FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Exploring the Feasibility of Using Large Language Models for Dark Web Threat Intelligence in Security and Defence

Ricardo Nunes Ribeiro

Mestrado em Engenharia Informática e Computação

Supervisor: Prof. Alexandra Mendes

July 30, 2025

Exploring the Feasibility of Using Large Language Models for Dark Web Threat Intelligence in Security and Defence

Ricardo Nunes Ribeiro

Mestrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

President: Prof. Carlos Baquero-Moreno

Referee: Prof. Nuno Santos

Supervisor: Prof. Alexandra Sofia Ferreira Mendes

July 30, 2025

Resumo

A Dark Web (DW) é uma parte oculta da Internet acessível apenas através de ferramentas específicas, como o navegador TOR. Ainda que fundamental para *whistle-blowing* e ativismo político, a Dark Web tornou-se igualmente num centro para atividades criminosas, desde tráfico de drogas e armas a roubo de identidade e cibercrimes. Esta crescente quantidade de dados representa um desafio crítico para as forças e serviços de segurança (*LEAs*) no que diz respeito à monitorização e extração de informação útil dos dados. Os métodos manuais de investigação acarretam um elevado consumo de recursos, frequentemente revelando-se ineficazes contra a natureza adaptativa e descentralizada das plataformas.

Em paralelo, os recentes progressos da Inteligência Artificial, sobretudo os Grandes Modelos de Linguagem (*LLMs*), evidenciam competências significativas na resposta a perguntas ou resumo de conteúdos. A presente dissertação investiga o potencial destas ferramentas para apoiar as *LEAs* na extração de inteligência útil e na geração de *insights* estratégicos para apoiar a formulação políticas e ações investigativas.

Decorrente da iniciativa Atlantic Security Awards 2024, realizámos uma série de experiências para esta investigação: modelação de tópicos para classificação de conteúdos, avaliação de relevância de conteúdos, identificação de drogas com base em listagens de mercados e extração de informações de forma estruturada e geração de recomendações estratégicas. Para o efeito, foram selecionados e avaliados oito LLMs de última geração da OpenAI, Google DeepMind e Mistral, utilizando três *datasets* da Dark Web (DUTA-10K, CODA e Nemesis Marketplace), com especial atenção as associações ao Oceano Atlântico. Os resultados indicam que, embora os LLMs demonstrem proficiência notável na classificação (atingindo pontuações F1 superiores a 80% na identificação de drogas), as tarefas que envolvem formulação de políticas continuam a ser mais desafiantes. Na análise de relevância, os modelos tendem a identificar conteúdo irrelevante, ao mesmo tempo que mostraram precisão limitada na deteção de pistas de investigação de acordo com os critérios dados. No entanto, a decomposição de tarefas e instruções personalizadas aumentaram o desempenho, sugerindo que uma engenharia de *prompts* bem desenhada pode melhorar a fiabilidade dos modelos neste contexto.

É proposta uma arquitetura de sistema modular para operacionalizar estas conclusões, delineando uma plataforma de monitorização e extração de inteligência da Dark Web em tempo real alimentada por LLM. O sistema incorpora componentes para ingestão de dados, processamento LLM multiagente e interfaces de interação para investigação, alertas estratégicos e desenvolvimento de políticas.

Esta dissertação demonstra que os LLMs podem ser ferramentas valiosas para aumentar as capacidades das LEA em investigações da Dark Web dando apoio ao processamento de grandes quantidades de dados. Com um planeamento cuidado e alinhado às necessidades operacionais, essas tecnologias são muito promissoras para o avanço do policiamento digital e da defesa cibernética.

Abstract

The Dark Web (DW) is a concealed portion of the Internet accessible only through specific tools, such as the TOR browser. While being instrumental for activities such as whistle-blowing and political activism, it has also increasingly become a hub for criminal enterprises, ranging from narcotics and weapons trafficking to identity theft and cybercrime. This increasing amount of activity poses a critical challenge to Law Enforcement Agencies (LEAs) in monitoring and extracting meaningful insights from this data. Manual investigation methods are resource-intensive and often ineffective against the adaptive and decentralised nature of Dark Web platforms.

Simultaneously, recent advances in Artificial Intelligence (AI), particularly Large Language Models (LLMs), have demonstrated powerful capabilities in question answering or content summarisation. The present dissertation investigates the potential of these models to support LEAs in extracting actionable intelligence and generating strategic insights to support policy formulation and investigative action.

In the context of the Atlantic Security Awards 2024 initiative, we conduct a series of experiments for this research: topic modelling for content classification, content relevance assessment, drug identification based on marketplace listings, and structured information extraction with strategic recommendation generation. For this purpose, eight state-of-the-art LLMs from OpenAI, Google DeepMind, and Mistral were selected and evaluated using three ethically sourced Dark Web datasets (DUTA-10K, CODA, and Nemesis Marketplace), giving a special focus on associations with the Atlantic Ocean when possible. The findings indicate that, whilst LLMs demonstrate notable proficiency in classification and domain-specific tagging tasks (attaining F1-scores exceeding 80% in drug identification), tasks involving policy-oriented generation remain more challenging. In relevance analysis, the models tended to recognise irrelevant content, while showing limited precision in detecting investigative leads according to the given criteria. However, task decomposition and tailored instructions increased performance, suggesting that thoughtful prompt engineering can improve model reliability in this context.

A modular system architecture is proposed to operationalise these findings, outlining an LLM-powered real-time Dark Web monitoring and intelligence extraction platform. This system incorporates components for data ingestion, multi-agent LLM processing, and interaction interfaces for investigation, strategic alerting, and policy development.

This dissertation demonstrates that LLMs can be valuable tools for augmenting LEA capabilities in Dark Web investigations by supporting the processing of vast amounts of Dark Web content. With careful design and alignment to operational needs, these technologies hold significant promise for advancing digital policing and strategic cyber defence.

UN Sustainable Development Goals

The United Nations Sustainable Development Goals (SDGs) provide a global framework to achieve a better and more sustainable future for all. It includes 17 goals to address the world's most pressing challenges, including poverty, inequality, climate change, environmental degradation, peace, and justice. Specifically, this dissertation's outcomes can help contribute to addressing SDG 16:

SDG 16 Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels

SDG 16 is subdivided in specific targets. The following list enumerates those in closer alignment with this work:

Target 16.2 End abuse, exploitation, trafficking and all forms of violence against and torture of children

Target 16.4 By 2030, significantly reduce illicit financial and arms flows, strengthen the recovery and return of stolen assets and combat all forms of organized crime

Target 16.a Strengthen relevant national institutions, including through international cooperation, for building capacity at all levels, in particular in developing countries, to prevent violence and combat terrorism and crime

SGD	Target	Contribution	Performance Indicators and Metrics
16	16.2	Enhance the detection and analysis of online content related to child exploitation on the dark web through advanced AI tools, supporting early intervention and prevention.	Number of relevant posts identified related to child abuse/exploitation leading to relevant actions.
	16.4	Directly addresses illicit trade (e.g., narcotics and weapons) by identifying and categorising dark web listings, enabling more effective disruption of organised crime networks.	Quantity of classified illicit trade instances (e.g., drugs, weapons) leading to relevant actions.
	16.a	Supports institutional capacity building by proposing a scalable, AI-powered system for automated monitoring and analysis of criminal content, suitable for adoption by LEAs.	Number of violent crimes investigated resulting from the interaction with the system

Acknowledgements

First and foremost, I would like to express my immense gratitude to my supervisor, Professor Alexandra Mendes, for having faith in my ability to conduct this research for my dissertation. I am deeply grateful for her continuous guidance and invaluable suggestions throughout this period.

I thank the Luso-American Development Foundation, the Atlantic Centre and the Portuguese National Defence Institute for acknowledging the value and merit of Professor Alexandra's proposal and helping make this project a reality. I would also like to acknowledge the contributions of the Portuguese Navy, Polícia Judiciária and Polícia Marítima in this research.

As I reflect on the last few years, I am incredibly thankful to my friends for their unwavering camaraderie and support throughout this journey's challenging and joyous moments.

Above all, I am profoundly thankful to my family. To my brother, for his wise counsel and constant encouragement. To my parents, whose selfless sacrifices and unconditional support were essential for this milestone.

Ricardo Nunes Ribeiro

"Go for greatness,
anything else is a waste of time"

Marianne Williamson

Contents

List of Figures

List of Tables

Listings

List of Acronyms

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
CODA	Corpus of Darknet Activities
CTI	Cyber Threat Intelligence
DUTA	Darknet Usage Text Addresses
DW	Dark Web
DWF	Dark Web Forum
FFN	Feed-Forward Network
GPT	Generative Pre-trained Transformer
LEA	Law Enforcement Agency
LLM	Large Language Model
LM	Language Model
ML	Machine Learning
NLP	Natural Language Processing
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RNN	Recurrent Neural Network
TOR	The Onion Router
WWW	World Wide Web

Chapter 1

Introduction

1.1 Context

The Internet serves as the foundation for the World Wide Web (*WWW*) [7], a system of interconnected documents and resources typically accessed through standard web browsers. Despite the common perception of the *WWW* as a unified structure, it is, in fact, composed of several distinct layers, each exhibiting varying degrees of accessibility and visibility. The most accessible and well-known layer is the Surface Web, where publicly indexed websites are readily searchable through engines such as Google.

The Deep Web, which comprises content not indexed by search engines, lies beneath the Surface Web. This category includes private databases, academic archives, password-protected sites, and internal networks. Despite common misconceptions, it is primarily composed of legitimate, non-public information.

The deepest layer of this hierarchy is the Dark Web (DW), a subset of the Deep Web that is only accessible through specialised software such as TOR (The Onion Router) [22]. The Dark Web has gained a reputation for facilitating illicit activities, including the operation of black markets for contraband of goods such as narcotics, weapons, and stolen data. An example of these markets was Silk Road [15]. However, it also serves as a platform for activities requiring higher levels of privacy, such as political activism, whistle-blowing, or disseminating censored information.

The substantial volume of transactions and data available on the Dark Web presents a considerable obstacle for law enforcement agencies (LEAs) attempting to analyse them. This can result in a significant delay in the investigation process or, in some cases, even prevent it from being carried out in its entirety. However, new technologies have arisen and, if well explored, can help address this.

In the field of artificial intelligence (AI), large language models (LLMs) [47, 59, 79] have emerged as a transformative technology, demonstrating the capacity to process and generate human language with high levels of accuracy and coherence.

A significant advancement in the development of LLMs was the introduction of the Transformer architecture [72]. LLMs are used across a broad spectrum of applications, ranging from

machine translation and sentiment analysis to automated question-answering. Additionally, they play a pivotal role in developing conversational agents, such as chatbots and virtual assistants, and in content generation for fields such as writing and coding.

1.2 Motivation

The technologies employed on the Dark Web platform enable users to benefit from robust encryption and effective identity concealment. This versatility allows the platform to be used for both legitimate and illicit purposes. Over time, a growing number of criminals, including both individual perpetrators and organised criminal networks, have been using the Dark Web for illicit activities [24].

In response, law enforcement agencies have intensified their efforts to disrupt the operations of Dark Web black markets and forums [36] [25]. However, the closure of a website does not necessarily result in the complete cessation of its activities. Instead, this void is often filled by the emergence of new websites [28].

Despite the efforts by law enforcement agencies to tackle illicit activities on the Dark Web, the vast amounts of available data still contain information that can further assist authorities. Finding ways to easily extract meaningful information from this data has the potential to significantly support authorities in their fight against crime. One promising direction to address these opportunities can be through using LLMs and their proven capabilities.

Researchers have been exploring how artificial intelligence can help authorities tackle this problem, primarily in the context of Cyber Threat Intelligence (CTI). Most existing work employs Natural Language Processing (NLP), mainly text classification models [3] [27] [33], to better understand and analyse the contents of the black markets and forums. With the rise of more capable and powerful LLMs, others have studied how these perform in classifying Dark Web textual data, both with out-of-the-box LLMs [39] and pre-trained models with Dark Web text corpora [38]. Both approaches achieved rather promising results, the latest showing a higher performance than previously recorded work.

Therefore, it is possible to say that NLP tools pose a solid foundation for CTI domain tasks and can be valuable in aiding authorities in the combat against Dark Web-based criminality. However, to our knowledge, these advances are not actively used to analyse and investigate criminal activities on the Dark Web by law enforcement. Providing insights on how authorities could easily and efficiently interact with these tools would be of great value for improving decision-making and policy formulation aimed at tackling the aforementioned issues.

In this context, a key component of this work is to gather and analyse concrete and objective requirements in collaboration with Law Enforcement Agencies, in the context of the Atlantic Security Award 2024, awarded by the Luso-American Development Foundation[1] (FLAD), the

[1] https://www.flad.pt/

Portuguese National Defence Institute[2], and the Atlantic Centre[3]. Given the scope of this award and initiative, the research will focus on the Atlantic Ocean whenever appropriate and enabled by the datasets used in this project.

1.3 Problem

One of the most significant challenges in addressing criminal activities on the Dark Web is the effective and timely analysis of the vast amount of data. The rapid proliferation of new forums, black markets, and other illicit platforms presents a significant challenge for law enforcement agencies attempting to monitor and intervene in such activities.

This research project seeks to address the challenge of analysing the large amounts of data available in the Dark Web using LLMs to assist law enforcement and governmental agencies in the process of decision-making and policy formulation and evaluate whether and how LLMs can be used to enhance the detection, analysis and mitigation of criminal activities related to the Dark Web, with a particular focus on regions of strategic importance.

The principal objective of this research is to develop a series of experiments integrating LLMs with data specific to the Dark Web to enhance the strategic planning capabilities of law enforcement agencies. The experiments and their outcomes can help support the development of defence policies and enable more informed decision-making processes when responding to cybercrime, illicit trade, and other security threats facilitated by the Dark Web.

1.4 Research Questions

With this work, we intended to tackle the following research questions:

- **RQ1:** Which Threat Intelligence (TI) tasks are current state-of-the-art LLMs better fit to carry out and how can they be leveraged for analytical and investigative purposes?

  We carry out a varied set of experiments in order to assess LLMs capabilities in analysing Dark Web content and report their performance.

- **RQ2:** How can LLMs be used to simplify law enforcement user interaction while maximising result accuracy and relevance in Dark Web TI?

  We analyse which methods of prompting LLMs are more effective, with a view to obtaining accurate and useful recommendations in the context of TI, and thus enhancing the effectiveness of law enforcement investigations.

- **RQ3:** How do different LLMs compare in identifying threats and offering actionable defence strategies from Dark Web text data?

[2] https://www.idn.gov.pt/pt
[3] https://www.defesa.gov.pt/pt/pdefesa/ac

We establish formal evaluation metrics and compare different LLMs' performance when prompted to, among others, provide suggestions for new policy formulation and courses of action.

1.5 Document Structure

This dissertation is structured into eight chapters, each corresponding to a distinct stage in investigating large language models' effectiveness for threat intelligence extraction from Dark Web content and their potential integration into law enforcement workflows.

- **Chapter 2 - Background:** Presents a technical overview of foundational concepts in language modelling, including the evolution of LLMs, the Transformer architecture, and the mechanisms underpinning attention-based models.
- **Chapter 3 - Literature Review:** Details the methodology and outcomes of a systematic literature review focusing on applications of LLMs in Dark Web content analysis grounded in the PRISMA framework.
- **Chapter 4 - State of the Art:** Based on the literature review from Chapter 3 and further research, provides a focused survey of contemporary approaches to threat intelligence, spanning traditional and machine learning-based methods across domains such as cybercrime, terrorism, and criminal marketplaces.
- **Chapter 5 - Experimental Procedures:** Describes the datasets, data preparation protocols, and LLM selection criteria. It then outlines the design of four experiments: topic modelling, relevance analysis, drug identification and structured information extraction and policy recommendation. The prompt construction, sampling, annotation, and evaluation procedures are also detailed to ensure reproducibility.
- **Chapter 6 - Results & Discussion:** Presents the results obtained from the experimental pipeline, including performance metrics, qualitative analyses, and confusion matrices. The chapter also discusses dataset-specific challenges and implications for real-world deployment in threat intelligence workflows.
- **Chapter 7 - Future Work:** Synthesises the experimental findings into designing a conceptual LLM-powered framework for monitoring and analysing Dark Web content. It defines both functional and non-functional requirements, presents a modular system architecture, and includes a usage scenario that illustrates a possible interaction between law enforcement users and system components.
- **Chapter 8 - Conclusions:** Concludes the dissertation by answering the research questions with a summary of the key findings and contributions, outlining the limitations of the current work, and concluding with directions for future research.

Chapter 2

Background

Since computers were invented, the research community has been focused on powering machines with the ability to understand and reply in human language [69]. Machines do not possess the inherent ability to understand and communicate like humans do. A significant contribution in providing machines with some reading, writing, and communication abilities is the advancements in language modelling [79]. Continuous investment in this field and increasing computing power capabilities combined with large amounts of data have culminated in the nowadays called Large Language Models [59]. This chapter explores the history and development of Language Models (LMs), including a brief overview of their evolution, the key architectures, current state-of-the-art models, and their primary applications and societal impacts.

2.1 Language Modelling Evolution

A language model is a general function that, given an input sequence of words, can output a set of probabilities for a missing word in the input sequence or a future word in the sentence [79]. In the early developments of language models during the '90s, models were based on statistical methods[1]. They were able to capture simple statistical patterns and correlations in the data, but they were not ideal for handling out-of-vocabulary words, nor did they excel at capturing semantic relationships between words. For instance, pronouns were seen as any other word in a sequence without the model "learning" to which noun it related.

Some years later, the introduction of the Recurrent Neural Network (RNN) [46] in the field allowed the creation of more complex models. These were better suited for capturing and modelling sequential data, like language, and generating. Nonetheless, they were still limited by short-term memory and extended training time requirements [59].

The significant advancement in LLMs took place upon the introduction of the Transformer by a Google research team in 2017 [72]. From then on, LLM architectures are predominantly based

[1] One example is the Bag-of-Words model. These work by creating a set of words from a given collection of documents and associating each word with its absolute frequency in the collection.

on the main Transformer's principle: (self) attention, which, leveraging computation parallelisation on GPUs, captures long-range dependencies within the input sequences and encodes multiple dimensions of relationships between words in the text. In 2018, Devlin et al. [21] introduced the Bidirectional Encoder Representations from Transformers (BERT), which was able to capture and encode word relationships from both the left and right of a given word, allowing a better understanding of contextual information. The model performed better than previous language models and effectively accomplished more advanced tasks such as question answering and text classification. Also in 2018, OpenAI introduced the concept of Generative Pre-trained Transformer (GPT) [58], the basis of their first model, GPT-1, which adopted a combination of unsupervised learning for the first training stage, resulting in a base model, and supervised fine-tuning for more specific tasks such as chatbots. Since then, the company has significantly advanced the field of NLP with the introduction of improved GPT model iterations [59].

2.2 Key Concepts and Principles

Modern large language models are primarily built upon the Transformer architecture. Before examining how Transformers function and how they are applied in contemporary models, it is necessary to introduce key concepts that support text processing. In particular, this section focuses on tokenisation and embedding, two core steps that enable the conversion of raw text into numerical representations suitable for computation. These concepts are essential for understanding how Transformers encode language and perform downstream tasks.

Tokenisation. Language models are not able to understand words as humans do. Moreover, as stated earlier, one of the issues in earlier language models was related to out-of-vocabulary words. Whenever a model was given a word not present in the training dataset, the model would struggle to predict the next word because it had no prior knowledge of which words usually appear next to this new word.

One way of solving this issue is to divide words into smaller unique character sequences, often called *tokens* through a *tokenisation* process [74]. This set of all possible tokens is converted into unique numerical representations called token IDs. One of the most common algorithms to achieve this is the Byte-Pair Encoding Algorithm (alternatives include UnigramLM or WordPiece) [56]. This algorithm works by merging the most frequent character sequences in the text. For instance, OpenAI created a tokeniser called *tiktoken*[2] that uses the Byte-Pair Encoding Algorithm [80]. However, for transformers and large language models, it is required to insert some special tokens, like beginning or end of sentence tokens: `bos_token` and `eos_token` [56].

Embeddings. Given a vocabulary V composed of a set of token IDs, computing the embedding of a token $V[i]$ consists of representing that token as a vector in a continuous vector space. In practice, the embedding of a token is implemented as a row in a trainable embedding matrix. Let

[2] https://github.com/openai/tiktoken

$E \in \mathbb{R}^{|V| \times d_{\text{emb}}}$ be the embedding matrix, where $|V|$ is the size of the vocabulary and d_{emb} is the embedding dimension, i.e. the size of the vector. For a given token $V[i]$, its embedding is given by

$$e_i = E[i] \in \mathbb{R}^{d_{\text{emb}}}, \tag{2.1}$$

which corresponds to the i-th row of E. This vector e_i serves as a continuous representation of the token in the embedding space, enabling operations such as dot-product comparisons to measure the similarity between tokens.

Example Let us consider the word "Internet". Using the *tiktoken* tool with the "o200k_base" tokenisation configuration, this word corresponds to the token "Internet" with ID 34831. To obtain the embedding for this token, the model looks up row 34831 of the embedding matrix E, retrieving the vector $e_{34831} \in \mathbb{R}^{d_{\text{emb}}}$. This vector is then passed to subsequent layers of the model as the input representations of the original word fragments.

2.3 The Transformer

The Transformer neural network model architecture was introduced by Vaswani et al. [72] in 2018. Nowadays, most language models are based on this architecture with some modifications and different performance optimisations. The main breakthrough was the introduction of the attention mechanism without using recurrence or convolution principles, addressing several disadvantages of RNNs regarding recurrence and sequential input analysis.

Figure 2.1 illustrates the structural organisation of the transformer, which follows the same encoder-decoder block approach as previous models of its kind [72]. The general idea is as follows:

1. The encoder (left side of the figure) receives an input sequence of token representations of length $\ell_x \in \mathbb{N}$, $x = (x_1, \ldots, x_{\ell_x})$ and produces an output sequence vector $z = (z_1, \ldots, z_{\ell_z})$, $\ell_z \in \mathbb{N}$, based on parameters defined during the training process.

2. The decoder (right side of the figure) generates a new token sequence $y = (y_1, \ldots, y_{\ell_y})$ of size $\ell_y \in \mathbb{N}$. At each iteration $i \leq \ell_y$, the decoder receives as input z (from the encoder block), and the previously generated sequence $y[:i-1]$ (subsequence of y from position 0 to $i-1$). This text generation process, based on previous outputs, is called auto-regression [30].

These blocks can be selected and used in different contexts, depending on specific needs and tasks, which will be detailed later in this chapter (see Encoder vs Decoder). Each block is composed of two layers: an Attention and a Position-wise Feed Forward Network layer. When both encoder and decoder blocks are used, the output of the encoder serves as input to a cross-attention sub-layer within the decoder.

All layers and sub-layers in the Transformer architecture operate on vector representations known as embeddings. These embeddings are fixed-length vectors, and their dimensionality is

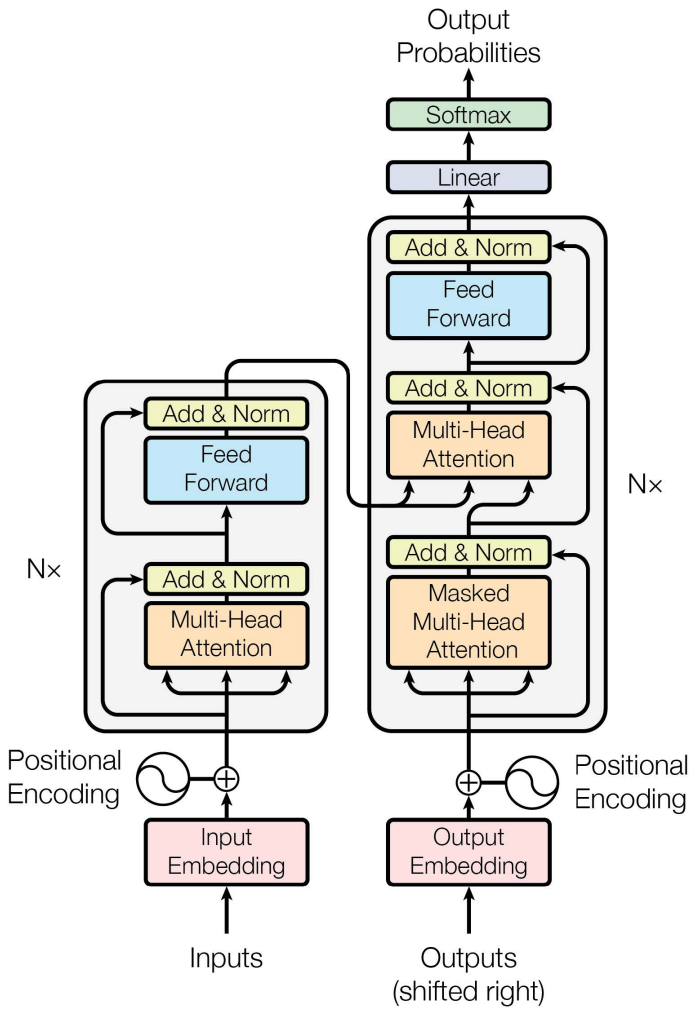


Figure 2.1: Transformer Architecture (extracted from [72])

denoted by d_{model}. This parameter defines the size of the embedding space and determines the width of the hidden layers throughout the network. Typical values of d_{model} in large models range from 512 to 4096, depending on model capacity.

2.3.1 Attention

The first layer of both the encoder and the decoder blocks is the attention layer. This layer is able to define input sequence representations by analysing each position and its relationships with other positions. The attention mechanism adopted by Vaswani et al. [72] implements the Query-Key-Value model [44]. An interpretation of the attention mechanism is to regard it as a dynamic aggregation method, whereby each token in a sequence synthesises information from every other token via a learned weighting scheme. In this framework, each token generates a query vector that is compared against a set of key vectors, each associated with corresponding value vectors, using a similarity measure (typically a scaled dot-product). The resulting scores are normalised to yield attention weights that modulate the contribution of each value to the final representation.

When analysing text sequences, the establishment of the relationships between words can be carried out in two different ways: self-attention and cross-attention, further explained in the following subsections.

2.3.1.1 Self-attention

In the case of self-attention, the relationships are defined regarding words in the same sequence: for each word in the sequence, we analyse how much this word correlates, or "attends", to all other words in the same sequence. Phuong and Hutter [56] intuitively explain this as having two sequences: a primary sequence X to which attention is applied and a secondary/context sequence Z. Recalling the previously defined input sequence x, we can define a matrix $X \in \mathbb{R}^{\ell_x \times d_x}$ by stacking all input representations (i.e. embeddings) row by row, where d_x is the embedding dimension of $x[i], \forall i \leq \ell_x$. Similarly, we can define a context matrix $Z \in \mathbb{R}^{\ell_z \times d_z}$. When applying self-attention to X, the primary and context sequences/matrices are the same ($X = Z$).

Unidirectional vs Bidirectional Self-Attention. Depending on the task the model performs, it may be more beneficial to apply attention to each token by considering the entire sequence as context. This is called bidirectional self-attention: for every sequence token, attention is calculated in relation to both the previous and following tokens. This is particularly useful in classification/sentiment analysis tasks, where words on the right of another word also matter for context.

However, when training a model for other tasks (e.g. text generation), the model should be able to predict the next word in a sequence based only on the previous words in the sequence, including itself. Therefore, the attention is performed only regarding the words on the left of a given token, i.e., unidirectionally. This is done through a process called masking and is only applied in the decoder block of the transformer [44].

Self-attention is clearly identifiable in Figure 2.1 in both blocks of the transformer. It corresponds to the single attention layer in the encoder and the bottom attention box in the decoder. The three incoming arrows of each box represent the query, keys and value vectors.

2.3.1.2 Cross-attention

There is a second attention sub-layer on the transformer's decoder block. In this case, the query vector inputs come from the preceding self-attention layer, whereas the key and value vectors are projected from the output of the encoder block [44]. Going back to Phuong and Hutter [56]'s notation, in cross-attention, $X \neq Z$.

2.3.2 High-level Attention Algorithm

As discussed above, attention uses a primary matrix X and a context matrix Z. The general mathematical equation for the attention calculation is the following:

$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{2.2}$$

where $Q \in \mathbb{R}^{\ell_x \times d_q}$, $K \in \mathbb{R}^{\ell_z \times d_k}$, and $V \in \mathbb{R}^{\ell_z \times d_v}$ represent the query, key and value matrices, respectively, $d_q = d_k$ are the dimensions of the query and key vectors, and d_v is the size of the value vectors.

The following list outlines the general steps for calculating attention and is summarised in Algorithm 1 (adapted from [56]):

1. Matrices Q, K, and V are obtained by multiplying X, Z, and Z with parameter trainable weight matrices $W^Q \in \mathbb{R}^{d_x \times d_q}$, $W^K \in \mathbb{R}^{d_z \times d_k}$, and $W^V \in \mathbb{R}^{d_z \times d_v}$, respectively.

   Each row $i \leq \ell_x$ in Q represents the query vector for a token $X[i]$, while, for each $i \leq \ell_z$, $K[i]$ and $V[i]$ represent the key and value vectors of the token at the i-th row in Z, $Z[i]$, respectively.

   When performing self-attention, Z will be equal to X.

2. The similarity scores between each query and all keys are then computed with QK^T. This results in a new matrix (let us call it $A \in \mathbb{R}^{\ell_x \times \ell_z}$) where each value $A[i,j]$ represents how similar the query vector $Q[i]$ is to the key vector $K[j]$ (the higher the value, the higher the similarity).

3. (Optional) In order to treat only left and current tokens as context for any given token, a *mask* must be applied. This mask is more commonly used in the decoder block and is defined by a matrix $M \in \{0, -\infty\}^{\ell_x \times \ell_z}$ which is added to A. M can be defined as

$$M[i,j] = \begin{cases} 0 & ,i \leq j, \\ -\infty & ,\text{otherwise.} \end{cases} \tag{2.3}$$

4. The softmax function is then applied to every row of A divided by $\sqrt{d_k}$. This division prevents gradients of the softmax functions from becoming too small [44]. This function is defined as:

$$\text{softmax}(A)[i,j] = \frac{e^{A[i,j]}}{\sum_k e^{A[i,k]}} \tag{2.4}$$

   This operation ensures that each row of the resulting matrix forms a probability distribution, where the sum of the entries in each row equals one.

The resulting matrix A is often called *attention matrix* [44]. Going back to the initial intuition of attention, this matrix holds the attention scores, i.e., the weights of the contribution of each value to the representation of each word/token.

5. The final step is to compute AV, resulting in a new matrix $\tilde{V} \in \mathbb{R}^{\ell_x \times d_v}$.

Algorithm 1: $\tilde{V} \leftarrow \texttt{Attention}(X, Z | \mathbb{W}_{qkv}, M)$

Input: Input sequence $X \in \mathbb{R}^{\ell_x \times d_x}$, context sequence $Z \in \mathbb{R}^{\ell_z \times d_z}$
Output: Output matrix $\tilde{V} \in \mathbb{R}^{\ell_x \times d_v}$
Parameters: Trainable weight matrices $\mathbb{W}^{qkv}$: $W^Q \in \mathbb{R}^{d_x \times d_q}$, $W^K \in \mathbb{R}^{d_z \times d_k}$, $W^V \in \mathbb{R}^{d_z \times d_v}$, Masking matrix $M \in \{0, -\infty\}^{\ell_x \times \ell_z}$

$Q \leftarrow XW^Q$
$K \leftarrow ZW^K$
$V \leftarrow ZW^V$
$A \leftarrow QK^T + M$
$A[i,j] \leftarrow \frac{e^{A[i,j]/\sqrt{d_k}}}{\sum_k e^{A[i,k]/\sqrt{d_k}}}$
return $\tilde{V} = AV$

2.3.3 Multi-Head Attention

In the original architecture of the transformer, the authors concluded it was beneficial to reduce the dimensionality of d_q, d_k and d_v, with $d_q = d_k$. In practice, the process is similar to the one described in the previous section but split across H *attention heads*. The process is detailed in the following list and summarised in Algorithm 2 (adapted from [44, 56, 72]):

1. Redefine d_q, d_k and d_v by dividing each value by H. In the original paper, $d_{\text{model}} = 512$ and $H = 8$, hence the values used were $d_q = d_k = d_v = d_{\text{model}}/8 = 64$.

2. For each attention head $head_i$, the corresponding trainable weight matrices are defined as $W_i^Q \in \mathbb{R}^{\ell_x \times d_q}$, $W_i^K \in \mathbb{R}^{\ell_z \times d_k}$, and $W_i^V \in \mathbb{R}^{\ell_z \times d_v}$, with the newly defined d_q, d_k and d_v from the previous step.

3. For each attention head $head_i$, follow steps 1-5 from the previous section, obtaining $\tilde{V}_i \in \mathbb{R}^{\ell_x \times d_v}$.

4. Concatenate all $\tilde{V}_i$: $Y \leftarrow [\tilde{V}_1; \ldots; \tilde{V}_H] \in \mathbb{R}^{\ell_x \times Hd_v}$

5. Given a trainable parameter matrix $W^O \in \mathbb{R}^{Hd_v \times \ell_x}$, $\tilde{Y} \leftarrow YW^O \in \mathbb{R}^{\ell_x \times Hd_v}$, where $Hd_v = d_{\text{model}}$.

Algorithm 2: $\tilde{Y} \leftarrow \texttt{MultiHeadAttention}(X, Z|\mathbb{W}, M)$

Input: Input sequence $X \in \mathbb{R}^{\ell_x \times d_x}$, context sequence $Z \in \mathbb{R}^{\ell_z \times d_z}$
Output: Output matrix $\tilde{Y} \in \mathbb{R}^{\ell_x \times d_{\text{model}}}$
Hyperparameters: Number of heads H
Parameters: Trainable weight matrices $\mathbb{W}$: output projection $W^O \in \mathbb{R}^{Hd_v \times d_{\text{model}}}$ and for each head $i \leq H$: $W_i^Q \in \mathbb{R}^{d_x \times d_q}$, $W_i^K \in \mathbb{R}^{d_z \times d_k}$, $W_i^V \in \mathbb{R}^{d_z \times d_v}$; Masking matrix $M \in \{0, -\infty\}^{\ell_x \times \ell_z}$

for $i \leftarrow 1$ ***to*** H **do**
  $\tilde{V}_i \leftarrow \texttt{Attention}(X, Z|\mathbb{W}_i^{qkv}, \text{Mask})$
end
$Y \leftarrow [\tilde{V}_1; \ldots; \tilde{V}_H]$
return $\tilde{Y} = YW^O$

2.3.4 Position-wise Feed Forward Network

The second main layer in each transformer block is a fully connected feed-forward network (FFN) [72]. This FFN with inner dimensionality $d_{\text{ffn}}, d_{\text{ffn}} > d_{\text{model}}$ applies further transformations to the output of the previous layer:

$$\text{FFN}(Y) = \text{ReLU}(YW_1 + b_1)W_2 + b_2 = \max(0,\ YW_1 + b_1)W_2 + b_2 \tag{2.5}$$

where $W_1 \in \mathbb{R}^{d_{\text{model}} \times d_{\text{ffn}}}$, $W_2 \in \mathbb{R}^{d_{\text{ffn}} \times d_{\text{model}}}$, $b_1 \in \mathbb{R}^{d_{\text{ffn}}}$ and $b_2 \in \mathbb{R}^{d_{\text{model}}}$ are trainable parameters, and ReLU is the Rectified Linear Unit activation function given as $\text{ReLU}(x) = \max(0, x)$ [44].

2.3.5 Positional encoding

The order of the words/tokens is crucial in text modelling. However, attention mechanisms are based on matrix multiplications; therefore, changing the order of tokens in the input sequence would produce the same attention results for the same token. There are several alternatives to incorporate this information into the input sequence, namely relative and absolute position representations [44].

In the original transformer, the authors use sine and cosine functions to compute absolute positional embeddings before each encoder and decoder stack, which are then added to the initial token embeddings, resulting in a new vector used as the actual input to the model. Let $X \in \mathbb{R}^{\ell_x \times d_{\text{model}}}$, where $X[i]$ is the embedding vector for token $x[i]$. The positional encoding matrix $P \in \mathbb{R}^{\ell_x \times d_{\text{model}}}$ would be defined as

$$P[i, 2j-1] = \sin\left(\frac{i}{n^{2j/d_{\text{model}}}}\right) \tag{2.6}$$

$$P[i, 2j] = \cos\left(\frac{i}{n^{2j/d_{\text{model}}}}\right) \tag{2.7}$$

for $0 < j <= d_{\text{model}}/2$, and $n \in \mathbb{R}$ (in the original paper, it was set to 10000). To incorporate the positions, P is then added to X.

2.3.6 Residual Connections & Layer Normalisation

The original Transformer architecture combines residual connections with layer normalisation to stabilise training and enable the construction of deeper models. In this design, the output of each sub-layer is computed as

$$Y = \text{LayerNorm}\big(x + \text{Sublayer}(x)\big), \tag{2.8}$$

where $\text{Sublayer}(x)$ represents the specific operation being performed (such as self-attention or the feed-forward network). To ensure that these residual connections integrate seamlessly, every sub-layer, as well as the embedding layers, produces representations of a fixed dimension d_{model}.

Layer normalisation is key to controlling the distribution of activations within the network. As noted by Phuong and Hutter [56], it standardises the mean and variance of each neuron's output, thereby contributing to training stability. They also point out that some Transformer variants adopt a more computationally efficient version - root mean square layer normalisation (RMSnorm) - by setting mean m and β as learnable parameters.

Residual connections themselves are crucial in mitigating issues like vanishing (token probabilities close to or equal to 0) or exploding gradients in deep networks. As described by Lin et al. [44], the original Transformer applies residual connections around each module. For instance, a typical Transformer encoder block can be expressed as:

$$H' = \text{LayerNorm}\big(\text{SelfAttention}(X) + X\big), \tag{2.9}$$

$$H = \text{LayerNorm}\big(\text{FFN}(H') + H'\big), \tag{2.10}$$

where $\text{SelfAttention}(\cdot)$ and $\text{FFN}(\cdot)$ denote the self-attention mechanism and the feed-forward network, respectively.

Later research has explored different placements for the layer normalisation step. The vanilla Transformer applies normalisation after the residual addition (post-LN), whereas subsequent implementations have proposed a pre-layer normalisation (pre-LN) variant—normalising the input within the residual branch prior to the main transformation. Although post-LN Transformers can experience initial instability (often requiring techniques such as learning rate warm-up), they are generally observed to outperform pre-LN variants once fully converged [44].

2.3.7 Encoder vs Decoder

The capabilities of the original transformer were first introduced for sequence-to-sequence tasks, more specifically, translation from English to German and English to French. This was done using the encoder-decoder attention architecture, where all positions in the sequence of the decoder attend to all positions of the input sequence in the encoder through cross-attention. Phuong and

Hutter [56] formally define this problem as a conditional probability distribution estimation. Given two text sequences x and z, the goal is to learn an estimate of $P(x|z)$. This estimation is usually decomposed as

$$\hat{P}(x|z) = \prod_{i=1}^{\ell_x} \hat{P}(x[i] \mid x[1:i-1], z) \tag{2.11}$$

In the example of a translation task from English to German, z would be the original English sentence and x the same sentence in German.

Another common implementation of the transformer model is by using solely the encoder block. Its self-attention layers create a deep representation of the input sequence. This is ideal for classification tasks or sentiment analysis. BERT is an example of an encoder-only transformer. BERT was trained with a dataset of sequences where some tokens were masked out, and the main goal was to assign a probability to other tokens in order to correctly recover the masked-out instance.

The most common implementations for sequence generation tasks, such as language models, rely on the decoder-only architecture. They are the foundation of most common large language models nowadays, the focus of this dissertation. In practice, a decoder-only transformer receives as input an initial prompt and is able to generate a probability distribution for the tokens it knows. Given a target sequence length ℓ_x, the main objective is to generate a sequence x, by learning the probability of a single token given its preceding context tokens:

$$\hat{P}(x) = \prod_{i=1}^{\ell_x} \hat{P}(x[i] \mid x[1:i-1]) \tag{2.12}$$

2.3.8 Training & Inference

Having defined the main components of both the encoder and the decoder blocks, these can be combined into a complete transformer. As the main focus of this dissertation is LLMs, which predominantly follow the decoder-only architecture, Algorithm 3 presents an overview of how a forward pass in the network functions (adapted from [56]). The decoder produces output vectors of dimension d_{emb}. Line 10 projects these vectors into a space with a dimension equal to the vocabulary size via the unembedding matrix $U \in \mathbb{R}^{d_{\text{emb}} \times |V|}$. The output of this linear layer is often called *logits*. These logits represent unnormalized scores for each token in the vocabulary, indicating the model's confidence for each possible next token in the sequence. The softmax layer then converts these logits into probabilities, ensuring they sum to 1. This allows the model to probabilistically select the next token in the sequence.

Once the model architecture is defined, the next step to create an LLM is to train the model. The general process' pseudocode is demonstrated in Algorithm 4 and follows the common training strategies applied in other deep learning models.

Algorithm 3: $P \leftarrow$ `DecoderTransformer`$(x|\theta)$

Input: Token IDs sequence $x \in V^*$ of size ℓ_x

Output: Probability matrix $P \in (0,1)^{\ell_x \times |V|}$, where the i-th row represents $\hat{P}_\theta(x[t+1]|x[1:t])$.

Hyperparameters: Number of decoder blocks L, embedding dimension d_e, FFN inner dimensionality d_{ffn}

Parameters: Set of parameters θ:

Embedding and positional encoding matrices $E \in \mathbb{R}^{|V| \times d_{\text{emb}}}$, $P \in \mathbb{R}^{\ell_x \times d_{\text{model}}}$

Unembedding matrix $U \in \mathbb{R}^{d_{\text{emb}} \times |V|}$

For each layer $l \leq L$:

Multi-head attention parameters $\mathbb{W}_l$

FNN parameters: $W_1^l \in \mathbb{R}^{d_{\text{ffn}} \times d_{\text{emb}}}$, $W_2^l \in \mathbb{R}^{d_{\text{emb}} \times d_{\text{ffn}}}$

$M[i,j] \leftarrow 0$ if $i \leq j$ else $-\infty$

for $l \in \ell_x : e_i \leftarrow E[x[i]] + P[i]$

$X \leftarrow [e_1, \ldots, e_{\ell_x}]$

for $l \leftarrow 1$ ***to*** L **do**

  $X \leftarrow X +$ `MultiHeadAttention`$(X|\mathbb{W}_l, M)$

  $X \leftarrow \text{LayerNorm}(X)$

  $X \leftarrow X + \text{FFN}(X)$

end

$X \leftarrow \text{LayerNorm}(X)$

return $P = \text{softmax}(XU)$

This training process results in a base model or a pretrained language model, outputting the learnt transformer weights (or parameters) later used to generate new tokens. To generate a text sequence of size ℓ_{gen}, an instance of the `DecoderTransformer` is created, passing the previously learnt weights. Then, a token is sampled from the resulting probability distribution. These steps are repeated ℓ_{gen} times, and the tokens are concatenated together, resulting in a new sequence.

Large Language Models are trained on extensive textual datasets comprising samples from crawled web pages, books, Wikipedia pages, public code repositories, and so forth. In order to enhance the capabilities of these base models further, such as in the creation of chatbots, a supervised fine-tuning process is required. This process relies on a dataset comprising questions and answers. The exposure of base models to these instructions has enabled LLMs to demonstrate their aptitude for various language-related tasks, including text synthesis, translation, summarisation, question answering, and sentiment analysis. This has been achieved by leveraging deep learning techniques and extensive datasets. The development of AI has been enabled by these capabilities, which have also facilitated the implementation of AI in a variety of fields, including but not limited to healthcare, education, law and finance [47, 79].

2.4 Summary

This chapter provides an overview of the evolution of language modelling, starting from the basic statistical methodologies, passing through the rise of RNNs, as well as some core principles in

Algorithm 4: $\hat{\theta} \leftarrow$ `DecoderTraining`$(x_{1:N}, \theta)$

Input: A set of N token sequences $\{x_n\}_{n=1}^{N}$
Input: Initial (randomly initialised) transformer parameters θ
Output: Trained transformer parameters $\hat{\theta}$
Hyperparameters: Number of training epochs $N_{\text{epochs}} \in \mathbb{N}$, Learning rate $\eta \in \mathbb{R}_0^+$

for $i = 1, 2, \ldots, N_{epochs}$ **do**
  for $n = 1, 2, \ldots N$ **do**
    $\ell \leftarrow \text{length}(x_n)$
    $P(\theta) \leftarrow$ `DecoderTransformer`$(x_n \mid \theta)$
    $\text{loss}(\theta) = -\sum_{t=1}^{\ell-1} \log P(\theta)[x_n[t+1], t]$
    $\theta \leftarrow \theta - \eta \cdot \nabla \text{loss}(\theta)$
  end
end
return $\hat{\theta} = \theta$

NLP, like tokenisation and word embeddings. One of the most important milestones in the history of NLP and Language Modelling was the introduction of the Transformer architecture [72], which nowadays is the building block of large language models. Section The Transformer outlines the main principles behind it, introducing the concepts of Attention and Position-wise Feed Forward Network. The chapter contains high-level descriptions and pseudocode for the main algorithms used in the transformer, namely the implementation of the decoder-only transformer and its training process.

Chapter 3

Literature Review

This chapter presents the systematic literature review conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework [53]. This method ensures a transparent and replicable review process, clarifying the motivations, methodological process, and selection criteria applied throughout the review.

3.1 Motivation and Objectives

The application of artificial intelligence, particularly through LLMs, for analysing Dark Web data holds transformative potential. While significant advancements have been made in using NLP and Machine Learning (ML) techniques to process large-scale unstructured data, the unique challenges of Dark Web data remain underexplored. In light of the high-stakes context in which this data is created, the need for timely and actionable intelligence is crucial for law enforcement and cybersecurity professionals.

This review aimed to identify, collect, and assess existing literature that employs Dark Web data and applies LLMs for Threat Intelligence. To ensure a rigorous and systematic selection process, the PRISMA methodology was adopted to structure the search, screening, and inclusion of relevant studies. Figure 3.1 presents the PRISMA flow diagram, providing a visual overview of each stage in the review process and detailing the progression of articles through identification, screening, and final inclusion stages.

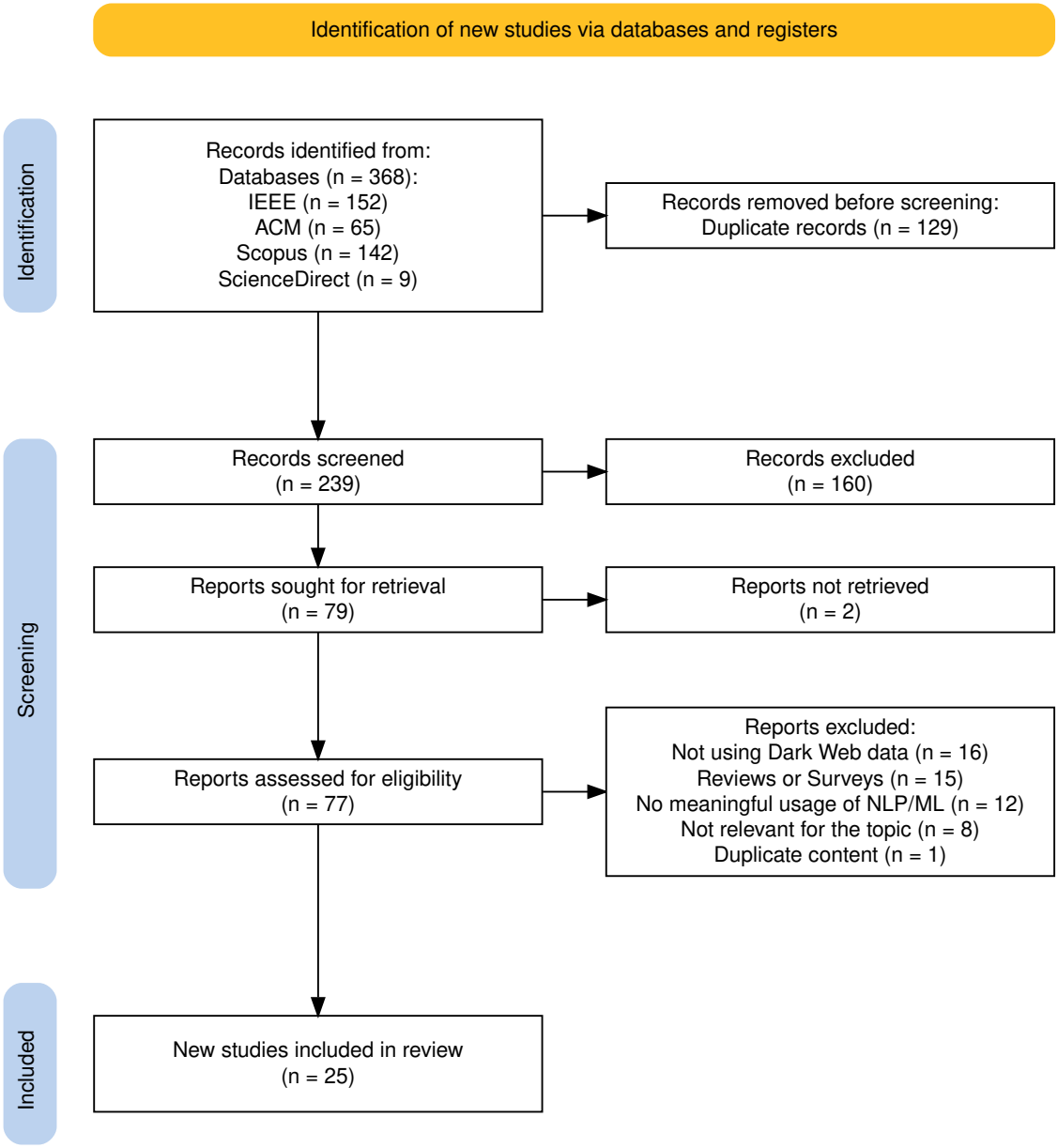


Figure 3.1: PRISMA Flow Diagram (created using [32])

3.2 Identification Process

The goal of the identification process of the PRISMA framework is to collect documents to be analysed in the Screening Phase. The following sections list the sources used, the process for creating the search queries, the filtering steps performed for each source and some quantitative insights about the results of this phase.

3.2.1 Sources

The selected sources were IEEE Xplore, ACM Digital Libraries, together with Scopus and ScienceDirect scientific bibliographic databases. The table below provides an overview of the primary information pertaining to these sources.

Table 3.1: Sources Details

Source	URL	Publisher	Access Date
IEEE Xplore	https://ieeexplore.ieee.org/	IEEE	24/out/2024
ACM Digital Library	https://dl.acm.org/	ACM	24/out/2024
Scopus	https://www.scopus.com/	Elsevier	24/out/2024
ScienceDirect	https://www.sciencedirect.com/	Elsevier	24/out/2024

3.2.2 Search Query Refinement Process

This literature review encompasses three main topics: Dark Web, Threat Intelligence, and Large Language Models. With the goal of broadening the results obtained and covering as much research as possible, the three topics were initially combined into three pairs, each involving two topics, resulting in three initial meta-query definitions:

- **Meta-query 1:** Dark Web AND Threat Intelligence
- **Meta-query 2:** Dark Web AND Large Language Models
- **Meta-query 3:** Threat Intelligence AND Large Language Models

A refinement process was then conducted against the IEEE Xplore Digital Library using the field *All Metadata* for each meta-query. IEEE was chosen for its broad topic coverage, making it particularly suitable for the initial identification process.

3.2.2.1 Search Query 1 Refinement

The initial translation of the meta-query 1 was to the search query *(“dark web”) AND (“threat analysis” OR “threat intelligence“)* (SQ1 V1). This yielded a total of 64 results. A review of the literature revealed that different authors use the terms “dark web”, “dark net”, “darkweb” and “darknet” as equivalent terms. Incorporating these three new terms increased the number of results to 107 documents. A further analysis revealed that some authors use the terms “threat detection” and “crime detection” as a synonym for “threat intelligence”. The inclusion of these terms in the query resulted in 20 more documents. The final query was *(“dark web” OR darkweb OR darknet OR “dark net”) AND (“threat analysis” OR “threat intelligence” OR “threat detection” OR “crime detection”)* (SQ1 V3). An overview of the process is detailed in Table 3.2.

Table 3.2: Search Query 1 Evolution

Version	Search Query	Results
SQ1 V1	(“dark web”) AND (“threat analysis” OR “threat intelligence”)	64
SQ1 V2	(“dark web” OR darkweb OR darknet OR “dark net”) AND (“threat analysis” OR “threat intelligence”)	107
SQ1 V3	(“dark web” OR darkweb OR darknet OR “dark net”) AND (“threat analysis” OR “threat intelligence” OR “threat detection” OR “crime detection”)	127

3.2.2.2 Search Query 2 Refinement

The first iteration of this second query combined the Dark Web refined terms from SQ1 with *large language models* (without quotation marks) (SQ2 V1). This first version outputs only 16 results. Consequently, following an analysis of the title and abstract of several documents, it became evident that the term *large* is somewhat limiting. Accordingly, the term was removed and both a singular and plural form (between quotation marks) for language models was added, as well as the term *transformer*, which is commonly associated with LLMs. This increased the number of results to 53. The full evolution of the second query is outlined in Table 3.3.

Table 3.3: Search Query 2 Evolution

Version	Search Query	Results
SQ2 V1	(“dark web” OR darkweb OR darknet OR “dark net”) AND (large language models)	16
SQ2 V2	(“dark web” OR darkweb OR darknet OR “dark net”) AND (“language model” OR “language models” OR transformer)	53

3.2.2.3 Search Query 3 Refinement

To convert meta-query 3 to a concrete search query, the insights derived from the two preceding refinement stages allowed for a direct formulation of the third search query: *(“threat analysis” OR “threat intelligence” OR “threat detection” OR “crime detection”) AND (“language model” OR “language models” OR transformer)*, yielding 123 results. It was noticeable, however, that a considerable number of articles were not, in fact, related to the Dark Web, which is the main topic of this review. For this reason, the search query was expanded to include this topic, resulting in a narrower list of results comprising 10 documents. The incremental steps of this process are illustrated in Table 3.4.

3.2.2.4 Final Queries

As a result of the refinement in the three preceding sections, the final queries to be executed in the remaining sources were as follows:

Table 3.4: Search Query 3 Evolution

Version	Search Query	Results
SQ3 V1	(“threat analysis” OR “threat intelligence” OR “threat detection” OR “crime detection”) AND (“language model” OR “language models” OR transformer)	123
SQ3 V2	(“dark web” OR darkweb OR darknet OR “dark net”) AND (“threat analysis” OR “threat intelligence” OR “threat detection” OR “crime detection”) AND (“language model” OR “language models” OR transformer)	10

- **SQ1:** (“dark web” OR darkweb OR darknet OR “dark net”) AND (“threat analysis” OR “threat intelligence” OR “threat detection” OR “crime detection”)
- **SQ2:** (“dark web” OR darkweb OR darknet OR “dark net”) AND (“language model” OR “language models” OR transformer)
- **SQ3:** (“dark web” OR darkweb OR darknet OR “dark net”) AND (“threat analysis” OR “threat intelligence” OR “threat detection” OR “crime detection”) AND (“language model” OR “language models” OR transformer)

3.2.3 Source Filtering Process

Each academic database source was queried and filtered based on specific criteria and the capabilities of individual search engines to ensure the retrieval of relevant results. All sources were filtered to return all documents from 2010 onwards. The remaining search and filtering process applied to each specific source, according to an evaluation of its capabilities, and with the goal of accelerating the process, is outlined below.

IEEE Xplore The search was conducted across all metadata. The content type was limited to “Conferences” and “Journals”. The results were further refined by selecting specific topics using IEEE’s checkbox filters:

- **Search Query 1:** Dark Web, Darknet, Threat Intelligence, Threat Detection, Cybercrime, Cybersecurity, Machine Learning.
- **Search Query 2:** Dark Web, Darknet, Language Model, Pre-trained Language Models, Transformer Model, Text Classification.
- **Search Query 3:** Dark Web, Threat Intelligence, Threat Detection, Machine Learning, Language Model, Pre-trained Language Models.

ACM Digital Library The search was limited to the Title and Abstract fields and expanded to The ACM Guide to Computing Literature. In this source, the Title and Abstract are individual fields and not merged by the search engine. Therefore, a Python script was created to facilitate

the construction of the advanced search string, accounting for all possible scenarios in which the search queries' *AND* terms appear in either one or both search fields. No further content or subject filters were applied within ACM.

Scopus In Scopus, the search was conducted across the article title, abstract and keywords fields. Filters were applied to restrict the results to "Conference paper" and "Article" content types within the "Computer Science" and "Engineering" subject areas. Furthermore, the language filter was set to "English".

ScienceDirect In the case of ScienceDirect, the search was conducted within the title, abstract, and author-specified keywords. Only "Research article" and "Review Articles" items within the "Computer Science" and "Engineering" subject areas were included.

The filtering process reduced the number of results from 446 to 368 documents, thereby reducing the volume of data for analysis. Table 3.5 presents the before and after filtering numbers for each source, demonstrating the effectiveness of the applied filtering criteria.

Table 3.5: Search Query Results Before and After Applying Filters

	IEEE		**ACM**		**Scopus**		**ScienceDirect**		**Total**	
	Before	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**
SQ1	127	100	37	36	136	107	5	5	305	248
SQ2	53	42	27	26	39	30	4	4	123	102
SQ3	10	10	3	3	5	5	0	0	18	18
Total	190	152	67	65	180	142	9	9	446	368

3.2.4 Duplicate Removal

Following the execution of search queries, each query result from each source was exported in BibTeX format and uploaded into Zotero[1] for organisation.

Duplicate removal was conducted using Zotero's built-in duplicate detection functionality, which identified 87 duplicate groups, encompassing 215 duplicates. Each identified duplicate was then verified by comparing the title, authors, and year to confirm its accuracy, and all identified items were validated as duplicates.

To ensure a swifter duplicate management process, the Zotero plugin Zoplicate[2] was used to enable the removal of duplicates in bulk. However, two groups of documents required manual merging due to discrepancies in the Zotero Item Type classification within the duplicate groups. In these instances, one item from each group, sourced from Scopus, was classified as a "Journal Article", whereas ACM, the original publisher, classified it as a "Conference paper". Manual

[1] https://www.zotero.org/
[2] https://github.com/ChenglongMa/zoplicate

merging was completed by designating the ACM entry as the primary item to maintain source accuracy.

Moreover, after manually scanning all remaining items, two items referring to the same article were identified, even though Zotero could not mark them as duplicates due to a different character in the titles. The manual merging of these two items increased the number of identified duplicate groups to 88, with a total of 217 duplicates. A total of 129 duplicated records were deleted.

Following the duplicate removal process, 239 unique items remain, including journal articles, conference papers, books, and one Ph.D. thesis, as detailed in Table 3.6.

Table 3.6: Item Type Overview as Identified by Zotero

Item Type	**Count**
Conference Papers	154
Journal Articles	79
Books	5
Thesis	1
Total	239

In summary, a rigorous identification process combining search query refinement, filtering, and duplicate removal resulted in 239 unique articles. These studies form the foundation for the subsequent screening phase.

3.3 Screening Process

The screening phase applied a systematic exclusion process to the initial 239 records identified through the search and filtering stages. This phase ensures that only the most relevant studies, those meeting the eligibility criteria and related directly to the primary research objectives, progress to subsequent stages of the review.

3.3.1 Eligibility Criteria

To ensure the review included solely the most pertinent studies that align with the subject under examination, and maintaining a focus on reliable and structured content, the filtering criteria, both exclusive and inclusive, were defined and enumerated in Table 3.7.

The review process output should only include documents that use data extracted from the Dark Web. The primary interest was on research that analyses extensive textual data from the Dark Web, for example, from anonymous marketplaces or community forums, and later applies LLM tools and techniques to analyse and extract relevant information for security and defence. Surveys were excluded to focus on empirical studies with actionable methodologies. However, an initial screening revealed that limiting the research to LLM usage would result in a significantly reduced collection of references. Therefore, traditional machine learning and smaller language models for text classification were also included.

Table 3.7: Eligibility Criteria for Exclusion and Inclusion in the Review

Criteria No.	Description
1	Results obtained from search queries in defined sources
2	Published after the year 2010
3	Written in English
4	Not duplicated
5	Document is a Journal Article or Conference Paper
6	Document is not a Survey or a Literature Review
7	Full text available
8	Uses data extracted from the Dark Web
9	Explores ML and/or LLMs for Threat Intelligence

3.3.2 Screening phase

After establishing clear and objective inclusion and exclusion criteria and removing duplicate work, the following steps were taken to identify the items that aligned with the scope of this review:

Exclusion of Non-Relevant Item Types Items classified as books and the single Ph.D. thesis were excluded due to irrelevance to the review's focus on journal articles and conference papers. Zotero classified six records as books and one as a thesis; these were identified through ACM, which does not allow content-type filtering in search results. Moreover, these items were inaccessible despite valid institutional credentials, further limiting their eligibility. This initial exclusion reduced the total records to 233.

Title and Abstract Analysis Following the removal of non-relevant item types, each remaining record's title and abstract were carefully reviewed to assess relevance to the review's primary focus on using dark web data and NLP or LLM techniques for extracting Threat Intelligence or classifying textual content. This analysis led to the exclusion of 154 items, categorised as follows:

- **30 records** were excluded for being only tangentially related to the review's topics, lacking explicit relevance to dark web data or NLP applications.
- **9 records** were excluded for using only surface web data (e.g., social media platforms such as Twitter) rather than data from the dark web.
- **115 records** were excluded due to their focus on dark web content that does not meet the eligibility criteria for data origin or analysis techniques. These studies primarily analyse traffic data, IP addresses, or IoT-related dark web traffic, rather than leveraging text-based data from anonymous marketplaces or applying NLP techniques.

After title and abstract screening, 79 records remained for further evaluation.

Full Text Access In this stage, Zotero's full-text search capability was used to obtain PDFs of the remaining studies. With this functionality, 36 articles were accessible directly through Zotero, while 41 were manually located and downloaded from the sources. Two items were not available for download and were, therefore, excluded.

This entire process resulted in the exclusion of 162 items from the initial 239, leaving 77 records for the next review phase.

3.4 Eligibility Phase

A total of 77 papers were fully read and assessed for eligibility. Through a systematic reading process, papers were excluded based on the criteria defined in Table 3.7. Following this exclusion process, only 25 papers remained for analysis.

For each eligible paper, an initial reading was conducted to extract general aspects of the paper (briefly summarised in A.1). Subsequently, a second round of reading and information extraction was performed to categorise the items more effectively. This step grouped documents based on the analysis topic, their technical approach, differentiating between traditional machine learning methods and transformer-based techniques, and the main task addressed by the paper, such as binary classification, clustering/categorisation, jargon detection, or entity recognition.

Table 3.8 presents the summarised phases of information collection.

Table 3.8: Information Collection Phases Summary

Phase	Focus of Analysis	Details
First Reading	General information	Problem statement Dataset collection Proposed solution Results Limitations
Second Reading	Detailed categorisation	Topic, subtopic and sub-subtopic Technical approaches Main task

3.5 Included Studies & Discussion

This section presents a brief analysis of the 25 resulting research papers identified throughout the review following the PRISMA framework based on the extracted information details listed in Table 3.8.

The analysis revealed a strong focus on addressing Dark Web-related crime. Specifically, three main areas of research were identified:

- **Cross-domain Threat Intelligence**: aggregates research without a specific topic of analysis. Researchers use a wide variety of datasets and sources, and often try to identify illegal

activities based on a discussion post in a Dark Web forum or a product listing in a Dark Web marketplace;

- **Terrorism Threat Intelligence**: two items have been identified as having the main objective of analysing terrorism and radical forums;
- **Cyber Threat Intelligence**: comprises a wide variety of content related to cybersecurity matters, namely hacking, software vulnerabilities or data leaks.

Regarding the main tasks addressed by the papers, a significant portion of the research focused on binary classification, where the primary objective was to classify text (e.g. "legal" vs "illegal"), or to detect whether the text mentioned specific topics such as illicit products or services. Similarly, clustering or categorisation was a prevalent task, where papers sought to group content into meaningful clusters or categories, such as types of illegal activities or discussion topics within forums. One study addressed jargon detection, aiming at identifying Chinese jargon terms commonly used in Dark Web marketplaces or forums, which can pose challenges for automated analysis. Another study concentrated on named entity recognition (NER) to extract key cybersecurity-related entities. Table 3.9 further specifies some topics of analysis and identifies which topics the papers focus on and the associated main task.

Table 3.9: Identified Threat Intelligence Topics and Respective References Grouped by the Main Task (sorted ascendingly by year)

Threat Intelligence Topic	**Main Task**			
	Binary Classification	Clustering	JD*	NER**
Cross-Domain	[52]	[20] [48] [12] [38] [62] [55] [61]	[42]	
Terrorism				
Islamic Radical Forums		[60]		
General	[9]			
Cyber				
Broad Cyber				
Data leaks/breaches	[1]			
Others	[23][5][70]	[14]		[71]
Vulnerabilities/Exploits				
At-risk system components		[50]		
Exploits		[77]		
Vulnerabilities	[66]			
Hacking				
Malicious hacking	[51]			
Hackers		[35][54]		
Malware	[41][40]			

* Jargon Detection, ** Named Entity Recognition

Furthermore, the collection can be divided into two main categories from a different and more technical perspective. On the one hand, 19 records applied/compared some kind of traditional machine learning algorithm. On the other hand, 12 records used more recent transformer-based technologies. Table 3.10 points out a trend in adopting both categories. From 2016 to 2021, only traditional ML algorithms were applied to analyse Dark Web text. However, from 2022 onwards,

the majority of the research studied the effectiveness of language models for text classification, BERT [21] being the distinguished preference.

Table 3.10: Main Approaches Used in the Selected Papers

Reference	Year	Transformers/LMs	Other ML Algorithms
[51]	2016		Co-training, LP, LR, RF, SVM
[50]	2018		DT, LR, NB, RF, SVM
[66]	2018		RF, SVM
[41]	2019		MLP
[77]	2019		DT, NB, RF
[1]	2020		GB, LR, RF
[23]	2020		BiLSTM, TSVM
[40]	2020		K-Means
[20]	2021		RF
[35]	2021		LDA
[48]	2021		DT, K-Means, LR, NB, RF, SVM
[12]	2022	BERT, RoBERTa, ULMFit	LSTM
[60]	2022	BERT, ELECTRA, ERNIE, XLNet	K-Means, SVM
[5]	2023	LongForm	
[14]	2023	ConGAN-BERT	
[38]	2023	DarkBERT	
[52]	2023	BERT, ELECTRA, ERNIE, XLNet	Neural Networks
[54]	2023		K-Means, LDA
[62]	2023	BERT	CNN, LSTM, RNN
[71]	2023	Fine-tuned BERT	
[9]	2024	BERT	DT, KNN, RF, SVM
[42]	2024	DC-BERT	
[55]	2024	BERTopic	
[61]	2024	BERT	GB, KNN, LR, LSTM, RF, VC
[70]	2024		DT, GB, LR, RF

BiLSTM: Bi-directional Long Short-Term Memory; CNN: Convolutionary Neural Network; DT: Decision Tree; GB: Gradient Boosting; KNN: *K*-nearest Neighbour; LDA: Latent Dirichlet Allocation; LP: Label Propagation; LR: Logistic Regression; MLP: Multi-Layer Perceptron; NB: Naive Bayes; RF: Random Forest; RNN: Recurrent Neural Network; SVM: Support Vector Machine; TSVM: Transductive SVM; UMLFit: Universal Language Model Fine-tuning; VC: Voting Classifier

3.6 Summary

This literature review has systematically examined the current state of research on the application of artificial intelligence to analyse Dark Web data. The review was conducted using the PRISMA framework [32] to ensure a transparent and replicable process, encompassing the identification, screening, and inclusion of relevant studies.

The identification process involved querying four academic databases, including IEEE Xplore and ACM Digital Library. A rigorous search query refinement and filtering process narrowed

the initial pool of 446 documents to 239 unique articles. These articles were then subjected to a systematic screening phase, applying strict eligibility criteria. This process resulted in 77 records, which were further evaluated for eligibility, ultimately selecting 25 papers for detailed analysis.

The included studies were categorised into three main areas of research: cross-domain TI, terrorism TI, and cyber TI. The primary tasks addressed by these studies included binary classification, clustering/categorisation, jargon detection, and named entity recognition. The analysis revealed a strong focus on addressing DW-related crime, with a significant portion of the research employing traditional machine learning algorithms. However, from 2022 onwards, there has been a noticeable shift towards the adoption of transformer-based technologies, particularly BERT, for text classification tasks.

Chapter 4

State of the Art

The analysis of Dark Web data has become increasingly critical in the fields of cybersecurity and law enforcement. The Dark Web, with its vast and often illicit content, presents unique challenges for data extraction and analysis. Recent advancements in machine learning and natural language processing have opened new avenues for effectively tackling these challenges. This chapter is based on the findings from the previous section, providing a comprehensive review focused on the main tools and techniques employed, grouping the work according to the primary topics of analysis.

4.1 Cross-Domain Threat Intelligence

The Dark Web is a complex and multifaceted environment where illicit activities often thrive, posing significant challenges for law enforcement and cybersecurity professionals. Several authors focus on a cross-domain threat intelligence, aimed at addressing these challenges by leveraging advanced analytical techniques to detect, monitor, and combat illegal activities across various Dark Web platforms. This section delves into the state of the art in cross-domain TI, exploring the traditional machine learning methods, transformer-based models, and combined approaches developed to extract actionable intelligence from the Dark Web.

4.1.1 Traditional Machine Learning Techniques

Traditional machine learning techniques have long been employed to analyse and extract insights from data. In 2018, the European Union provided financial support for the ANITA project (Advanced tools for fighting oNline Illegal TrAfficking)[1], which comprised a suite of tools designed to facilitate the monitoring and combating of illicit criminal activities on the Dark Web [20]. The project involved 17 international participants from both the academic and enterprise sectors[1], who introduced various Big Data analytical tools for analysing data extracted from illegal marketplaces. The objective was to detect trends and extract actionable information on purchasing habits, transaction patterns, and user behaviours.

[1] https://cordis.europa.eu/project/id/787061

The ANITA project had two main objectives: firstly, to enhance and accelerate the investigative processes of law enforcement agencies and, secondly, to enhance their operational capabilities through the utilisation of innovative tools designed to address the challenges associated with online illegal trafficking, including online data source analysis, blockchain analysis, big data analytics, and knowledge modelling. As illustrated in the diagram in Figure 4.1, the ANITA system comprises five main components: (1) Source Analysis Toolbox to detect and collect potential sources; (2) Big Data Neural Network-Based Analysis Box for automatic categorisation, entity extraction and illegal trafficking trend analysis; (3) Knowledge Management and Reasoning Box for semantic inference tasks and criminal network construction; (4) Investigation Application Box for real-time data analysis and results providing; (5) Integration of Human Factors into the Analysis Loop that incorporates user feedback and aims at decreasing the training process of new investigators.

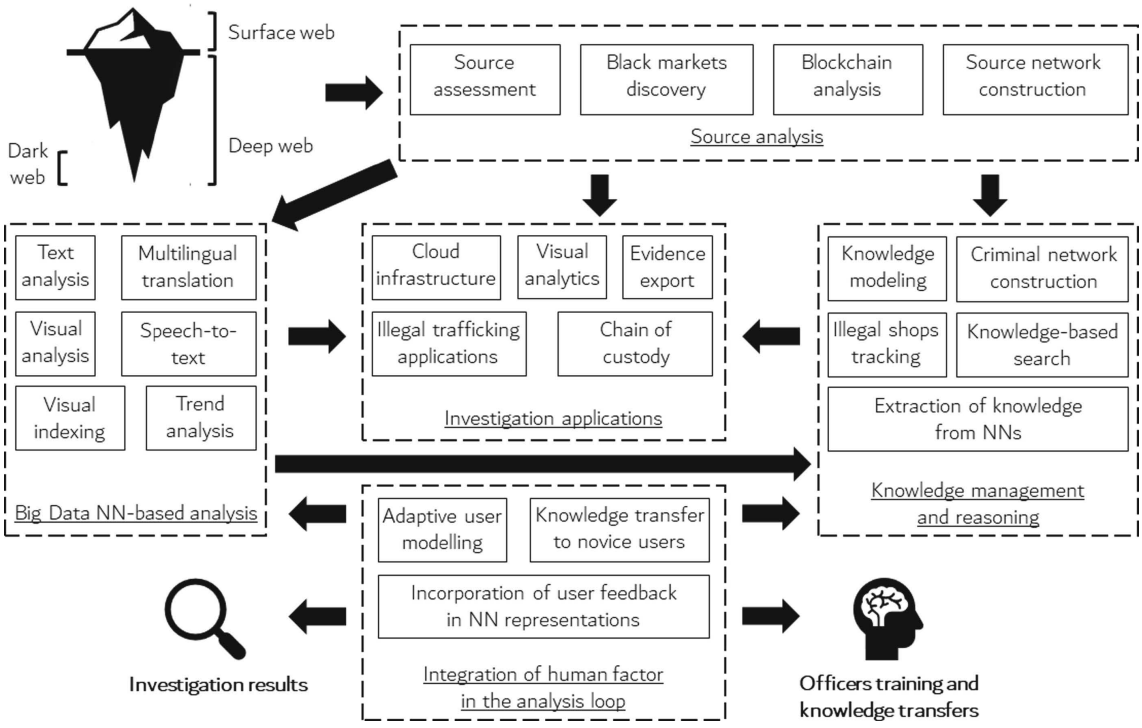


Figure 4.1: Graphical Representation of ANITA [20]

The ANITA system demonstrated an accuracy of 81.66% in predicting web pages containing illegal activities, employing Random Forest for the classification of activities into three categories: *Suspicious*, *Unknown*, and *Normal*. Furthermore, the system attains an accuracy rate of 66.25% in classifying 26 specific activity types, including drugs, hacking, forums, social networks, violence, fraud, and counterfeit money.

The majority of methods for analysing Dark Web forums are dependent on the continuous labelling of data, which presents significant challenges due to the high costs involved and the infeasibility of obtaining such labelling from the intrinsic nature of the data itself [48]. To address these limitations, Nazah et al. [48] proposed an unsupervised model that combines the K-means

clustering algorithm with Decision Tree methods to identify and characterise forums on the Dark Web. Their approach achieved an F1-score of 98%.

4.1.2 Transformers and Language Models

Transformers and language models represent a significant advancement in natural language processing, offering powerful tools for analysing complex and diverse data. However, a significant challenge in Dark Web data analysis is the extreme lexical and structural diversity of the language used, which differs considerably from that of the Surface Web [39]. This divergence renders the construction of effective representations for Dark Web data a challenging endeavour, thereby necessitating specialised models. To address this, Jin et al. [38] introduced DarkBERT in 2023, a language model that has been explicitly pretrained on Dark Web data.

The authors selected RoBERTa [45] as the base architecture, choosing this model because it excludes the Next Sentence Prediction task, which is typically part of BERT pretraining but may not be as helpful for the Dark Web, where sentence structures are often less conventional. DarkBERT adopts the same Byte Pair Encoding tokenisation vocabulary as RoBERTa, with each page in the pretraining corpus being separated using RoBERTa's separator token (</s>).

DarkBERT was evaluated across a range of threat intelligence tasks, including detecting underground activities, and outperformed existing pretrained language models. In classification tasks, DarkBERT exhibited superior results in the Dark Web domain relative to other general-purpose models, including BERT and RoBERTa.

In the context of ransomware leak site detection, the model demonstrated an ability to accurately identify pages associated with ransomware groups. With regard to noteworthy thread detection, DarkBERT exhibited superior performance in terms of precision, recall, and F1-scores compared to other models. Nevertheless, its practical deployment demonstrates the potential for further enhancements by incorporating supplementary training data and attributes, such as authorship information. In the context of threat keyword inference, DarkBERT exhibited superior performance in predicting specific terminology, outperforming models such as $BERT_{Reddit}$. The model demonstrated proficiency in identifying subtle nuances, such as drug-related euphemisms (e.g., "Tesla" or "Champagnewhich"), which are often overlooked in surface web datasets.

Jargon Detection The use of jargon in criminal conversations is prevalent in underground markets, particularly as a strategy to evade law enforcement surveillance, creating significant challenges for criminal investigations [42]. In response to this issue, Ke et al. [42] designed an unsupervised cross-domain adaptation framework for Chinese jargon detection, known as the CJD-Framework, which is integrated with a pretrained BERT model. The authors collect a dataset of Chinese darknet corpus (DC-dataset) from six underground Chinese markets and reported a detection accuracy of 91.5%. They justify the success with a new word discovery technology based on information entropy and mutual information.

In 2024, Pastor-Galindo et al. [55] put forth a comprehensive end-to-end scalable architecture designed for the early detection of emerging TOR onion services. The authors' work was based

on the model BERTopic [31] to extract topics from the textual content of 72,045 onion services. BERTopic is an unsupervised machine learning model that uses transformer-based embeddings for topic modelling. This approach enables the identification of coherent and interpretable topics in extensive textual corpora, making it especially suited to analysing the diverse and dynamic content found in onion services. The system was able to connect to and characterise 90% of new Dark Web websites into 35 distinct low-level clusters and 11 high-level topics.

4.1.3 Combined Approaches

With the purpose of (1) detecting illicit content on the Dark Web, and (2) identifying various types of drugs, Cascavilla, Catolino, and Sangiovanni [12] compiled an heterogeneous dataset consisting of 113,995 onion sites and Dark Web marketplaces. Subsequently, the authors compared a number of pretrained transferable models, including ULMFit (Universal Language Model Fine-tuning) [34], BERT, and RoBERTa, with traditional text classification models, such as Long Short-Term Memory (LSTM) neural networks. BERT outperformed all other models in both tasks, achieving an average 94% accuracy in the two classification tasks.

In 2023, to investigate the potential of syntax-based neural networks as an alternative to Transformer models for tasks involving previously unseen sentences, Onorati et al. [52] showed that syntax-based models can rival Transformers, even after fine-tuning and domain adaptation. The authors argue that, while Transformer-based models like BERT can achieve high accuracy on pretrained tasks with extreme domain adaptation, syntax-based models offer greater transparency, fewer parameters and perform well on novel data without the need for extensive fine-tuning.

The results indicate that syntax-based and lexical-based neural networks outperform Transformer models, even those that have been fine-tuned, in tasks involving sentences that have never been seen before by the models (such as classifying legal and illegal content on the Dark Web).

In [62], the researchers evaluate the efficacy of four distinct architectures (RNN, CNN, LSTM, and BERT) in classifying illicit activities related to cybercrimes on Dark Web forums. The study uses the Agora dataset from the DarkNet Market Archive [11], which enumerates 109 activities classified into three categories: “Cybercrime”, “Not Cybercrime”, and “Can’t say if cybercrime”. The BERT-based model achieved a high accuracy of 96%, significantly outperforming the other models. CNN achieved the lowest accuracy of 73%, while RNN scored 88%.

In addition, Sangher, Singh, and Pandey [61] employ data from six marketplaces to analyse interactions in the form of user remarks and comments, with the objective of identifying popular social media platforms and the intent behind discussions. The study reports that BERT and LSTM models achieved high accuracy, with LSTM performing best with 91.35% accuracy for multi-class classification of social media usage intentions in hackers’ discussions. Subsequently, latent topics were extracted from the text and cybercrime instances were divided into four categories: (1) hacking, (2) drug, (3) fraud, and (4) phishing.

4.2 Terrorism Threat Intelligence

Terrorism threat intelligence involves the analysis of data to identify and mitigate potential terrorist activities. Ranaldi et al. [60] addressed the challenge of authorship attribution (AA) within the context of an English-language Islamic forum that hosts discussions on Islam and its global implications by proposing an AA task that analyses user posts' syntactic and stylistic features. They tested a variety of transformer-based models, including BERT, XLNet [75], ERNIE [78], and ELECTRA [16], alongside stylistic and syntactic classifiers such as KERMIT, a kernel-inspired encoder utilising parse trees [76]. Their findings demonstrate that syntactic and stylistic models can outperform transformers on that specific dataset. This outcome was attributed to the partial stylistic nature of the dataset.

In 2024, Biagio et al. [9] presented MARPLE, an AI-driven threat intelligence framework that leverages deep learning techniques to assist intelligence analysts in identifying terrorist threats on social media, blogs, and forums. The framework is multilingual and capable of detecting and classifying malicious posts in multiple languages (English, Polish, Slovenian, Italian, Spanish/Basque). The research integrates BERT as the core text analysis algorithm. In one experiment, a dataset of 1,100 posts gathered from various platforms and validated by law enforcement agencies was used to simulate real-world conditions. The highest performance was achieved with 99% accuracy and 95% recall.

4.3 Cyber Threat Intelligence

Cyber threat intelligence (CTI) is crucial for identifying and responding to cyber threats. This section provides an overview of the techniques and models used to analyse cyber threat data, including the use of traditional machine learning and advanced language models. It discusses the importance of CTI in protecting against cyberattacks and the ongoing efforts to enhance the accuracy and timeliness of threat detection.

4.3.1 Broad Cyber Threat Intelligence

Traditional Machine Learning Techniques Broad cyber threat intelligence encompasses a wide range of activities and data sources to provide a comprehensive view of the cyber threat landscape. The application of traditional ML techniques has been demonstrated to yield effective results in the identification and analysis of cyber threats. This provides analysts with valuable insights into the patterns and attack vectors associated with such threats. For example, Adewopo, Gonen, and Adewopo [1] conducted an exhaustive examination of over 10 billion breached records from 8,000 cases documented in the United States between 2005 and 2019, relying on data from the Privacy Rights Clearinghouse (PRC). The study used three classification algorithms, namely Random Forest, Logistic Regression, and Gradient Boosting classifiers. The RF algorithm demonstrated the most optimal performance, achieving an F1-score of 0.81. The findings suggest that

cyberattacks are becoming increasingly prevalent in medical institutions and business organisations, with the healthcare sector bearing the greatest financial burden.

Building on these efforts, Ebrahimi, Nunamaker, and Chen [23] combined Transductive SVM (TSVM) paired with an LSTM network for classification. The methodology was evaluated on 79,000 product listings from underground marketplaces, achieving an F1-score of 89.55%. The study provided examples of threats such as carding tools and breached bank account data, which the model successfully identified with high confidence.

Additionally, Varbanov [70] explored the use of advanced ML algorithms for classifying cyber threats within datasets derived from hacker forums and the Agora Dark Web Marketplace dataset. Following the implementation of TF-IDF vectorisation, the researchers evaluated four classification models, namely Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting. They concluded that Logistic Regression had the highest F1-score of 0.87.

Transformers and Language Models In light of the challenges inherent in the efficient use of open source threat intelligence feeds (OTIFs) and Dark Web data for enhanced cybersecurity, recent research has demonstrated the significant potential of transformer-based models in threat identification and actionable intelligence extraction. For instance, Balasubramanian, Seby, and Kostakos [5] extended transformer-based methodologies by integrating LASER embeddings[2] and FAISS indexing[3] into a semantic-driven focused crawler. The system effectively retrieved relevant web pages containing CTI, with an average harvest rate of 0.87 and an irrelevance ratio of 0.13. The data was classified using a fine-tuned Longformer model [6], yielding a classification accuracy of 96%.

In a related development, Cherqi et al. [14] proposed ConGAN-BERT, which builds upon the GAN-BERT [18] framework. This proposed enhancement incorporates self-supervised contrastive learning to address class overlap in threat classification tasks. ConGAN-BERT demonstrated superior performance compared to BERT and GAN-BERT, achieving an F1-score improvement of 3–12% across diverse OTIF datasets.

Similarly, Varghese, S, and Kb [71] fine-tuned a BERT model on the DNRTI dataset [73] with the objective of performing named entity recognition on crawled Dark Web data. The system identified entities such as hacker IDs, tools, and organisations with high precision, recall, and F1-scores. BERT's contextual understanding enabled it to distinguish semantically similar terms, such as "bank" in different contexts, which traditional word embeddings could not achieve. The fine-tuned model achieved substantial improvements over traditional NER systems, enabling actionable insights like identifying new malware or data breaches.

4.3.2 Exploits and Vulnerabilities

Recent research has identified several techniques for extracting actionable intelligence on exploits and vulnerabilities from data sources on the Dark Web. This subsection presents three notable

[2] https://github.com/facebookresearch/LASER
[3] https://github.com/facebookresearch/faiss

contributions to this field, focusing on applying machine learning and reasoning systems to Dark Web forums and marketplaces.

At-Risk Systems Components In their study, Nunes, Shakarian, and Simari [50] put forth a reasoning system that integrates Defeasible Logic Programming with machine learning classifiers. This system is designed to identify at-risk systems based on hacker discussions observed on Dark Web platforms. The study conceptualises this task as a multi-label classification problem and evaluates the performance of a range of classifiers. A significant challenge was posed by the imbalance of label representation, whereby vendors and products with limited training data resulted in misclassification. Nevertheless, by utilising a reduced label set, the authors achieve considerable enhancements in precision, ranging from 15% to 57%, while maintaining comparable recall.

Exploit Detection In [77], the authors propose descriptive and predictive analytics to extract cyber threat intelligence from Dark Web forum posts. The study uses IBM Watson Analytics[4] for data trends and WEKA machine learning tools [26] for classification to identify the types of exploits discussed in forum posts. Among the three machine learning models evaluated, Random Forest achieves the highest accuracy of 78.27%, outperforming Random Tree (67.47%) and Naive Bayes (52.81%). Notably, Random Forest excels in detecting system exploits with 97.87% accuracy.

Vulnerabilities In 2018, Tavabi et al. [66] introduced DarkEmbed. This neural language modelling approach incorporates distributed representations of Dark Web and Deep Web discussions to predict whether vulnerabilities will be exploited by capturing linguistic patterns in hacker discussions. The F1-score achieved by DarkEmbed on the exploit prediction task was 0.74.

4.3.3 Hacking

Identifying and analysing malicious hacking activities and the influential individuals who perpetrate them on Dark Web forums are critical for developing cyber threat intelligence. This subsection reviews three notable contributions in this field, focusing on leveraging machine learning, social network analysis, and semantic analysis to address these challenges.

In 2016, Nunes et al. [51] focused on classifying relevant products and forum topics as a binary classification task by evaluating supervised methods, including Naive Bayes, Random Forest, Support Vector Machine, and Logistic Regression. The results highlighted the superiority of SVM with a linear kernel in the supervised setting, achieving 87% recall and 85% precision for relevant products.

In 2021, an alternative approach was presented by Huang et al. [35], a framework for automatically identifying key hackers on underground forums by combining content and social network analysis. Using Latent Dirichlet Allocation, the study extracted users' topic preferences and

[4] https://www.ibm.com/watson

integrated this information with an improved Topic-specific PageRank algorithm. HackerRank outperformed methods relying solely on CA or SNA, increasing the coverage rate across five underground forums by 3.14% and 16.19%, respectively.

With a similar goal to [35], in 2023, Paracha, Arshad, and Khan [54] proposed the INSPECT framework to identify influential users on Dark Web forums. The framework employs a combination of feature engineering, social network analysis, semantic analysis, and K-means clustering to calculate an influencer score for each user, based on their significance within the network. Evaluation of the INSPECT framework using the CrimeBB dataset demonstrates its effectiveness in identifying dominant users with high prestige and domain knowledge.

4.3.4 Malware

The recent developments in the techniques employed in cyberattacks have rendered the implementation of effective defensive measures increasingly challenging. To address this, Kadoguichi et al. [41] proposed a machine learning approach for the extraction of critical posts, specifically those related to malware offers, from Dark Web forums, The methodology involved the use of doc2vec and a Multi-Layer Perceptron for classification. The dataset was collected from the Sixgill platform and analysed using a stratified K-fold cross-validation. The model achieved a mean accuracy of 93.9% on validation data but dropped to 79.4% on unseen datasets. This was attributed to two factors: firstly, the limited training data and secondly, the challenges in distinguishing ambiguous posts.

In addition, in 2020, Kadoguchi et al. [40] introduced a more efficient method using Deep Clustering, which combines auto-encoders and K-means to automatically classify Dark Web forum posts without manual labelling. Features extracted with doc2vec were clustered, and experiments focused on posts related to malware offers. Additionally, data collected from Sixgill was used with this technique, which significantly reduced the effort required for labelling and improved the efficiency of threat intelligence extraction.

4.4 Application of LLMs in Crime Analysis and Threat Intelligence

After conducting the systematic review presented in the previous sections, recurrent searches for new applications of these tools were conducted with the goal of identifying novel applications of LLMs in the same scope of this project. This section outlines some of the most relevant results in the context.

In recent years, LLMs have been the focus of experimentation in the field of smart policing, with the objective of assisting law enforcement agencies in the collection and analysis of both physical and digital evidence. Sarzaeim, Mahmoud, and Azim [63] developed a crime prediction framework based on three state-of-the-art language models (BART, GPT-3, and GPT-4) focusing on crime classification. The authors curated a domain-specific instruction dataset based on crime incident reports from San Francisco and Los Angeles, serving as a basis for fine-tuning LLMs. The authors demonstrate that GPT-3 and GPT-4 can outperform traditional ML models in crime

classification tasks using zero-shot prompting, i.e., without including examples on how to carry out a given task in the prompt. The results were enhanced when particular examples were incorporated within the prompt (an input and the anticipated output), a method referred to as few-shot prompting. Despite the authors' emphasis on the challenges associated with attaining uniform performance across both datasets, they argue that users, such as police officers, could enhance their comprehension of the propensity of criminal activities in a specific location, thereby facilitating optimal resource allocation and assignment.

Another tool designed to impact forensic investigations presented by Nikolakopoulos et al. [49] is the Investigation Enhancement Model (IEM). The system is founded upon a Retrieval Augmented Generation (RAG) LLM that has been trained with data collected from criminal incidents. RAG [43] is a knowledge augmentation technique that includes supplementary and relevant information in a prompt, allowing the LLM to better answer the initial query. The authors posit that the system could engage in a rational conversation with the user regarding a specific crime, generate incident reports, and provide a chronological account of the events.

Bokkena [10] explored the usage of LLMs in IT security, particularly the early detection of zero-day exploits. The research's objective was to analyse LLMs' capabilities in identifying linguistic and non-linguistic patterns from IT security logs and threat databases. The primary focus of the study was on network logs, phishing emails, malware samples and security incident reports, with the achievement of an accuracy rate of 95% being a key finding.

Porlou et al. [57] propose a pipeline to extract structured information from firearm-related listings posted in Dark Web Marketplaces. The authors address limitations in approaches that rely on HTML tags and pattern recognition to identify and automate extracting valuable structured information. This assists LEAs in addressing the proliferation of arms trafficking in the Dark Web. The authors combine prompt engineering techniques and evaluation techniques to test open-source models (Llama3 [29] and Gemma [67]). The 70 billion parameter version of the Llama3 series demonstrated superior performance, achieving 81% and 38% accuracy scores when evaluated with similarity score and exact match metrics, respectively. Notwithstanding the encouraging outcomes, the authors underscore the occurrence of sporadic challenges in extracting information from the data. This is primarily attributable to the training of most models to refrain from responding to prompts or requests that encompass content deemed to be illicit or of a sensitive nature, a category that encompasses firearm advertisements.

Also in the realm of Dark Web Threat Intelligence, Chen et al. [13] use some of the most common datasets associated with Dark Web research in combination with LLMs for text classification and zero-shot learning in the Dark Web. The authors demonstrate that decoder-only models can outperform encoder-only models in some scenarios. These results were further improved when combined with BERT-like models. The research focuses on two main tasks: (1) text classification/modelling, and (2) relevant Threat Intelligence textual data identification. The first involves categorising the contents of a website into one category. The second, defined as a binary classification problem, consists of determining which websites contain valuable intelligence, such as leaked confidential information or any other noteworthy content. Results prove GPT-like models,

such as Llama or Mistral [37], can outperform BERT-like models. Two Dark Web datasets were used to evaluate model performance: CODA [39] and DUTA-10K [4]. CODA (Corpus of Darknet Activities) is a publicly available dataset of 10,000 web documents tailored for text-based Dark Web analysis. It offers higher content quality and is easier to use due to its preprocessing, making it a better benchmark dataset than DUTA-10K. In contrast, DUTA-10K (Darknet Usage Text Addresses) contains over 10,000 samples across 25 main classes but includes much raw content from Dark Web pages, which poses challenges in establishing standard preprocessing protocols. During the experiments, it was also noted that general-purpose LLMs are not particularly effective at handling NLP tasks in Dark Web research. To enhance the performance of LLMs in this domain, additional pretraining or fine-tuning, along with the development of new datasets, is required. The results reveal that GPT-like models can surpass current state-of-the-art BERT-like models in supervised and zero-shot settings.

In the same scope, Shah and Madisetti [64] introduce MAD-CTI (Multi-Agent Dark Web Cyber Threat Intelligence), a novel framework employing a multi-agent LLM architecture to extract and classify cyber threat intelligence from Dark Web sources. MAD-CTI leverages OpenAI's GPT-4o mini with Microsoft AutoGen to orchestrate autonomous agents that sequentially perform distinct tasks - scraping, translation, semantic analysis, relevancy determination, and threat category classification. Each agent specialises in its function and communicates with others through a shared memory architecture, enabling context-aware, scalable analysis of complex, multilingual, and often obfuscated Dark Web content. The system's efficacy was empirically validated using 650 cyber-related annotated samples from the CODA dataset. Performance evaluation yielded an overall relevancy classification accuracy of 68.01% and category labelling accuracy of 63.52%, with notable improvement over both single-agent baselines and prior state-of-the-art models such as DarkBERT and ThreatCrawl. In particular, MAD-CTI achieved an 18.90% higher F1-score than DarkBERT in identifying relevant texts.

4.5 Summary & Critical Review

The field Dark Web data analysis has seen significant advancements in recent years, particularly through machine learning and natural language processing techniques. This review has highlighted the main tools and techniques currently employed and the primary topics of analysis within the field.

The review identifies several key methodologies used in the extraction and analysis of Dark Web data. The efficacy of traditional machine learning techniques, as exemplified by the ANITA project, has been demonstrated in the monitoring of illicit activities. The project achieved notable accuracy in predicting illegal activities using Big Data analytical tools. Additionally, unsupervised models, which combine K-means clustering and decision trees, have shown impressive results in characterising Dark Web forums.

Transformers and language models represent a significant leap forward in this field. DarkBERT, a language model pretrained specifically on Dark Web data, has outperformed general-purpose models like BERT and RoBERTa in various threat intelligence tasks, highlighting the need for a tailored analysis given the nuances and linguistic characteristics of Dark Web language and text corpora. Furthermore, combined approaches have been investigated, which integrate traditional machine learning methods with modern transformer-based models.

The literature under review addresses three principal areas of analysis: cross-domain threat intelligence, terrorism threat intelligence, and cyber threat intelligence. Cross-domain TI encompasses research aimed at identifying illegal activities across various forums and marketplaces. Terrorism TI focuses on analysing forums related to terrorism and radicalisation, while cyber TI addresses cybersecurity issues such as hacking, software vulnerabilities, and data leaks.

The overarching goal of this state of the art review was to identify literature employing large language models as the primary technique for extracting information from the Dark Web. At the time of writing, the findings indicate that up-to-date research has few references to applications of LLMs in the field of Dark Web Threat Intelligence and policy formulation. There is an absence of research exploring the integration of Large Language Models into full-spectrum intelligence pipelines - ones that not only extract relevant information but also generate strategic insights, formulate policy recommendations, and assist in guiding investigations.

It is precisely this gap that this dissertation seeks to address. By designing and executing a series of controlled experiments, ranging from topic modelling and relevance analysis to drug identification and structured policy formulation, this work offers a comprehensive assessment of LLMs in a variety of Dark Web intelligence tasks. Furthermore, synthesising these findings into a concrete system architecture tailored for law enforcement use, the dissertation directly responds to the operational void identified in the literature. It demonstrates the potential of LLMs to serve as analytical assistants and provides a practical framework for their deployment in real-world scenarios. Through this contribution, the research advances the current state of the art and offers a strategic direction for future work in Dark Web intelligence.

Chapter 5

Experimental Procedures

5.1 Data Collection and Processing

The data collection process from the Dark Web involves establishing secure connections to onion services via the TOR network and implementing enhanced security measures to preserve user anonymity and system security. However, collecting and processing data from the Dark Web introduces a range of complex ethical and legal challenges. These include concerns around the potential exposure to/of illicit or harmful content, the inadvertent collection of personal or sensitive data, and the lack of informed consent from users whose posts are analysed.

While these issues are critically important for responsible research, a thorough exploration of their implications was not feasible within the time and scope constraints of this project. Nevertheless, future work should prioritise a formal ethical assessment, including risk mitigation strategies and compliance with relevant legal standards, especially if the data are to be operationalised in real-world threat intelligence systems.

In light of these challenges, the experiments conducted within the scope of this project are based on three ethically sourced datasets: DUTA-10K, CODA, and a dataset from the Nemesis Marketplace [19], the latter kindly provided by Nicolas Christin and Alejandro Cuevas from Carnegie Mellon University. This section discusses the acquisition and preprocessing steps to prepare the datasets for use in the Design of Experiments section.

5.1.1 DUTA-10K

The DUTA-10K dataset [4] is an extension of the DUTA dataset [3]. The original dataset comprised 6,831 pages, collected between May and July 2016, and manually classified across 26 different classes/categories (seven of which were subdivided into 22 subclasses) based on their contents. Approximately one year later, and for the same period, the same authors expanded this dataset to 10,367 distributed snapshots of Dark Web forums across 25 redefined categories/activities and 20 subclasses, in comparison to the original collection. These websites are written in

41 different languages, with 84.74% of the records being in English. We requested the dataset directly from the author's website[1].

The dataset was analysed and processed, with the aim of ensuring the highest possible data quality for use in the experimental process. To guarantee the reproducibility of the experiments, the following list enumerates the steps followed:

1. Subclass labels were dropped to focus only on the general topics, and because most classes were not divided into subclasses.
2. The dataset contained 28 unique class values, despite the paper reporting 25. This discrepancy was due to trailing spaces in some labels and inconsistencies in the merger of the *Fraud* and *Leaked-Data* classes observed in the provided dataset. As such, trailing spaces were removed, and all *Leaked-Data* instances were relabelled to *Fraud*, resolving the inconsistency.
3. Non-English records were removed to ensure compatibility with LLMs, which are predominantly trained on English corpora.
4. Records classified as *Empty* were excluded.
5. Four records lacking textual content were discarded.
6. All extraneous white spaces, blank lines, and URLs were removed from the textual contents.
7. Texts were tokenised using the GPT-4o encoder, and entries exceeding 64,000 tokens were excluded to ensure compatibility with standard LLM context windows.

After this process, 7,114 records remained, with their class distribution presented in table 5.1. For simplicity, DUTA-10K is from here onwards only referred to as DUTA.

5.1.2 CODA

The CODA (Corpus of Darknet Activities) dataset [39] comprises 10,000 records of onion website contents, which were downloaded from Hugging Face[2]. A request was formulated to the creators of the dataset, including a justification for access and agreement to the Terms of Use. The public version of the dataset was extensively anonymised: 18 different types of user identifiers have been masked by the authors, ranging from IP addresses, usernames, email and cryptocurrency wallet addresses to drug weights, file sizes and other number tokens. With respect to linguistics, the distribution of 51 different languages present in the dataset, with 88.55% of English records.

Each record includes a label regarding the topic it falls under, including a more abstract *Others* category, accounting for nearly 30% of the documents.

Our processing steps of the raw dataset included the removal of blanks, spaces, and URLs from text; filtering out non-English records; deleting duplicate records; and tokenising and removing text entries with more than 64,000 tokens. Table 5.2 shows the 10 topic-based classification

[1] https://gvis.unileon.es/datasets-duta-10k/
[2] https://huggingface.co/datasets/s2w-ai/CoDA

Table 5.1: DUTA-10K Dataset Class Distribution After Processing

Class	**Count**	**Percentage**
Hosting	2224	30.4%
Cryptocurrency	847	11.58%
Down	755	10.32%
Locked	682	9.32%
Personal	419	5.73%
Counterfeit Credit-Cards	392	5.36%
Social-Network	293	4.0%
Drugs	290	3.96%
Services	284	3.88%
Porno	226	3.09%
Marketplace	189	2.58%
Hacking	182	2.49%
Forum	128	1.75%
Violence	87	1.19%
Counterfeit Money	81	1.11%
Counterfeit Personal-Identification	60	0.82%
Cryptolocker	50	0.68%
Fraud	36	0.49%
Library	30	0.41%
Casino	26	0.36%
Religion	16	0.22%
Art	14	0.19%
Human-Trafficking	3	0.04%
Politics	2	0.03%
Total	7114	100%

distribution of the remaining 8,839 items after processing. It is noteworthy that the proportion of the *Others* category was reduced from 30% in the original data to 24%.

5.1.3 Nemesis Marketplace

The Nemesis Marketplace was founded in 2021 and functioned as an encrypted drug trafficking network facilitator, namely for fentanyl, but also offered digital items, hacking services, and false identification documents. On March 24th, 2025, the U.S. Department of the Treasury's Office of Foreign Assets Control (OFAC) imposed sanctions on the administrator of Nemesis, Behrouz Parsarad [68]. Authorities estimate the market facilitated the sale of around $30 million between 2021 and March 20th, 2024, the date law enforcement agencies from the United States, Germany and Lithuania seized Nemesis' servers in a joint operation.

The third dataset used in this work is related to this marketplace. As outlined by Cuevas and Christin [19], the data was scraped from November 18th, 2022 to February 1st, 2023 and subsequently compiled in a database. This database contained information regarding users, items being sold and sales feedback from users [19]. The authors estimated a total revenue of $6,388,411

Table 5.2: CODA Dataset Class Distribution After Processing

Class	Count	Percentage
Others	2131	24.11%
Porn	1171	13.25%
Drugs	967	10.94%
Financial	956	10.82%
Gambling	756	8.55%
Crypto	744	8.42%
Hacking	630	7.13%
Arms	597	6.75%
Violence	467	5.28%
Electronic	420	4.75%
Total	8839	100%

for the specified period. This estimation was based on data from 372 vendors and 18,794 feedbacks left on the vendors' profile pages. These feedbacks were linked to the item to which the message refers, thus enabling the acquisition of data regarding sold items.

The dataset was made available in an SQL database format and comprises 4 tables: 'marketplace', 'items, 'sales_and_feedbacks' (renamed from 'feedbacks') and 'users'. The 'marketplace' table was dropped as it only contained one row and several aggregate columns derived from the preceding works of the authors (e.g. 'total_sales' or 'last_30' which represented sales for the last 30 days). The remaining tables were normalised until the third normal form[3], at which point any rows that did not adhere to the foreign and primary key constraints were also removed. For instance, within the 'sales_and_feedbacks' table, there were items for which there was no matching entry in the 'items' table, or products in the 'items' table for which there was no seller information in the 'users' table.

The authors of the dataset designed a machine learning classifier to predict the category of each item [65]. By applying this classifier to the Nemesis Marketplace dataset, the items are categorised into 11 distinct categories. The distribution of the categories is outlined in Table 5.3 in terms of frequency and revenue, sorted from the most to the least purchased category.

5.1.3.1 Drugs and Atlantic Region Focus

An in-depth analysis of Table 5.3 reveals that all drug-related categories (all categories with the exception of *Other*, *Digital Goods* and *Misc*) account for 93.14% of the market's total revenue (89.44% of which is attributable to all items from *Stimulants* up to and including *Benzos*, with the remaining 3.7% being accounted for by *Prescription* drugs). On this basis, the dataset was further processed to exclude items from non-drug-related categories. The resultant list comprised 2,422 drug products.

[3]The Third Normal Form [8] ensures a database contains no partial dependencies between attributes and all non-key attributes depend on the primary key, i.e. are not dependent on any other non-key attribute.

Table 5.3: Revenue per Category in the Nemesis Market

Category	**Count**	**Count %**	**Revenue**	**Revenue %**	**Cumulative**	**Cumulative %**
Stimulants	461	10.2%	$1,836,036.08	28.85%	$1,836,036.08	28.85%
Cannabis	788	17.44%	$1,538,948.22	24.18%	$3,374,984.30	53.03%
Ecstasy	233	5.16%	$620,222.37	9.75%	$3,995,206.67	62.78%
Opioids	144	3.19%	$548,382.72	8.62%	$4,543,589.39	71.4%
Psychedelics	313	6.93%	$536,641.35	8.43%	$5,080,230.74	79.83%
Dissociatives	119	2.63%	$309,333.15	4.86%	$5,389,563.89	84.69%
Benzos	173	3.83%	$302,183.89	4.75%	$5,691,747.78	89.44%
Digital Goods	1588	35.14%	$275,923.40	4.34%	$5,967,671.18	93.78%
Prescription	191	4.23%	$235,617.78	3.7%	$6,203,288.96	97.48%
Misc	403	8.92%	$86,486.35	1.36%	$6,289,775.31	98.84%
Other	106	2.35%	$73,673.42	1.16%	$6,363,448.73	100.0%
Total	4519	100%	$6,363,448.73	100%		

Furthermore, whenever an item is advertised by a seller on the market, they may choose to specify from which country or region of the world their products are shipped from or to. This information is of great value in understanding the level of prevalence of the drug trade in each region. For the context of this research, the decision was made to evaluate the capabilities of LLMs exclusively based on data pertaining to the Atlantic Region. This decision was made to focus on items for which an origin and destination can be identified and for which shipment across or through the Atlantic Region is feasible.

The list of regions mentioned across the 2,422 items contains 21 countries and three regions: the World Wide, the European Union, and North America. From the list of unique locations, a list of regions belonging to the Atlantic Region was compiled. The list of regions included the following countries and regions: the United States, the United Kingdom, Germany, the Netherlands, the European Union, Australia, France, Spain, Canada, North America, Belgium, Slovenia, Argentina, the Netherlands Antilles, Mexico, and Albania. The European Union was included in the study because it may account for countries connected to the Atlantic but not in the list, such as Portugal.

In the filtering process, all those items that did not satisfy these criteria were eliminated, i.e. those with both shipping origin and destination outside the Atlantic Region, which reduced the number from 2,422 to 2,271 items. Consequently, items shipped from countries outside the aforementioned list, but to a country on the list (or vice-versa), are retained. Table 5.5 illustrates the number of items shipped to and from each country according to the filtering process described.

5.1.4 Sampling

In light of the substantial size of the processed datasets, sampling was employed to expedite experimental procedures and reduce computational costs associated with LLM inference. The objective was to guarantee that the samples retained statistical representativeness of their respective populations whilst facilitating efficient experimentation and annotation in the experiments presented in

Table 5.4: Nemesis Market Revenue per Drug Category in the Atlantic Region

Category	Count	Count %	Revenue	%	Cumulative	Cumulative %
Stimulants	430	18.93%	$1,774,451.08	31.05%	$1,774,451.08	31.05%
Cannabis	730	32.14%	$1,525,616.26	26.69%	$3,300,067.34	57.74%
Ecstasy	221	9.73%	$578,876.42	10.13%	$3,878,943.76	67.87%
Psychedelics	301	13.25%	$530,391.87	9.28%	$4,409,335.63	77.15%
Opioids	133	5.86%	$493,948.08	8.64%	$4,903,283.71	85.79%
Dissociatives	116	5.11%	$306,943.59	5.37%	$5,210,227.30	91.16%
Benzos	166	7.31%	$273,191.68	4.78%	$5,483,418.98	95.94%
Prescription	174	7.66%	$232,230.82	4.06%	$5,715,649.80	100.0%
Total	2271	100%	$5,715,649.80	100%		

the next section.

The calculation of the optimal sample size was conducted in accordance with the prevailing statistical methodology, Cochran's Formula [2, 17], with a confidence level of 95% and a margin of error (MoE) of 5%. The initial sample size was calculated under the assumption of an infinite population using the following formula:

$$n = \left\lceil \frac{z^2 \cdot p \cdot (1-p)}{\text{MoE}^2} \right\rceil = \left\lceil \frac{1.96^2 \cdot 0.5 \cdot (1-0.5)}{0.05^2} \right\rceil = 385 \tag{5.1}$$

where $z = 1.96$ is the z-score corresponding to a 95% confidence level, $p = 0.5$ is the assumed population proportion (maximising variability), and $\text{MoE} = 0.05$ is the desired margin of error. The ceiling function ($\lceil\ \rceil$) ensures the result is rounded up to the nearest integer.

Since the processed datasets have known and finite sizes, a correction for finite populations can be applied using:

$$n' = \frac{n}{1 + \frac{n-1}{N}} \tag{5.2}$$

where n is the sample size for an infinite population and N is the total population size. Applying this formula yields adjusted sample sizes of 365 for DUTA, 368 for CODA, and 328 for Nemesis. However, given that these values are close to the infinite population estimate, and to maintain consistency across datasets, a conservative fixed sample size of 385 was adopted in all cases.

Two sampling strategies were employed:

1. **Random Sampling**: For the Nemesis Marketplace dataset, a completely random sample of 385 items was extracted. For DUTA and CODA, random samples of 385 items were drawn, limited to documents with fewer than 1,000 tokens to facilitate manual labelling processes.

2. **Stratified Sampling**: In order to ensure that the samples reflected the underlying topic distribution in each dataset, stratified samples were also generated. For DUTA and CODA, the stratified samples preserved the class proportions present in the original datasets. In the case of DUTA, the *Human-Trafficking* and *Politics* categories were excluded from the

Table 5.5: Drug Origin and Destination in the Nemesis Market in the Atlantic Region

Country	**From**	**To**	**Country**	**From**	**To**
United States	745	677	North America	0	18
United Kingdom	568	362	Czech Republic⋆	9	0
World Wide⋆	0	716	Belgium	5	0
Germany	337	62	Slovenia	4	0
Netherlands	323	7	Austria⋆	3	0
European Union	0	278	Argentina	3	0
Australia	92	91	Albania	0	1
France	99	52	Switzerland⋆	1	0
Spain	41	0	Netherlands Antilles	1	0
Unknown Location⋆	26	0	Mexico	1	0
Canada	13	7			

⋆ Not in the Atlantic Region List

stratified sample due to insufficient representation. For the Nemesis dataset, the stratified sample maintained the distribution of drug categories.

5.2 LLMs

The present study investigates the capabilities of LLMs for Dark Web text analysis, owing to their demonstrated ability to comprehend, generate, and reason with natural language, as outlined in the Background chapter. LLMs have revealed state-of-the-art performance across a range of tasks in natural language processing, particularly when scaled to billions of parameters and trained on diverse datasets. Their proficiency in zero-shot and few-shot scenarios makes them potentially suitable for tasks involving unstructured or heterogeneous text.

5.2.1 Selection Criteria

To ensure comprehensive evaluation and generalisability of the results, our study employed a diverse set of LLMs from multiple providers and of varying parameter sizes. The selection was guided by four key principles:

1. **Diversity in Model Providers**
   The LLMs selected were drawn from three leading organisations in the field: OpenAI[4], Google DeepMind/Gemini[5], and Mistral[6]. This cross-provider selection enables an analysis of whether and how differences in training methodology, dataset composition, and model architecture influence performance. Each provider has adopted distinct approaches to pre-training data, alignment techniques, and model optimisation, which can result in variations in output quality, interpretability, and robustness.

[4] https://openai.com/
[5] https://deepmind.google/
[6] https://mistral.ai/

2. **Variation in Model Sizes**
   The models span a range of parameter counts, from small-scale models optimised for latency and cost-efficiency, to large-scale models designed for maximum accuracy and generalisation. This stratified selection facilitates the investigation of the relationship between model size and task performance, providing empirical evidence on whether smaller models can approximate the effectiveness of their larger counterparts in the target application domain. The findings presented here are of crucial importance when assessing trade-offs in real-world deployments, where computational efficiency is a consideration.

3. **Inclusion of State-of-the-Art Architectures**
   The selection of models encompasses the latest versions of LLMs, thereby ensuring that the study captures the current state of the art in the field of LLM development. This temporal alignment ensures that performance comparisons are not confounded by obsolescence or uneven access to advancements in model architecture or fine-tuning techniques.

4. **Capability for Structured Output Generation**
   Many of the tasks in this study require precise, structured outputs, such as JSON-formatted responses or schema-conformant representations. Such capabilities are essential for applications involving automated reasoning, system integration and machine parsing of the results. Including models with these features ensures that the evaluation captures both generative fluency and adherence to format and controllability of responses.

5.2.2 List of Models

Following the criteria listed above, Table 5.6 details the chosen models and some of their details, including the snapshot/version and the knowledge cut-off of the models, i.e. the latest date for which the model contains information. This approach facilitates insights into both the technical characteristics and practical trade-offs associated with LLM deployment.

Table 5.6: Selected LLMs and Details

Provider	Name	Snapshot	Knowledge Cut-off
Google	Gemini 2.0 Flash	February 05th, 2025	August 2024
Google	Gemini 2.0 Flash Lite	February 25th, 2025	August 2024
Mistral	Mistral Large	November 2024	August 2024
Mistral	Mistral Medium	May 2025	August 2024
Mistral	Mistral Small	March 2025	August 2024
OpenAI	GPT-4.1	14th April 2025	June 1st, 2024
OpenAI	GPT-4.1 Mini	14th April 2025	June 1st, 2024
OpenAI	GPT-4.1 Nano	14th April 2025	June 1st, 2024

5.3 Design of Experiments

This section details the experiments conducted with the purpose of evaluating the ability of LLMs to analyse and detect relevant content on the Dark Web, and to provide insights into the potential applications of LLMs in the formulation of policy recommendations for security and defence.

5.3.1 Topic Modelling

The central aim of the topic modelling experiments was to determine whether LLMs can perform reliable topic classification of Dark Web data and to investigate their potential contribution to security and defence policy formulation.

Topic modelling is a fundamental task in the analysis of Dark Web content, serving as a basis for content understanding, automated monitoring, and intelligence extraction. Given the fragmented and frequently chaotic nature of Dark Web forums, marketplaces, and portals, the identification of dominant topics enables the structuring of information, enhances situational awareness, and supports law enforcement or policy actors in understanding criminal ecosystems.

These experiments focus on the application of zero-shot prompting techniques using pretrained LLMs, without fine-tuning, to simulate a plausible operational setting where models are applied directly to raw or pre-processed text derived from anonymous online platforms.

The DUTA and CODA datasets were selected for this experiment due to the presence of ground truth topic labelling, allowing for quantitative evaluation of model predictions. Each dataset defines a set of class values associated with each document (i.e., Dark Web page). For the purpose of the topic modelling task, these class labels were used as candidate categories in zero-shot classification prompts, enabling an assessment of the LLMs' ability to infer the correct label based solely on textual content.

5.3.1.1 Zero-Shot Classification without Class Descriptions

In this initial experiment, each record from the DUTA and CODA stratified samples was used to query an LLM with the goal of obtaining a predicted topic label from a predefined list of possible classes. The task was formulated as a zero-shot classification problem, with each of the models being prompted to decide the most appropriate class label based on the entire text contents of the record.

Initial attempts to perform classification without first establishing a system-level role resulted in inconsistent behaviour, including refusals to respond to queries due to the sensitive nature of the content. To mitigate this issue and guide the model to adopt a professional and analytical stance, a system prompt was introduced to simulate the perspective of a specialised analyst:

Listing 5.1: System Prompt

```
1 You are a highly specialized dark web intelligence analyst.
```

The user prompt was then designed to present the classification task, list the available classes, and request a response that adheres strictly to the list of categories while also including a brief justification for the chosen label. The template is as follows:

Listing 5.2: Website Classification Prompt Template

```
Classify a darkweb website based on its text contents from the following
    ↪ list of possible main classes:

{class_values}

Here are the contents of the website:

{website_contents}

Please classify the website into one of the main classes above, do not
    ↪ deviate. It is mandatory that the replied class is contained in the
    ↪ list of possible categories.
Moreover, provide a brief explanation for your classification.
```

To facilitate structured responses and enable consistent parsing and evaluation of the model's output, two response schemas were developed using the Pydantic library[7], which enables robust data validation and modelling using Python type hints. When incorporated in the request to the model, the schemas ensure that the classification output is valid, i.e. limited to the allowed labels, and accompanied by a textual explanation. Below is the definition of the response schemas for both DUTA and CODA:

Listing 5.3: DUTA and CODA Topic Modelling Pydantic Response Schemas

```
from typing import Literal
from pydantic import Field, BaseModel

class DUTAClassification(BaseModel):
    class_prediction: Literal[tuple(DUTA_CLASS_VALUES)] = Field(..., description="
        The predicted class of the darkweb website")
    explanation: str = Field(..., description="The explanation for the
        classification")

class CODAClassification(BaseModel):
    class_prediction: Literal[tuple(CODA_CLASS_VALUES)] = Field(..., description="
        The predicted class of the darkweb website")
    explanation: str = Field(..., description="The explanation for the
        classification")
```

[7] https://docs.pydantic.dev/

The `class_prediction` field is constrained to one of the listed values in the respective dataset, thereby avoiding hallucinated or out-of-scope responses. The accompanying `explanation` field provides clarity into the model's decision-making process, which is essential for downstream analysis of threat intelligence workflows.

5.3.1.2 Classification with Class Descriptions (CODA Only)

The second experiment builds upon the first by enriching the input with textual descriptions of each class. This additional semantic context is drawn from the original CODA dataset publication, which provides concise explanations for each topic category. The motivation behind this enhancement is to assess whether LLMs benefit from more explicit category semantics, thereby improving classification accuracy and reducing ambiguity.

The same system prompt as in the first experiment is retained, maintaining the simulated intelligence analyst persona. However, the user prompt was modified to include the descriptions associated with each class. An illustrative version of the modified prompt is provided below:

Listing 5.4: CODA Classification Prompt with Class Descriptions Template

```
Classify a darkweb website based on its text contents from the following list of possible main classes (for each class, a brief description is provided):
- "Pornography": general/child pornography and other explicit content;
- "Drugs": various types of legal/illegal drugs such as medications, steroids, pain killers, viagra, cannabis, hashish, meth, benzos, ecstasies, opioids, and psychedelics;
- "Financial": counterfeit/cloned/stolen money or identifications (e.g., bills, credit cards, certificates, passports), money transfers (e.g ., PayPal), fiat money, ATM skimmers, magnetic card readers, etc.;
- "Gambling": any type of gambling, betting, casinos, lotteries, etc.;
- "Cryptocurrency": cryptocurrency-specific services or technologies such as wallets, generators, mining, laundering, mixing, multiplying, doubling, scamming, and escrow;
- "Hacking": hacking tools, hacking guides, hacking groups, hacking services, ransomware, malware, exploits, DDoS attacks, cracking, botnet, etc.;
- "Arms": any type of non-lethal/lethal weapons such as guns, ammunition, explosives, knives, missiles, and chemical weapons;
- "Violence": human trafficking, hitman, kidnapping, poisoning, torture, extortion, sextortion, sex slavery, blackmail, etc.;
- "Electronics": sale of or information on (stolen/hacked) mobile phones, laptops, tablet computers, etc.;
```

```
- "Others": all other content that does not fit the above categories,
    ↪ including log-in pages, error messages, etc.

Here are the contents of the website:

{website_contents}

Please classify the website into one of the main classes above, do not
    ↪ deviate and do not include the description in your answer. It is
    ↪ mandatory that the replied class is contained in the list of
    ↪ possible categories.
Moreover, provide a brief explanation for your classification.
```

This experimental design allows for a comparative evaluation between two prompting strategies: one based solely on class labels (minimal semantic context) and another using both labels and human-readable definitions (rich semantic context). The outcome informs whether additional contextual grounding contributes to classification performance, which may have implications for the design of LLM-based threat intelligence pipelines, especially in scenarios involving ambiguous or novel content.

5.3.2 Relevance Analysis

This experiment investigates the feasibility of using LLMs as a foundational component in an automated flagging system capable of identifying potentially criminal content within Dark Web data. The primary objective is to determine whether LLMs can support the initial screening of unstructured text to identify material that may indicate past or future criminal activity, thereby reducing the burden of manual triage for law enforcement personnel.

The underlying task is framed as a content relevance assessment, aiming to identify whether a document contains actionable intelligence that could contribute to criminal investigations. While the problem could initially be cast as a binary classification task - distinguishing *Relevant* from *Irrelevant* content - preliminary annotation trials revealed that such dichotomous labelling was insufficiently granular. In many cases, content that was not immediately indicative of a threat nonetheless contained leads or contextual elements valuable for intelligence gathering. Consequently, a more nuanced four-class flagging scheme was adopted:

- **Irrelevant (IRR)**: Content with no discernible threat or actionable lead.
- **Investigative Lead (IL)**: Content lacking explicit threats but containing potential indicators of past crimes or identifiers useful for investigation.
- **Threat No Lead (TNL)**: Content indicating an active or imminent threat without sufficient specificity to aid in investigation.

- **Threat With Lead (TWL)**: Content containing both threats and specific, actionable leads that may aid in prevention or response.

5.3.2.1 Formal Relevance Criteria

A set of formal criteria was developed to operationalise this classification scheme. The United States Department of Homeland Security defines a *threat* as "a statement or action indicating an intention to harm or cause damage". Expanding upon this, the definition in our study encompasses broader forms of criminal intent, coordination, and solicitation, along with the advertisement or sale of illegal goods, services, and tools that facilitate unlawful acts.

The relevance assessment follows a two-step process:

Step 1: Threat Identification A document is marked as a *threat* if it exhibits at least one of the following features:

- **Criminal Service Advertisement**: Promotion or sale of illicit services such as assassination, human trafficking, counterfeiting, hacking, etc.
- **Planning or Coordination**: Discussions indicating preparation for a crime (e.g., scheduling of attacks, target identification). Crime examples: drug trafficking, cyberattacks, weapons sales, and murder-for-hire.
- **Threat of Violence**: Direct or indirect expressions of intent to cause harm.
- **Solicitation or Recruitment**: Attempt to recruit, delegate, or request help/collaboration from participants for criminal activity.
- **Illegal Marketplace Listings**: Examples include weapons, drugs, child exploitation materials, and forged documents.
- **Hacking Tools or Data Dumps**: Sale/distribution of malware, exploits, credentials, or surveillance tools.
- **Incitement to Crime**: Posts encouraging others to commit criminal acts.
- **Tools for Crime**: Instructions on how to carry out specific attacks (e.g. DIY bombs, how to code a Trojan horse, and how to bypass a security system/password)

Step 2: Lead Identification A document is considered to contain a *lead* if it provides clues that may assist in linking the content to actors, targets, or events. These include:

- **Target Identification**: Real-world individuals, companies, organisations, institutions, infrastructures, usernames, and email addresses that could point towards the target of a crime.
- **Actor Identification**: Phone numbers, credit card numbers, and political affiliations that could point towards the perpetrators of a crime.

- **Timelines**: Specific dates, timeframes, or schedules.
- **Geographical Clues**: Cities or smaller regions, public places (e.g. train/bus stations, airports, shopping malls), physical addresses, GPS coordinates, etc.
- **Financial Identifiers**: E.g. cryptocurrency wallet addresses, transaction logs.
- **Communication Channels**: E.g. Telegram handles, Discord, and Signal.
- **Proofs of Crime**: E.g. screenshots/photos, videos, text messages, testimonies, victim data, leaked documents, payments, transaction records, source code.
- **Technical Artefacts**: Malware hashes, C2 URLs, source code samples, etc.

It is acknowledged that some elements, such as usernames, onion email addresses, or blockchain wallets, do not directly identify individuals. Nonetheless, these artefacts can be instrumental when cross-referenced with other investigations, thereby warranting a lead classification under the principle of prioritising investigative value in ambiguous cases.

5.3.2.2 Annotation Procedure

From the random samples of 385 entries extracted from both the DUTA and CODA datasets, each item was manually annotated based on the two-step relevance criteria described above. To assist and help consolidate this process, the two LLMs from Google's model suite, Gemini Flash 2.0 and Gemini Flash 2.0 Lite, were tasked with performing the same classification. Each model received instructions to:

1. Evaluate the presence of a threat based on predefined criteria.
2. Assess whether any actionable leads were present.
3. Provide detailed justifications and textual evidence for each decision.
4. Assign one of the four relevance tags.

The initial prompt is replicated in Listing 5.5, where placeholders, such as `website_contents`, `threat_criteria`, and `leads_criteria`, were populated with the appropriate values. Moreover, the models' response schema for this task is detailed in Listing 5.6.

Listing 5.5: Initial Relevance Prompt

```
1 Analyse the contents of the following \dw{} page to help law enforcement
    ↪ agencies in criminal investigation and prediction.
```

```
This analysis is part of an initial screening and filtering process to identify relevant information for law enforcement agencies. In the phase, only pages which contain actionable information are selected for further analysis: only pages that contain information that can be used to identify threats, targets, locations, involved parties, single or collective entities, etc.
To achieve this, follow the two-step annotation process below to determine whether the page contains a threat and/or lead. Base your judgement strictly on the provided Formal Relevance Criteria.

If the information is relevant, provide a detailed analysis of the information contained in the website and how it can be useful for law enforcement agencies.

Here are the contents of the website:
{website_contents}

Please provide a detailed analysis and not deviate from the task.

Step 1: Threat Detection
Determine whether the page contains a threat by checking if any of the following indicators are present. If any of the following list apply and the threat is clearly identifiable, classify as a "Threat". Otherwise, classify as "No Threat".
Threat Criteria Checklist (classify as "Threat" if one or more of the following apply):

{threat_criteria}

Decision:
Threat / No Threat

Justification:
Explain your reasoning and cite the matching content indicators or phrases from the page.

Step 2: Lead Detection
If a threat is found, check if there are leads indicating present or future crimes.
If no threat, check for leads indicating past crimes.
If any lead criterion is met, classify as "With Lead", otherwise "No Lead".

Lead Criteria Checklist (classify as "With Lead" if one or more of the following apply):
```

```
30
31 {leads_criteria}
32
33 Decision:
34 With Lead / No Lead
35
36 Justification:
37 List the types of leads detected and quote relevant content if available.
38
39
40 Final Relevance Tag
41 Based on the combination of Step 1 and Step 2, assign one of the following
       ↪ tags:
42 - Irrelevant (No Threat + No Lead)
43 - Investigative Lead (No Threat + Lead)
44 - Threat No Lead
45 - Threat With Lead
46
47 Note: Follow the principle of "highest priority whenever in doubt":
       ↪ ambiguous identifiers (e.g., usernames, onion emails) should be
       ↪ flagged as leads if they could potentially be cross-referenced in
       ↪ other investigations.
```

Listing 5.6: Relevance Response Schema

```
1 from typing import Literal
2 from pydantic import Field, BaseModel
3
4 class Relevance(BaseModel):
5     relevance_prediction: Literal["Irrelevant", "Investigative Lead", "Threat No
          Lead", "Threat With Lead"] = Field(..., description="The predicted
          relevance of the darkweb website")
6     explanation: str = Field(..., description="The explanation for the
          classification")
```

Each model's output was manually reviewed. If an LLM's classification and reasoning were deemed valid, the corresponding ground truth label was updated. This review led to 39 label modifications in the DUTA sample and 24 in the CODA sample. The label distribution for each dataset before and after the relabelling process is detailed in Table 5.8.

The overall preliminary results post-relabelling process are presented in Table 5.7, and Figure 5.1 shows the confusion matrices for Gemini Flash 2.0, the best model in each dataset. An initial analysis showed that the model excelled in detecting threats and identifying the corresponding leads. However, they show an inability to disregard content that should not be classified as

non-leads. Regarding the CODA dataset, the results are substantially worse than those of DUTA. The main reason has to do with the high level of anonymity processing of the dataset, as explained in the CODA introduction section. For instance, the mask `ID_BTC_ADDRESS` would be considered as an actual Bitcoin address, even though the address was not actually present and, therefore, could not be considered a lead. An inspection of the initial results revealed several systematic misinterpretations by the LLMs:

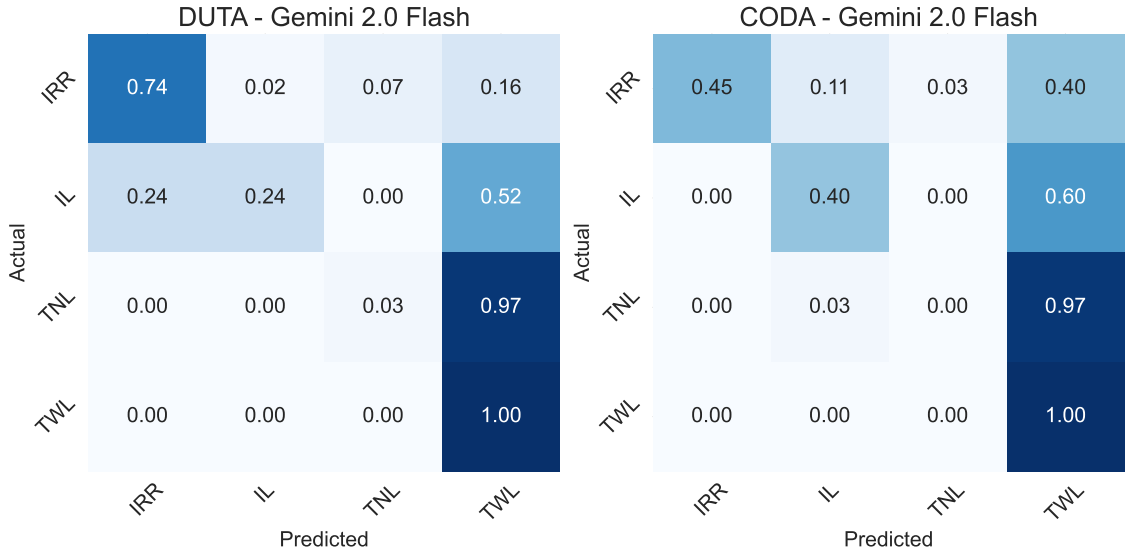


Figure 5.1: Preliminary Relevance Confusion Matrices for Gemini Flash 2.0

Table 5.7: Preliminary Relevance Experiment Results

Dataset	Provider	Model	Accuracy	Precision	Recall	F1
DUTA	Google	Gemini 2.0 Flash	67.27%	43.44%	50.25%	40.17%
DUTA	Google	Gemini 2.0 Flash Lite	64.68%	39.77%	56.91%	40.54%
CODA	Google	Gemini 2.0 Flash	37.40%	30.31%	46.28%	24.77%
CODA	Google	Gemini 2.0 Flash Lite	38.70%	35.46%	41.45%	24.46%

Table 5.8: Relevance Labels Distribution Before & After Relabelling

Label	CODA Before	CODA After	DUTA Before	DUTA After
Irrelevant (IRR)	250	235	328	296
Investigative Lead (IL)	7	5	12	21
Threat No Lead (TNL)	102	109	27	35
Threat With Lead (TWL)	26	36	18	33

DUTA Observations

- Bitcoin mixers[8] were classified as illegal despite their legality in some jurisdictions.
- The presence of login forms was mistaken as evidence of leads (e.g., the mere reference to passwords would be considered as a lead).
- Geographical clues were incorrectly inferred from currency acronyms (e.g., USD, BTC) and generic references like "USA".
- Mere mentions of PGP keys or attacks on websites were flagged as threats or leads, even in the absence of relevant context.

CODA Observations

- Bitcoin generators were flagged as illegal.
- Masked identifiers, like `ID_EMAIL`, were misclassified as actual leads.
- Fictional discussions (e.g., in crime books) were incorrectly labelled as real threats or leads.

To mitigate these misclassifications, the prompt was refined by removing the note regarding the "highest priority" principle in Line 47 of the initial prompt and by appending a set of disambiguating notes. These notes clarified that masking strings, fictional scenarios, generic currencies, and indirect references should not be considered valid indicators unless accompanied by concrete, context-specific content.

Moreover, adding these criteria also served as a test to infer the capabilities of LLMs to follow strict criteria and/or exceptions. In a real-world situation, LEAs may want to restrict the analysis of content or create custom system filtering strategies. One example could be the case of Bitcoin mixers/generators exclusion or inclusion, according to the focus of an investigation or specific jurisdiction/local law.

Listing 5.7: Adjusted Relevance Prompt

```
1 (...)
2
3 {leads_criteria}
4
5 NOTES:
```

[8] A Bitcoin mixer is a tool that blends a user's Bitcoin with those of other users, making it difficult to trace the origin or destination of the funds. Mixers work by pooling together coins from many participants and then redistributing them in such a way that the transaction trail between sender and recipient is obscured, enhancing privacy and anonymity (source: https://www.coinbase.com/en-gb/learn/your-crypto/what-is-a-bitcoin-mixer).

```
1. Masking strings/expressions for identity concealment purposes should not be considered as a lead as they are not real identifiers. For example, ID_NUMBER, ID_GENERAL_MONEY, ID_EMAIL, ID_CRYPTO_MONEY, ID_BTC_ADDRESS, ID_ONION_URL, ID_FILENAME, ID_WEIGHT, ID_TIME etc. are not real identifiers and should not be considered as leads. There may, however, be some other leads in the text and the present of these strings should not prevent the identification of other leads.
2. Currency acronyms (e.g., USD, EUR, BTC, ETH, etc.) should not be considered as geographical clues.
3. If you find a page that contains a login form, do not consider the login form itself as a lead as it is not a real identifier. However, if the page contains other leads (e.g., usernames, email addresses, etc.), you should consider those as leads.
4. The mere reference to a PGP key or a bitcoin address is not enough to consider it a lead, only consider if they are actually present in the text.
5. Don't consider a mention to an attack on the website a threat, as it is not a threat to law enforcement agencies or society in general. However, if the page contains other threats (e.g., planning an attack, inciting violence, etc.), you should consider those as threats.
6. Some websites may be discussing a story in a violent book of fiction, in this case, do not consider the story as a threat or lead.

Decision:
With Lead / No Lead

(...)
```

Addressing the Impact of Complex Instructions with a Two-Step Strategy

Large Language Models have a limited context length, which can limit their ability to process lengthy inputs effectively. In scenarios involving extensive website content and, consequently, lengthier prompts, critical details within the prompt, particularly instructions or classification criteria, may be overlooked, potentially degrading response quality and task performance.

To address this, a two-step classification strategy was tested. Instead of combining all instructions into a single prompt, the process was split into two independent stages:

- **Step 1 - Threat Identification**: The LLM assessed whether the content satisfied the criteria for a threat. The response was limited to *Threat* or *No Threat*.
- **Step 2 - Lead Identification**: A second, separate prompt evaluated whether the same content contained any actionable leads, again restricted to *Lead* or *No Lead*.

Each stage used a streamlined version of the original prompt (see Listing 5.7), with Pydantic schemas tailored to the task-specific outputs. This separation aimed to isolate the cognitive load associated with each decision, thereby potentially reducing misclassifications due to overlooked criteria.

This two-step method may not only serve to mitigate the performance drop from long inputs, but also offer a more modular, interpretable design. In practice, such modularity could enable LEAs to adapt or fine-tune sub-components of the pipeline independently, for instance, excluding certain types of threats based on jurisdictional scope.

5.3.3 Drug Identification

The purpose of this experiment was to evaluate the capability of LLMs to accurately identify drug categories based on real-world Dark Web marketplace listings. The stratified sample extracted from the Nemesis Marketplace dataset was used for this task, incorporating product descriptions, vendor metadata, and buyer feedback.

This experiment focused on the extent to which LLMs could interpret domain-specific slang, acronyms, and coded terminology commonly used to obfuscate drug-related content. Such linguistic patterns are less frequently encountered in traditional corpora and present a unique challenge for general-purpose language models. Therefore, successful classification in this context would suggest that the models are not only capable of handling non-standard language but also potentially viable for operational use in narcotics threat intelligence pipelines.

Given the obfuscation strategies employed by vendors (e.g., use of chemical codes, street names, or euphemistic descriptions), a constrained classification approach was adopted. The model was instructed to assign each instance to one of a predefined set of drug categories, as outlined in Table 5.4. This controlled vocabulary included the most prevalent classes in the dataset, balancing semantic coverage with the need for disambiguation.

To enhance response reliability, the prompt explicitly mandated selection from the given list, disallowing deviations or invented labels. Additionally, models were required to justify each prediction through a brief, free-text rationale as formalised in the structured prompt in Listing 5.8.

Listing 5.8: Drug Identification Prompt

```
1 Analyse the contents of the following \dw{} marketplace drug product,
    ↪ vendor's details and buyer's feedbacks and identify the drug
    ↪ category they refer to from the following list of drugs types:
2 {drugs_list}
3
4 Here are the contents of the website:
5 {website_contents}
6
```

```
Please classify the items into one of the main categories above, do not deviate. It is mandatory that the replied class is contained in the list of possible drugs categories.
Moreover, provide a brief explanation for your categorisation.
```

To enforce schema consistency and reduce ambiguity in response interpretation, a strict response format was adopted, as defined by the Pydantic schema below:

Listing 5.9: Drug Identification Response Schema

```python
from typing import Literal
from pydantic import Field, BaseModel

class DrugIdentification(BaseModel):
    drug_prediction: Literal["Stimulants", "Cannabis", "Ecstasy", "Opioids", "Psychedelics", "Dissociatives", "Benzos", "Prescription"] = Field(..., description="The type of drug identified from the darkweb website")
    explanation: str = Field(..., description="The explanation for the drug identification")
```

The resulting outputs were subsequently assessed for both categorical accuracy and interpretive coherence, allowing a nuanced understanding of LLM capabilities in this highly domain-specific application area. Detailed performance metrics analysis of the model outputs are discussed in the next chapter.

5.3.4 Information Extraction and Policy Formulation

This experiment sought to assess the capacity of LLMs to extract structured, actionable intelligence from Dark Web content and to generate informed policy recommendations for law enforcement agencies. The experimental design was grounded in practical requirements gathered through a collaborative engagement with law enforcement representatives by means of a survey sent to the main entities involved in the project. Both the questions and the obtained answers are listed in Appendix B.

Although this interaction proved essential in refining the investigative priorities, it was, unfortunately, only feasible during the later stages of the project due to difficulties in obtaining responses from the relevant partner institutions. Despite the temporal constraints, the obtained input posed a solid foundation towards the formulation of domain-relevant extraction targets and contributed to aligning the experimental scope with possible operational needs and workflows in the field.

During the consultation, several pressing challenges were identified by LEAs in relation to the monitoring and analysis of Dark Web environments. The prevailing among these were: (1) the frequent use of obfuscated and non-standard language, which complicates semantic interpretation; (2) the overwhelming volume and velocity of data, which impairs manual review; (3) and the

limited reliability of user profile information, which complicates concrete attribution and linkage analysis. In response to these challenges, the experiment was structured to explore the extent to which LLMs could serve as analytical aides in overcoming these limitations.

Based on the requirements elicited from the law enforcement partners, a comprehensive prompt was designed to guide the models in the extraction of pertinent details. These requirements encompassed diverse categories of information, such as the identification of communication and payment methods, shipment logistics, buyer-seller dynamics, user identifiers (e.g., usernames and email addresses), geographic indicators, keywords denoting illicit activities, and potential leads for criminal investigations. Furthermore, the prompt instructs the models to infer strategic intelligence and formulate operational or policy-level recommendations that could assist LEAs in prioritising investigative resources, enhancing cross-border cooperation, and ultimately contributing to successful law enforcement actions, such as arrests and seizures.

The prompt used in the experiment was the following:

Listing 5.10: Information Extraction and Policy Formulation Prompt

```
Analyse the contents of the following \dw{} website and formulate policy or recommendations for law enforcement agencies. The objective is to extract actionable intelligence and formulate detailed recommendations for law enforcement agencies.

Focus your analysis to address the following key dimensions:

1. Emerging Threats and Trends
  - Identify novel or recurring criminal trends observed on the site.

2. Criminal Activities and Illicit Services
  - Identify the crimes present in the website contents (e.g., drug trafficking, human trafficking, etc)
  - Identify services being offered (e.g., hacking services, murder-for-hire, etc)

3. Potential Targets and Victims
  - Identify individuals, groups, sectors, or institutions likely to be targeted by a crime.

4. Evasion Techniques
  - Detail any detectable method used by users to evade detection or prosecution (e.g., anonymisation tools, encrypted communications, operational security practices).

5. Geographical and Supply Chain Information
```

```
  - Indicate any relevant geographical locations (origins or destinations of goods/services, crime locations, etc).
  - Describe distribution channels and methods of shipment for illicit goods/services.

6. Communication Methods and Tools
  - Describe the tools, platforms, or technologies used to facilitate communication (e.g., forums, encrypted messaging, PGP).
  - Note behavioural patterns in communication that may aid profiling.

7. Payment and Financial Transactions
  - Identify the financial methods used (e.g., cryptocurrencies, escrow).
  - Highlight transaction patterns and any laundering techniques.

8. Feedback, Reviews, and Reputation Systems
  - Summarise buyer/seller feedback and review mechanisms.
  - Discuss how such systems may be used to detect fraud, deception, or trust-building practices.

9. Indicators of Criminal Activity
  - Extract specific keywords, jargon, or linguistic patterns that signify illicit intent.
  - Identify user handles, usernames, or aliases potentially relevant for ongoing investigations.

10. Overall Intelligence Assessment
   - Provide an overarching evaluation of the site's relevance to law enforcement.
   - Assess its impact, risk level, and strategic intelligence value.

11. Policy and Operational Recommendations
   - Suggest guidelines, frameworks, or procedures to enhance:
      - Threat detection
      - Investigative effectiveness
      - Resource prioritisation
      - International cooperation

Focus only on information that is actually present in the website contents, do not make assumptions or extrapolations beyond the provided text.
For instance, if a payment method is not mentioned, do not assume its existence.

Ensure that the analysis is detailed, structured, and suitable for informing national or international law enforcement efforts.
```

```
53
54 Here are the contents of the website:
55 {website_contents}
```

To ensure consistency and enable structured downstream analysis, model responses were required to adhere to a predefined Pydantic schema. This schema defines the expected fields of extracted intelligence based on the listed requirements, including criminal activity types, evasive behaviours, user identifiers, keywords, communication patterns, and strategic recommendations. The inclusion of detailed and categorical output was mainly based on the survey responses and intended to mirror potential types of intelligence reports routinely produced by threat analysts and to test whether LLMs could serve as force multipliers in this analytical process.

Listing 5.11: Information Extraction and Policy Formulation Pydantic Response Schema

```python
from typing import Literal
from pydantic import Field, BaseModel

class Analysis(BaseModel):
    # Direct Information Extraction Fields
    crimes: List[str] = Field(..., description="Descriptions of identifiable
        criminal activities observed on the website (e.g., drug trafficking,
        identity theft, weapons sales).")
    services: List[str] = Field(..., description="Details of detected illicit
        services offered (e.g., hacking services, document forgery, murder-for-hire
        ).")
    leads: List[str] = Field(..., description="Actionable concrete information that
         is present in the website and may support further investigations (e.g.,
        usernames, addresses, linked aliases, cross-postings, linked marketplaces).
        ")
    geographical_locations: List[str] = Field(..., description="Locations mentioned
         in relation to origin, destination, or operational hubs for products/
        services.")
    targets: List[str] = Field(..., description="Identified or implied victims or
        targeted entities (e.g., individuals, institutions, industries).")
    usernames: List[str] = Field(..., description="Usernames explicitly found on
        the website potentially linked to illicit activity.")
    other_user_identifiers: List[str] = Field(..., description="Additional user-
        specific identifiers such as email addresses, PGP keys, aliases, or avatars
        .")
    justice_evasion: List[str] = Field(..., description="Tactics and tools used to
        avoid detection or prosecution (e.g., VPNs, Tor, tumblers, OPSEC guides).")
    communication_channels: List[str] = Field(..., description="Platforms, tools,
        or mediums used for interaction (e.g., encrypted chats, forums, email).")
    communication_patterns: List[str] = Field(..., description="Observed linguistic
        , behavioural, or interactional trends among users (e.g., formal
        transaction structures, slang usage).")
```

```
    payment_methods: List[str] = Field(..., description="Monetary systems used (e.g., Bitcoin, Monero, escrow services, prepaid cards).")
    shipment_methods: List[str] = Field(..., description="Methods used for delivering illicit goods (e.g., dead drops, courier disguises, postal service manipulation).")
    keywords: List[str] = Field(..., description="Significant phrases, jargon, or encoded terms indicative of criminal activity.")

    # Analytical Insights and Recommendations
    feedbacks_analysis: List[str] = Field(..., description="Insights derived from reviews or reputation systems, including credibility assessments and dispute indicators.")
    criminal_investigations_suggestions: List[str] = Field(..., description="Suggestions for actionable investigative steps based on the site's content (e.g., tracing user trails, targeting admin accounts).")
    policy_recommendations: List[str] = Field(..., description="Strategic recommendations to inform law enforcement policy, resource allocation, or legislative improvement.")
    overall_website_analysis: List[str] = Field(..., description="Summary assessment of the website's significance, threat level, and intelligence value.")
```

The response schema in Listing 5.11 contains 17 fields. The majority (13) are fields corresponding to more direct information identification and extraction, such as `keywords`, `crimes` or `usernames`. The remaining four fields focus on deeper analytical and recommendation insights, namely feedback analysis and policy and investigative suggestions, as well as a summary of the website and its main details and contents.

Overall, this experiment constituted a critical step towards the analysis of the feasibility of integrating LLMs into an analytical pipeline for Dark Web-based investigations. It represents a synthesis of technical innovation and operational relevance, where LLMs are assessed not merely for linguistic capabilities but for their potential to inform real-world policing strategies. The results, discussed in the next chapter, offer insights into both the strengths and limitations of current models when applied to high-stakes, intelligence-driven contexts.

5.4 Summary

This chapter presented the experimental framework designed to assess the capabilities of LLMs in analysing Dark Web content for law enforcement and intelligence purposes. The experiments were grounded in the need to identify relevant content, extract actionable intelligence, and support policy formulation. To this end, four distinct experiments were developed: topic modelling, relevance analysis, drug identification, structured information extraction and policy generation.

Three ethically sourced datasets - DUTA-10K, CODA, and Nemesis Marketplace - were selected and preprocessed to ensure compatibility with LLMs and reliability of the analyses. This included standardising class labels, removing duplicates and non-English content, and applying

statistical sampling techniques to preserve representativeness. For the Nemesis dataset, a targeted filtering strategy was applied to focus on drug-related listings in the Atlantic region.

Eight LLMs were chosen from three major providers (OpenAI, Google DeepMind, and Mistral), spanning a range of model sizes and architectures. Selection was based on diversity, scale, support for structured outputs, and alignment with the state of the art. This enabled the evaluation of performance across different computational and design trade-offs, while ensuring that the models could handle both general classification tasks and structured analytical outputs.

Each experiment was carefully designed to simulate realistic operational scenarios. Topic classification was tested both with and without class descriptions; relevance analysis employed a four-class threat-lead taxonomy; and drug identification focused on domain-specific slang interpretation. The final experiment incorporated inputs from law enforcement to extract targeted intelligence fields and generate strategic recommendations. Together, these procedures provide a rigorous foundation for assessing the practical utility of LLMs in security and defence applications.

Chapter 6

Results & Discussion

This chapter presents the empirical findings and analyses from a series of experiments that were conducted with the objective of evaluating the capabilities of large language models in four critical tasks related to the extraction of intelligence from the Dark Web. The four tasks were: topic modelling, relevance analysis, drug identification and information extraction and policy formulation. The evaluation includes comparisons across eight diverse LLMs, spanning different model sizes and families.

All predictions were obtained via inference using the respective model providers' APIs. For Google models, individual API calls were issued per input using the Lanchain framework[1]. In contrast, batch processing was used for Mistral and OpenAI models, with a separate batch file created for each combination of model, dataset, and task. Each request included a custom system prompt (see Listing 5.1) alongside the corresponding task-specific user prompt. All models were run using their default configuration and parameters, with the exception of a fixed maximum response length of 250 tokens.

Performance metrics such as accuracy, precision, recall, and F1-score were used to quantify effectiveness, with additional insights drawn from confusion matrices. The chapter also includes discussions of common error patterns, dataset-specific behaviours, and practical implications for operational deployment.

6.1 Topic Modelling

6.1.1 Zero-Shot Classification without Class Descriptions

This first experiment involved prompting each of the eight LLMs without providing them with class descriptions, simulating a zero-shot scenario with minimal guidance. All LLMs were asked to assign a topic label to each document without being explicitly told what each topic meant.

The results in Table 6.1 indicate that all models consistently yielded higher performance on the CODA dataset. This is likely caused by the higher number of classes in the DUTA dataset (24 versus 10), as having too many classes increases the chance of misclassification in case the text

[1] https://www.langchain.com/

contains references to multiple topics, which is easily observable in Figure 6.1. For instance, the *Fraud* class is half of the time confused with *Counterfeit Personal-Identification*.

The worst-performing model on CODA (GPT-4.1 Nano at 69.05% F1-score) still outperformed the best-performing model on DUTA (GPT-4.1at 51.88%) by approximately 17 percentage points. Interestingly, the top two models for CODA were Mistral Small and Mistral Large, while for DUTA, the best two models were GPT-4.1 and GPT-4.1 Mini, making it impossible to extract the influence of model size for this task.

Table 6.1: First Topic Modelling Experiment Results (sorted descendingly by F1-Score)

Dataset	Provider	Model	Accuracy	Precision	Recall	F1
CODA	Mistral	Mistral Small	87.01%	87.52%	88.82%	87.23%
CODA	Mistral	Mistral Large	83.12%	85.08%	87.86%	84.87%
CODA	Google	Gemini 2.0 Flash Lite	84.68%	85.41%	85.85%	84.77%
CODA	OpenAI	GPT-4.1 Mini	85.71%	86.21%	84.62%	84.61%
CODA	Google	Gemini 2.0 Flash	84.16%	84.65%	85.42%	84.38%
CODA	Mistral	Mistral Medium	81.04%	83.92%	87.47%	83.40%
CODA	OpenAI	GPT-4.1	82.60%	82.98%	83.67%	82.27%
CODA	OpenAI	GPT-4.1 Nano	70.39%	70.47%	72.24%	69.05%
DUTA	OpenAI	GPT-4.1	64.94%	56.27%	62.72%	51.88%
DUTA	OpenAI	GPT-4.1 Mini	61.82%	53.97%	62.92%	50.47%
DUTA	Google	Gemini 2.0 Flash	62.34%	60.20%	57.79%	49.70%
DUTA	Mistral	Mistral Medium	61.56%	50.79%	60.26%	48.97%
DUTA	Mistral	Mistral Small	60.52%	49.83%	59.68%	47.88%
DUTA	Mistral	Mistral Large	61.30%	49.88%	59.77%	47.70%
DUTA	Google	Gemini 2.0 Flash Lite	55.58%	55.88%	53.35%	45.32%
DUTA	OpenAI	GPT-4.1 Nano	59.48%	48.98%	50.32%	40.51%

Figure 6.2a shows the confusion matrix for the best-performing model, Mistral Small, on the CODA dataset. The most difficult class to predict was *Others*. This is likely due to the inherently broad and ambiguous nature of the category, whose content may semantically overlap with multiple other topics. For example, documents containing generic security warnings (e.g., about JavaScript errors or login forms) were often misclassified as *Hacking*.

6.1.2 Classification with Class Descriptions (CODA Only)

In the second experiment, class descriptions for the CODA dataset were provided to the models during inference. This change resulted in improvements for several models, particularly the GPT-4.1 Nano. However, an 8.4 percentage points decrease in F1-Score was observed for Gemini Flash 2.0, while Gemini Flash 2.0 Lite, GPT-4.1 Mini, and Mistral Small experienced smaller decreases in performance (see Table 6.2).

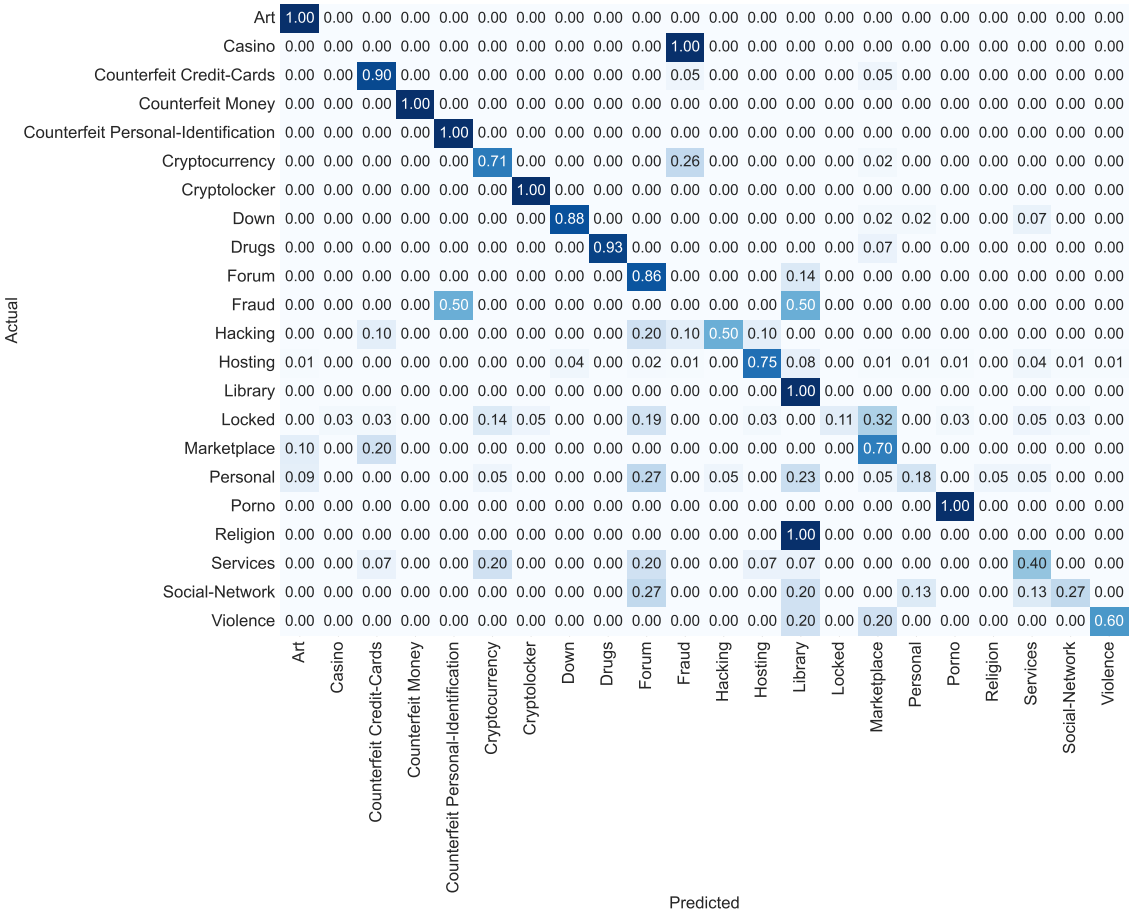
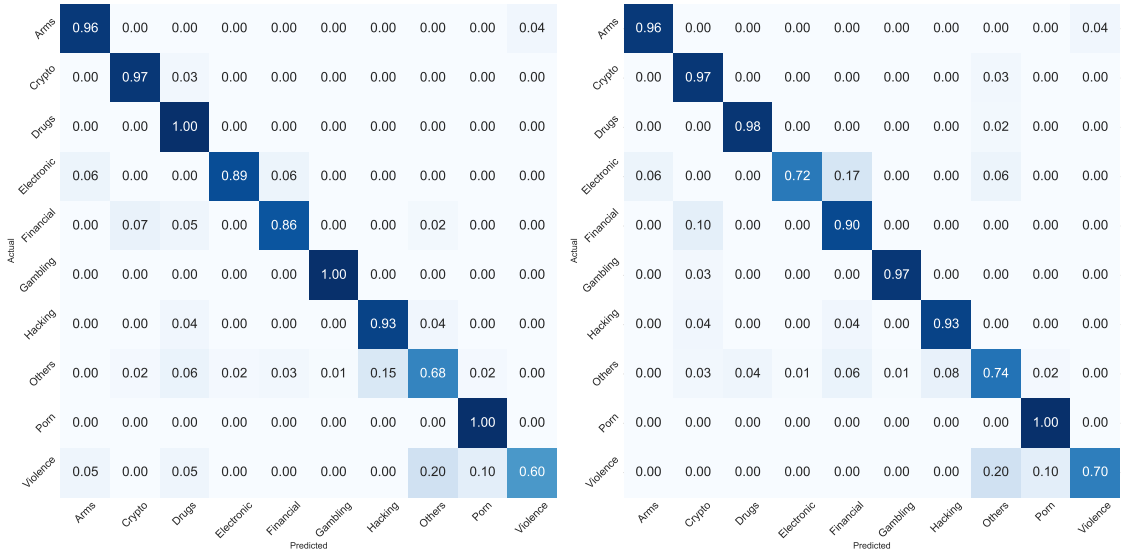


Figure 6.1: GPT-4.1 First Topic Modelling Experiment Confusion Matrix for DUTA Dataset

Figure 6.2b shows the confusion matrix of Mistral Small for the second experiment. This model performed similarly across both experiments with a one percentage point F1-score decrease. The most notable difference was the accuracy related to the *Electronic* class, decreasing from 89% to 72%. However, the accuracy for the *Violence* class improved by 10 percentage points, and misclassifications involving the *Others* category decreased by approximately six percentage points.

These findings indicate that class descriptions could help disambiguate vague categories and support some models in more accurately distinguishing between overlapping classes. However, the results also reveal no definitive trend linking model size to improved performance. In the first experiment, the best-performing models were smaller versions from each provider. Conversely, in the second experiment, Mistral Large achieved the highest F1-score. This suggests that, while smaller models can be effective in certain zero-shot scenarios, the addition of structured guidance (i.e., class descriptions) may help larger models realise their full potential.



(a) First Experiment (b) Second Experiment

Figure 6.2: Mistral Small Topic Modelling Confusion Matrices for CODA Dataset

6.2 Relevance Analysis

This experiment assessed the ability of LLMs to classify Dark Web content according to a predefined four-label relevance scheme: *Irrelevant (IRR)*, *Investigative Lead (IL)*, *Threat No Lead (TNL)*, *Threat With Lead (TWL)*. The models were evaluated using accuracy, precision, recall, and F1-score, as presented in Table 6.3, with additional confusion matrix insights depicted in Figure 6.3.

Across both datasets, GPT-4.1 achieved the strongest performance, obtaining the highest accuracy on DUTA (85.19%) and CODA (75.32%). Notably, these were also the only two configurations to exceed an F1-score of 50% - 60.53% on DUTA and 52.57% on CODA. These values highlight the considerable difficulty of the relevance classification task for the used relevance criteria, particularly given the imbalance in label distributions (see Table 5.8) and the subtle distinctions between categories.

Performance was, in general, higher on the DUTA dataset than on CODA. For example, GPT-4.1 Mini yielded an F1-score of 46.84% on DUTA but only 40.42% on CODA. Similarly, Mistral Small scored 48.75% on DUTA versus 32.12% on CODA. This discrepancy may stem from differences in topic distribution, writing style, or the level of anonymity in the CODA dataset. To assess if the presence of masks in the CODA dataset was impacting the performance, we replaced some of the most relevant masks for lead detection with fictitious values according to Table C.1 and updated the lead label from *Lead* to *No Lead*, causing *Irrelevant* records to change to *Investigative Lead* and *Threat No Lead* items to *Threat With Lead*. We found that this process did not improve the results, as can be verified in Figure C.1. With the exception of GPT-4.1 Nano (which was able to maintain or slightly improve results for the final relevance label), all models obtain lower F1-scores. However, these results do not allow for a definite conclusion on the reasons be-

Table 6.2: Second Topic Modelling Experiment Results (sorted descendingly by F1-Score)

Provider	Model	Accuracy	Precision	Recall	F1
Mistral	Mistral Large (▲1)	89.09%	88.46%	90.69%	89.00% (▲4.11%)
Mistral	Mistral Small (▼1)	88.31%	88.94%	88.74%	88.23% (▼1%)
Mistral	Mistral Medium (▲3)	87.79%	87.68%	89.90%	88.09% (▲4.69%)
OpenAI	GPT-4.1 (▲3)	85.97%	87.36%	84.72%	85.26%(▲2.99%)
OpenAI	GPT-4.1 Mini (▼1)	84.16%	86.19%	81.89%	82.41% (▼2.2%)
Google	Gemini 2.0 Flash Lite (▼3)	84.42%	86.72%	81.65%	81.91% (▼2.86%)
Google	Gemini 2.0 Flash (▼2)	82.34%	75.22%	77.97%	75.98% (▼8.4%)
OpenAI	GPT-4.1 Nano (=)	76.88%	76.30%	74.47%	74.20% (▲5.15%)

▼/▲: in comparison with CODA dataset's results in Table 6.1

hind the lower performance of the models regarding the CODA dataset, as we are still considering synthetic data.

The confusion matrices reveal a recurring pattern: *Threat No Lead* posts are frequently misclassified as *Threat With Lead*, while *Investigative Lead* posts are often confused with other categories. This behaviour could undermine the practical utility of these models in operational settings, where distinguishing between *Threat With Lead* and *Threat No Lead* may be critical for triaging intelligence and allocating investigative resources. Given that the datasets were annotated using the same relevance criteria provided to the LLMs, the results imply that the models struggled to apply the criteria definition to the given contents effectively.

Nonetheless, the relatively strong performance in identifying *Irrelevant* posts - achieving near-perfect accuracy in many cases (e.g., GPT-4.1 on DUTA) - is valuable. In real-world deployments, this could substantially reduce the burden on analysts by allowing LLMs to filter out unimportant content.

In general, the overall low F1-scores and error patterns indicate that current LLMs are not yet reliable substitutes for expert human judgment in relevance classification of Dark Web content. Practical implementations of this kind of system should involve careful definition of relevance criteria and thorough testing and evaluation before real-world deployments.

Two Steps Relevance Analysis

As discussed in Section 5.3.2.2, a two-step classification approach was carried out to assess the impact of reducing the number of instructions in a single prompt. The results of the experiments are plotted in Figures 6.4 and 6.5 for the CODA and DUTA datasets, respectively (red bars under the "Relevance 2 Steps" label). The charts show the F1-scores for the two subtasks, Threat and Lead Identification, the F1-score obtained when combining both steps in a single prompt, and the F1-score for the relevance obtained when merging both sub-classifications. Overall, the results demonstrate the superior effectiveness of prompting the models to carry out a single step of the process rather than including both steps at once. On average, the models performed 8.43 percentage points better on the CODA dataset and 6.4 percentage points on the DUTA dataset.

Table 6.3: Relevance Experiment Results

Dataset	Provider	Model	Accuracy	Precision	Recall	F1
DUTA	OpenAI	GPT-4.1	85.19%	62.13%	66.61%	60.53%
CODA	OpenAI	GPT-4.1	75.32%	53.73%	54.20%	52.57%
DUTA	Mistral	Mistral Small	75.58%	46.69%	55.34%	48.75%
DUTA	OpenAI	GPT-4.1 Mini	82.08%	48.66%	52.44%	46.84%
DUTA	Google	Gemini 2.0 Flash	71.43%	46.42%	59.25%	46.36% (▲6.19%)*
DUTA	Mistral	Mistral Large	74.81%	41.89%	52.85%	44.00%
DUTA	Mistral	Mistral Medium	72.21%	42.01%	54.69%	43.33%
DUTA	OpenAI	GPT-4.1 Nano	76.10%	42.05%	44.87%	42.71%
DUTA	Google	Gemini 2.0 Flash Lite	69.35%	39.20%	59.33%	42.44% (▲1.9%)*
CODA	OpenAI	GPT-4.1 Mini	59.74%	44.27%	49.22%	40.42%
CODA	Mistral	Mistral Small	52.21%	39.21%	45.49%	32.12%
CODA	OpenAI	GPT-4.1 Nano	52.47%	34.98%	39.87%	31.93%
CODA	Mistral	Mistral Large	45.19%	37.68%	44.51%	29.92%
CODA	Google	Gemini 2.0 Flash Lite	39.74%	29.71%	46.06%	25.51% (▲0.74%)*
CODA	Google	Gemini 2.0 Flash	38.70%	29.65%	46.81%	24.79% (▲0.02%)*
CODA	Mistral	Mistral Medium	38.70%	31.24%	37.27%	23.89%

* In comparison with Table 5.7

6.3 Drug Identification

This experiment evaluated the linguistic robustness and domain adaptation of large language models in identifying categories of illicit substances within Dark Web marketplace listings. Despite the pervasive use of drug slang, like code/street names, in such environments, the models demonstrated strong classification capabilities when guided by a constrained prompt referencing a fixed drug taxonomy (see Listing 5.8).

As shown in Table 6.4, the models achieved notably high performance. Seven out of the eight evaluated LLMs surpassed an F1-score of 80%. The best-performing models - GPT-4.1 Mini, Mistral Small, and Gemini Flash 2.0 Lite - yielded F1-scores between 84.00% and 84.50%, with comparable precision and recall values. This suggests that these smaller, more efficient versions retain substantial contextual understanding of domain-specific language despite their reduced parameter counts.

Table 6.4: Drug Identification Results (sorted descendingly by F1-Score)

Provider	Model	Accuracy	Precision	Recall	F1
OpenAI	GPT-4.1 Mini	90.13%	84.92%	86.07%	84.50%
Mistral	Mistral Small	88.31%	83.67%	85.32%	84.12%
Google	Gemini 2.0 Flash Lite	90.13%	85.67%	86.36%	84.00%
Mistral	Mistral Large	89.35%	84.04%	86.11%	83.79%
Mistral	Mistral Medium	88.05%	82.80%	85.66%	83.61%
OpenAI	GPT-4.1	88.57%	82.91%	85.66%	83.22%
Google	Gemini 2.0 Flash	88.83%	83.63%	84.72%	82.60%
OpenAI	GPT-4.1 Nano	83.64%	78.47%	80.66%	77.92%

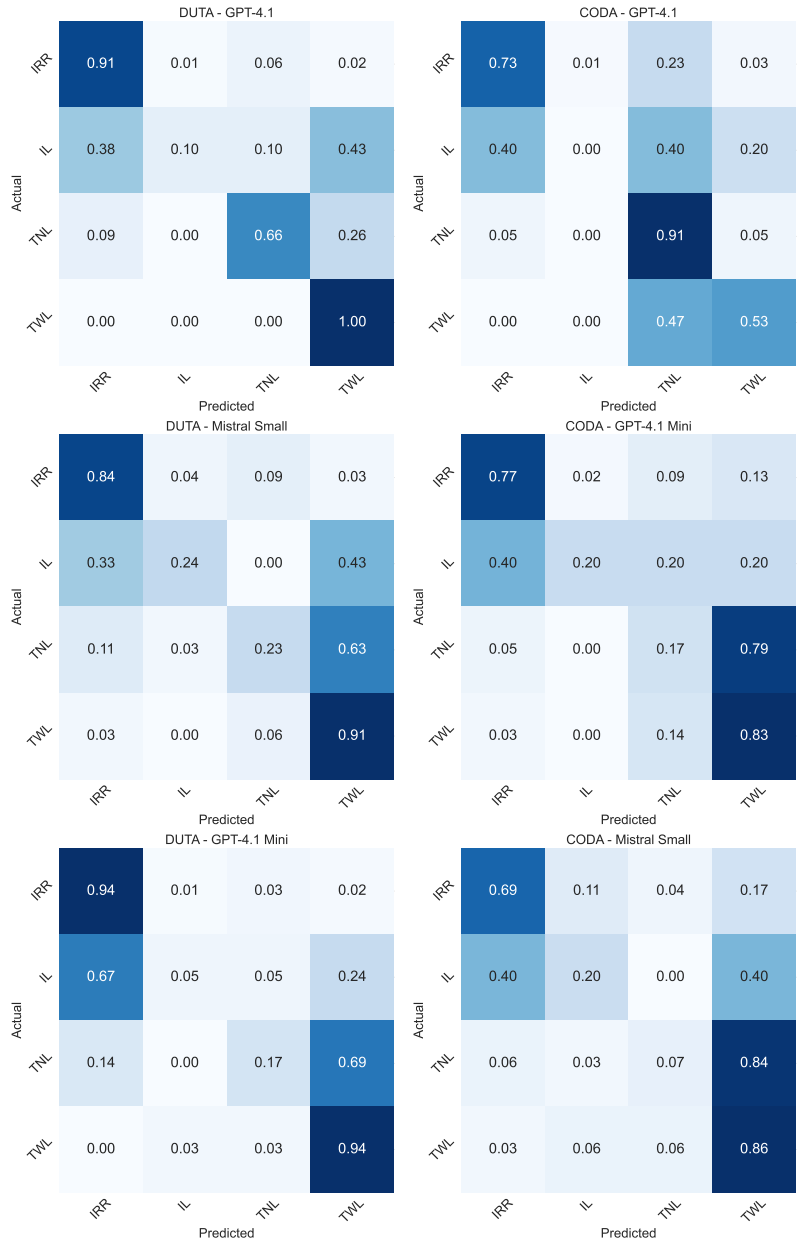


Figure 6.3: Top-3 Models Relevance Confusion Matrices for DUTA and CODA

The overall results indicate a strong consistency across models, with minor performance variations. Accuracy scores ranged from 83.64% to 90.13%, reflecting the models' robust generalisation over the Nemesis dataset.

Analysis of confusion matrices, such as the one shown in Figure 6.6 for GPT-4.1 Mini, reveals a mostly uniform performance across drug categories. Nonetheless, two categories stood out as more challenging: *Opioids* and *Prescription* drugs. *Prescription* drugs, in particular, were frequently misclassified, being accurately identified in only 31% of the cases. This confusion can be attributed to the heterogeneous nature of prescription medications, which span several pharmacological categories, like alprazolam, a benzodiazepine commonly used to treat anxiety,

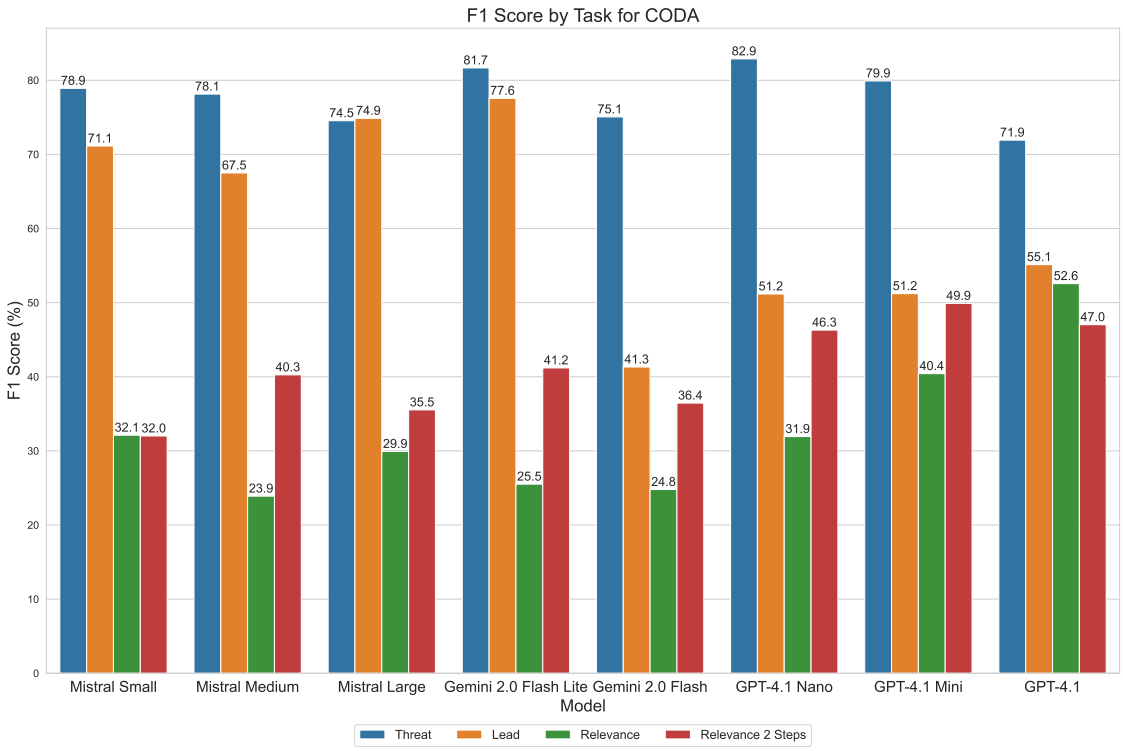


Figure 6.4: Relevance F1-score Results For CODA

and synthetic opioids for pain relief, like oxycodone.

Despite these challenges, the ability of LLMs to handle obfuscated drug-related terminology demonstrates their practical applicability to narcotics intelligence extraction.

6.4 Information Extraction and Policy Formulation

This section presents the results of the Information Extraction and Policy Formulation experiment. Building upon the earlier description of the methodological setup, it aims to qualitatively assess the extent to which state-of-the-art LLMs can extract actionable intelligence and formulate law enforcement recommendations based on marketplace content. Each of the eight selected models was prompted with five randomly chosen listings from the Nemesis Marketplace corpus. Each sample comprised the product listing, the vendor profile (if available) and the user feedback. The models were instructed to populate a structured schema consisting of fields for direct information extraction and analytical recommendations (see Listing 5.11).

Due to time constraints and later-than-anticipated feedback from law enforcement partners, only a subset of five samples from the Nemesis Marketplace dataset was used for this experiment. Although a broader evaluation would have offered a more comprehensive basis for analysis, this focused approach was considered sufficient to capture trends in model performance across the selected dimensions.

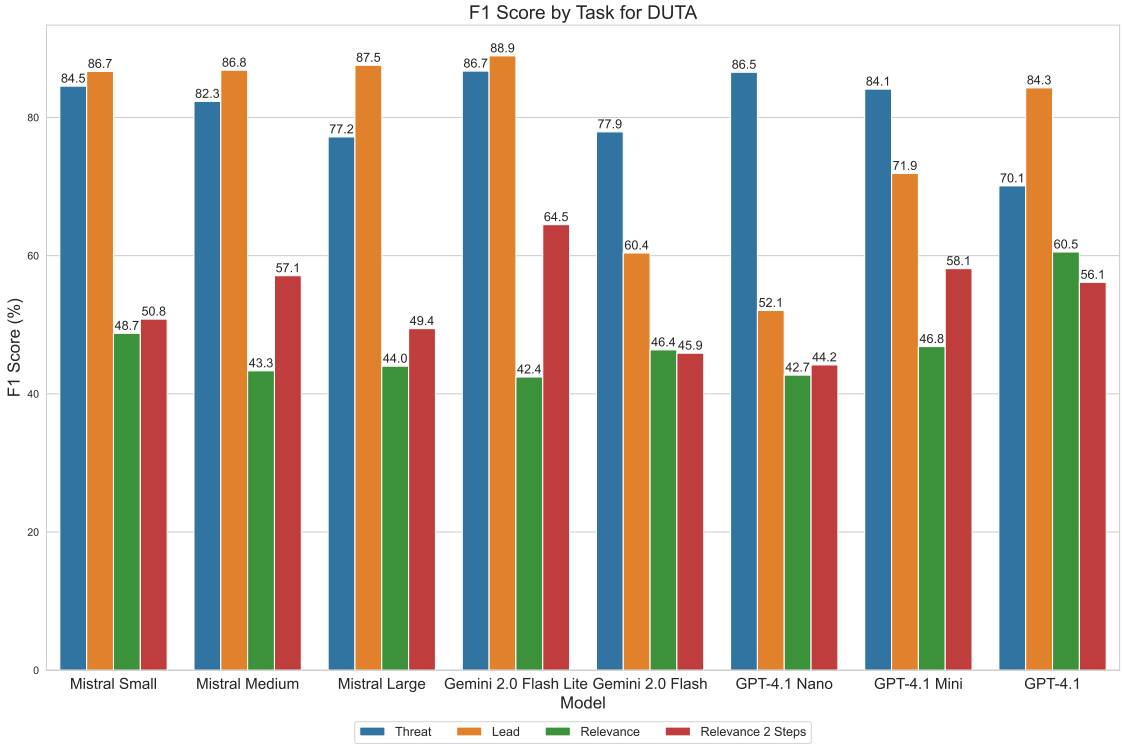


Figure 6.5: Relevance F1-score Results for DUTA

Moreover, given the number of models and fields analysed, and the infeasibility of involving law enforcement agencies in the manual evaluation process within the available timeframe, a manual labelling system (Table 6.5) was formulated, accounting for correctness, completeness, relevance, and conciseness. Each model's output was evaluated field-by-field per item, providing a detailed overview of the models' strengths and weaknesses. The evaluation results, grouped by schema field, are summarised in Table 6.6. Each sequence of numbers in this table represents the evaluation of all models' responses to the corresponding field in the same order as the order of the models in Table 5.6 (the first value is the response from Gemini Flash 2.0, the third from Mistral Large and the sixth from GPT-4.1).

Information Extraction

Extracting surface-level information, such as the type of criminal activity or the nature of the products being sold, was generally successful across most models. In almost all of the five samples, the majority of the models correctly identified drug trafficking as the main illicit activity. Even smaller models, such as Gemini Flash 2.0 Lite and GPT-4.1 Mini, demonstrated consistent precision in populating the `crimes` and `keywords` fields. However, GPT-4.1 Nano left three entries blank.

Nevertheless, there was consistent confusion between the fields `services` and `crimes`. It was frequently observed that models repeatedly used the term "drug trafficking" in the `services` field, thus overlooking the distinction intended by the schema and the criteria outlined in the

Figure 6.6: Drug Identification Experiment Confusion Matrix For GPT-4.1 Mini

prompt. A recurring pattern was also observed in the `targets` field, where instead of identifying potential victims or institutions of a crime (which, in this case, were none), several models treated the prospective buyer of a drug as the "target" of the website.

Furthermore, models demonstrated a tendency to hallucinate or overgeneralise in a number of instances. For instance, despite the absence of such elements in the original data, there were multiple references to encrypted communication channels or anonymisation tactics in the `justice_evasion` field.

Handling of Missing Data

The handling of empty fields deviated largely across models. The better-performing models yielded truly empty responses when a specific piece of information was absent. However, other

Table 6.5: Evaluation Labelling System

Code	Name	Description
0	Wrongly Empty	The field is blank, but should contain relevant information that was clearly extractable
1	Incorrect	Content is factually wrong, hallucinated, or misaligned with the input
2	Valid but Useless	The content is true in general but generic, not actionable, without value or insufficiently precise
3	Valid but Wordy or Incomplete	The response is correct and useful, but either overly verbose or unfocused OR incomplete
4	Correct and Useful	The content is precise, actionable, and meaningfully contributes to analysis.
5	Correctly Empty	The field is blank, and that is appropriate given the input.
6	Empty Message	The field should be empty but contains a placeholder like 'Not specified'

models included default placeholders such as "Not specified in the provided content". The majority of models correctly left fields like `usernames` and `other_user_identifiers` empty when such data was not present.

Policy Recommendations and Strategic Insights

The most critical differentiator in this experiment was the ability of the models to generate high-level assessments and actionable recommendations.

The overall performance in these domains was modest. The content was frequently repetitive and lacking in operational specificity. Common entries included general suggestions such as "enhance international cooperation", "improve monitoring systems", or "collaborate with postal services to track shipments". While not incorrect, these statements were often disconnected from the actual content of the listings. This may suggest that the models were extrapolating from learned patterns rather than reasoning from the input.

With regard to the `overall_website_analysis` field, the majority of the models were able to offer coherent summarisations of the content, indicating the presence of illegal narcotics, vendor credibility signals, and the general level of risk or threat implied by the listing. Despite occasional verbosity, these assessments were, in general, well-aligned with the input.

Concluding Remarks

This experiment reveals both the potential and constraints of using LLMs as instruments for intelligence extraction and policy formulation in the context of Dark Web investigations. On a positive note, most models demonstrated a solid ability to identify illicit activity and summarise structured information in a schema-compliant manner. When evaluated qualitatively, the models' performance in fields such as `crimes`, `keywords`, and `overall_website_analysis` consistently demonstrated high accuracy and relevance.

Table 6.6: Response Evaluation by Label

Label	Item 1	Item 2	Item 3	Item 4	Item 5
crimes	4,4,4,4,4,4,4,0	4,4,4,4,4,4,4,0	4,4,4,4,4,4,4,4	4,4,4,4,4,4,4,0	4,4,4,4,4,4,4,4
services	1,5,5,1,5,1,1,1	1,5,5,5,5,1,1,1	1,5,5,1,1,1,1,1	1,5,5,1,5,1,1,5	1,5,5,1,1,1,1,1
leads	5,5,1,1,5,5,5,1	5,5,5,2,5,2,2,5	1,1,2,1,1,1,2,1	5,1,1,1,1,1,5,5	4,4,3,4,4,4,4,4
geo_loc	5,5,5,6,5,5,5,5	5,5,5,5,5,1,5,5	4,4,4,4,4,4,4,4	5,5,5,6,5,6,6,5	4,4,4,4,4,4,4,4
targets	5,5,5,1,5,1,1,1	5,5,5,5,5,2,2,5	1,1,5,1,1,1,1,1	5,5,1,1,1,1,1,5	5,5,1,1,1,1,1,1
usernames	5,5,5,6,5,5,5,5	5,5,5,5,5,5,5,5	5,5,5,5,5,5,5,5	5,5,5,6,5,5,5,5	4,4,4,4,4,4,4,4
oth_user	5,5,5,6,5,5,5,5	5,5,5,5,5,5,5,5	5,5,5,5,5,5,5,6	5,5,5,6,5,5,5,5	5,1,5,6,1,5,5,1
just_eva	5,5,5,2,5,5,5,5	5,5,5,5,5,5,5,5	5,5,5,1,1,1,1,1	2,2,5,2,2,2,2,5	4,2,4,4,4,4,3,4
com_cha	5,5,5,2,5,5,5,5	5,5,5,5,5,5,5,5	5,5,5,2,2,2,2,2	5,5,5,2,2,2,2,5	5,2,2,2,2,2,2,2
com_pat	5,5,5,1,5,1,1,5	5,5,5,5,5,1,5,5	5,5,5,1,1,1,1,1	5,5,5,2,5,2,2,5	5,2,2,2,2,2,2,2
pay_met	5,5,5,6,5,5,5,5	5,5,5,5,2,5,5,5	5,5,5,5,5,5,5,6	5,5,5,6,5,6,5,5	5,5,5,6,5,5,5,6
ship_met	2,0,2,3,2,2,0,0	5,5,5,5,5,5,5,5	2,2,2,2,2,2,2,2	2,2,2,2,2,2,2,2	4,4,4,4,4,4,4,4
keywords	4,4,4,4,4,4,4,4	4,4,4,4,4,4,4,4	4,4,4,4,4,4,4,4	4,4,4,4,4,4,4,0	4,4,4,4,4,4,4,4
feed_ana	4,4,1,4,4,4,4,4	4,4,4,3,3,4,4,4	2,4,1,2,1,2,1,2	4,4,4,4,4,4,4,4	4,4,4,4,4,4,4,4
inv_sug	2,0,0,2,2,2,2,0	2,0,2,2,0,2,2,2	2,2,2,2,2,2,2,2	2,2,2,2,2,2,2,2	3,4,3,4,4,3,4,4
pol_rec	2,2,2,2,2,2,2,0	2,2,2,2,2,2,2,0	2,2,2,2,2,2,2,2	2,2,2,2,2,2,2,2	4,4,4,4,4,4,4,4
web_ana	4,4,3,4,4,4,4,4	4,4,4,4,3,4,4,4	4,4,4,4,4,4,4,4	4,4,4,4,4,4,4,4	4,4,4,4,4,4,4,4

geo_loc: geographical_locations; **oth_user**: other_user_identifiers; **just_eva**: justice_evasion;
com_cha: communication_channels; **com_pat**: communication_patterns;
pay_met: payment_methods; **ship_met**: shipment_methods;
feed_ana: feedback_analysis; **inv_sug**: criminal_investigations_suggestions;
pol_rec: policy_recommendations; **web_ana**: overall_website_analysis

However, strategic and policy-level insights were frequently superficial, highlighting limitations in applying LLMs to more abstract reasoning tasks under loosely defined conditions. Additionally, common confusions between schema elements (e.g., `services` vs `crimes`, `targets` vs buyers) indicate that further prompt engineering or model fine-tuning may be necessary to improve reliability in high-stakes operational environments.

6.5 Summary

This chapter presented and analysed the results of four experiments evaluating LLM performance on Dark Web intelligence tasks: topic modelling, relevance classification, drug identification, and structured information extraction and policy formulation.

In the topic modelling task, models performed better on datasets with fewer classes and clearer labels, such as CODA. The inclusion of class descriptions highlighted how prompt formulation can significantly influence performance. Relevance classification proved more difficult. Although most models achieved high overall accuracy, this was driven largely by their ability to detect irrelevant content. However, a two-step formulation by assessing threat and lead separately improved model reliability, suggesting that decomposing complex tasks can enhance LLM consistency.

The drug identification task demonstrated strong model performance, with most achieving F1-scores above 80%. Conversely, in the structured extraction task, models reliably identified surface-level details. As for policy formulation capabilities, while some models could generate plausible summaries, most lacked the legal or operational depth needed for policy development.

Chapter 7

Future Work

This chapter introduces a conceptual design for an LLM-powered system, building on the experimental findings discussed in the previous chapters. The system is designed to assist LEAs in real-time monitoring, analysis and extraction of intelligence from Dark Web content. While experiments have revealed that LLMs have the capacity to contribute to relevance detection, content classification or policy formulation, the absence of a unified, scalable and customisable platform imposes limitations on their operational applicability. The proposed system is intended not merely as a passive analytic tool but as an active intelligence and policy-support environment, addressing this gap by envisioning a modular architecture capable of integrating Dark Web crawling, large-scale content analysis, and investigative assistance in a coherent and user-driven workflow.

It is worth noting that the present dissertation did not involve the implementation of this system due to temporal constraints and the limited availability of requirements from LEA partners. Nevertheless, the plan presented here is designed to provide a strong base for future work and possible development projects in this area.

7.1 Requirements

This section presents the functional and non-functional requirements that the system must satisfy in order to be reliable and adaptable to real-world investigative contexts. The following requirements were derived from a combination of the analysis of the empirical evidence gathered throughout the LLM experiments detailed in the previous two chapters of this document and late-stage interactions with LEAs. These were further extended with additional practical and technical considerations grounded in possible operational needs, system design best practices, and the specific constraints posed by Dark Web monitoring.

7.1.1 Functional Requirements

The functional requirements of the system are centred on the delivery of operational value to intelligence analysts. This is achieved by automating key stages of Dark Web monitoring, enhancing

situational awareness, and providing structured support for decision-making. These requirements are categorised into core capabilities that the system should fulfil.

Dark Web Crawling and Scraping. The system must autonomously access onion services via the TOR network, dynamically discover internal and external links, and download the relevant content of each visited page. This content, which may include marketplace listings, forum discussions, vendor profiles, and user feedback, should be parsed and normalised into a structured representation that facilitates downstream analysis. The parser must support diverse text formats, like HTML or Markdown, while also performing deduplication and basic fingerprinting to avoid redundancy.

LLM-Powered Information Extraction. The backbone of the analytical layer is an ensemble of LLM-driven modules that classify, extract, and analyse the ingested data. These modules must be capable of:

- Classifying content by type (e.g., listing, discussion).
- Determining relevance using a fine-grained multi-label taxonomy (e.g., irrelevant, investigative lead, threat with lead).
- Identifying references to criminal services and goods and extracting contact methods, shipment practices, and financial modalities.
- Extracting named entities and indicators relevant to investigations, such as usernames, PGP keys, cryptocurrency addresses, geographical locations, targets, and potential victims.
- Suggesting high-level insights such as policy actions or operational strategies.

Cross-Market Surveillance. The system should incorporate mechanisms to facilitate the linking of users, listings, and behaviours across multiple Dark Web platforms, despite the use of aliases or slightly modified handles.

Customisability and Analyst-Guided Configuration. The framework must also provide a high degree of personalisability. For example, an investigator might configure the system to prioritise content related to a particular marketplace or type of illicit product. These preferences should influence both the relevance criteria and the subsequent extraction strategies employed by the system. Such configurations would guide the behaviour of system agents, enabling LLMs to adapt prompts and execution strategies accordingly.

Continuous Monitoring and Alerting. The system should support near-real-time crawling and analysis pipelines, enabling automated alerts when newly ingested content matches user-defined relevance criteria or threat patterns. These alerts should include justifications and links to the original content to enable fast triage.

Data Model and Storage Layer. The underlying data layer should support, among other things, the storage of raw and processed data, semantic similarity searches, and maintenance of detailed system logs for transparency and auditability.

7.1.2 Non-Functional Requirements

Beyond functional aspects, the system must satisfy a range of non-functional properties:

Security and Privacy. All data must be handled following legal and ethical standards for sensitive information. Communications should be encrypted, access-controlled, and auditable. Anonymised handling of potentially identifying content is required, especially when dealing with victim-related information.

Scalability. The system must be capable of horizontal scaling in order to accommodate an expanding number of crawled sites and queries without degradation in responsiveness or analytical accuracy.

Robustness. Given the volatility of Dark Web services, the system must properly handle inaccessible pages, non-responsive sites, or malformed content, continuing execution without data loss or pipeline collapse.

Modularity and Extensibility. The architecture must allow easy replacement or augmentation of components, such as swapping an LLM with a newer model or integrating image-based analysis modules in the future.

Explainability. LLM outputs must include interpretable rationales for decisions and extractions, allowing human analysts to evaluate the trustworthiness and coherence of model responses.

Real-Time Responsiveness. In scenarios of high-stakes monitoring (e.g., threats of violence or attacks), the system must deliver outputs with minimal latency from crawl to classification to alert.

7.2 Architecture

This section presents a conceptual four-layer architecture that operationalises the above requirements, succinctly depicted in Figure 7.1.

Data Collection and Ingestion Pipeline The entry point of the system is a TOR-compatible crawler, capable of following links recursively across onion services. It then parses the websites' raw HTML, keeping only text or other media types like images and videos.

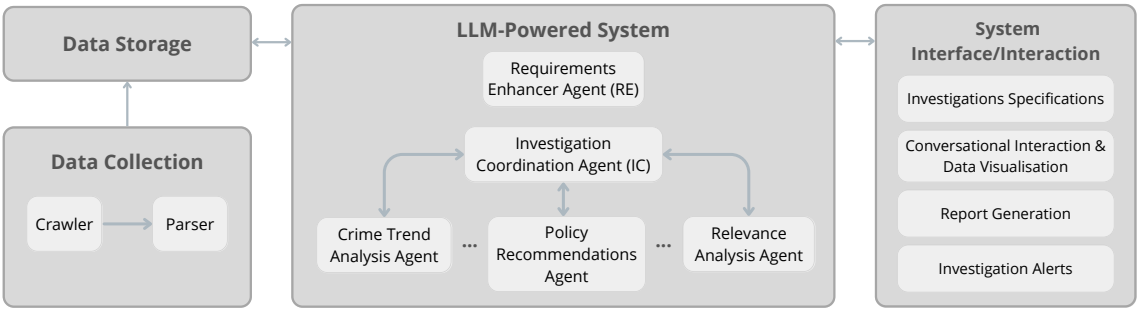


Figure 7.1: High Level System Architecture and Components

Data Storage Parsed and annotated data could be stored in a relational database (e.g., PostgreSQL[1]) with vector support for similarity search (e.g., through the pgvector extension[2]). The system should also log all LLM-agent and user interactions.

LLM-Based Analysis Agents The analytical core of the system is structured as a multi-agent ecosystem, where each agent performs a distinct but complementary role. Coordination is handled through a central Investigation Coordinator agent, as detailed below.

- **Requirements Enhancer Agent (RE).** Upon user interaction, the RE agent assists in formulating structured and precise investigative goals. It helps translate informal requests into system-compatible configurations and constraints, enabling guided execution. For instance, it may help transform "Track fentanyl sales in Germany" into a formal agent plan involving geolocation filters and drug listings detectors.
- **Investigation Coordinator Agent (IC).** This agent receives directives from the RE and orchestrates a plan of action. It determines which Operator Agents to invoke and in what sequence. The IC is responsible for synthesising the outputs into a coherent report that aligns with the user's investigative intent.
- **Operator Agents.** Each operator agent is specialised for a task and interacts directly with the underlying data and models. Examples could include a *Crime Trend Analysis Agent* that executes database queries to detect shifts in product offerings or pricing patterns over time; a *Relevance Analysis Agent* to flag relevant contents, or a *Policy Recommendations Agent* for website content summarisation or strategic investigation suggestions.

User Interaction and Interface The user interface should be designed to facilitate intuitive and flexible interaction with the system while supporting the analytical workflows of law enforcement professionals.

The primary front-end may include a dynamic dashboard that presents curated views of crawled content, vendor profiles, extracted intelligence, and risk assessments. Users can filter and explore

[1] https://www.postgresql.org/
[2] https://www.postgresql.org/about/news/pgvector-050-released-2700/

listings, review associated metadata or and access timeline-based activity visualisations to track evolving threats.

Another key feature would be integrating a conversational interface powered by a natural language understanding layer, allowing users to interact with the system through queries such as "Show all listings offering stealth shipping to the UK" or "Identify sellers active in both Marketplace A and Marketplace B". This enables a non-technical access to complex querying and reasoning over structured data.

The interface may also support the configuration of real-time alerting mechanisms. Users may define monitoring rules, such as the appearance of specific keywords, geographies, product types, or vendor names, and subscribe to corresponding alerts. When a match is detected during crawling and analysis, the system dispatches a notification, optionally including a summary and prioritised relevance score, to designated personnel or dashboards.

This multimodal interface design prioritises usability without sacrificing depth, enabling both strategic oversight and tactical interrogation of the Dark Web intelligence corpus.

7.3 Usage Example

Let us consider a scenario in which an LEA officer seeks to investigate patterns of cocaine trafficking across the Atlantic Ocean. The officer begins by creating a new investigative module within the system interface. Through a conversation with the *Requirements Enhancer Agent*, the officer specifies the module's focus: identifying content related to cocaine shipments and associated vendor activity between South America, West Africa, and Europe. The officer may further define the expected outputs, such as threat summaries, lead identifications, and suggestions for cross-market entity resolution.

The *Investigation Coordinator Agent* receives these requirements and schedules a routine execution cycle. It instructs the relevant operator agents to filter incoming content for thematic relevance (e.g., cocaine, ports, shipment), identify emergent user handles, and generate a summary report synthesising findings, trends, and strategic suggestions. Upon completion, the system alerts the officer and provides a downloadable report containing structured findings and policy implications.

7.4 Summary

This chapter presented a comprehensive proposal for an LLM-powered system to support law enforcement agencies in continuously monitoring, analysing, and exploiting Dark Web data. Grounded in the experimental findings and informed by operational insights gathered from LEA interactions, the chapter articulated a coherent set of functional and non-functional requirements that such a system must satisfy.

The proposed architecture introduces a modular and extensible framework comprising Dark Web ingestion pipelines, a multi-agent LLM analysis layer, and a user interface for investigation, reporting, and policy support. Particular emphasis was placed on customisability, real-time alerting, investigative traceability, and the integration of strategic recommendation modules - all essential for practical deployment in high-stakes law enforcement contexts.

Although the system was not implemented due to project constraints and the later-than-anticipated timing of partner feedback, this design establishes a solid conceptual foundation for future research and development efforts. It may help bridge the gap between experimental LLM capabilities and real-world operational needs, proposing a path toward scalable, interpretable, and domain-aligned intelligence systems for Dark Web surveillance.

Chapter 8

Conclusions

8.1 Final Remarks

This dissertation investigated the applicability of Large Language Models (LLMs) to the analysis of Dark Web data in support of Law Enforcement Agencies. Motivated by the growing complexity of cybercriminal ecosystems and the vast, unstructured nature of Dark Web content, the research aimed to determine whether LLMs could be used to assist in extracting relevant and policy-relevant intelligence. Through a series of four experiments, the study explored the ability of several state-of-the-art models to perform key Threat Intelligence tasks while also assessing their practical value for investigative workflows. As such, we were able to answer the formulated research questions as follows:

- **RQ1: How do current state-of-the-art LLMs perform at Threat Intelligence (TI) tasks, and how can they be leveraged for analytical and investigative purposes?**

  The experiments demonstrated that current LLMs can provide useful analytical and investigative support across a range of TI tasks, including topic modelling, relevance classification (namely detecting *Irrelevant* content), and structured information extraction. While larger models can achieve higher accuracy and consistency, smaller open-source models, like Mistral Small and GPT-4.1 Mini, also exhibited strong performance in specific tasks at a more efficient ratio between computation cost and performance. Notably, models were able to identify key concepts or generate structured outputs with minimal task-specific training data, underscoring their potential for rapid prototyping in intelligence contexts.

  However, limitations were observed in generalisation across domains, especially when semantic ambiguity or topic overlap was present (e.g., in topic modelling or drug identification). Despite having lower training on Dark Web-specific content, the models were still able to extract operationally relevant information, albeit with occasional irrelevant associations. This suggests that, while LLMs hold promise for investigative augmentation, their outputs should be validated or monitored, particularly when deployed in real-world intelligence pipelines.

- **RQ2: How can LLMs be used to simplify law enforcement user interaction while maximising result accuracy and relevance in Dark Web TI?**

  Several design choices were found to significantly influence model output quality and usability for non-technical users, namely the positive impact of a custom system prompting strategy, as well as the importance of clear instructions and benefits of complex task decomposition for improved performance, particularly in the relevance analysis task.

- **RQ3: How do different LLMs compare in identifying threats and offering actionable defence strategies from Dark Web text data?**

  Relevance Analysis, including Threat and Lead Identification, proved a challenging task for most models, achieving modest results in F1-score (between $\sim$24% and $\sim$61%), even though some models achieved high accuracies for the sampled datasets.

  Across other experiments, performance differences between models were not always correlated with model size. While larger models such as GPT-4.1 and Mistral Large exhibited superior generalisation and abstraction capabilities, smaller models like GPT-4.1 Mini and Mistral Small achieved comparable results in more structured tasks like drug name identification or label-based classification, respectively. This reinforces the notion that model scale alone is insufficient for policy reasoning tasks and that fine-tuning or domain-specific retrieval may be necessary for such applications.

Overall, these findings support the idea that LLMs can serve as user-aligned tools for law enforcement, provided that the interaction layer is designed to scaffold model understanding and constrain ambiguity.

8.2 Limitations and Future Work

Despite promising results, this study is subject to several limitations that can affect:

- **Limited exposure to domain-specific data:** None of the evaluated models were explicitly trained on dark web content or threat intelligence corpora. As such, many outputs lacked domain precision or context sensitivity.

- **Lack of policy-specific grounding:** Particularly in the policy extraction task, models demonstrated a limited understanding of legal, procedural, or security frameworks. This likely stems from a lack of training on real-world policy documents, defence guidelines, or law enforcement protocols.

- **Indeterministic outputs:** The experiments relied on single-response evaluations per prompt due to computational constraints. As LLMs are inherently non-deterministic, this introduces response variability that could have been mitigated with multiple generations and aggregated scores.

- **Dataset and sampling constraints:** The experiments were conducted on subsets of larger datasets, which may not capture the full distributional complexity of dark web content. Full-scale evaluations were not performed due to cost and time considerations.

To address these limitations and further explore the potential of LLMs in dark web intelligence applications, several avenues for future work are proposed:

- **Fine-tuning on TI and policy data:** A key direction is to fine-tune LLMs on curated corpora that include defence policies, security guidelines, and real-world TI reports. This could improve their ability to generate grounded, actionable recommendations and avoid hallucinations in critical tasks.
- **Larger-scale and multi-run evaluations:** Expanding experiments to include the full datasets or newer collected data and running inference multiple times per prompt would offer more statistically robust insights into model performance and consistency.
- **Law enforcement collaboration:** Partnering with law enforcement agencies could enable (1) more targeted data collection, (2) better task formulation according to specific requirements and (3) identification of new TI use cases with prompt engineering aligned with real investigative workflows.

References

[1] Victor Adewopo, Bilal Gonen, and Festus Adewopo. "Exploring Open Source Information for Cyber Threat Intelligence". In: *2020 IEEE International Conference on Big Data (Big Data)*. Dec. 2020, pp. 2232–2241. DOI: 10.1109/BigData50022.2020.9378220. URL: https://ieeexplore.ieee.org/document/9378220 (visited on 10/31/2024).

[2] Sirwan Khalid Ahmed. "How to choose a sampling technique and determine sample size for research: A simplified guide for researchers". In: *Oral Oncology Reports* 12 (Dec. 2024), p. 100662. ISSN: 2772-9060. DOI: 10.1016/j.oor.2024.100662. URL: https://www.sciencedirect.com/science/article/pii/S2772906024005089 (visited on 06/20/2025).

[3] Mhd Wesam Al Nabki et al. "Classifying Illegal Activities on Tor Network Based on Web Textual Contents". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 35–43. URL: https://aclanthology.org/E17-1004 (visited on 09/27/2024).

[4] Wesam Al Nabki et al. "ToRank: Identifying the Most Influential Suspicious Domains in the Tor Network". In: *Expert Systems with Applications* 123 (June 2019). DOI: 10.1016/j.eswa.2019.01.029.

[5] Prasasthy Balasubramanian, Justin Seby, and Panos Kostakos. "Semantic-Driven Focused Crawling Using LASER and FAISS: A Novel Approach for Threat Detection and Improved Information Retrieval". In: *2023 IEEE 22nd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*. ISSN: 2324-9013. Nov. 2023, pp. 1598–1605. DOI: 10.1109/TrustCom60117.2023.00218. URL: https://ieeexplore.ieee.org/document/10538633 (visited on 11/01/2024).

[6] Iz Beltagy, Matthew E. Peters, and Arman Cohan. *Longformer: The Long-Document Transformer*. arXiv:2004.05150 [cs]. Dec. 2020. DOI: 10.48550/arXiv.2004.05150. URL: http://arxiv.org/abs/2004.05150 (visited on 01/27/2025).

[7] Timothy J Berners-Lee. *Information Management: A Proposal*. Geneva, 1989. URL: https://cds.cern.ch/record/369245 (visited on 10/02/2024).

[8] Philip A. Bernstein. "Synthesizing third normal form relations from functional dependencies". In: *ACM Trans. Database Syst.* 1.4 (Dec. 1976), pp. 277–298. ISSN: 0362-5915. DOI: 10.1145/320493.320489. URL: https://dl.acm.org/doi/10.1145/320493.320489 (visited on 06/20/2025).

[9] Marco San Biagio et al. "MARPLE: A Framework for Social Media Threat Intelligence". In: *2024 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA)*. Feb. 2024, pp. 1–6. DOI: 10.1109/ACDSA59508.2024.10467738. URL: https://ieeexplore.ieee.org/document/10467738 (visited on 11/01/2024).

[10] Bhargava Bokkena. "Enhancing IT Security with LLM-Powered Predictive Threat Intelligence". In: *2024 5th International Conference on Smart Electronics and Communication (ICOSEC)*. Sept. 2024, pp. 751–756. DOI: 10.1109/ICOSEC61587.2024.10722712. URL: https://ieeexplore.ieee.org/document/10722712 (visited on 02/21/2025).

[11] Gwern Branwen et al. *Dark Net Market archives, 2011–2015*. https://gwern.net/dnm-archive. dataset. Accessed: DATE. July 2015. URL: https://gwern.net/dnm-archive.

[12] Giuseppe Cascavilla, Gemma Catolino, and Mirella Sangiovanni. "Illicit Darkweb Classification via Natural-language Processing: Classifying Illicit Content of Webpages based on Textual Information". English. In: *Proceedings of the International Conference on Security and Cryptography*. Vol. 1. ISSN: 21847711 Type: Conference paper. Science and Technology Publications, Lda, 2022, pp. 620–626. ISBN: 978-989-758-590-6. DOI: 10.5220/0011298600003283. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85178502961&doi=10.5220%2f0011298600003283&partnerID=40&md5=fd6a0ee04f7401f3fc93a81d92bb16d1.

[13] Hongfan Chen et al. "Decode the Dark Side of the Language: Applications of LLMs in the Dark Web". In: *2024 IEEE 9th International Conference on Data Science in Cyberspace (DSC)*. Aug. 2024, pp. 224–231. DOI: 10.1109/DSC63484.2024.00037. URL: https://ieeexplore.ieee.org/document/10858984 (visited on 02/21/2025).

[14] Othmane Cherqi et al. "Enhancing Cyber Threat Identification in Open-Source Intelligence Feeds Through an Improved Semi-Supervised Generative Adversarial Learning Approach With Contrastive Learning". In: *IEEE Access* 11 (2023), pp. 84440–84452. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2023.3299604.

[15] Nicolas Christin. "Traveling the silk road: a measurement analysis of a large anonymous online marketplace". en. In: *Proceedings of the 22nd international conference on World Wide Web*. Rio de Janeiro Brazil: ACM, May 2013, pp. 213–224. ISBN: 978-1-4503-2035-1. DOI: 10.1145/2488388.2488408. URL: https://dl.acm.org/doi/10.1145/2488388.2488408 (visited on 09/23/2024).

[16] Kevin Clark et al. *ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators*. arXiv:2003.10555 [cs]. Mar. 2020. DOI: 10.48550/arXiv.2003.10555. URL: http://arxiv.org/abs/2003.10555 (visited on 06/29/2025).

[17] William Gemmell Cochran. *Sampling techniques*. john wiley & sons, 1977.

[18] Danilo Croce, Giuseppe Castellucci, and Roberto Basili. "GAN-BERT: Generative Adversarial Learning for Robust Text Classification with a Bunch of Labeled Examples". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 2114–2119. DOI: 10.18653/v1/2020.acl-main.191. URL: https://aclanthology.org/2020.acl-main.191/ (visited on 01/27/2025).

[19] Alejandro Cuevas and Nicolas Christin. "Does Online Anonymous Market Vendor Reputation Matter?" en. In: *Proceedings of the 33rd USENIX Security Symposium (USENIX Security'24). Philadelphia, PA. To appear*. 2024, pp. 4641–4656. ISBN: 978-1-939133-44-1. URL: https://www.usenix.org/conference/usenixsecurity24/presentation/cuevas (visited on 05/31/2025).

[20] Daniel De Pascale et al. “Services Computing for Cyber-Threat Intelligence: The ANITA Approach”. English. In: *Communications in Computer and Information Science* 1360 (2021). ISBN: 978-303071905-0 Publisher: Springer Science and Business Media Deutschland GmbH Type: Conference paper, pp. 166–172. ISSN: 18650929. DOI: 10.1007/978-3-030-71906-7_15. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85103522080&doi=10.1007%2f978-3-030-71906-7_15&partnerID=40&md5=b4d07edbb6568bdaaf8149d255aa5c30.

[21] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423/.

[22] Roger Dingledine, Nick Mathewson, and Paul Syverson. “Tor: The {Second-Generation} Onion Router”. en. In: 2004. URL: https://www.usenix.org/conference/13th-usenix-security-symposium/tor-second-generation-onion-router (visited on 10/02/2024).

[23] Mohammadreza Ebrahimi, Jay F. Nunamaker, and Hsinchun Chen. “Semi-Supervised Cyber Threat Identification in Dark Net Markets: A Transductive and Deep Learning Approach”. English. In: *Journal of Management Information Systems* 37.3 (2020). Publisher: Routledge Type: Article, pp. 694–722. ISSN: 07421222. DOI: 10.1080/07421222.2020.1790186. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85096159504&doi=10.1080%2f07421222.2020.1790186&partnerID=40&md5=e55577c0692394b19268d1c05bda3a10 (visited on 11/01/2024).

[24] Europol. *IOCTA, internet organised crime threat assessment 2024*. Publications Office of the European Union, Luxembourg, 2024.

[25] Europol. *Massive blow to criminal Dark Web activities after globally coordinated operation*. en. July 2017. URL: https://www.europol.europa.eu/media-press/newsroom/news/massive-blow-to-criminal-dark-web-activities-after-globally-coordinated-operation (visited on 10/03/2024).

[26] Eibe Frank, Mark A. Hall, and Ian H. Witten. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*. https://www.cs.waikato.ac.nz/ml/weka/book.html. Online Appendix. 2016.

[27] Shalini Ghosh et al. “ATOL: A Framework for Automated Analysis and Categorization of the Darkweb Ecosystem”. In: 2017. URL: https://www.semanticscholar.org/paper/ATOL%3A-A-Framework-for-Automated-Analysis-and-of-the-Ghosh-Porras/7d12d8336e48396663c0aa511a4533838d03a393 (visited on 07/23/2024).

[28] Sean E. Goodison et al. *Identifying Law Enforcement Needs for Conducting Criminal Investigations Involving Evidence on the Dark Web*. en. Tech. rep. RAND Corporation, Oct. 2019. URL: https://www.rand.org/pubs/research_reports/RR2704.html (visited on 10/02/2024).

[29] Aaron Grattafiori et al. *The Llama 3 Herd of Models*. arXiv:2407.21783 [cs]. Nov. 2024. DOI: 10.48550/arXiv.2407.21783. URL: http://arxiv.org/abs/2407.21783 (visited on 04/18/2025).

[30] Alex Graves. *Generating Sequences With Recurrent Neural Networks*. arXiv:1308.0850 [cs]. June 2014. DOI: 10.48550/arXiv.1308.0850. URL: http://arxiv.org/abs/1308.0850 (visited on 02/15/2025).

[31] Maarten Grootendorst. *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv:2203.05794 [cs]. Mar. 2022. DOI: 10.48550/arXiv.2203.05794. URL: http://arxiv.org/abs/2203.05794 (visited on 01/27/2025).

[32] Neal R. Haddaway et al. "PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis". In: *Campbell Systematic Reviews* 18.2 (June 2022). Publisher: John Wiley & Sons, Ltd, e1230. ISSN: 1891-1803. DOI: 10.1002/cl2.1230. URL: https://doi.org/10.1002/cl2.1230 (visited on 05/06/2022).

[33] Siyu He, Yongzhong He, and Mingzhe Li. "Classification of Illegal Activities on the Dark Web". In: *Proceedings of the 2nd International Conference on Information Science and Systems*. ICISS '19. New York, NY, USA: Association for Computing Machinery, Mar. 2019, pp. 73–78. ISBN: 978-1-4503-6103-3. DOI: 10.1145/3322645.3322691. URL: https://doi.org/10.1145/3322645.3322691 (visited on 07/23/2024).

[34] Jeremy Howard and Sebastian Ruder. *Universal Language Model Fine-tuning for Text Classification*. arXiv:1801.06146 [cs]. May 2018. DOI: 10.48550/arXiv.1801.06146. URL: http://arxiv.org/abs/1801.06146 (visited on 01/27/2025).

[35] Cheng Huang et al. "HackerRank: Identifying key hackers in underground forums". English. In: *International Journal of Distributed Sensor Networks* 17.5 (2021). Publisher: SAGE Publications Ltd Type: Article. ISSN: 15501329. DOI: 10.1177/15501477211015145. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85105530510&doi=10.1177%2f15501477211015145&partnerID=40&md5=c55fe5827755e74ef2a0304684bd9d6c.

[36] Federal Bureau of Investigation. *Ross Ulbricht, the Creator and Owner of the Silk Road Website, Found Guilty in Manhattan Federal Court on All Counts — FBI*. en-us. Press Release. Feb. 2015. URL: https://www.fbi.gov/contact-us/field-offices/newyork/news/press-releases/ross-ulbricht-the-creator-and-owner-of-the-silk-road-website-found-guilty-in-manhattan-federal-court-on-all-counts (visited on 10/03/2024).

[37] Albert Q. Jiang et al. *Mistral 7B*. arXiv:2310.06825 [cs]. Oct. 2023. DOI: 10.48550/arXiv.2310.06825. URL: http://arxiv.org/abs/2310.06825 (visited on 04/18/2025).

[38] Youngjin Jin et al. "DarkBERT: A Language Model for the Dark Side of the Internet". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 7515–7533. DOI: 10.18653/v1/2023.acl-long.415. URL: https://aclanthology.org/2023.acl-long.415 (visited on 09/27/2024).

[39] Youngjin Jin et al. "Shedding New Light on the Language of the Dark Web". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 5621–5637. DOI: 10.18653/v1/2022.naacl-main.412. URL: https://aclanthology.org/2022.naacl-main.412 (visited on 09/27/2024).

[40] Masashi Kadoguchi et al. "Deep Self-Supervised Clustering of the Dark Web for Cyber Threat Intelligence". English. In: *2020 IEEE International Conference on Intelligence and Security Informatics (ISI)*. Place: Arlington, VA. IEEE Press, 2020, pp. 1–6. ISBN: 978-1-72818-800-3. DOI: 10.1109/ISI49825.2020.9280485. URL: https://doi.org/10.1109/ISI49825.2020.9280485 (visited on 10/31/2024).

[41] Masashi Kadoguichi et al. "Exploring the Dark Web for Cyber Threat Intelligence using Machine Leaning". English. In: *2019 IEEE International Conference on Intelligence and Security Informatics (ISI)*. Place: Shenzhen, China. IEEE Press, 2019, pp. 200–202. ISBN: 978-1-72812-504-6. DOI: 10.1109/ISI.2019.8823360. URL: https://doi.org/10.1109/ISI.2019.8823360 (visited on 07/22/2024).

[42] Liang Ke et al. "A novel cross-domain adaptation framework for unsupervised criminal jargon detection via pre-trained contextual embedding of darknet corpus". English. In: *Expert Syst. Appl.* 242.C (May 2024). Place: USA Publisher: Pergamon Press, Inc., p. 122715. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2023.122715. URL: https://doi.org/10.1016/j.eswa.2023.122715.

[43] Patrick Lewis et al. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. arXiv:2005.11401 [cs]. Apr. 2021. DOI: 10.48550/arXiv.2005.11401. URL: http://arxiv.org/abs/2005.11401 (visited on 04/18/2025).

[44] Tianyang Lin et al. *A Survey of Transformers*. arXiv:2106.04554 [cs]. June 2021. DOI: 10.48550/arXiv.2106.04554. URL: http://arxiv.org/abs/2106.04554 (visited on 02/11/2025).

[45] Yinhan Liu et al. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv:1907.11692 [cs]. July 2019. DOI: 10.48550/arXiv.1907.11692. URL: http://arxiv.org/abs/1907.11692.

[46] Tomáš Mikolov et al. "Recurrent neural network based language model". In: 2010, pp. 1045–1048. DOI: 10.21437/Interspeech.2010-343. URL: https://www.isca-archive.org/interspeech_2010/mikolov10_interspeech.html (visited on 02/11/2025).

[47] Humza Naveed et al. *A Comprehensive Overview of Large Language Models*. arXiv:2307.06435 [cs]. Oct. 2024. DOI: 10.48550/arXiv.2307.06435. URL: http://arxiv.org/abs/2307.06435 (visited on 01/27/2025).

[48] Saiba Nazah et al. "An Unsupervised Model for Identifying and Characterizing Dark Web Forums". English. In: *IEEE Access* 9 (2021). Publisher: Institute of Electrical and Electronics Engineers Inc. Type: Article, pp. 112871–112892. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3103319. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85114001890&doi=10.1109%2fACCESS.2021.3103319&partnerID=40&md5=4e066fffd6240efcdc7a89e39419a88e (visited on 07/23/2024).

[49] Anastasios Nikolakopoulos et al. "Large Language Models in Modern Forensic Investigations: Harnessing the Power of Generative Artificial Intelligence in Crime Resolution and Suspect Identification". In: *2024 5th International Conference in Electronic Engineering, Information Technology & Education (EEITE)*. May 2024, pp. 1–5. DOI: 10.1109/EEITE61750.2024.10654427. URL: https://ieeexplore.ieee.org/document/10654427 (visited on 02/21/2025).

[50] Eric Nunes, Paulo Shakarian, and Gerardo I. Simari. "At-risk system identification via analysis of discussions on the darkweb". English. In: *2018 APWG Symposium on Electronic Crime Research (eCrime)*. Vol. 2018-May. ISSN: 2159-1245. IEEE Computer Society, May 2018, pp. 1–12. ISBN: 978-1-5386-4922-0. DOI: 10.1109/ECRIME.2018.8376211. URL: https://ieeexplore.ieee.org/document/8376211/?arnumber=8376211 (visited on 10/31/2024).

[51] Eric Nunes et al. "Darknet and deepnet mining for proactive cybersecurity threat intelligence". English. In: *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*. Place: Tucson, AZ, USA. IEEE Press, 2016, pp. 7–12. ISBN: 978-1-5090-3865-7. DOI: 10.1109/ISI.2016.7745435. URL: https://doi.org/10.1109/ISI.2016.7745435.

[52] Dario Onorati et al. "The Dark Side of the Language: Syntax-Based Neural Networks Rivaling Transformers in Definitely Unseen Sentences". In: *2023 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. Oct. 2023, pp. 111–118. DOI: 10.1109/WI-IAT59888.2023.00021.

[53] Matthew J. Page et al. "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews". en. In: *BMJ* 372 (Mar. 2021). Publisher: British Medical Journal Publishing Group Section: Research Methods & Reporting, n71. ISSN: 1756-1833. DOI: 10.1136/bmj.n71. URL: https://www.bmj.com/content/372/bmj.n71 (visited on 10/25/2024).

[54] Anum Atique Paracha, Junaid Arshad, and Muhammad Mubashir Khan. "S.U.S. You're SUS!—Identifying influencer hackers on dark web social networks". English. In: *Computers and Electrical Engineering* 107 (2023). Publisher: Elsevier Ltd Type: Article. ISSN: 00457906. DOI: 10.1016/j.compeleceng.2023.108627. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85148330075&doi=10.1016%2fj.compeleceng.2023.108627&partnerID=40&md5=105fa38d50eaf4e7e90477f2f

[55] Javier Pastor-Galindo et al. "A Big Data architecture for early identification and categorization of dark web sites". English. In: *Future Gener. Comput. Syst.* 157.C (July 2024). Place: NLD Publisher: Elsevier Science Publishers B. V., pp. 67–81. ISSN: 0167-739X. DOI: 10.1016/j.future.2024.03.025. URL: https://doi.org/10.1016/j.future.2024.03.025.

[56] Mary Phuong and Marcus Hutter. *Formal Algorithms for Transformers*. arXiv:2207.09238 [cs]. July 2022. DOI: 10.48550/arXiv.2207.09238. URL: http://arxiv.org/abs/2207.09238 (visited on 02/11/2025).

[57] Chaido Porlou et al. "Optimizing an LLM Prompt for Accurate Data Extraction from Firearm-Related Listings in Dark Web Marketplaces". In: *2024 IEEE International Conference on Big Data (BigData)*. ISSN: 2573-2978. Dec. 2024, pp. 2821–2830. DOI: 10.1109/BigData62323.2024.10825446. URL: https://ieeexplore.ieee.org/abstract/document/10825446/keywords#full-text-header (visited on 02/10/2025).

[58] Alec Radford et al. "Improving Language Understanding by Generative Pre-Training". en. In: (June 2018).

[59] Mohaimenul Azam Khan Raiaan et al. “A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges”. In: *IEEE Access* 12 (2024). Conference Name: IEEE Access, pp. 26839–26874. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2024.3365742. URL: https://ieeexplore.ieee.org/document/10433480 (visited on 01/29/2025).

[60] Leonardo Ranaldi et al. “Shedding Light on the Dark Web: Authorship Attribution in Radical Forums”. English. In: *Information (Switzerland)* 13.9 (2022). Publisher: MDPI Type: Article. ISSN: 20782489. DOI: 10.3390/info13090435. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85138699501&doi=10.3390%2finfo13090435&partnerID=40&md5=921423e9e9dce74b7c496efb5317802b.

[61] Kanti Singh Sangher, Archana Singh, and Hari Mohan Pandey. “LSTM and BERT based transformers models for cyber threat intelligence for intent identification of social media platforms exploitation from darknet forums”. English. In: *International Journal of Information Technology (Singapore)* 16.8 (2024). Publisher: Springer Science and Business Media B.V. Type: Article, pp. 5277–5292. ISSN: 25112104. DOI: 10.1007/s41870-024-02077-5. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85201187585&doi=10.1007%2fs41870-024-02077-5&partnerID=40&md5=237fe360bed34d140ba453daf8243872 (visited on 11/01/2024).

[62] Kanti Singh Sangher et al. “Towards Safe Cyber Practices: Developing a Proactive Cyber-Threat Intelligence System for Dark Web Forum Content by Identifying Cybercrimes”. English. In: *Information (Switzerland)* 14.6 (2023). Publisher: MDPI Type: Article, p. 349. ISSN: 20782489. DOI: 10.3390/info14060349. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85163762697&doi=10.3390%2finfo14060349&partnerID=40&md5=bd18f4b29037ac2d44d59894fc39d64b (visited on 09/23/2024).

[63] Paria Sarzaeim, Qusay H. Mahmoud, and Akramul Azim. “A Framework for LLM-Assisted Smart Policing System”. In: *IEEE Access* 12 (2024). Conference Name: IEEE Access, pp. 74915–74929. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2024.3404862. URL: https://ieeexplore.ieee.org/document/10538107 (visited on 02/21/2025).

[64] Sayuj Shah and Vijay K. Madisetti. “MAD-CTI: Cyber Threat Intelligence Analysis of the Dark Web Using a Multi-Agent Framework”. In: *IEEE Access* 13 (2025), pp. 40158–40168. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2025.3547172. URL: https://ieeexplore.ieee.org/document/10908603 (visited on 05/31/2025).

[65] Kyle Soska and Nicolas Christin. “Measuring the longitudinal evolution of the online anonymous marketplace ecosystem”. In: *24th USENIX security symposium (USENIX security 15)*. Read_Status: New Read_Status_Date: 2024-10-19T14:30:03.683Z. 2015, pp. 33–48. URL: https://www.usenix.org/conference/usenixsecurity15/technical-sessions/presentation/soska (visited on 09/23/2024).

[66] Nazgol Tavabi et al. “DarkEmbed: exploit prediction with neural language models”. English. In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI’18/IAAI’18/EAAI’18. Place: New Orleans, Louisiana, USA. AAAI Press, 2018, pp. 7849–7854. ISBN: 978-1-57735-800-8. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85102203426&partnerID=40&md5=39ce20a7fe5fc9b52f7e84443d9b78dd.

[67] Gemma Team et al. *Gemma: Open Models Based on Gemini Research and Technology*. arXiv:2403.08295 [cs]. Apr. 2024. DOI: 10.48550/arXiv.2403.08295. URL: http://arxiv.org/abs/2403.08295 (visited on 04/18/2025).

[68] *Treasury Sanctions Head of Online Darknet Marketplace Tied to Fentanyl Sales*. en. Feb. 2025. URL: https://home.treasury.gov/news/press-releases/sb0040 (visited on 05/31/2025).

[69] A. M. Turing. "Computing Machinery and Intelligence". In: *Mind, New Series* 59.236 (1950), pp. 433–460. URL: http://www.jstor.org/stable/2251299.

[70] Velizar Varbanov. "Advancing Cyber Threat Intelligence through Machine Learning Algorithms". English. In: *Proceedings of the Cognitive Models and Artificial Intelligence Conference*. AICCONF '24. event-place: undefinedstanbul, Turkiye. New York, NY, USA: Association for Computing Machinery, June 2024, pp. 111–115. ISBN: 9798400716928. DOI: 10.1145/3660853.3660879. URL: https://dl.acm.org/doi/10.1145/3660853.3660879 (visited on 10/31/2024).

[71] Varsha Varghese, Mahalakshmi S, and Senthilkumar Kb. "Extraction of Actionable Threat Intelligence from Dark Web Data". English. In: *2023 International Conference on Control, Communication and Computing (ICCC)*. Type: Conference paper. Institute of Electrical and Electronics Engineers Inc., May 2023, pp. 1–5. ISBN: 979-835033412-8. DOI: 10.1109/ICCC57789.2023.10165477. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85165130454&doi=10.1109%2fICCC57789.2023.10165477&partnerID=40&md5=4ebc398390b60a273e4df3bfb972c7f1 (visited on 11/01/2024).

[72] Ashish Vaswani et al. "Attention is All you Need". In: June 2017. URL: https://www.semanticscholar.org/paper/Attention-is-All-you-Need-Vaswani-Shazeer/204e3073870fae3d05bcbc2f6a8e263d9b72e776 (visited on 10/02/2024).

[73] Xuren Wang et al. "DNRTI: A Large-Scale Dataset for Named Entity Recognition in Threat Intelligence". In: *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*. Guangzhou, China: IEEE, Dec. 2020, pp. 1842–1848. ISBN: 978-1-6654-0392-4. DOI: 10.1109/TrustCom50675.2020.00252. URL: https://ieeexplore.ieee.org/document/9343158/ (visited on 01/27/2025).

[74] Jonathan J. Webster and Chunyu Kit. "Tokenization as the Initial Phase in NLP". In: *COLING 1992 Volume 4: The 14th International Conference on Computational Linguistics*. 1992. URL: https://aclanthology.org/C92-4173/ (visited on 02/18/2025).

[75] Zhilin Yang et al. *XLNet: Generalized Autoregressive Pretraining for Language Understanding*. arXiv:1906.08237 [cs]. Jan. 2020. DOI: 10.48550/arXiv.1906.08237. URL: http://arxiv.org/abs/1906.08237 (visited on 06/29/2025).

[76] Fabio Massimo Zanzotto et al. "KERMIT: Complementing Transformer Architectures with Encoders of Explicit Syntactic Interpretations". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 256–267. DOI: 10.18653/v1/2020.emnlp-main.18. URL: https://aclanthology.org/2020.emnlp-main.18/ (visited on 06/29/2025).

[77] Azene Zenebe et al. “Cyber Threat Discovery from Dark Web”. en. In: *EPiC Series in Computing*. Vol. 64. ISSN: 23987340 Type: Conference paper. EasyChair, 2019, pp. 174–163. DOI: 10.29007/nkfk. URL: https://easychair.org/publications/paper/MK31 (visited on 10/31/2024).

[78] Zhengyan Zhang et al. *ERNIE: Enhanced Language Representation with Informative Entities*. arXiv:1905.07129 [cs]. June 2019. DOI: 10.48550/arXiv.1905.07129. URL: http://arxiv.org/abs/1905.07129 (visited on 06/29/2025).

[79] Wayne Xin Zhao et al. *A Survey of Large Language Models*. arXiv:2303.18223 [cs]. Oct. 2024. DOI: 10.48550/arXiv.2303.18223. URL: http://arxiv.org/abs/2303.18223 (visited on 02/04/2025).

[80] Vilém Zouhar et al. *A Formal Perspective on Byte-Pair Encoding*. en. arXiv:2306.16837 [cs]. Sept. 2024. DOI: 10.48550/arXiv.2306.16837. URL: http://arxiv.org/abs/2306.16837 (visited on 02/11/2025).

Appendix A

State of the Art Table

Table A.1: Summary of Extracted Information from Selected Literature Review Documents

Work	Problem statement	Dataset collection	Proposed solution	Results	Limitations
[51]	Identify actionable cyber threat intelligence from DWFs and DWMs from a high volume of unstructured data	Crawled 27 DWMs and 21 DWFs (12K products and 240K posts, with 13% and 19% relevance to hacking, respectively. Only trained with data from 10 DWMs	Combined supervised and semi-supervised ML classifiers	Achieved 92% recall for hacking-related products and 80% for relevant forum topics	Relies on manually curated sites and expert-labelled datasets
[50]	Identify systems at risk from hacker discussions on the dark web	Crawled 302 DWFs and DWMs (over 6M posts)	Combining DeLP and ML classifiers for multi-label classification to identify at-risk platforms, vendors, and products.	Improved precision by 15%-57% over baseline models while maintaining comparable recall	Limited representation of certain components in training data and inability to address newly identified products
[66]	Better predict exploited vulnerabilities effectively by capturing context from text	Crawled 200 sites relating to malicious hacking and/or online financial fraud.	DarkEmbed: paragraph embeddings and SVM to predict and classify vulnerabilities' exploitability	DarkEmbed achieved a 0.74 F1-score in the exploiting prediction task	Relies on explicit CVE mentions, excluding unstructured exploit discussions

[41]	Extracting actionable CTI from dark web forum posts	500 critical and 2,500 non-critical posts labeled as "malware offers" or not from Sixgill	Employs doc2vec and Multi-Layer Perceptron to classify posts as critical or non-critical	Achieved a mean accuracy of 93.9% on training data	Limited dataset size, reducing accuracy to 79.4% on unseen data
[77]	Address the limitations of reactive cybersecurity measures by adopting a proactive approach	DWFs posts spanning 2003-2016 collected from the University of Arizona's AI Lab	Descriptive and visual analytics to identify trends in exploit types and key actors, and ML classifiers	Overrepresentation of system exploits biased classifiers	Dataset lacked adequate examples of less common exploit types
[1]	Correlate surface and DW data to identify cyber threats	0.5M tweets and 128k DWF posts + 8k records of data breaches from the Privacy Rights Clearinghouse	RF to integrate and classify cyber threat data from diverse sources	0.81 F1-score; Random Forest outperformed LR and GB	Manual annotation for initial training and accurate classification
[23]	Address the challenges of limited labelled data and high data variability in DNMs	80K product listings from 7 DWMs, with 3% manually labelled and the remainder processed using TSVM	TSVM for semi-supervised labelling and BiLSTM for deep sequence classification	Achieved an F1-score of 89.55% and a 2% increase in the area under the ROC curve compared to baseline supervised methods	Struggles to account for evolving data patterns
[40]	Reduce need for manually labeling of data in DW content monitoring tools	Collected 850 relevant and 850 non relevant malware offers DWFs	doc2vec for feature extraction and deep clustering with an autoencoder and K-means to classify critical posts without manual labeling	The model classified 57% of posts with 97% accuracy during training, but 89% of unseen data were misclassified into mixed clusters	Small dataset size, errors in unseen data classification, manual hyperparameter tuning
[20]	Assist LEAs fighting illegal trafficking	Scraped multiple black markets pages	ANITA: systems to detect criminal activities, trends, and networks	A RF classifier achieved 81.66% accuracy in detecting illegal activities	Real-time monitoring scalability, manual parameter tuning

[35]	Identify key hackers in underground forums	Crawled 5 DWMs, gathering threads, posts, and user interactions	Content analysis and social network analysis using an improved PageRank algorithm to automatically identify and rank key hackers	Outperformed standalone CA or SNA methods with an average coverage rate increase of 3.14% and 16.19%, respectively	Doesn't address cross-forum identity linkage
[48]	Reduce need for manually labelling of data	Dataset by Gwern Branwen (focused on the Agora market)	Unsupervised clustering and DT for identification and characterisation	0.98 F1-score	Requires manual cluster labelling
[12]	Classify illicit content to support LEAs	DUTA-10K + 148 DW sites	Compared LSTM, ULMFit, BERT, and RoBERTa	BERT achieved the highest accuracy of 96.08% for general illicit activity classification and 91.98% for specific drug type identification	Identified overfitting challenges in the models
[60]	Identify and trace users in radical Islamic forums on the dark web	The dataset contains 91,874 posts from the Islamic Network forum, focusing on the 10 most radical users, with 1,400 posts per user.	Compared lexical, syntactic, and transformer models for authorship attribution	POS-based TF-IDF achieved 84.9% accuracy for 3 authors, BERT underperformed due to poor adaptation to unseen data	The results for transformer models are limited by the lack of dark web-specific pretraining
[5]	Identify semantic context in textual data to reduce relying on keyword analysis	Crawled 1,759 web pages, 59% Surface Web and 41% from the Dark Web	Combine LASER embeddings and FAISS indexing to rank pages based on semantic similarity. Fine-tuned Longformer model to classify crawled pages as relevant or irrelevant	Achieved a harvest rate of 87% and an irrelevance ratio of 13%, with the Longformer classifier reaching 96% accuracy in page classification.	The system faces latency issues due to embedding generation and pruning
[14]	Address limited availability of annotated data	Crawled data from Open Threat Exchange, Exploits' Database, GDELT, and DWMs	ConGAN-BERT: semi-supervised GANs with contrastive learning to improve threat classification	Up to a 16-point accuracy improvement using limited annotated data	The maximum F1-score was only 55.36% using 2% of annotated data

[38]	Address low exposure to linguistic characteristics of DW data of the traditional NLP models	Crawled 6.1M DW pages, filtered to approximately 5.83 GB of text	DarkBERT: Pretrained RoBERTa with gathered data	DarkBERT achieved higher F1-scores (80% for DUTA, 94% for CoDA)	Specific tasks still require labeled datasets and fine-tuning for optimal performance
[52]	Classify content as legal or illegal	DUTA10K + eBay drug posts	Proposed lexical and syntactic-based NNs compared to 4 transformers with FT and extreme domain adaptation	NNs are preferred to LM's with extreme domain adaptation	Good results require extreme domain adaptation
[54]	Identify influential users in DWFs by analysing user profiles, interactions, and activities	BeginnerHacking forum, contained in the CrimeBB database by the Cambridge Cyber Crime Center	INSPECT framework: Social Network Analysis, Semantic Analysis using LDA, and ML to calculate user influence scores	The framework effectively identified influential users, filtered noise, and provided accurate influence scores based on SNA	High score does not guarantee user influence; can be deceived by detecting advanced bots or deceptive profiles.
[62]	Identify and categorise cybercrimes in DW forums	Agora DWM dataset: 109,000+ annotations entries about drugs, forgeries, and hacking services	Compared RNN, CNN, LSTM, and BERT to classify activities as "Cybercrime", "Not Cybercrime", or "Can't Say"	BERT achieved the highest accuracy of 96% and strong F1-scores	Struggle with vague descriptions
[71]	Identify meaningful cybersecurity-related entities	Fine-tuned with DNRTI dataset. Tested with crawled DW data	Fine-tuned BERT to identify and classify entities as HackerIDs, tools, organisations, or offensive activities	Achieved an F1-score close to 0.9	No actual evaluation was performed using dark web data
[9]	Identify terrorism suspicious content to support LEA's in surpressing threats	The proposed framework collected both surface and DW data	Use BERT to extract features from multilingual texts and binarily classify posts	Achieved 93% F1-score and 99% accuracy	Relatively small testing DW dataset (1100 posts)
[42]	Detect chinese jargon on the Dark Web	Gathered 44,720 posts from 6 chinese DWM	Compare semantic similarity by using contextual embeddings with DC-BERT	91.5% accuracy ascertaining valid jargon	Classification is restricted to 10 crime categories

[55]	Early identification and DW content analysis in near-real time	72,045 HTML DW documents	BERTopic for topic extraction and document clustering	Characterized 90% of the site correctly	Manual labelling and cluster merging
[61]	Categorise DW criminal activities in social media platforms	235,668 reviews from 6 DWFs (Gwern Brandwen dataset)	Compare LR, RF, GB, KNN, GB, VC, LSTM and BERT classification of cybercrime intents	LSTM achieved an accuracy of 91.24% and BERT 90.12% in categorising cybercrimes into hacking, fraud, drugs, and phishing	The classification was restricted to four broad categories, lacking deeper subcategories or relationships
[70]	Identify relevant and emergent threads in dark web forums and marketplaces	Data from 3 public DW datasets	Apply LR, DT, GB, RF to classify text as relevant or not for CTI	F1-score ranging from 0.73 (DT) to 0.86 (LR)	Relatively small dataset

Appendix B

Authorities Survey

The following questions were shared with relevant Portuguese Authorities by means of an online survey, having obtained two answers from two authorities, later identified as A1 and A2.

What type of information is most relevant when analysing data from various types of websites on the Dark Web?

A1 - Exposed Credentials; Confidential Organization Data; Early Threat Detection

A2 - The way to access a specific site on the Darknet (there are sites that require registration or another type of access control), the various .onion addresses that refer to that site, information contained in the profile of the markets/sellers.

Is this information dependent on the type of platform (e.g., online marketplaces vs. forums)? If so, how do they differ from each other?

A1 - ***Did Not Answer***

A2 - Yes, each site/marketplace has its own specificities, and they determine how access is provided. For example, discussion forums often require only the creation of an account to access content. Some marketplaces are freely accessible, while others require registration and other control methods determined by the marketplace.

What types of platforms are most relevant for analysis?

A1 - Online Marketplaces

A2 - Online Marketplaces and Discussion Forums

Of the following data types, which would be most relevant for analysis (one or both)? (1) Products (namely drugs, weapons, etc.) listed on various online marketplaces (Silk Road, Agora, Nemesis, etc.), as well as the seller's username, country of origin, user sales, etc. (2) Textual content from websites (buyer feedback regarding the product/seller they purchased, user profile, etc). To what extent are each of the points relevant, and how could they be explored within the scope of this project?

A1 - Textual content of websites.

A2 - Both. Regarding the first point, this allows for the collection of information regarding a seller's "specialty" and possibly the origin of the product being sold, as well as the locations to which orders are shipped, since some marketplaces ship orders anywhere in the world, while others are limited to specific geographic areas. Furthermore, it is common for sellers to be present in multiple marketplaces with the same user, which makes it easier to locate them. Regarding the second point, the textual content of websites is essential for understanding how the marketplace/seller operates. Through the information contained in their profile, it is possible, for example, to identify the contact, purchase, payment, and shipping methods, which can ultimately lead to the complete identification of the seller. Also, by analyzing feedback provided by buyers, it is possible to gather information about the product and its seller. In discussion forums, it is common for several buyers to share their experiences with specific products, sellers, and marketplaces, sometimes sharing crucial information for identifying them. Within the scope of this project, this data could be exploited to create a general profile of a given seller or marketplace, with access to all their essential information, allowing a global view of their mode and method of action. It would also be essential to create a tool to track posts with specific keywords, e.g., "cocaine", "white", "brown", or even a specific username, to facilitate the search and collection of information. This will contribute to developing investigation and police action strategies, with the aim of identifying those responsible for the seller or marketplace profile, leading to future seizures and/or arrests.

What databases are currently used for investigations?

A1 - Internal to the organization. OSINT

A2 - PJ databases, none containing darkweb data.

Are there any public databases or databases that, although not public, could be shared so that it would be possible to carry out a comparative analysis through LLMs of the content extracted from the Dark Web with previously identified references/entities? (a.) Lists of identified entities. (b.) Lists of names/codes and their meaning. (c.) Others? If so, please tell us how we can access the databases and documents.

A1 - Lists of names/codes and their meaning.

A2 - No

What are the main obstacles when analysing Dark Web data? (a.) Language (multiple languages, jargon, coded language, etc.)? (b.) Quantity and volume of data? (c.) Others?

A1 - a.

A2 - Language, quantity and volume of data, difficulty in accessing information contained on websites, lack of tools to support research on the dark web, difficulties in tracking IP, the fact that the data contained in profile creations is unconfirmed data.

How would the implementation of an LLM system be helpful in analysing the data? What (types of) questions would you like to see answered by these models?

A1 - Smuggling and Illegal Trade; Prevention; Incident Response; Law Enforcement; Crime-as-a-Service

A2 - An LLM system would facilitate the collection of essential information relating to a given profile/market, contributing to the easy identification of markets and profiles of interest within the scope of a given investigation.

What kind of suggestions or action policies would be useful for the system to provide (e.g., suggestions for investigations, resource allocation, geographic and/or crime areas to prioritise over others)?

A1 - Investigation suggestions; geographic areas; Cooperation with Security Forces

A2 - Identification of geographic area, identification of crimes, identification of profiles/markets of interest for investigation and general information about them.

Appendix C

Relevance Analysis CODA Unmasked Values and Results

Table C.1: Fictitious Values Used to Replace CODA Masks

Mask	Value
ID_IP_ADDRESS	198.255.111.158*
ID_EMAIL	gy4xbtp05w@bltiwd.com**
ID_ONION_URL	gy4xbtp05w.onion
ID_NORMAL_URL	https://www.bltiwd.com
ID_BTC_ADDRESS	mq2bZG2oaJxoohMB6BFDKsMbEhZQAKWiDT***
ID_ETH_ADDRESS	0x1C38E03643578685D4B9D89CA7F6B682524236B0***
ID_LTC_ADDRESS	mkwVX7d4qBEHj2ACwpHk9a4oL6L6X1gn2C***

*https://www.ipvoid.com/random-ip/, ** https://temp-mail.io/en,
***https://randommer.io/bitcoin-address-generator

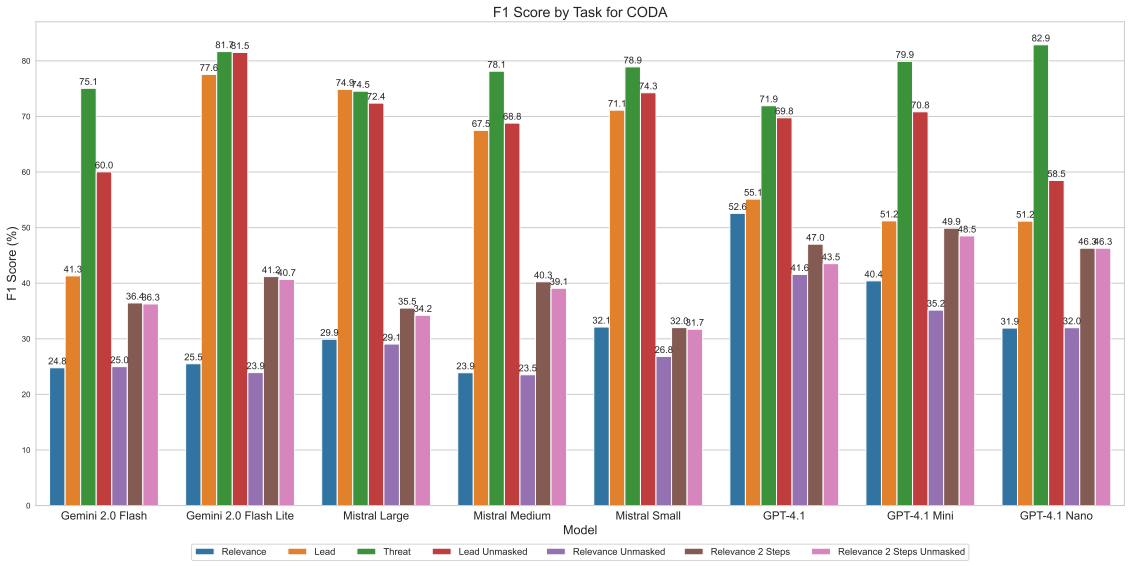


Figure C.1: Relevance F1-score Results for CODA Unmasked