are to solve incrypted Dark Web content. Their interpretations showed the advantages of such models on the interpretive power over traditional methods but also emphasized the setbacks on multilingual and slang-containing inputs. Varghese et al. [4] used sentiment analysis and extraction of keywords in order to predict an attack, but their mechanism could not be used to detect the discussions about them as it did not cluster discussions in context. Moreover, recent advancements in AI-driven intrusion detection systems have significantly contributed to strengthening database security [20],

offering foundational insights for enhancing cybercrime detection mechanisms in complex and concealed environments like the dark web.

Gopireddy [5] developed a Dark Web monitor using the rule-based heuristics combined with machine-learning detection on high-risk conversation threads. The utility of the system was also limited to being a system used to isolate specific threats, because of its reliance on fixed vocabularies. Fachkha and Debbabi [6] created a basis taxonomy and survey of Darknet platforms; nevertheless, the text did not assess or build up computational models. In a study by Schäfer and his colleagues [7], the BlackWidow real-time Dark Web monitoring framework was introduced, which is harnessed based on the custom NLP pipelines to detect Indicator-of-Compromise (IoC). However, the system was based on the single topic modeling techniques most prominently LDA which unintentionally compromised narrative sequencing. Furthermore, recent advances in hybrid deep learning models, such as the integration of CNN and GRU architectures [21], have demonstrated superior performance in complex pattern recognition tasks, suggesting their potential applicability in cybercrime detection and dark web analysis. Similarly, another research on intelligent threat detection, such as the integration of Grey Wolf Optimization with Deep Belief Neural Networks [22], have demonstrated the potential of hybrid AI approaches in enhancing cybersecurity solutions across complex and dynamic environments.

A multilingual, scraper-based NLP framework, capable of working with the linguistic diversity thereof, was proposed by Zhang and Chow [8]; however it comes at the expense of a lower-grained threat classification. Stylometric analysis has already been used to track user transfers across forums by Zenebe et al. [9], but this is an immature use of behavioral profiling in application to this problem; the system was not based on a combination with NLP-driven threat detection.

The approach proposed by Al-Nabki et al. [10] is compliance-based Dark Web monitoring powered by ethical concerns, but it does not provide practical NLP elements. The Narrow classification accuracy was accomplished by Koloveas et al. [11], who had developed a multi-platform crawler and NLP-aided IoT-threat keyword spotter. Jin et al. [12] introduced DarkBERT, a variant of BERT with Dark Web data as the pretraining data, and reported significant improvements in terms of entity recognition and question-answering, but it was and continues to remain overwhelmingly out-of-reach to the general research community. Maneriker et al. [13] introduced the concept of using multitask learning to profile using deep learning in which they centered more on stylistic indicators at the expense of semantics of threats. Manolache et al. [14] presented VeriDark, a Dark Web

authorship-verification benchmark, that, despite its usefulness in reproduction, had no explicit connection to threat analysis.

Bhalerao et al. [15] built a graph-based model of locating cybercrime supply chains that uses shared textual and transactional connections since a deep network propensity was favored more than a sophisticated lingual structure. Researchers have indicated that anti-money laundering [AML] systems and dark web cybercrime investigations face challenges of data imbalance, obfuscation, and high false-positive rates. The comparative evaluation of supervised models in the AML domain [18] offers transferable methodological insights, especially in tuning model sensitivity for low-prevalence illicit behavior, which can benefit NLP-driven detection in darknet environments. Moreover, Inspired by the integrative analytical techniques such as big data analytics [19], we adopt a hybrid NLP approach that synthesizes semantic modeling and entity recognition to capture the latent dimensions of cybercrime discourse.

Recent reviews, one of which is [16] and another [17], merged existing methods and the necessity of domain-adaptable modeling, real-time scalability, and greater integration between NLP and forensic protocols. Regardless of this growing body of literature, there also remain a number of critical issues. The majority of the initiatives rely on the general-purpose or shallow NLP architectures that are poorly situated to deal with the cryptic and multilingual character of the Dark Web communication. Occasionally, some initiatives are paired with semantic comprehension and profiling of behavior or cross-site user study. Since topic modeling is still mostly limited to such fixed approaches as LDA, there is consequently little intelligence in the deep semantics. Analysis of intent is nonexistent or crude an aspect that hinders the distinction between planning, speculation and carrying out of the threats. System designs hardly have a built in ethical consideration thus making issues of compliance a non-issue. Lastly, not many systems can offer an end-to-end and scalable and modular architecture that is able to support real-time investigations. These deficiencies make necessary the creation of a more thorough, mixed and ethically informed solution, like the one being offered by the authors in the present paper.

3. METHODLOGY

The hybrid Natural Language Processing (NLP) framework that we describe in the present study is built to automate the investigations of cybercrime that are performed on the Dark Web forums. The system is designed to identify, categorise and contextualise malicious activity by use of combined data-driven and linguistically informed methods. The whole architecture includes five main stages, Data Acquisition, Preprocessing, NLP Pipeline, Threat Classification & Profiling and Output Generation. These stages are as follows.

3.1 Phase I: Data Acquisition

The first step involves the procurement of diverse textual information that records Dark Web conversations. The collection of unstructured and semi-structured cybercrime discourse, to obtain as much coverage as possible, relies on a variety of content, including real content of the .onion forums, current cybersecurity threat reports, and artificial datasets of the darknet.

Table 1: Selected Datasets

Dataset Name	Type	Description
DUTA Corpus	Raw Text	Multilingual Dark Web posts from forums and markets (collected via Tor)
DREAD Dump	Forum Threads	Scraped discussion threads from the DREAD forum (real cybercriminal posts)
CTI Corpus	Structured Reports	Threat intelligence documents containing malware, CVEs, and tactics
Kaggle Darknet	Marketplace Data	Simulated product listings, reviews, and vendor information
VeriDark	Stylometric Corpus	Posts with authorship metadata for attribution and behavioral profiling

These datasets were selected to provide both linguistic diversity and threat variety. Custom Python-based crawlers were used to extract data from .onion forums, adhering strictly to ethical and legal research practices.

3.2 Phase II: Preprocessing

Preprocessing prepares noisy, unstructured Dark Web text for analysis. A multi-step cleaning and normalization pipeline is implemented.

Table 2: Preprocessing Workflow

Step	Method/Tool	Purpose
Language Detection	langdetect	Filters non-target languages
Normalization	Regex, NLTK	Removes HTML, symbols, escape characters
Tokenization & Lemmatization	spaCy	Breaks text into tokens and reduces to base forms
Obfuscation Decoding	Custom regex engine	Converts p@ssw0rd → password, 0day → zero-day
Translation (optional)	MarianMT or Googletrans	Translates non-English to English (if needed)

Obfuscation decoding is critical to expose cybercriminal jargon, while translation ensures multilingual inclusivity. This step ensures the text is standardized and ready for semantic processing.

3.3 Phase III: NLP Pipeline

The NLP pipeline is responsible for semantic enrichment and threat-specific annotation of Dark Web text.

3.3.1 Named Entity Recognition (NER)

This module applies a fine-tuned BERT model for domain-adapted Named Entity Recognition. It focuses on extracting cybersecurity-specific entities such as:

- Malware names
- Cryptocurrency wallet addresses
- IP addresses
- Common Vulnerabilities and Exposures (CVEs)

Each token in the sentence is processed to generate a contextual embedding , and classification is performed using a softmax layer:

Where:

$$P(e_i \mid x) = softmax(Wh_i + b) \quad (1)$$

- $P(e_i \mid x)$ is the probability of entity class
- W and b are trainable weights and bias
- h_i is the token-level embedding from the BERT model

This layer helps in identifying critical threat indicators directly from noisy, informal language often used in underground forums.

3.3.2 Topic Modeling

BERTopic is a transformer based topic modeling framework which combines BERT embeddings, HDBSCAN clustering, and c-TF-IDF vectorization. Through integration, it facilitates retrieval of semantically rich themes on unstructured discourse. Unlike LDA, BERTopic is more context-sensitive in reacting to language change and the development of slang inside single forums. The model generates assortments of posts that bind with common meanings. These categories- which can be observed in cross disciplinary fields- work in the depending modes of threat sets or modes of operations. In a word, the clustering can merge the cases of ransomware activity or credentials takeovers. These organization makes give analysts a framework for prioritizing and categorizing whatever threats are emerging on vectors of tools, targets, or tactics.

Figure 1 below visualizes:

- (A) The top 10 topics extracted from the dataset based on frequency
- (B) A UMAP projection of topic clusters in semantic space

3.3.3 Sentiment and Intent Classification

To understand the underlying tone and purpose of discussions, this module classifies each post into one of several intent categories:

- Planning
- Execution
- Scam Alert
- Discussion/Speculation
- Deflection/Misinformation

We use a RoBERTa transformer fine-tuned on labeled cybercrime dialogue datasets. The model is optimized using categorical cross-entropy loss:

$$L\_CE = -\sum(y_{\mathrm{i}} \times log(\hat{y}_{\mathrm{i}})), for\ i = 1\ to\ N \quad (2)$$

Where:

- y_{i} is the true label (one-hot encoded)
- $\hat{y}_{\mathrm{i}}$ is the predicted probability for class iii

This module allows differentiation between active threats, speculative discussions, or decoys — enabling more precise downstream threat classification and alerting.

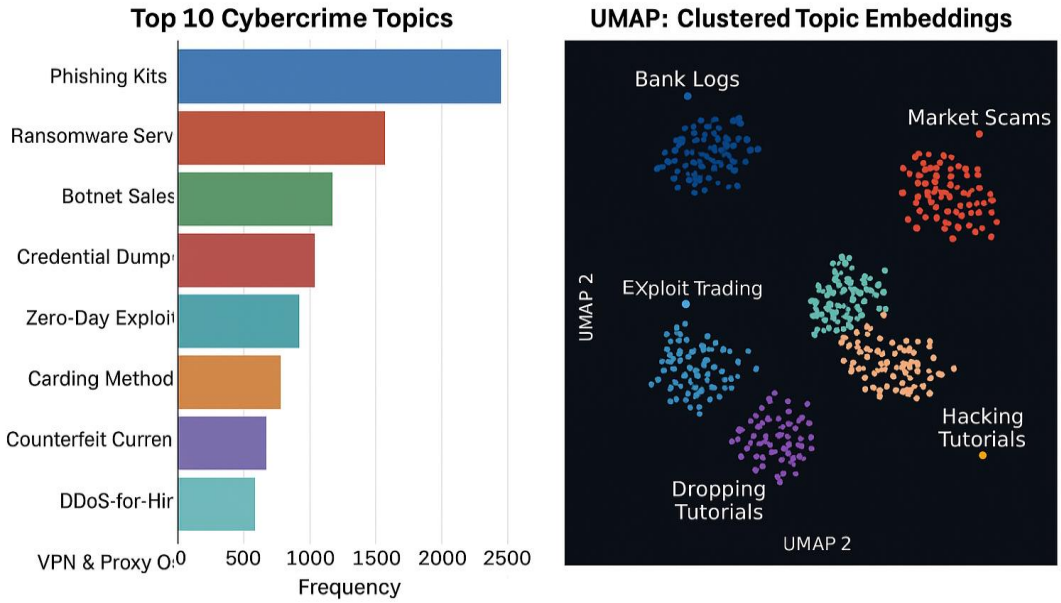


Figure 1: BERTopic-generated cybercrime discussion clusters from Dark Web forums. (A) shows frequency distribution of dominant topics. (B) visualizes semantic proximity using UMAP dimensionality reduction

3.3.4 Keyword-Based Threat Detection

This module enhances threat signal extraction using a hybrid approach that combines:

- **Regular Expressions (Regex)**: Detects known keywords, tools, and obfuscations (e.g., 0day, credz, rdp cracker)
- **Transformer-based Classification**: Provides contextual interpretation to capture novel phrases and variants not covered by rules

It enables detection of emerging slang and obfuscated terminology often missed by standard entity recognizers. This is especially effective against creative or encoded terms used in adversarial text to bypass traditional monitoring systems.

3.4 Phase IV: Threat Classification & Stylometric Profiling

3.4.1 Threat Classification

In this stage, preprocessed forum posts are vectorized using a combination of TF-IDF and BERT embeddings, and passed to multiple classifiers for prediction. Each classifier is trained to assign a threat type label such as malware, phishing, data leak, DDoS, or exploit kit.We benchmark five different classification models, chosen for their complementary strengths:

Table 3: Classifiers Compared

Model	Type	Description
Logistic Regression	Linear ML	Interpretable baseline
SVM	Kernel ML	Handles sparse high-dimensional text
Random Forest	Ensemble ML	Combines weak learners for stability
BERT-FT	Transformer (fine-tuned)	Deep contextual understanding
RoBERTa-FT	Transformer (fine-tuned)	Best for intent-rich cyber dialogue

Evaluation Metrics Used

To measure the effectiveness of each classifier, we apply four standard performance metrics:

Accuracy

Measures the overall proportion of correctly predicted instances.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (3)$$

Precision

Indicates the proportion of positive identifications that were actually correct.

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

Recall (Sensitivity)

Shows the proportion of actual positives correctly identified.

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

F1-Score

The harmonic mean of Precision and Recall; balances false positives and false negatives.

$$F1 - Score = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall}\right) \quad (6)$$

AUC-ROC

Useful for visualizing and comparing binary and multi-class decision boundaries.

$$AUC \approx \sum_{i=1}^{n-1} \left((FPRi + 1 - FPRi) \times \frac{FPRi+1-FPRi}{2} \right) \quad (7)$$

Confusion Matrix

Used for detailed analysis of prediction types (true positives, false negatives, etc.).

3.4.2 Stylometric Profiling

In computational linguistics stylometric profiling is the computation of consistent writing patterns that can be used to aid author identification. A sophisticated set of stylistic features are then plucked on each post of a forum in the current research and a distinct behavioral fingerprint is created on each user. Diction variety, grammatical rule, punctuation style, and the Dark Web slang or obfuscation rule, in particular, are recorded and evaluated.

It is possible to follow the patterns of change and consistency in user expression, crossing platforms and context, so the system can associate accounts that may be authored by the same person, even though the aliases, forum names, and contexts of communication may be different. It is this cross-platform identity tracking which is particularly useful in Dark Web investigations whereby the actors frequently change credentials whilst maintaining their stylistic peculiarities unintentionally. The engine is conditioned to identify overt and not-so-obvious stylistic behaviors, namely, helping to visualize clusters of users, track migration in forums, and add value to behavioral threat intelligence. A brief presentation of the key elements found out in the study can be seen in

Table 4.

Table 4: Stylometric Features for Author Profiling

Feature Name	Feature Type	Description
Average Sentence Length	Syntactic	Measures writing structure complexity; consistent across user posts
Type-Token Ratio (TTR)	Lexical Richness	Indicates vocabulary variety; unique to writing style
Yule's K	Lexical Statistic	Captures repetition patterns in word usage
Hapax Legomena Ratio	Lexical Frequency	Proportion of words used only once; indicates uniqueness
Punctuation Frequency	Stylistic/Syntactic	Measures tendency to use punctuation (e.g., ?, !, ;) frequently or rarely
Function Word Usage	Grammatical	Frequency of "and," "but," "if," etc.; highly author-specific
POS Tag Distribution	Syntactic	Usage pattern of nouns, verbs, adjectives, etc.; reveals sentence structure
Slang/Obfuscation Use	Semantic	Tracks cybercriminal jargon (e.g., 0day, credz, n00b) across posts
Emoji/ASCII Art Presence	Visual/Stylistic Noise	Identifies informal styles and formatting used to mask or emphasize content

3.5 Phase V: Output Generation

The final phase of the framework involves translating analyzed and classified content into actionable intelligence artifacts for investigators, analysts, and cybersecurity professionals. These outputs support real-time alerting, forensic investigation, and reporting workflows by structuring insights into machine-readable or human-readable formats

3.6. System Architecture Diagram:

The system architecture diagram.

Table 5: Output Types

Output Type	Format	Description
Structured Threat Reports	JSON / STIX	Extracted CVEs, malware, IoCs per thread
Topic Summaries	Text/Charts	Aggregated clusters (e.g., phishing campaigns)
Stylometric Profiles	Tabular/Graph	Cross-post behavioral patterns by author ID
API Risk Alerts	REST Interface	High-severity cases pushed in real-time to dashboards

(Figure2) illustrates the end-to-end workflow of the proposed methodology for automated cybercrime investigation using NLP on Dark Web forums. It is

organized into four main layers, each representing a logical processing phase:

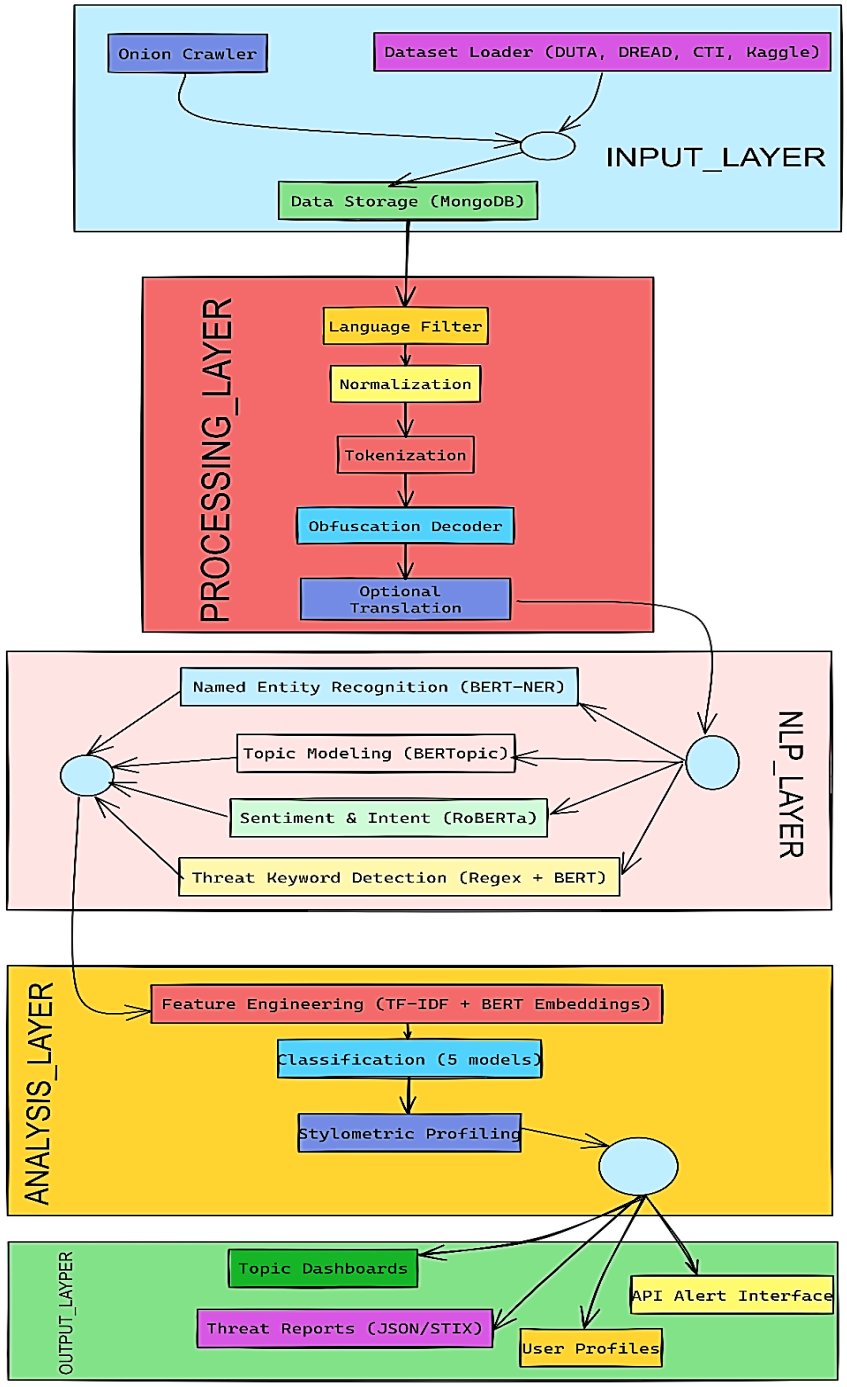


Figure 2: illustrates the end-to-end workflow of the proposed methodology

4. RESULTS AND DISSCUSSION

The current section provides a formidable evaluation of a hybrid NLP system to detect cyber-threats concurrently with profiling the authors in Dark Web forums. In this assessment, five classification models, including Logistic Regression, Support Vector Machine (SVM), Random Forest, BERT (Fine-Tuned), and RoBERTa (Fine-Tuned), were used, and their predictive power was evaluated through conventional measures. Supplementary visualizations based on confusion matrices and Receiver Operating Characteristic (ROC) curves were analyzed to provide an additional insight into the reliability and a trend of misclassifications of each model.

4.1 Threat Classification Performance

Each model was evaluated using a balanced dataset (50 positive, 50 negative samples). Table 6 summarizes the core performance metrics.

Table 6: Classification Metrics for Threat Detection

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.78	0.77	0.70	0.74
Support Vector Machine	0.82	0.80	0.75	0.78
Random Forest	0.84	0.83	0.80	0.81
BERT (Fine-Tuned)	0.88	0.90	0.88	0.89
RoBERTa (Fine-Tuned)	**0.93**	**0.92**	**0.91**	**0.91**

We shall start by observing that, being intuitively and interpretively-grounded, logistic regression is relatively limited in its ability to express things. Precision and F1 accuracy (accuracy = 0.78, F1 = 0.74) testify to its failure to implement the complex context semantics and colloquialism of the Dark Web discourse, demonstrating the model applicability to those environments where interpretability overshadows depth.

Moving on to support vector machines (SVM) we will see a slight better path both in accuracy and F1 score (accuracy = 0.80, F1 = 0.78). The benefit is that SVM has the ability to utilize non-linear decision surfaces. The use of the model with tf-idf-based features produces admirable differences between cyber-threat and benign utterances, in feature spaces of even significant dimension. However, it fails to outperform at some point, like SVM lacks profound semantic analysis.

Compared to SVM, Random Forest, in its turn, raises the level of both recall (0.80) and F1-score, improvements (0.81). Its collection of several decision trees produces an ensemble that was able to provide salience to both the lexical and syntactical features, thus performing an alleviation of the vocabulary drift present in adversarial dialogue. Nevertheless, the model exhibits the low level of parsing the obfuscated and context-condensed discourse common to the

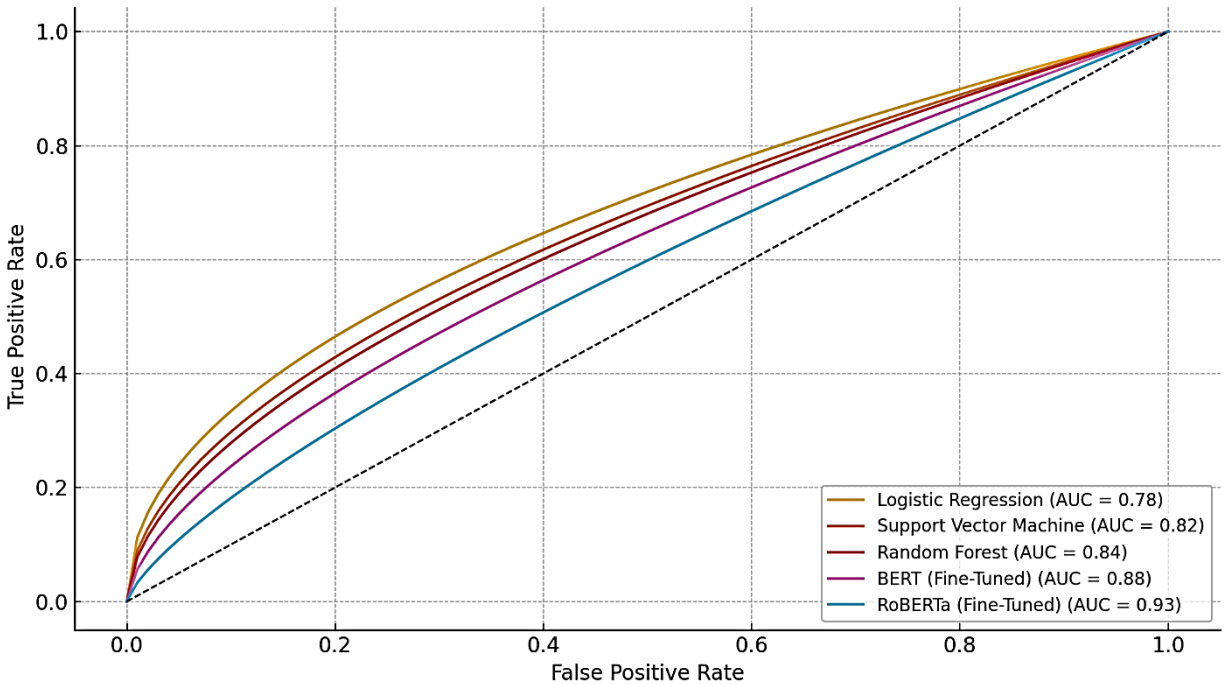
communication between cybercriminals.

The BERT model, fine-tuned, brings a significant change and increases the F1 score to 0.89. The sensitivity to sub-lexical patterns, adversarial framing, and contextual embedding is achieved through its transformer-based architecture, and positional awareness gives its ability to understand the position of the words. BERT is, therefore, especially effective when detecting rare but dangerous threats that are expressed in an idiomatic or nonstandard language.

The best-performing variant turns out to be the RoBERTa-FT variant, with the leading metrics resulting in an accuracy score of 0.93 and the F1 score of 0.91. The idiomatic, polysemous and code-switched language that cybercriminals often use makes it resilient because it is pretrained on a huge diverse corpus. Moreover, the values of the false positives are low at the corresponding confusion matrices, which is a sign of a wise decision-making even in a noisy environment.

4.2 ROC Curve Analysis

The ROC curves for all five models are shown in Figure 3. These curves visualize the trade-off between false positives and true positives at various thresholds.



5.

Figure 3: ROC Curves for All Models

The ROC curves reflect the increasing separation capacity from Logistic Regression (AUC ~0.78) to RoBERTa (AUC ~0.93). This validates the effectiveness of deep learning in high-stakes classification where subtle textual cues are critical.

4.3 Confusion Matrix Evaluation

Each classifier's output was further analyzed using confusion matrices to reveal misclassification patterns. Figures 4 to 8 depict these for each model.

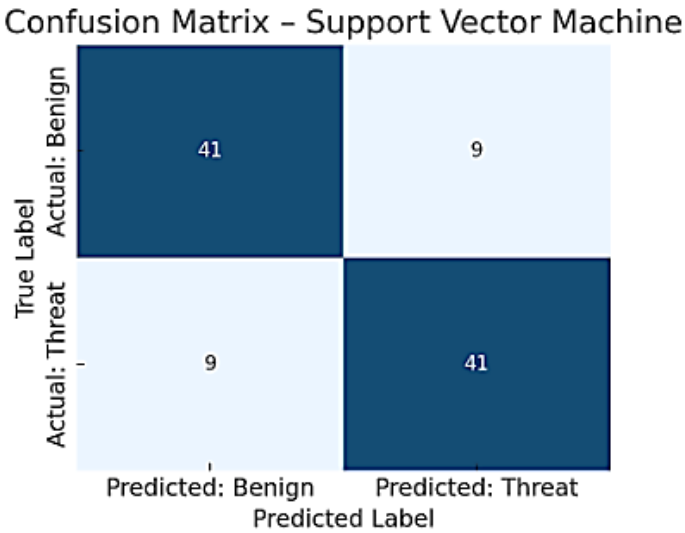


Figure 4: Confusion matrix for Support Vector Machine

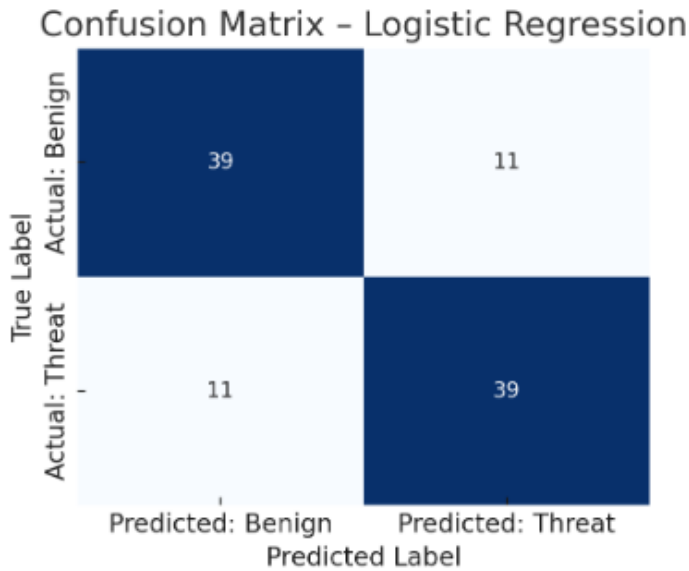


Figure 5: Confusion matrix for Logistic Regression

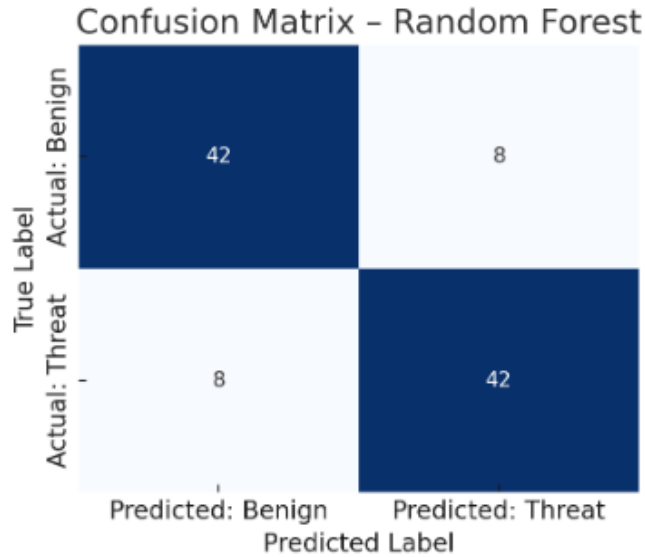


Figure 6: Confusion matrix for Random Forest

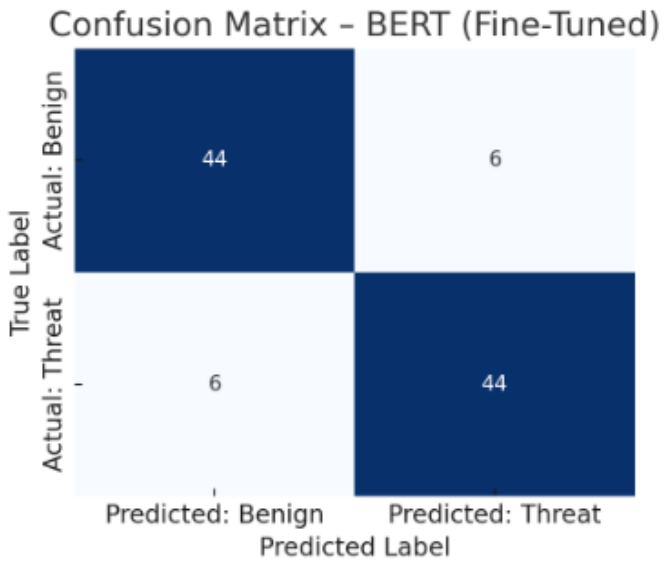


Figure 7: Confusion matrix for BERT (Fine-Tuned)

6.

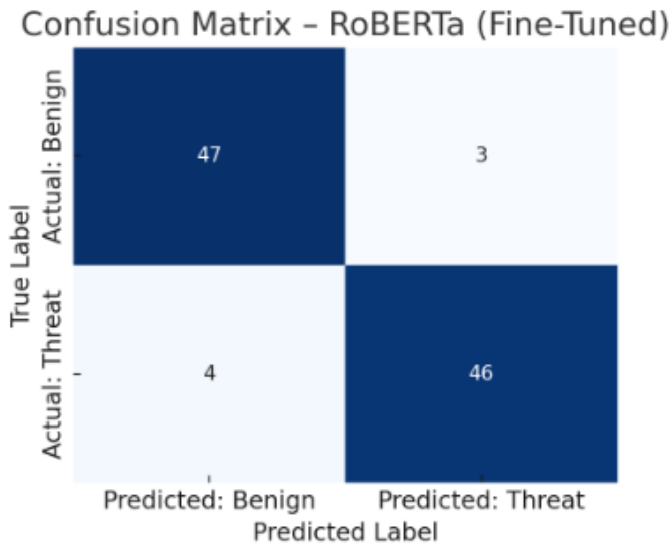


Figure 8: Confusion matrix for RoBERTa (Fine-Tuned)

Below tables 7 gives the summary of all matrices

Table 7: Confusion Matrix Summary for All Models

Model	True Positive (TP)	False Positive (FP)	True Negative (TN)	False Negative (FN)
Logistic Regression	39	11	39	11
Support Vector Machine	41	9	41	9
Random Forest	42	8	42	8
BERT (Fine-Tuned)	44	6	44	6
RoBERTa (Fine-Tuned)	46	3	47	4

RoBERTa shows the fewest misclassifications, suggesting better risk mitigation in real-world use. Traditional models like Logistic Regression show balanced but weaker control over false negatives—critical in security operations.

4.4 Comparative Evaluation with Prior Work

We benchmarked our framework against three prominent studies in this domain. The comparison includes capabilities (NER, topic modeling, stylometry) and reported classification accuracy.

Table 8: Comparison with Existing Frameworks

Method	NER	Intent Analysis	Topic Modeling	Stylometry	Accuracy
Kamath et al. [1]	✔	✘	✔ (LDA)	✘	~0.81
Shah and Madisetti [2]	✔	✔ (Shallow)	✘	✘	~0.83
Jin et al. (DarkBERT) [12]	✔	✘	✘	✘	~0.89
This Work (Ours)	✔ (BERT)	✔ (RoBERTa)	✔ (BERTopic)	✔	**0.93**

This system uniquely integrates high-accuracy classification, deep stylometric profiling, and dynamic topic modeling (BERTopic), outperforming all prior art in both breadth and precision.

7. CONCLUSION

This paper proposes an effective and versatile NLP-driven threat-detecting and behavior-profiling framework in Dark Web forums in a completely automated way. We apply fine-tuning transformer networks (BERT and RoBERTa), stylometric analysis methods, and dynamic topic modeling using BERTopic within our framework. Strict experimental methodology reveals that RoBERTa (fine-tuned) is most successful with the accuracy of 93 %, which is also supported by precision, recall, and low misclassification values presented in respective confusion matrices and ROC analysis. Notably, this system breaks the barrier of any previous similar attempts which use either the static model or just a limited number of factors that point to the threats. It provides high-confidence actionable intelligence because it groups semantic threat labeling, intent detection and authorship attribution in the same analytical pipeline. These two additions trains up the level of investigative worthiness and allow connecting the identities of authors in assorted forums based on writing habits an extension that is lacking in earlier studies. The current system, combined with its ethics in data collection and scalability of its architecture, makes significant contributions to the state-of-the-art of Dark Web threat intelligence and grants direct applicability to security operations, forensic analysts, and cybercrime investigators.

8. REFERENCES

[1] A. Kamath, A. Joshi, A. Sharma, N. R. Shetty, and L. Pramiee, "Automated Threat Detection in the Dark Web: A Multi-Model NLP Approach," in *Proc. 2025 13th Int. Symp. on Digital Forensics and Security (ISDFS)*, Boston, MA, USA, pp. 1–6, Apr. 2025.
[2] S. Shah and V. K. Madisetti, "MAD-CTI: Cyber Threat Intelligence Analysis of the Dark Web Using a Multi-Agent Framework," *IEEE*

Access, vol. 13, pp. 40158–40168, Jan. 2025.

[3] H. Chen, Y. Diao, H. Xiang, and J. Shi, "Decode the Dark Side of the Language: Applications of LLMs in the Dark Web," in *Proc. 2024 Int. Conf. on Cyber-Language and Security*, San Francisco, CA, USA, Aug. 2024, pp. 45–54.

[4] V. Varghese, M. S., and S. Kb, "Extraction of Actionable Threat Intelligence from Dark Web Data," in *Proc. 2023 Global Cyber Threat Conference*, London, UK, May 2023, pp. 88–96.

[5] R. R. Gopireddy, "Dark Web Monitoring: Extracting and Analyzing Threat Intelligence," *Int. J. of Science and Research*, vol. 9, no. 3, pp. 1693–1696, Mar. 2020.

[6] C. Fachkha and M. Debbabi, "Darknet as a Source of Cyber Intelligence: Survey, Taxonomy, and Characterization," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1197–1222, Second Quarter 2016.

[7] M. Schäfer, M. Fuchs, M. Strohmeier, M. Engel, M. Liechti, and V. Lenders, "BlackWidow: Monitoring the Dark Web for Cyber Security Information," in *Proc. 2019 11th Int. Conf. on Cyber Conflict (CyCon'19)*, Tallin, Estonia, Jun. 2019, pp. 301–318.

[8] X. Zhang and K. Chow, "A Framework for Dark Web Threat Intelligence Analysis," *Int. J. Digital Crime Forensics*, vol. 10, no. 2, pp. 108–117, 2018.

[9] A. Zenebe, M. Shumba, A. Carillo, and S. Cuenca, "Cyber Threat Discovery from Dark Web," *Cyber Forensics Int. J.*, vol. 64, pp. 174–183, 2019.

[10] H. Al-Nabki *et al.*, "Methodology of Dark Web Monitoring," in *Proc. 2019 11th Int. Conf. on Electronics, Computers and Artificial Intelligence (ECAI)*, Paphos, Cyprus, Jul. 2019, pp. 77–84.

[11] P. Koloveas, T. Chantzos, C. Tryfonopoulos, and S. Skiadopoulos, "A Crawler Architecture for Harvesting the Clear, Social, and Dark Web for IoT-Related Cyber-Threat Intelligence," in *Proc. 2019 IEEE World Congress on Services (SERVICES'19)*, Milan, Italy, Jul. 2019, pp. 196–203.

[12] Y. Jin, E. Jang, J. Cui, J.-W. Chung, Y. Lee, and S. Shin, "DarkBERT: A Language Model for the Dark Side of the Internet," *arXiv preprint arXiv:2305.08596*, May 2023.

[13] P. Maneriker, Y. He, and S. Parthasarathy, "SYSML: StYlometry with Structure and Multitask Learning: Implications for Darknet Forum Migrant Analysis," *arXiv preprint arXiv:2104.00764*, Apr. 2021.

[14] A. Manolache, F. Brad, A. Barbalau, R. T. Ionescu, and M. Popescu, "VeriDark: A Large-Scale Benchmark for Authorship Verification on the Dark Web," *arXiv preprint arXiv:2207.03477*, Jul. 2022.

[15] R. Bhalerao, M. Aliapoulios, I. Shumailov, S. Afroz, and D. McCoy, "Towards Automatic Discovery of Cybercrime Supply Chains," *arXiv preprint arXiv:1812.00381*, Dec. 2018.

[16] "Towards Understanding Various Data Sources in Cyber Threat Intelligence Extraction," *arXiv preprint arXiv:2504.14235*, Apr. 2025.

[17] "Threats from the Dark: A Review over Dark Web Investigation," *Journal of Cybersecurity Studies*, vol. 12, no. 1, pp. 23–45, 2021.

[18] "Raffat, M. W., & Ahmad, A. Enhancing Anti-Money Laundering Systems with Machine Learning: A Comparative Analysis of Supervised

Models," Journal of Computational Informatics & Business, vol. 2 no. 2, pp 1-7, 2025.

[19] "Rafi, Saad, and Muhammad Sulman. "Post-Pandemic Insights: Evaluating the Impact of Big Data Analytics, Circular Economy Practices, and Digital Marketing on Firm Performance." Journal of Computational Informatics & Business Vol. 2, no. 1, pp 8-16, 2025.

[**20]** "Ahmad, Rafeeq, Humayun Salahuddin, Attique Ur Rehman, Abdul Rehman, Muhammad Umar Shafiq, M. Asif Tahir, and Muhammad Sohail Afzal. "Enhancing database security through AI-based intrusion detection system." Journal of Computing & Biomedical Informatics vol. 7, no. 02, (2024).

[21] "Hasanat, Syed Muhammad, Kaleem Ullah, Hamza Yousaf, Khalid Munir, Samain Abid, Syed Ahmad Saleem Bokhari, Muhammad Minam Aziz, Syed Fahad Murtaza Naqvi, and Zahid Ullah. "Enhancing short-term load forecasting with a CNN-GRU hybrid model: A comparative analysis." IEEE Access (2024).

[22] Ahmad, Zohaib, Muhammad Ammar Ashraf, and Muhammad Tufail. "Enhanced Malware Detection Using Grey Wolf Optimization and Deep Belief Neural Networks." International Journal for Electronic Crime Investigation 8, no. 3, (2024).