Cairo University

Faculty of Computer Science and Artificial Intelligence

Artificial Intelligence Department

NeoSilk: AI-Enhanced Dark Web Monitoring

By:

Abdelhamid Mahmoud Ahmed **20210192**

Adham Tarek Abdelaziz **20210051**

Mohammed Hassanien Sayed **20210333**

Mostafa Aly Hashem **20210394**

Nour El-Dien Mostafa Mohammed **20210435**

Supervisor: ***Dr. Dina Tarek Mohammed***

IT Department

Abstract

With the increasing volume of illicit activity occurring on dark web marketplaces, there is a growing need for automated systems capable of extracting, analyzing, and presenting cyber threat intelligence.Our graduation project, **NeoSilk: AI-Enhanced Dark Web Monitoring**, presents a full pipeline that integrates dark web data collection, advanced natural language processing (NLP), machine learning (ML), and visualization. Data was collected from both CAPTCHA-protected and non-CAPTCHA '.onion' websites using a Tor-enabled crawler. CAPTCHA bypassing was implemented using a semi-automated solver to facilitate access to authenticated content. The collected data — including product names, descriptions, prices, vendor details, and customer feedback — was cleaned and processed for downstream tasks. The AI phase includes three core NLP tasks: (1) **Product Classification** into darknet-related categories using transformer-based models (e.g., BERT, RoBERTa, DarkBERT), (2) **Sentiment Analysis** on user feedback leveraging product ratings, and (3) exploration of a **Retrieval-Augmented Generation (RAG)** module to enable question answering over darknet listings. Following model training and evaluation, an extensive data analysis was conducted to explore category distributions, shipping behaviors, vendor activity, and pricing trends. These insights were integrated into a fully interactive dashboard to assist cybersecurity professionals in visual threat monitoring. The final system serves as a practical, AI-driven solution that transforms raw dark web content into actionable intelligence. It highlights the potential of combining crawling, NLP, and dashboard visualization for cyber threat detection and darknet monitoring.

Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Full Form
AI	Artificial Intelligence
API	Application Programming Interface
BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolutional Neural Network
CRNN	Convolutional Recurrent Neural Network
CSS	Cascading Style Sheets
CSV	Comma-Separated Values
DarkBERT	Domain-Specific BERT Trained on Dark Web Text
DAX	Data Analysis Expressions
DDoS	Distributed Denial of Service
ECE	Expected Calibration Error
EDA	Exploratory Data Analysis
ETL	Extract, Transform, Load
FAISS	Facebook AI Similarity Search
GloVe	Global Vectors for Word Representation
GPT	Generative Pretrained Transformer
GPU	Graphics Processing Unit
HTML	HyperText Markup Language
HTTP	Hypertext Transfer Protocol
KPI	Key Performance Indicator
LDA	Latent Dirichlet Allocation
LSVC	Linear Support Vector Classifier
LSTM	Long Short-Term Memory
MCE	Maximum Calibration Error

Abbreviation	Full Form
MFA	Multi-Factor Authentication
ML	Machine Learning
NLL	Negative Log-Likelihood
NLP	Natural Language Processing
OCR	Optical Character Recognition
PGP	Pretty Good Privacy
QA	Question Answering
RAG	Retrieval-Augmented Generation
RoBERTa	Robustly Optimized BERT Pretraining Approach
SHAP	SHapley Additive exPlanations
SIEM	Security Information and Event Management
SQL	Structured Query Language
SVM	Support Vector Machine
TextCNN	Text Convolutional Neural Network
Tor	The Onion Router
ULMFiT	Universal Language Model Fine-tuning
URL	Uniform Resource Locator
XAI	Explainable Artificial Intelligence

Chapter 1

Introduction

1 CHAPTER 1 Introduction

The rise of the dark web has introduced an expansive ecosystem for illicit transactions, posing critical challenges to cybersecurity efforts worldwide. Traditional monitoring techniques struggle to keep pace with the dynamic, unstructured, and highly anonymized nature of darknet marketplaces. As a result, there is a growing need for automated, intelligent solutions that can collect, interpret, and act on data from these hidden networks. This chapter introduces the problem landscape, outlines the objectives of the NeoSilk project, and presents the structure of the report, which is designed to guide readers through each phase of the developed system—from data collection to AI modeling and final visualization.

1.1 Problem Definition

The dark web serves as a hub for a wide array of illicit activities, including identity theft, financial fraud, drug trafficking, cybercrime, and the distribution of malicious software. These operations thrive under strong anonymity, making it extremely difficult for cybersecurity professionals and law enforcement agencies to monitor or mitigate emerging threats. One of the primary challenges lies in the absence of efficient, automated mechanisms to continuously crawl, extract, and analyze the vast volumes of unstructured data within dark web platforms. Traditional security frameworks are often limited in scope, lack scalability, and fail to provide real-time threat insights from hidden services. This gap leaves organizations vulnerable to cyberattacks, data breaches, and other darknet-driven threats. Therefore, there is an urgent need for an AI-powered solution capable of intelligently monitoring the dark web, extracting meaningful intelligence, and supporting proactive threat mitigation strategies.

1.2 Project Objectives

The **NeoSilk: AI-Enhanced Dark Web Monitoring** project aims to develop an intelligent system for monitoring and analyzing cyber threats on darknet marketplaces. The system will automatically collect data from .onion sites via the Tor network, overcoming access barriers such as CAPTCHAs.[1] and session controls to ensure scalable data extraction. Collected data will undergo preprocessing and feature engineering, focusing on structuring product listings, vendor profiles, and user reviews for natural language processing (NLP) tasks. Using NLP and machine learning models like BERT.[2], RoBERTa.[3], and DarkBERT.[4], the system will classify illicit products (e.g., drugs, malware), analyze review sentiment, and enable question answering through retrieval-augmented generation.[5]. Interactive dashboards, built with tools like Power BI, will visualize vendor activity, pricing, shipping trends, and anomalies to support cybersecurity professionals. The modular pipeline is designed to support future enhancements, including real-time crawling, anomaly detection, and automated alerts for suspicious listings.

1.3 Project Organization

NeoSilk proposes an AI-driven system that automates the process of collecting, analyzing, and visualizing darknet content to support cybersecurity operations. The system integrates multiple components into an end-to-end pipeline: dark web crawling (with/out CAPTCHA-handling), intelligent scraping, text preprocessing, NLP-based classification (e.g., product categorization and sentiment analysis), and a final visualization layer through a dynamic dashboard. Advanced models such as DarkBERT, BERT, and GPT.[6] have been employed to classify and interpret illicit listings, while exploratory work in RAG and anomaly detection further enhances threat detection capabilities.

CHAPTER 1 Introduction

Defines the problem of monitoring dark web marketplaces, outlines the objectives of NeoSilk and presents the overall project structure.

CHAPTER 2 Dark Web

This chapter introduces the nature of the dark web, differentiates between types of '.onion' marketplaces, and explains the categories found there. It also discusses practical challenges encountered when scraping these sites, such as CAPTCHA.[7] defenses, URL volatility, and site accessibility limitations, finally a sestion that presents and compares previous research efforts in dark web threat detection. It highlights studies that developed crawling and classification systems, explored dark web security mechanisms, or introduced methods for automating dark content labeling and CAPTCHA handling, and papers on thereat detection visualizing using dashboards.

CHAPTER 3 Data Collection and NLP Modelling

This chapter details the end-to-end AI pipeline used in the project. It begins with the data collection strategies across both CAPTCHA-protected and open-access marketplaces, then explains the preprocessing steps, NLP model development for classification and sentiment analysis, integration of Explainable AI (XAI), and the experimental exploration of Retrieval-Augmented Generation (RAG).

CHAPTER 4 Data Visualization

This chapter outlines the dashboards built for both marketplaces (Hidden Market and MGM Grand), showing how scraped data was analyzed and presented. It walks through visuals highlighting product trends, vendor activity, category distribution, and location-based shipping insights.

CHAPTER 5 Results

This chapter reports the performance of the trained models on category classification and sentiment analysis tasks. It presents accuracy and F1-score comparisons across different transformer models (BERT, RoBERTa, and DarkBERT), along with interpretations of model behavior.

CHAPTER 6 Conclusion and Future Work

This chapter summarizes the key achievements of the system and outlines potential extensions. These include automation of data cleaning and dashboard updates, integration of real-time alerts, and expansion toward multilingual data sources and dark web forums.

Chapter 2

Dark Web

2 CHAPTER 2 Dark Web

2.1 Introduction

The Internet is a huge and complex network that encompasses various layers of content and services. Most users are familiar with the **surface web**, the portion of the Internet indexed by search engines, where websites and social media platforms are easily accessible. However, beyond the surface web lies the **Deep Web**, which consists of data not indexed by conventional search engines, and furthermore the **Dark Web**.[8], a small, often misunderstood segment of the Internet that is intentionally hidden and accessible only through specialized software like Tor.

Table 2.1: Comparison of Surface Web, Deep Web, and Dark Web.

Feature	Surface Web	Deep Web	Dark Web
Indexing	Indexed by search engines	Not indexed (private/dynamic)	Not indexed (hidden on purpose)
Access	Public, no login needed	Login or query required	Requires Tor/I2P software
Content	Public sites, blogs, news	Banking, academic DBs, private services	Markets, forums, whistleblowing platforms
Anonymity	Low	Moderate	High
Usage	General browsing, shopping	Secure access to private data	Anonymous communication, illicit trade

The dark Web is frequently associated with anonymity and privacy, providing a space where users can engage in activities that are often protected from observation. Although this anonymity can be used for legitimate purposes, it is also known to harbor illegal activities, including black markets, illicit services, and cybercrime, and it is known to be the famous side.

There are many dark Web sites, but we will focus on Onion sites that sites that end with the.onion domain are considered some of the most famous and widely

recognized websites on the Dark Web. Why onion sites? will onion sites provide some of feature not available in other sites. Her why Onion Sites have gained prominence:

- **Anonymity and Privacy:** Onion sites are hosted on the Tor network, which stands for "The Onion Router." The Tor network is designed to anonymize users and websites by routing traffic through multiple nodes around the world. This makes it extremely difficult to trace the location or identity of either the user or the website host. This heightened privacy is one of the primary reasons for the popularity of onion sites on the Dark Web.

- **Popular Among Privacy Advocates:** Because of their focus on anonymity and security, onion sites attract privacy-conscious users who are seeking to protect their personal data. Many users prefer these sites for secure communication, file sharing, and browsing without leaving digital traces.

2.2 Accessing Dark Web Sites

Accessing the dark web requires careful preparation and adherence to security measures to ensure privacy and safety [9]. The Tor network provides anonymity by routing internet traffic through multiple encrypted relays, making it difficult to trace users [10]. However, studies show that using Tor alone does not guarantee complete security, necessitating additional precautions [11]. The following steps outline the process of accessing dark web sites securely:

- **Installing Tor Browser:** The primary gateway to the dark web is the Tor Browser, which should be downloaded exclusively from the official Tor Project website [12]. Research indicates that obtaining software from untrusted sources significantly increases malware infection risks [8]. Installation should follow the official guidelines for the user's specific operating system.

- **Connecting to the Tor Network:** When launched, the Tor Browser establishes a secure circuit through multiple relay nodes, a process that typically takes

15-30 seconds [13]. This multi-hop routing is fundamental to Tor's anonymity guarantees [10].

- **Navigating Onion Sites:** .onion sites require direct URL entry as they are not indexed by conventional search engines [9]. Recent studies classify onion site directories into three categories: curated lists, community forums, and search engines [8]. Users should exercise extreme caution as [11] found that 32% of sampled onion sites contained malicious content.

- **Ensuring Security and Anonymity:** While Tor provides network-layer anonymity, [14] demonstrates that application-layer leaks remain a significant threat. Users must avoid:

  - Personal information disclosure [12]
  - Browser plugin usage [10]
  - Maximizing browser windows (can leak screen resolution) [13]

Critical Security Enhancements:

- **VPN Configuration:** The ongoing debate between VPN-over-Tor and Tor-over-VPN is analyzed in [14], with the latter recommended for most users. This configuration hides Tor usage from the ISP [12].

- **Download Precautions:** [8] identified that 68% of executable files on dark web marketplaces contained malware. If downloads are necessary:

  - Use disposable virtual machines [11]
  - Verify file hashes when available [9]
  - Never open documents directly [12]

Tor's Privacy Mechanisms:

- **Onion Routing:** As described in [10], traffic passes through at least three encrypted hops (entry, middle, and exit nodes), with each node only aware of its immediate neighbors.
- **Traffic Analysis Resistance:** [13] demonstrates how Tor's fixed-size cells (512 bytes) and constant packet rates help thwart timing analysis attacks.
- **Circuit Isolation:** Modern Tor implementations create separate circuits for different domains as shown in [9], preventing cross-site correlation.

Current research [8], [14] suggests that while Tor provides robust anonymity, users must complement it with operational security practices. [12] provides a comprehensive checklist for maintaining security during dark web sessions.

2.3 Categories of Onion Sites

As we mentioned before Dark Web has various types of sites, we will try to cover the important ones and mention some sights about each type and each site. We can first divide the types for three or four large types:

- **Sites that help us to find onion site and direct us to it:** Search Engines and Link Lists.
- **Commercial Sites**: Drugs, Weapons, and Currency.
- **Service Sites:** Hacking and Murder.

2.3.1 Search Engines

- torch: `http://xmh57jrknzkhv6y3ls3ubitzfqnkrwxhopf5aygthi7d6rplyvk3noyd.onion/cgi-bin/omega/omega/`

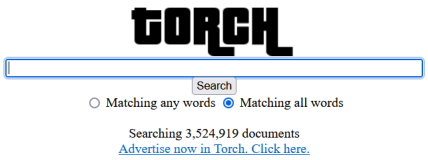

Figure 2.1: Screenshot of the Torch site.

- Ahmia: http://juhanurmihxlp77nkq76byazcldy2hlmovfu2epvl5ankdibsot4csyd.onion/

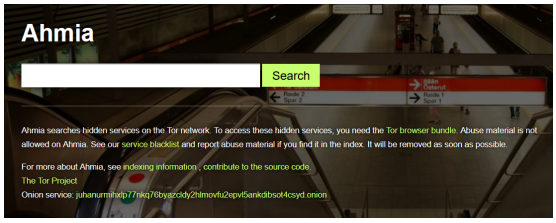

Figure 2.2: Screenshot of the Ahmia site.

2.3.2 Link Lists

- Deep Links Dump: http://deepqelxz6iddqi5obzla2bbwh5ssyqqobxin27uzkr624wtubht.onion/

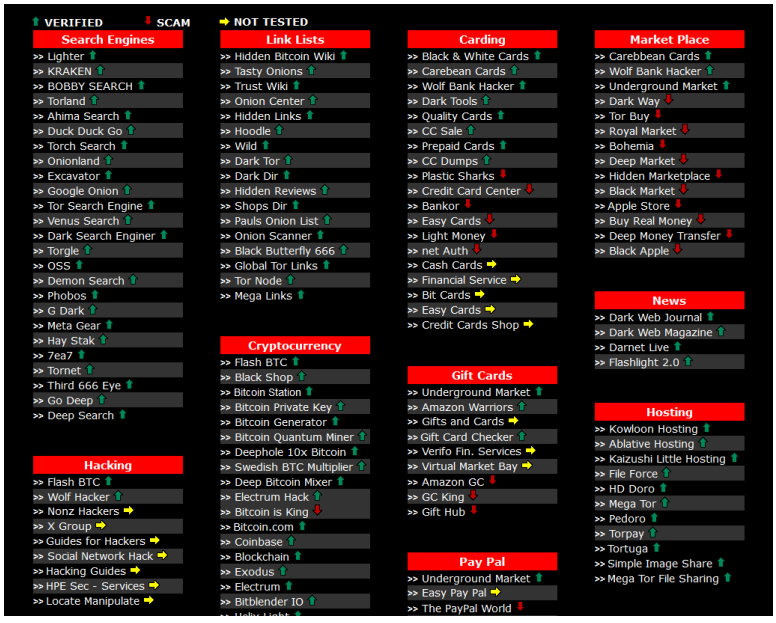

Figure 2.3: Screenshot of the Deep Links Dump site.

As we see that site provide us very useful service, which is the sate of the site is it verified or scam?

2.3.3 Forums

- Hidden Answers: http://lp2fkbyfmiefvscyawqvssyh7rnwfjsifdhebp5me5xizte3s47yu onion/

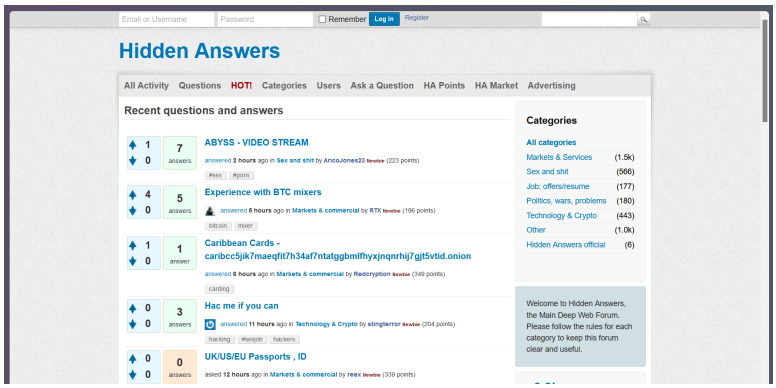

Figure 2.4: Screenshot 1 of the Hidden Answers site.

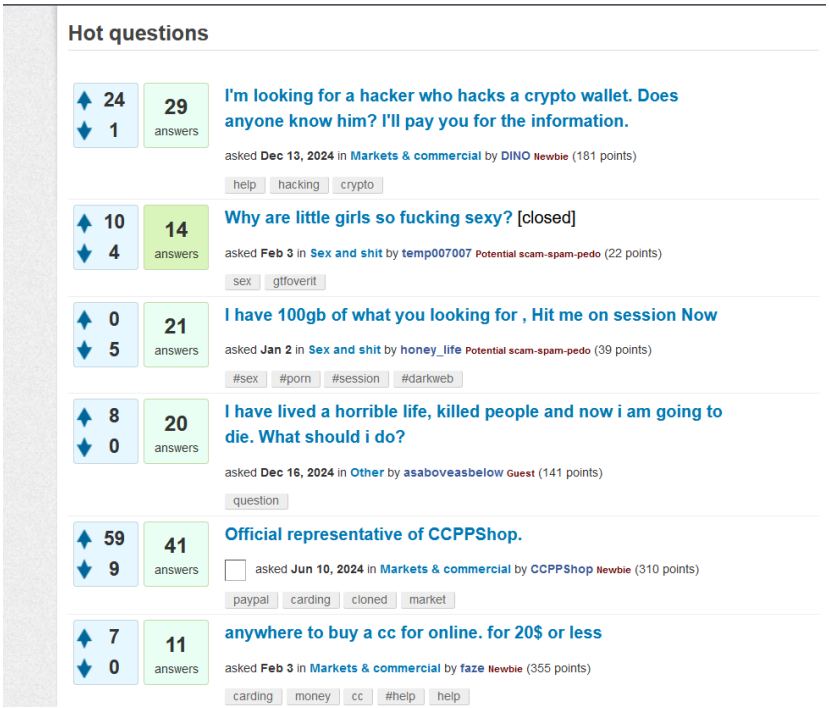

Figure 2.5: Screenshot 2 of the Hidden Answers site.

2.3.4 Cryptocurrency and Cards

- Carebean Cards: http://cardpl74ltmwe4o7pgpefljcng6qr36cnn7gzer2wermedxz3volxl onion/

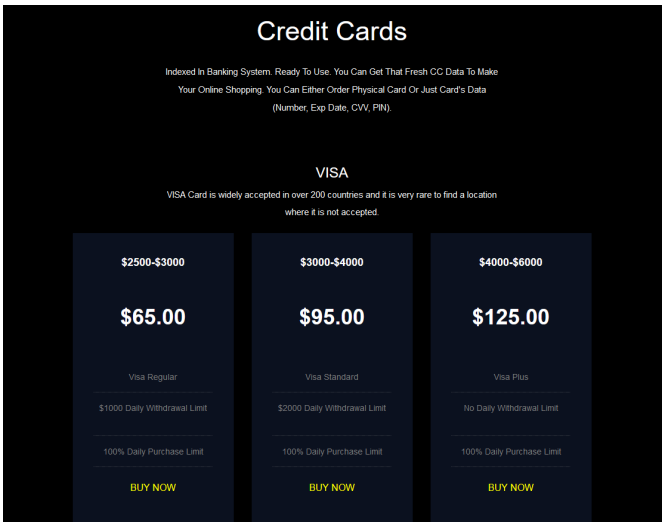

Figure 2.6: Screenshot of the Carebean Cards site.

- Underground Market: http://undeb6m465pjocdl6kvyiwefj5xxzcu3hgzngpfe5eolw764s onion/

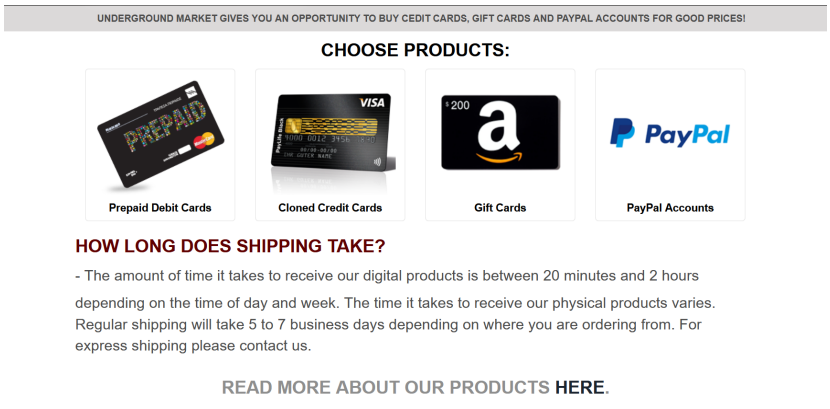

Figure 2.7: Screenshot of the Underground Market site.

2.3.5 Drugs

- Drughub: http://drughub666py6fgnml5kmxa7fva5noppkf6wkai4fwwvzwt4rz645aqd.onion/

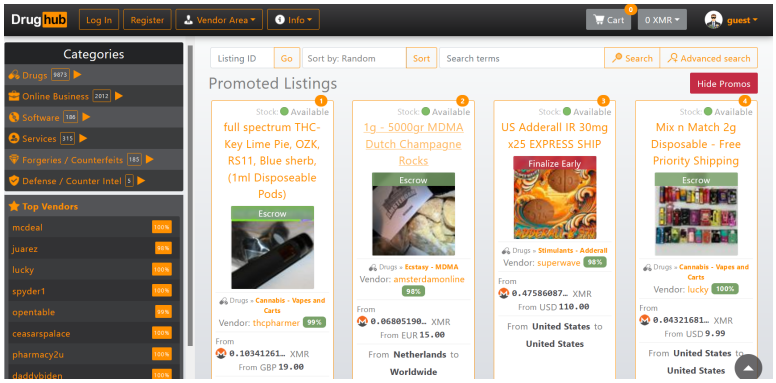

Figure 2.8: Screenshot 1 of the Drughub Cards site.

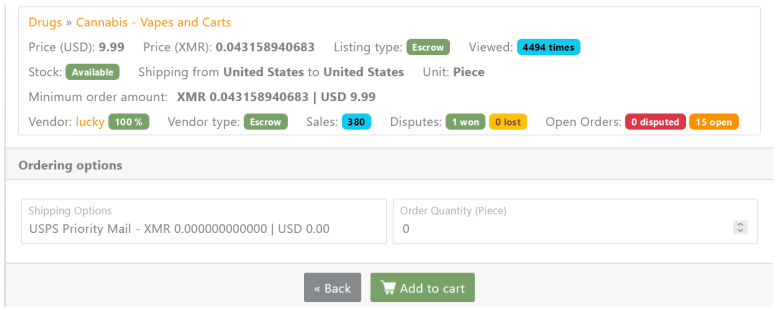

Figure 2.9: Screenshot 2 of the Drughub Market site.

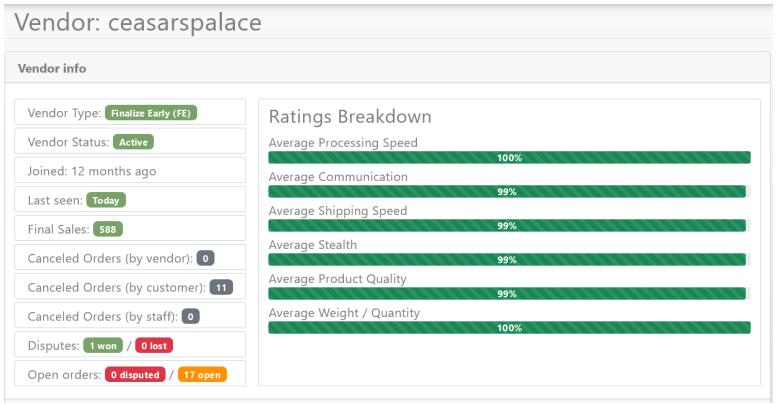

Figure 2.10: Screenshot 3 of the Drughub Market site.

2.3.6 Hacking

- HackingProgs: http://hackltxlmapssd5u6wro4ms7cn3tjbtyij5iuhfgaoxjmpwcc224iby onion/

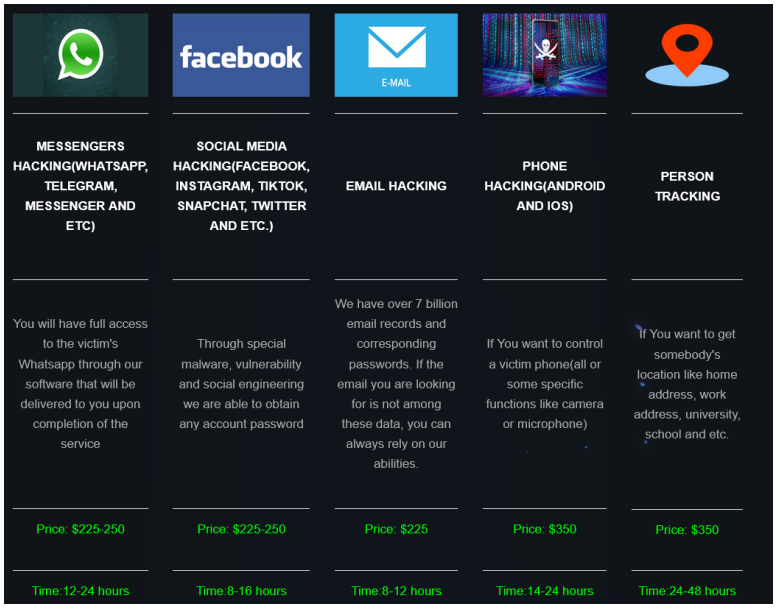

Figure 2.11: Screenshot 1 of the HackingProgs Cards site.

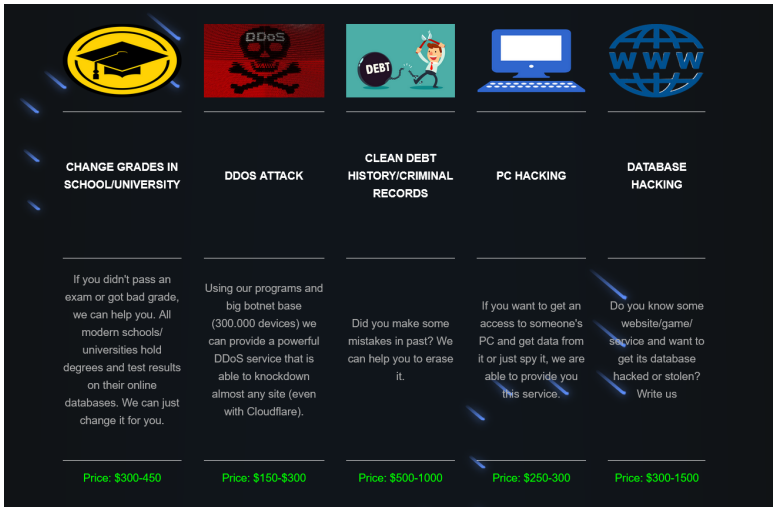

Figure 2.12: Screenshot 2 of the HackingProgs Market site.

2.3.7 Murder and Rape

- Red Room 1: http://redroomed5y77jcxvkejc4bm4qwdo2fghjanbip2iy6ihwnsea4g5gid onion/

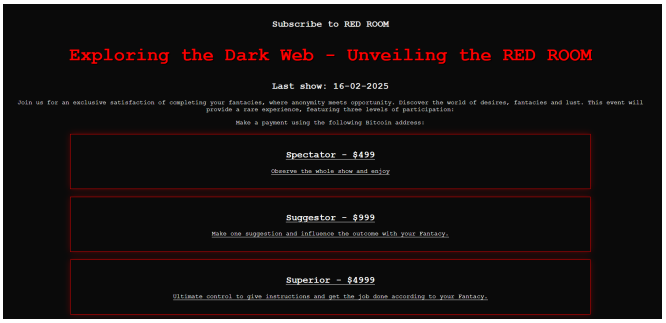

Figure 2.13: Screenshot 1 of the Red Room site.

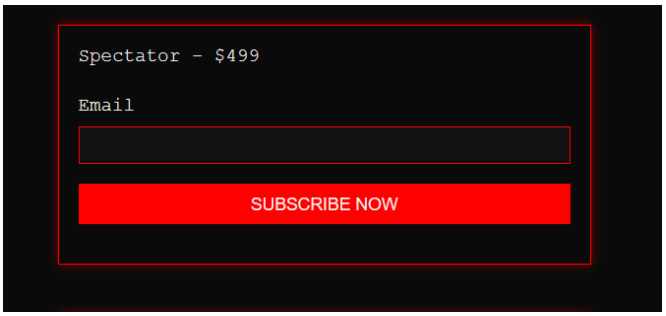

Figure 2.14: Screenshot 2 of the Red Room site.

- Red Room 2: http://redroomx57wri7sgl5qwhxdetmmhhv7fzjybwleypyzjycdxnhkr4sqd onion/

Figure 2.15: Screenshot 1 of the Red Room site.

- Hitmen Crop: http://45omf2f6zxc6ioo2wgvxnts3twjofza6l4ayk6hzd5aufeimzzvmxpi onion/

Figure 2.16: Screenshot 1 of the Hitmen Crop site.

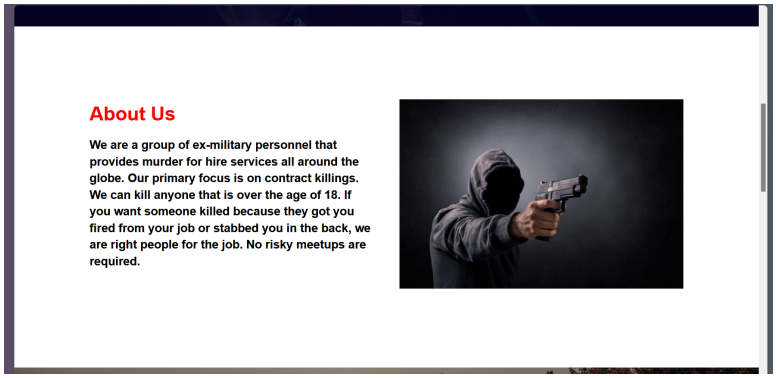

Figure 2.17: Screenshot 2 of the Hitmen Crop site.

Figure 2.18: Screenshot 3 of the Hitmen Crop site.

2.4 Onion Sites Problems

One of the most significant challenges when working with Dark Web sites is their inherent instability. These sites frequently change URLs, become inaccessible, or face seizure by authorities, making long-term access unpredictable. This volatility poses a challenge for researchers and cybersecurity analysts who rely on these platforms for data collection and analysis.

For instance, after successfully developing a CAPTCHA-solving mechanism, two of the targeted sites became unavailable due to sudden closures. This unexpected dis-

ruption emphasized the risks associated with relying on unstable dark web platforms and underscored the importance of building adaptable and resilient data collection systems that can quickly pivot to new sources or marketplaces.

2.5 Related Works

2.5.1 Crawling and Classification Systems

2.5.1.1 Dalvi et al. (2022) Dalvi et al. [15] proposed a dark web crawling and classification system aimed at detecting illicit activity across various '.onion' domains. The system employs a custom crawler that begins with seed URLs collected from Ahmia (a Tor search engine) and recursively explores hyperlinks to collect additional '.onion' pages. In total, the crawler collected over 3,100 pages, storing both URLs and HTML content in a MongoDB database. Their approach emphasizes page-level content collection, where each web page is treated as a single document. After automatic data cleaning, documents were classified into five categories: *Drugs*, *Fake ID*, *Hacking*, *Weapons*, and *Others*. For classification, the authors evaluated several machine learning models, with Linear Support Vector Classifier (LSVC) achieving the best performance (accuracy of 92%). The work contributes a scalable crawling framework and automatic labeling approach but lacks focus on individual product details or category substructure.

Comparison with NeoSilk:

While Dalvi et al.'s work provides a foundational approach for crawling and classification of dark web pages, our project **NeoSilk** builds upon and advances this idea through a more granular and modular pipeline. The comparison is summarized in Table 2.2.

Table 2.2: Comparison between Dalvi et al. (2022) and NeoSilk

Aspect	**Dalvi et al. (2022)**	**NeoSilk**
Granularity of Data	Page-level scraping (entire HTML treated as one sample)	Structured field-level extraction (product name, description, comments, etc.)
Source Pages	General '.onion' domains (via Ahmia)	Specific darknet marketplaces (e.g., Hidden Market, MGM Grand)
Scraping Scope	No CAPTCHA-handling or site-specific scraping logic	Includes advanced CAPTCHA-solving (alphanumeric, clock, association)
Categories	5 high-level categories (e.g., Drugs, Hacking)	50+ detailed product categories and subcategories
ML Models	LSVC, Naïve Bayes, Random Forest	BERT, RoBERTa, **DarkBERT**, GPT, and RAG
NLP Tasks	Classification only	Classification, Sentiment Analysis, RAG-based QA
Explainability (XAI)	Not addressed	SHAP-based explainability module (planned)
Visualization	Not included	Tableau/Power BI dashboard with trends, vendors, pricing

This comparison highlights NeoSilk's advancement in both technical depth and domain coverage. By targeting specific product-level fields, leveraging advanced NLP models, and incorporating real-time visualization, NeoSilk offers a more robust and scalable system for dark web threat intelligence.

2.5.1.2 Ramalingam et al. (2023)

Ramalingam et al. (2023) . [16] focuses on building an automated system for **dark web content classification**. The system aims to crawl '.onion' sites, clean the collected data, label it automatically, and classify it into high-level threat categories such as Drugs, Fake ID, Hacking, Weapon, and Others.

Main Contributions and Methodology:

- **Data Collection:** A custom crawler is used to start from seed '.onion' URLs and iteratively explore the dark web. Over 3,500 links were collected and stored in a MongoDB database for further analysis.
- **Preprocessing:** The authors employed HTML tag removal, lowercase normalization, stopword elimination, and tokenization to clean the raw content for ML analysis.
- **Automatic Labeling:** A keyword-based method is proposed to match documents with predefined class terms, reaching a 90% labeling accuracy and eliminating the need for manual annotation.
- **Classification Models:** Three models were evaluated—Linear SVC (91%), Random Forest (89%), and Naive Bayes (81%). SVC showed the best performance.
- **Evaluation Metrics:** Accuracy, precision, recall, and F1-score were used. LSVC achieved an F1-score of 88%, making it the most suitable classifier.

Comparison with NeoSilk

Unlike NeoSilk, which uses BERT-based deep learning models like DarkBERT and RoBERTa for semantic-level classification, Ramalingam et al. rely on traditional machine learning and keyword-based automatic labeling. NeoSilk also incorporates **fine-grained NLP tasks** such as *sentiment analysis* and *RAG-based question answering*, which are absent in this work. Moreover, NeoSilk focuses on visual analytics and dashboard construction, whereas this paper is limited to backend classification.

2.5.1.3 Wang et al. (2024)

Wang et al. (2024). [17]conducted a **security-focused investigation** into the defenses used by darknet marketplaces. The paper analyzes twelve marketplaces over four months and categorizes protective mechanisms into **web security** and **account security**.

Main Contributions and Observations:

- **Web Security:** Explores CAPTCHAs (text, image, math-based), DDoS protection, waiting queues, anti-phishing tools, and rate-limiting systems. Many sites use advanced CAPTCHA variations like color-based or interactive puzzles.
- **Account Security:** Highlights authentication policies including PGP-based MFA, kill-switch features, mnemonic recovery phrases, and variable password/PIN rules.
- **Crawling Barrier:** Most data had to be collected manually due to these strong defenses. Few sites offered API access, making scraping infeasible.
- **Ethical Guidelines:** Advocates for observer-only approaches, discouraging intrusion or active manipulation.

Comparison with NeoSilk:

Wang et al. do not implement an AI pipeline but focus solely on documenting **defensive mechanisms**. NeoSilk, on the other hand, actively attempts to bypass such defenses—most notably by solving **alphanumeric CAPTCHAs**. The works complement each other: one provides a landscape of barriers, and the other demonstrates technical solutions to overcome them.

2.5.2 NLP-Based Classification and Illicit Content Detection

Recent studies have leveraged deep learning and NLP to analyze dark web content. Here, we detail two notable works and contrast them with our pipeline, **NeoSilk**.

2.5.2.1 Al-Naser et al. (2021) – TextCNN with LDA-Based Labeling

Al-Naser et al. [18] propose a hybrid framework combining unsupervised topic modeling (LDA) and convolutional neural network classification (TextCNN). They aim to classify '.onion' pages into categories such as forums, blogs, marketplaces, and news. Specifically:

- **Dataset**: Collected approximately 6,000 '.onion' pages across dark web directories.
- **Labeling Process**: Applied LDA to discover latent topics; manual validation refined labels into 4–6 classes.
- **Model**: Employed TextCNN with pre-trained embeddings (e.g., GloVe) to capture local patterns in text.
- **Results**: Achieved 88

Limitations: - Focused on broad structural categories rather than illicit content. - Labeling depends on manual topic review. - Does not address product-level content or downstream tasks like sentiment.

2.5.2.2 Bhayani et al. (2022) – Transformer-Based Illicit Content Classification

Bhayani et al. [19] performed large-scale multi-class classification of illicit categories (drugs, porn, hacking, weapons) using transformer architectures:

- **Dataset**: Over 114,000 scraped '.onion' pages, manually or semi-automatically labeled across illicit themes.
- **Models**: Evaluated LSTM, ULMFiT, BERT, and RoBERTa.
- **Evaluation**: BERT delivered highest performance: 96
- **Contribution**: Demonstrated the effectiveness of pre-trained transformers for large-scale illicit content detection in dark web data.

Limitations:

- Operates at document rather than product level

- Does not incorporate sentiment, summarization, or explainability.

Comparison with NeoSilk

Al-Naser et al. (2021) and Bhayani et al. (2022) focused on general or document-level classification of dark web pages using traditional and transformer-based models. Al-Naser used LDA for topic modeling and TextCNN for classification on a small dataset of 6,000 pages, while Bhayani applied models like BERT and RoBERTa on over 114,000 illicit-themed documents. In contrast, **NeoSilk** operates at the product level within darknet marketplaces, using metadata-driven labels and incorporating advanced NLP tasks including classification, sentiment analysis, and RAG-based QA. It further integrates explainability tools and interactive dashboards, offering a more fine-grained, multi-task, and actionable threat intelligence pipeline.

While **Al-Naser et al.** and **Bhayani et al.** provide foundational approaches for dark web NLP, **NeoSilk** advances further by:

1. **Product-level focus**: Centered on marketplaces and specific listings rather than generic pages.

2. **Multiple NLP tasks**: Integrates category classification, sentiment analysis, summarization, and retrieval-augmented QA.

3. **Explainability and visualization**: Implements XAI techniques and dashboards for gaining informative insights that helps in threat detection.

2.5.3 Dashboard and Visualization for Threat Intelligence

2.5.3.1 Koven et al. (2021) – Cybersecurity Visualization and Dashboard Design Koven et al. [20] present a comprehensive study on designing cybersecurity dashboards that prioritize clarity, real-time monitoring, and actionable insights for security analysts. The system integrates multiple threat intelligence sources—such

as network traffic logs, SIEM feeds, and dark web content—and visualizes key threat indicators to support decision-making in SOC environments. **Dashboard Features:** Their proposed dashboard employs diverse visual elements:

- **Bar Charts** to show the frequency of detected threat types (e.g., malware, phishing, DDoS).
- **Sankey Diagrams** to depict flow between entities such as threat actors, victim IPs, and malware families.
- **Geographic Maps** highlighting the source and target of cyberattacks, aiding geographical threat tracking.
- **Dark UI Design** with vibrant contrasting colors (e.g., red, neon green) to enhance visual clarity, suitable for dark environments.

Interactivity and Integration:

The dashboard supports real-time filtering based on threat category and time range. It connects to SIEM systems and databases like Elasticsearch to stream live alerts and historical records. This modular backend architecture aligns well with large-scale threat intelligence platforms.

Data Analysis and Preprocessing:

The system includes automated cleaning of unstructured dark web data using Python-based ETL pipelines. These pipelines handle tasks such as noise removal, entity normalization (e.g., stripping price units), and timestamp formatting. For analytics, it leverages Random Forest and anomaly detection techniques to classify threats and detect outliers.

Visualization Tools and Deployment:

The dashboard is implemented using Power BI and leverages DAX (Data Analysis Expressions) for real-time metric calculation and dynamic visual updates. This makes it highly responsive and adaptable to various cyber monitoring contexts.

Comparison with NeoSilk:

- **Scope and Focus:** While Koven et al. target general threat intelligence—including network-level data—NeoSilk focuses on dark web marketplaces and vendor-product relations.
- **Dashboard Features:** Both systems use Power BI and support filtering and dynamic updates. However, NeoSilk emphasizes vendor trends, category distributions, and shipping patterns rather than network attacks.
- **Data Sources:** Koven et al. incorporate network logs and SIEM systems, while NeoSilk scrapes '.onion' sites and uses product-level data (e.g., description, rating).
- **AI Models:** NeoSilk integrates deep NLP models (DarkBERT, BERT, RoBERTa) for product classification and feedback sentiment analysis, whereas Koven et al. use Random Forest for structured cyber threat indicators.
- **Future Integration:** NeoSilk plans to incorporate alert systems and anomaly detection, closely aligned with Koven et al.'s real-time alerting framework.

Chapter 3

Data Collection and NLP Modeling

3 CHAPTER 3: Data Collection and NLP Modeling

3.1 Introduction

In this chapter, we outline the technical methodology and architectural design behind **NeoSilk**, our AI-powered system for dark web threat intelligence. The entire pipeline is structured to facilitate the automated collection, structuring, modeling, and visualization of data extracted from darknet sources—specifically, high-activity marketplaces. Unlike many earlier studies that target general **.onion** sites such as discussion forums, news portals, or blogs, our project intentionally focuses on **darknet marketplaces** as the primary data source. This focus is motivated by several critical observations. First, marketplaces constitute the core of illicit economic activity on the dark web, hosting a wide range of products from drugs and counterfeit IDs to hacking tools, malware, and fraudulent services. These platforms also offer well-defined product listings and vendor information, which provides structure necessary for machine learning applications. Marketplaces are particularly valuable because:

- They contain **highly structured data**, including fields like product name, description, price, category, vendor profile, availability, and user feedback.
- They allow us to track **vendor behavior, user interaction, and shipping routes**, which are crucial for threat modeling.
- They offer a rich basis for **natural language processing (NLP)** tasks due to their descriptive text content.

Within these marketplaces, we gave special analytical priority to the **Drugs** category. This decision is based on the observation that drug-related listings represent the majority of entries in most active markets, both in volume and vendor engagement. Targeting this high-density category allows us to maximize the signal-to-noise ratio and extract more actionable cyber threat patterns. Furthermore, this chapter

details how the system was built to handle both simple and complex access scenarios—including sites protected by CAPTCHA and anti-bot mechanisms. It also explains how we integrated modern AI tasks such as classification, sentiment analysis, and retrieval-augmented generation (RAG), along with visual intelligence via interactive dashboards.

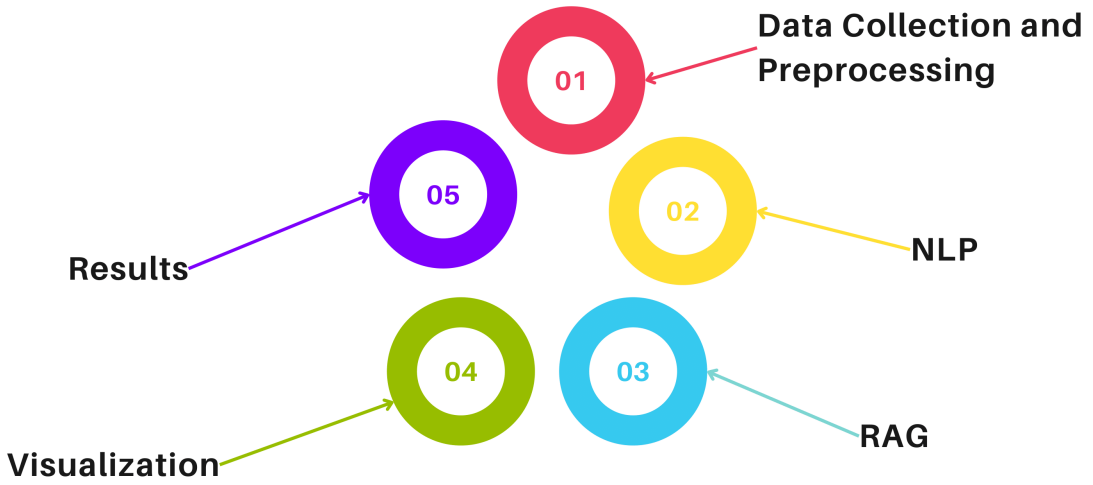

Figure 3.1: Dark Web Data Analysis and AI Modeling Pipeline

4 Data Collection

4.1 Scraping Approaches: Onion Sites with and without CAPTCHAs

Accessing and collecting data from dark web marketplaces poses various challenges, particularly in how sites handle automated requests. Broadly, onion sites fall into two categories based on their access barriers:

- **Onion Sites Without CAPTCHA:** These sites allow automated scraping directly without requiring visual verification. A straightforward crawling and scraping mechanism can be used to navigate internal links and extract structured content. This approach was applied to the Hidden Marketplace [21], [22].

- **Onion Sites With CAPTCHA:** These sites implement CAPTCHA mechanisms to block bots. Scraping them requires additional steps such as solving CAPTCHA

images, managing authentication tokens, and maintaining session integrity. A specialized scraper was developed to handle these complexities, particularly for the MGM Grand Marketplace [23], [24].

In the following sections, we describe in detail the scraping workflows and technical implementations used for both site types.

4.1.1 Scraping Onion Sites Without CAPTCHA

The Hidden Marketplace served as the primary case study for scraping onion sites without CAPTCHA protection. Its open access structure enabled a complete end-to-end scraping process involving three main phases for key categories: **Drugs**, **Digital**, and **Tutorials**. Each category was processed independently through the same three-stage pipeline to ensure thorough extraction of product information [25]–[27].

4.1.1.1 Phase 1: Extracting Undetailed Product Features

In the first phase, the scraper targeted category listing pages that show a preview of each product. These outer pages contain basic metadata such as:

- **Product Name**
- **Description**
- **Rating (%)**
- **Views**
- **Purchased Items**
- **Price**

A session was created over the Tor network using Python's `requests` and `BeautifulSoup`, looping through all paginated category pages. For each product preview card, the above information was extracted and stored incrementally in a CSV file. This phase produced a dataset referred to as the Undetailed Data.

4.1.1.2 Phase 2: Collecting Product URLs

To access more in-depth information about each product, a second phase was required to collect the individual product page URLs. The scraper revisited each category page and extracted all valid hyperlinks that follow the pattern `/product/XYZ`. These URLs were stored in a CSV file to be used in the final scraping phase.

4.1.1.3 Phase 3: Scraping Detailed Product Features

In the third phase, the previously collected product URLs were visited one by one. Each product page contains more granular details, including:

- **Comments (User Feedback)**
- **Seller Name**
- **Seller Location**
- **Ships to (Seller & Product)**
- **Category**
- **Quantity in Stock**
- **Dead Drop**
- **Availability**

Additionally, the product name and URL were also re-extracted for integrity. All fields were then saved to another CSV file referred to as the **Detailed Data**.

4.1.1.4 Final Datase

After all three phases, the final product dataset was constructed by performing an **inner join** between the Undetailed and Detailed data based on the `Product Name`. This resulted in a complete dataset combining all available structured information for products across selected categories. This multi-phase scraping methodology

enabled a highly structured, scalable approach to data collection from uncensored onion markets, forming the foundation for further NLP and ML tasks in the project.

4.1.2 Scraping Onion Sites With CAPTCHA

Unlike open-access onion marketplaces, some dark web platforms implement CAPTCHA verification to block automated scraping. In our project, the MGM Grand marketplace represented a real-world case of this challenge. To extract product data from it, we first examined the CAPTCHA types it uses, then developed a targeted scraping solution for its pages.

4.1.3 Types of CAPTCHAs Encountered

During our exploration of dark web marketplaces, we encountered three primary CAPTCHA types:

- **Alphanumeric CAPTCHAs:** Images showing five-character combinations from 62-character space (a–z, A–Z, 0–9).

Figure 4.1: Alphanumeric Captcha Example

- **Clock CAPTCHAs:** Image-based tests that require selecting the correct clock time from visual cues.

Figure 4.2: Clock Captcha Example

- **Association CAPTCHAs:** Require matching an image to a semantically relevant label or concept (e.g., Eiffel Tower → France).

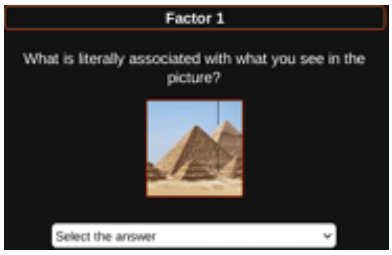

Figure 4.3: Association Captcha Example

4.1.4 Scraping MGM Grand Market with Alphanumeric CAPTCHA

To extract structured product listings from MGM Grand, a marketplace protected by Alphanumeric CAPTCHA at login, we designed a scraping pipeline composed of three main phases.

4.1.4.1 CAPTCHA-Secured Login

The MGM Grand platform required solving a CAPTCHA on the login page. To do this, we manually labeled a dataset of alphanumeric CAPTCHA images and trained an OCR-based model capable of recognizing the 5-character codes. The automated scraper integrated this model and submitted CAPTCHA text with login credentials to authenticate successfully.

4.1.4.2 Navigating Marketplace Listings

After login, we navigated the product category pages using a custom session-based scraper configured with Tor proxies. The scraper looped through paginated listings and extracted product summaries visible on the main listing page. These included: **Name, Price, Description, Rating, Views, and Purchased Count**.

4.1.4.3 Extracting Full Product Details

We then collected all individual product URLs and visited them one by one. The detailed product pages contained additional metadata such as: **Vendor, Vendor Level,**

Location, Shipping Info, Stock, Escrow, Availability, and User Comments. The scraper parsed and stored these fields into a structured dataset using `BeautifulSoup`, with results written to CSV format for later preprocessing and modeling. The final dataset after merging product summaries with full detail pages served as the input for downstream NLP tasks. Compared to scraping open-access markets, the CAPTCHA integration step added considerable engineering complexity to the MGM Grand scraper, but resulted in a rich dataset with high-value threat intelligence signals.

4.1.5 Problems of Integrating Captcha Solutions for Web Scraping

Integrating CAPTCHA-solving solutions into our web scraping workflow has proven to be a challenging task, as we have encountered various obstacles. These challenges have significantly impacted the scraping process, and resolving them is critical to ensuring the success of the project. The key issues are outlined as follows:

4.1.5.1 Accessibility of Captcha Images

One challenge wasn't just the dynamic nature of CAPTCHA images, but our inability to access the 'div' elements where they are embedded—particularly with association and clock CAPTCHAs. These CAPTCHAs are dynamically rendered via JavaScript, making them difficult to scrape using traditional methods such as parsing 'img' tags. Attempts to access the images programmatically were unsuccessful due to limitations in rendering caused by the dark web browser environment. As a result, we resorted to manually interacting with the onion site to capture the necessary images for analysis.

4.1.5.2 Authorization Issues After Captcha Resolution

Even after solving a CAPTCHA, another major issue arises: the system often redirects the user back to the login page, indicating failed authentication. This is most frequently encountered with alphanumeric CAPTCHAs. Despite correct resolution, session or token mismatches cause the system to deny access, suggesting improper session management or incomplete authentication handshake.

This complicates scraping, as it requires repeated CAPTCHA solving while maintaining session continuity—an issue especially challenging in stateless scraping scripts.

4.1.5.3 Variability of CAPTCHA Types Across Onion Sites

Across the tested marketplaces, CAPTCHA types and structures vary significantly. Even when targeting CAPTCHAs with similar visual designs (e.g., alphanumeric), we encountered different validation rules, presentation layers, or formats. Some sites implement new CAPTCHA types like:

- Image-based puzzles
- Sliding blocks
- Timed verification

This variability, combined with frequent CAPTCHA updates, hinders the development of a generalized solving strategy.

4.1.5.4 Access Denial Due to Phishing Warnings

In some cases, we faced access denials or warnings even before reaching CAPTCHA-protected pages. This issue was particularly observed during our scraping attempts on the **MGM Grand** darknet marketplace. As shown in Figure, some browsers—especially privacy-focused ones or those with enhanced security plugins—flagged certain '.onion' links as suspected phishing threats.

This introduced an additional obstacle in the scraping pipeline. Even though the site was accessible via the Tor network, scraping tools using standard HTTP headers or browser emulation were blocked by such warnings. This prevented automation scripts from loading the actual site content and made it difficult to consistently bypass such browser-level security checks.

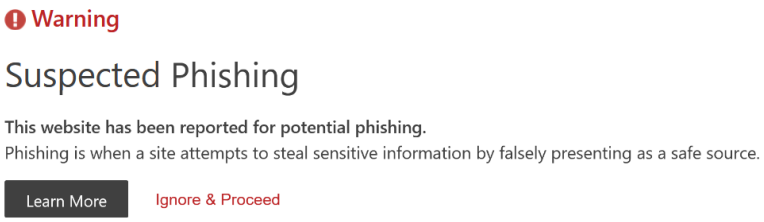

Figure 4.4: Phishing Warning Page Encountered on MGM Grand Market

These challenges illustrate the volatile and adversarial nature of dark web scraping, where automated access is frequently obstructed by CAPTCHA mechanisms, session handling traps, and browser-level defenses. Addressing these issues remains an ongoing process that requires a careful blend of web automation techniques, deep learning models, and human supervision.

4.2 Data Preprocessing

A significant portion of preprocessing was inherently handled during the scraping phase itself. Since the scrapers were designed to extract data based on specific HTML structures (e.g., targeting product name, rating, comments, price), the scraped data was already segmented into well-defined fields. This reduced the need for complex parsing or field extraction, making the remaining preprocessing tasks relatively straightforward. Most of the work focused on cleaning minor inconsistencies, handling nulls, and standardizing formats across the dataset. Following the data cleaning phase, the next crucial step was to preprocess the dataset to ensure it was ready for effective analysis and modeling. The preprocessing phase aimed to improve the dataset's consistency, usability, and relevance by refining its structure and removing any residual inconsistencies. We adopted a dual-tool approach combining **Python scripting** and **Power BI's Power Query editor**:

- **In Python:** Scripts were used to handle initial processing such as merging data sources (detailed and undetailed features), formatting text fields (e.g., standard-

izing case and spacing), filtering out extreme outliers (e.g., price anomalies above $10,000), and exporting clean results into structured CSVs.

- **In Power Query:** Power BI's built-in Power Query editor was used extensively during exploratory data analysis (EDA) to:
  - Remove null or incomplete entries.
  - Eliminate duplicate rows.
  - Convert data types (e.g., text to numeric for price, views, etc.).
  - Extract and reformat fields such as vendor level or country names.

In addition, **text columns** (such as `Product Description`, `Product Name`, and `Comments`) were cleaned to remove emojis, symbols, and non-ASCII characters. For example, regex and string normalization techniques were applied to preserve meaningful text content while eliminating noise. This preprocessing ensured the data was standardized and suitable for both machine learning tasks and dashboard visualizations. By the end of this phase, we had a clean and structured dataset containing valid entries for all key features—ready for NLP modeling and data-driven threat analysis.

4.3 NLP Modelling

Natural Language Processing and Machine Learning techniques were utilized to convert raw text data into structured representations suitable for classification. Preprocessing steps included tokenization, normalization, and noise removal. Pre-trained transformer models were used to generate dense contextual embeddings from the input text. For modeling, transformer-based encoders were selected due to their superior performance on a wide range of NLP tasks. Specifically, BERT, RoBERTa, and Dark BERT were employed to extract meaningful features from the text. These

representations were then passed into task-specific classification heads to predict output classes.

4.3.1 Model Selection

The decision to use BERT-based models instead of GPT-based models was guided by architectural, functional, and practical considerations relevant to classification tasks.

4.3.1.1 Architectural Considerations

BERT-based models use bidirectional attention, meaning they consider both preceding and succeeding words in a sentence simultaneously. This enables a deeper understanding of context, which is critical for tasks like sentiment analysis and topic classification. In contrast, GPT models use unidirectional (left-to-right) attention due to their causal masking. This design is effective for text generation but less suited for tasks requiring holistic understanding of the input.

4.3.1.2 Pre-training Objectives

BERT is pre-trained using Masked Language Modeling (MLM), where the model predicts randomly masked tokens using context from both directions. This encourages the development of rich, context-aware embeddings suitable for classification. GPT, on the other hand, is trained using Causal Language Modeling (CLM), predicting the next token in a sequence. While effective for generation tasks, it is less optimal for understanding entire input sequences, which is essential in classification.

4.3.1.3 Fine-tuning Simplicity and Model Variants

BERT models are designed with classification in mind. Fine-tuning typically involves adding a classification head on top of the encoder and using the `[CLS]` token embedding. This makes adaptation to new tasks straightforward and efficient. GPT-based models require prompt engineering or reformulating classification as a generative task, which can be complex and computationally demanding.

4.3.1.4 Models Used

The models selected for this study include DarkBERT, BERT, and RoBERTa, which were chosen for their demonstrated superior performance compared to GPT models. These models incorporate enhanced training methodologies, including larger training datasets, dynamic masking techniques, and extended training durations. Notably, **DarkBERT** provides specialized performance through domain-specific training that is specifically tailored to capture nuanced linguistic patterns in specialized textual domains, making it particularly suitable for **NeoSilk**.

4.3.2 Category Classification Models

The classification component involved fine-tuning the encoder models for specific tasks. A fully connected classification layer was added on top of the encoder output. During training, this layer learned to map the contextualized token representations (typically from the `[CLS]` token) to output labels using a softmax activation function. The models were evaluated on standard classification metrics, including accuracy, precision, recall, and F1-score. Fine-tuning was performed using cross-entropy loss and the Adam optimizer.

4.3.2.1 Drug Category Classification

The overview of data collected from scraping onion site **Hidden Market** comprises 13,945 labeled samples across nine pharmaceutical categories: Stimulants, Cannabis, Psychedelics, Benzodiazepines, Ecstasy, Opioids, Prescription Medications, Dissociatives, and Anabolic Steroids. Stimulants (3,254 samples) and Cannabis (3,074 samples) form the largest categories, while Anabolic Steroids (401 samples) is the smallest. This pronounced class imbalance reflects real-world substance use trends and introduces complexity in training and evaluation.

4.3.2.2 Digital Category Classification

This task uses a curated dataset of 3,108 samples from scraped onion site **Hidden**

Market, categorized into six classes: Accounts, Application Software, Documents, Hacks, Pornographic Content, and Social Security Numbers. The class distribution is highly skewed, with Pornographic Content accounting for over 56% of samples, while Application Software makes up only 3.8%. This imbalance mirrors the prevalence of illicit content types in real-world marketplaces and poses significant challenges for robust classification.

4.3.2.3 Tutorial Category Classification

This task utilizes a curated dataset of 1,341 samples from scraped onion site **Hidden Market**, categorized into five classes: eBooks, Carding, Dump, Drugs Production, and Theft. The class distribution is highly imbalanced, with eBooks comprising approximately 71.7 of the samples (962 instances), while Drugs Production and Theft each represent only 3.2 (43 instances each). This skewed distribution reflects the varying prevalence of illicit content types in hidden marketplaces and presents challenges for effective classification, particularly for underrepresented classes.

4.3.3 Model Training, Evaluation, and Calibration

4.3.3.1 Data Partitioning and Preprocessing

Stratified sampling was used to split both datasets into training (80%), validation (10%), and test (10%) sets while preserving class proportions. Tokenization was performed using model-specific tokenizers, with a maximum sequence length of 128 tokens. Dynamic padding was applied during batch processing via `DataCollatorWithPadding` to improve memory efficiency.

4.3.3.2 Training Configuration and Hyperparameters

All models were trained for five epochs using a learning rate of 3×10^{-5} and a batch size of 16. Weight decay was set to 0.01 to prevent overfitting. The AdamW optimizer with linear learning rate decay (no warmup) was used. Macro F1-score was the primary evaluation and model selection criterion due to class imbalance.

4.3.3.3 Evaluation Metrics

Performance was assessed using accuracy, macro F1-score, and weighted F1-score. Macro F1-score was prioritized to ensure fair evaluation across all classes. Detailed classification reports were generated to analyze precision, recall, and F1-score per class.

4.3.3.4 Expected Calibration Error (ECE)

ECE measures the average difference between predicted confidence and observed accuracy across all predictions. Predictions are grouped into confidence bins; within each bin, the average predicted confidence is compared to the actual accuracy. A lower ECE indicates better calibration.

4.3.3.5 Maximum Calibration Error (MCE)

MCE reports the worst-case calibration error—the largest gap between confidence and accuracy in any confidence bin. While ECE provides an overall summary, MCE highlights where the model's confidence is most unreliable.

4.3.3.6 Brier Score

The Brier Score captures both the accuracy and calibration of probabilistic predictions. It computes the mean squared difference between predicted probabilities and actual binary outcomes. Overconfident incorrect predictions are penalized more than uncertain ones.

4.3.3.7 Negative Log-Likelihood (NLL)

NLL evaluates how well the model assigns high probabilities to the correct class. A lower NLL indicates that the model reliably associates high probabilities with true outcomes.

4.3.3.8 Reliability Diagrams

Reliability diagrams visualize calibration by plotting predicted confidence against

observed accuracy. A perfectly calibrated model lies on the diagonal where confidence equals accuracy. These diagrams help identify areas of overconfidence or underconfidence.

4.3.4 XAI in Category Classification

In this subsection, we explore how Explainable Artificial Intelligence (XAI) techniques are used to interpret the outputs of deep learning models trained for content classification in dark web marketplaces. The focus is on using LIME (Local Interpretable Model-Agnostic Explanations) in conjunction with domain-specific language models across multiple classification categories.

4.3.4.1 Classification Domains

The XAI framework has been implemented across three primary classification domains within dark web marketplace analysis: **Drug Classification:** The system classifies drug-related content into nine distinct categories including *Benzos, Cannabis, Dissociatives, Ecstasy, Opioids, Prescription, Psychedelics, Steroids*, and *Stimulants*. Each category represents a specific class of substances commonly traded in these marketplaces. **Digital Goods Classification:** The framework extends to digital marketplace content, categorizing various digital products and services such as software, databases, accounts, and digital tools commonly found in dark web marketplaces. **Tutorial Classification:** Educational and instructional content is classified to identify various types of tutorials and guides distributed through dark web platforms, including technical guides, operational instructions, and educational materials.

4.3.4.2 Model Setup and Architecture

We utilized fine-tuned versions of DarkBERT for each classification domain. DarkBERT is a specialized language model trained on dark web content, providing enhanced understanding of marketplace terminology and context compared to general-purpose language models. The models are loaded using the Hugging Face transform-

ers library, along with their corresponding tokenizers, and mapped to custom label schemas specific to each classification domain. This setup allows for tokenization of input samples and obtaining predictions with high confidence using softmax probabilities. The domain-specific nature of the models ensures accurate classification across the varied terminology and contexts found in dark web marketplaces.

4.3.4.3 Prediction Function and Processing Pipeline

To enable LIME to interact with the models, prediction wrappers are defined for each classification domain. These wrappers accept raw text and return class probabilities after preprocessing with the tokenizer and inference using the respective models. The preprocessing includes standardized padding, truncation, and length normalization to ensure consistent model input formatting. The prediction pipeline handles tokenization, device management, and output probability extraction while ensuring compatibility with LIME's explanation generation requirements and maintaining computational efficiency across all classification domains.

4.3.4.4 LIME Integration and Configuration

We use the LimeTextExplainer configured with domain-specific parameters optimized for dark web content analysis. The explainer is instantiated with class names corresponding to each classification domain for better visualization and interpretation. Key configuration elements include preservation of word order, optimized neighborhood sampling through kernel width adjustment, and feature analysis for word-level importance extraction. For each text sample across all classification domains, LIME highlights the most influential words contributing to the classification outcome through a systematic process of perturbation analysis, local model training, and feature importance ranking.

4.3.4.5 Explanation Generation Process

The explanation generation follows a consistent methodology across all classification

domains: **Perturbation Analysis:** Creating variations of the input text by masking or replacing words to understand feature influence on model predictions. **Local Model Training:** Fitting linear models to approximate the complex model's behavior in the local neighborhood of each sample. **Feature Importance Ranking:** Identifying and ranking the most influential words for classification decisions with numerical importance scores. **Multi-Class Analysis:** Providing confidence scores and explanations across multiple probable categories for comprehensive understanding.

4.3.4.6 Visualization and Interpretability

The system generates comprehensive explanations that include top predicted classes with associated confidence scores, feature importance scores for individual words, color-coded visualization of influential terms, and interactive displays compatible with notebook environments for detailed analysis. This setup helps visualize which keywords in darknet listings influence the AI models' predictions across drug, digital goods, and tutorial classifications, improving transparency and trust in automated systems monitoring illicit content.

This comprehensive XAI implementation provides essential transparency and interpretability for dark web content classification across drug, digital goods, and tutorial categories, supporting both operational requirements and research objectives while maintaining high standards of accuracy and reliability.

4.3.5 Sentiment Analysis

4.3.5.1 Dataset Description and Structure

The sentiment analysis task was conducted on pharmaceutical feedback data collected from user reviews. The dataset was available in two formats to support different model.

4.3.5.2 Preprocessing and Label Engineering

Initial preprocessing involved removing entries with missing rating values to preserve

dataset quality. Numerical user ratings were mapped to sentiment categories (e.g., positive, neutral, negative) in line with standard pharmaceutical feedback sentiment conventions. Column names were standardized—"Filtered Rating" was renamed to `label` and "Feedback" to `text`—to align with typical transformer input schema.

4.3.5.3 Data Partitioning and Input Handling

The dataset was partitioned using stratified sampling to ensure balanced sentiment representation across splits: 80% for training, and 10% each for validation and testing. All models used tokenizer-specific preprocessing, with padding and truncation enabled. The maximum sequence length was set to 512 tokens to support variable-length user feedback without loss of critical information.

4.3.5.4 Training Configuration and Implementation Details

A unified training framework was applied to all models to ensure comparability. Models were trained for three epochs with a learning rate of 2×10^{-5}, batch size of 16, and weight decay of 0.01. Evaluation was performed every 500 steps using standard Hugging Face utilities. The `AutoModelForSequenceClassification` architecture was employed for each model.

DistilBERT and DistilRoBERTa were used with default configurations and pre-trained weights from Hugging Face. No architecture-level modifications were required beyond standard classification head initialization.

DistilGPT-2, being a generative model, required additional setup for classification. Since it lacks a native padding token, `pad_token_id` was manually set to `tokenizer.eos_token_id` to support batch-based training.

4.3.5.5 Implementation and Reproducibility Considerations All experiments were conducted in Google Colab using NVIDIA T4 GPUs. Reproducibility was ensured by setting a fixed random seed (42), maintaining consistent data splits, and leveraging widely adopted libraries: `transformers`, `datasets`, `pandas`, and `scikit-learn`.

This setup facilitates reproducibility and replicability in both academic and production contexts.

4.4 Retrieval-Augmented Generation

This section presents a Retrieval-Augmented Generation (RAG).[28] system that answers user queries using structured darknet product listings. The system integrates dense vector search with a language model for grounded and context-aware generation, enabling intelligent querying of marketplace data through natural language interactions.

4.4.1 Data Preparation and Preprocessing

The RAG system utilizes a comprehensive dataset of darknet marketplace listings, focusing on essential product attributes to ensure comprehensive coverage of marketplace characteristics. The data preparation process involves careful selection and preprocessing of key attributes including product name, pricing information, seller details, geographical shipping constraints, category classification, inventory levels, and availability status. The preprocessing pipeline implements robust data cleaning procedures, removing entries with missing critical information to ensure data quality and consistency. Each product listing is transformed into a structured, descriptive format that captures all relevant marketplace information in a standardized representation suitable for semantic search and retrieval operations. The transformation process converts tabular data into natural language descriptions, creating coherent textual representations that preserve the structured information while enabling effective semantic matching during retrieval operations. This approach ensures that the retrieval system can effectively match user queries with relevant product listings based on semantic similarity rather than exact keyword matching.

4.4.2 Embedding Generation and Vector Indexing

The system employs advanced sentence embedding techniques using the Sentence-BERT architecture, specifically utilizing a large-scale general text embedding model optimized for semantic understanding across diverse domains. The embedding model processes the prepared textual descriptions to generate high-dimensional dense vector representations that capture semantic meaning and contextual relationships within the product listings. Vector normalization is applied during the embedding generation process to ensure consistent similarity measurements and optimal retrieval performance. The resulting embeddings maintain semantic relationships between similar products while distinguishing between different categories and characteristics. For efficient similarity search operations, the system implements FAISS (Facebook AI Similarity Search) indexing with inner product similarity measurement. The index structure is optimized for real-time retrieval operations, enabling rapid identification of the most relevant product listings for any given user query. The indexing approach supports scalable operations across large datasets while maintaining sub-second response times for interactive applications.

4.4.3 Language Model Integration and Generation Pipeline

The generation component utilizes an instruction-tuned causal language model specifically designed for multi-turn conversational interactions. The model selection prioritizes both performance and efficiency, enabling deployment in resource-constrained environments while maintaining high-quality generation capabilities. The language model is integrated through a comprehensive pipeline that handles tokenization, model loading, and generation configuration. The pipeline implements temperature-based sampling strategies to balance response diversity with factual accuracy, ensuring that generated answers remain grounded in the retrieved context while providing natural, conversational responses. Model configuration includes

optimized parameters for maximum token generation, temperature settings for controlled randomness, and sampling strategies that promote coherent and contextually appropriate responses. The pipeline design ensures consistent performance across different query types and complexity levels.

4.4.4 Retrieval-Augmented Query Processing

The core RAG functionality implements a sophisticated query processing pipeline that seamlessly integrates retrieval and generation components. When processing user queries, the system first generates query embeddings using the same embedding model employed for document indexing, ensuring semantic consistency between queries and indexed content. The retrieval process employs configurable top-k selection, allowing for flexible balance between context richness and processing efficiency. Retrieved documents are ranked by semantic similarity and concatenated into structured context blocks that provide comprehensive information while maintaining readability and coherence. The prompt engineering component constructs carefully designed prompts that clearly delineate context information, user questions, and expected response format. This structured approach ensures that the language model can effectively utilize the retrieved information to generate accurate, contextually appropriate answers that directly address user inquiries.

4.4.5 Interactive Interface Development

The system features a comprehensive interactive interface developed using modern web application frameworks, providing users with intuitive access to the RAG capabilities. The interface design prioritizes user experience through clean, responsive layouts that accommodate various query types and response formats. The interface implementation includes real-time query processing, displaying results with minimal latency while providing visual feedback during processing operations. Users can submit natural language questions through text input fields and receive formatted

responses that clearly present the generated answers along with relevant context information. The interface architecture supports both local deployment and public sharing capabilities, enabling flexible deployment scenarios for different use cases. The system includes comprehensive error handling and user guidance features to ensure smooth operation across different user skill levels and query complexity.

This comprehensive RAG implementation provides powerful capabilities for intelligent querying of darknet marketplace data, combining advanced retrieval techniques with sophisticated language generation to deliver accurate, contextual responses to user inquiries while maintaining system efficiency and usability.

Chapter 4

Data Visualization

5 CHAPTER 4: Data Visualization

5.1 Introduction

This chapter presents the visual analysis and dashboard construction process that transforms raw darknet data into actionable intelligence. Leveraging the structured datasets extracted from Hidden Market and MGM Grand marketplaces, we developed interactive dashboards that summarize key marketplace behaviors, trends, and risk signals. The dashboards were designed to support cybersecurity professionals by offering a visual overview of product categories, vendor activity, pricing anomalies, and shipping flows. The analysis was conducted using Power BI and Tableau, with each visualization tailored to highlight patterns and relationships relevant to darknet threat monitoring. In particular, visuals such as bar charts, funnel plots, pie charts, and geo-mapping provide insights into sales performance, product distribution, and potential anomalies across regions. Each marketplace is analyzed separately, followed by a presentation of the full dashboard view and key interpretations that support decision-making in cyber threat intelligence workflows.

5.2 Marketplace 1: Hidden Market

5.2.1 Dashboard Overview

The Hidden Market dashboard was developed using Power BI to facilitate interactive exploration of product-level data. It contains both numerical KPIs and categorical visualizations to highlight activity patterns across product categories, vendors, and shipping destinations. The primary focus of this dashboard is to monitor the popularity, stock levels, vendor contributions, and geographical reach of the products listed in the marketplace. The dashboard is organized into the following components

- **KPI Cards:**

Table 5.1: Dashboard KPI Metrics

KPI Metric	Value
Total Number of Products	18.233K
Total Number of Sellers	881
Total Purchased Units	107K
Total Views	6M
Average Product Price	$614.52

- **KPI Gauge Charts:**

Table 5.2: Dashboard Performance Metrics

KPI Metric	Current Value	Target
Percentage of Products in Stock	0.97	0.9
Average Rating	17.62%	70%
Average Units Sold per Product	5.89	50
Conversion Rate (Purchased ÷ Views)	0.02	0.1

The analysis of **Hidden Market** reveals key operational characteristics:

- Market Overview:

With over 18,000 products and nearly 900 sellers, the marketplace shows strong vendor diversity. However, the 107K purchases from 6M views indicate high interest but low conversion.

- Inventory Strength:

An impressive 97% of products are in stock, suggesting efficient inventory handling or inflated availability to boost seller credibility.

- Performance Gaps:

Critical KPIs fall short: average rating is 17.6% (vs. 70% target), conversion rate is 2% (vs. 10%), and units sold average only 5.9 (vs. 50 target). These point to poor engagement, trust issues, or ineffective product strategies.

- Economic Focus:

The average price is $614.52, indicating a high-value market segment. Yet low purchase volumes suggest pricing or trust barriers that limit buyer commitment. Overall, while the Hidden Market attracts attention and maintains strong inventory, it suffers from low buyer conversion and engagement, reflecting common challenges in darknet commerce.

5.2.2 Main Visuals

A variety of categorical charts were created to analyze product behavior and market dynamics. These visuals were embedded into the dashboard and directly support actionable threat intelligence.

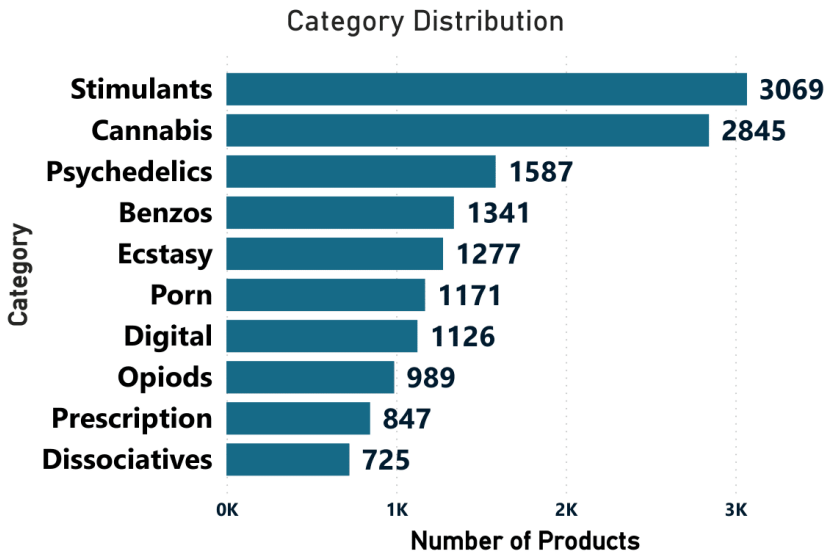

Figure 5.1: Category Distribution

- **Category Distribution**

  Displays the number of distinct products per category. This reveals the most

saturated or active categories in Hidden Market which is **Drugs**.

Figure 5.2: Top Vendors by Quantity Sold

- **Top Vendors by Quantity Sold** Ranks vendors by total units sold. here the figure indicates most active vendor in the market with max number of products sold **danielvitor61** which can lead to further analysis specifically on this vendor as a threat detection aspect.

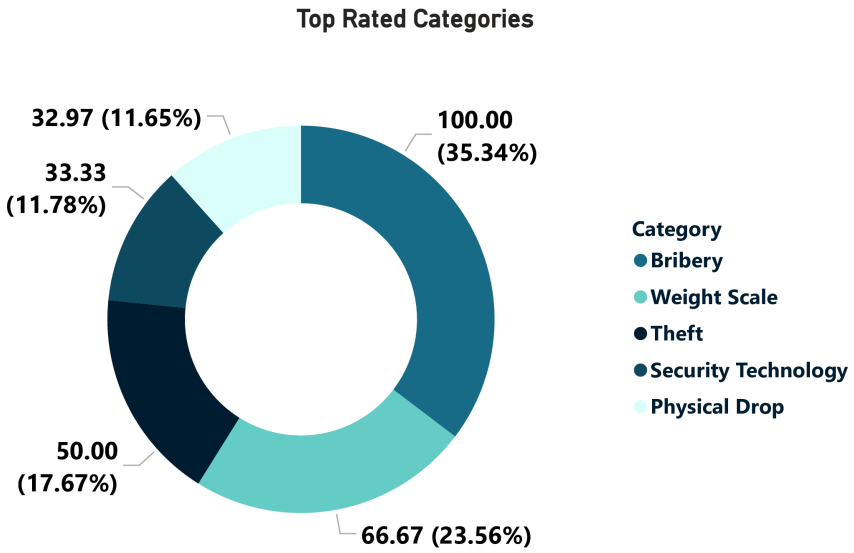

Figure 5.3: Top-Rated Categories

- **Top-Rated Categories** – Shows average product rating per category. Helps identify quality trends across product domains. The analysis reveals that Bribery leads with 35% of top ratings, followed by Weight Scale at 23%, and Theft at 17%. Security Technology and Physical Drop categories show lower ratings at 11.78% and 11.65% respectively. This distribution suggests that service-based categories (Bribery) and equipment-related products (Weight Scale) receive higher customer satisfaction ratings compared to digital security tools, potentially indicating more reliable delivery and product quality in these segments.

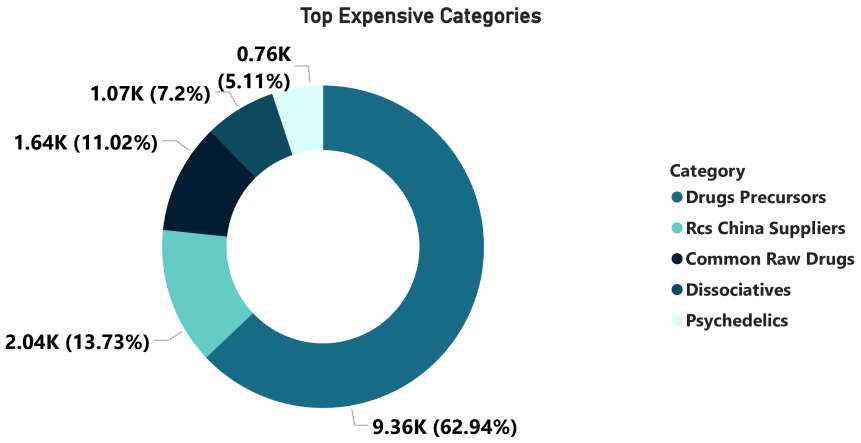

Figure 5.4: Top $ Categories

- **Top $ Categories** – Displays categories with the highest average price per product. here the most expensive category belongs to drugs with also different types of it(Dissociatives,Psychedelics) and other categories

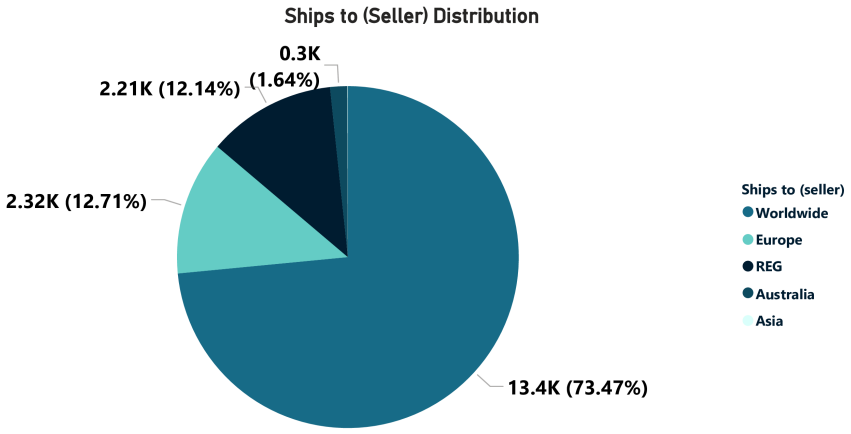

Figure 5.5: Ships To (Seller) Distribution

- **Ships To (Seller) Distribution**Indicates seller-defined shipping regions. Useful for analyzing supply chain origin.The most of Hidden Market Vendors claim that their products ship to everywhere world wide but that will be based on product itself as product related.

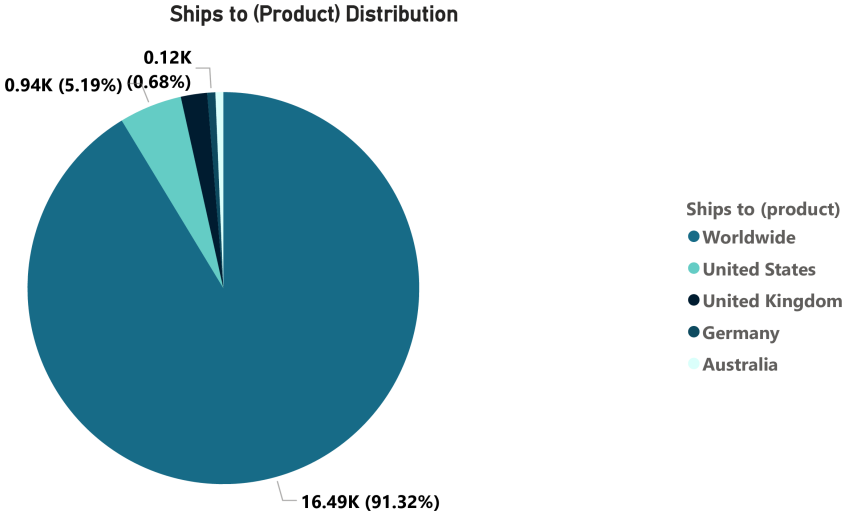

Figure 5.6: Ships To (Product) Distribution

- **Ships To (Product) Distribution** Based on individual product listings, shows intended delivery destinations. This supports demand-side geolocation insights.Most of vendors actually ship their products ship to everywhere all around the world worldwide, but Most Known country that recieves products is U.S.A, comes after it the U.K.

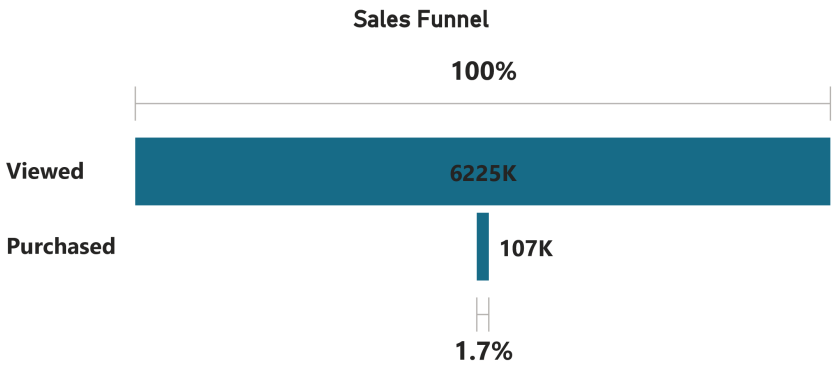

Figure 5.7: Sales Funnel: Views to Purchases

- **Sales Funnel (Views → Purchased)** – A funnel chart visualizing the drop-off rate from product views to actual purchases. This chart informs conversion efficiency and market interest.here can conclude that most of users joined **Hidden Market** just for watching products not actually purchase it .

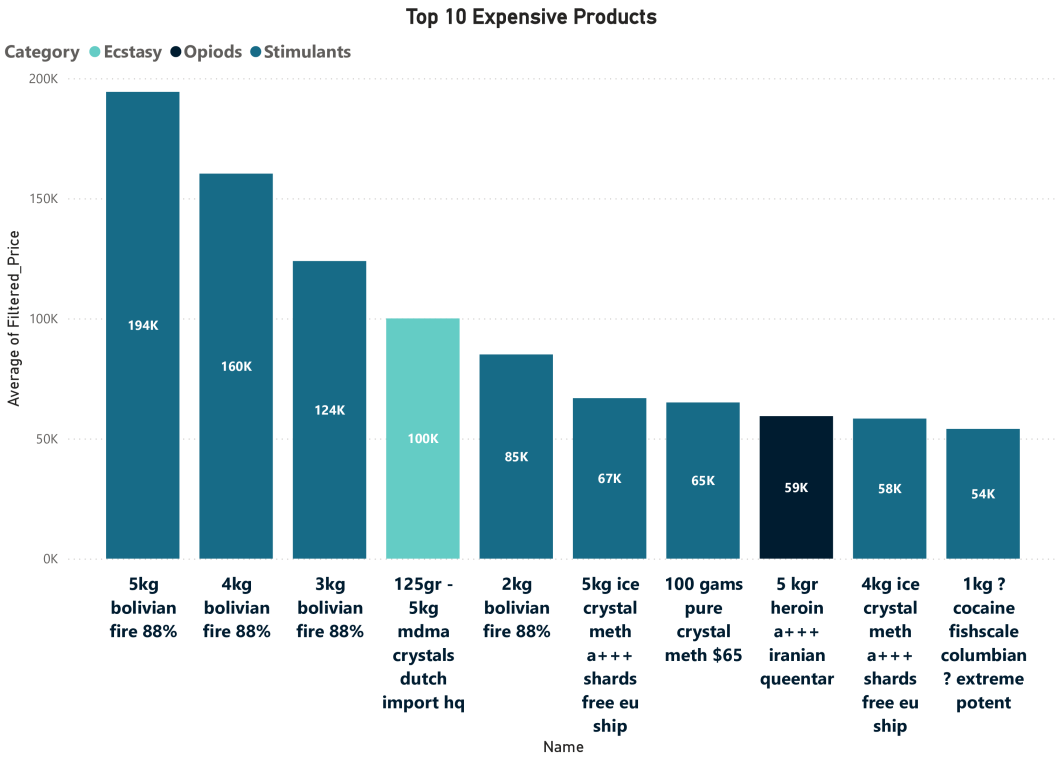

Figure 5.8: Top 10 Most Expensive Products

- **Top 10 Most Expensive Products** – Highlights the highest-priced listings. These can be signal high-risk items as shown most all of these products are drugs

with different categories.

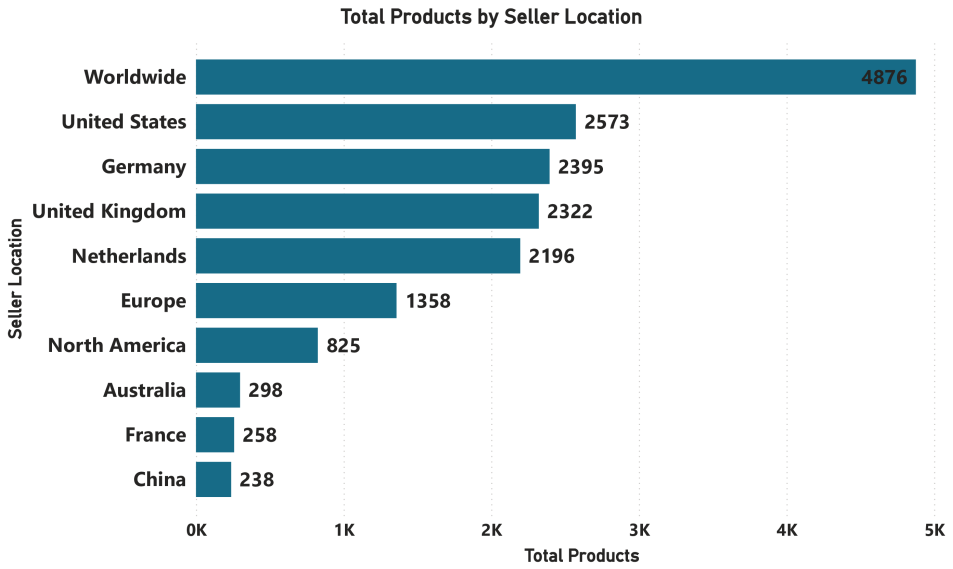

Figure 5.9: Seller Location vs Distinct Product Count

- **Seller Location vs Product Count**Compares number of distinct products per seller country.Here the most are worldwide for security purposes, then comes U.S.A as most known country for vendors.

5.2.3 Final Dashboard Snapshot

A full snapshot of the final Power BI dashboard is provided in Figure 5.17, summarizing all interactive visuals, filters, and KPIs used for monitoring the Hidden Market. The dashboard presents a comprehensive analysis of 14,017 products distributed across ten distinct categories, with Stimulants (3,069 products) and Cannabis (2,845 products) representing the largest market segments. The interface provides real-time insights through dynamic visualizations, enabling stakeholders to track category distributions, market trends, and performance metrics. Interactive filtering capabilities allow users to drill down into specific timeframes and cross-category relationships, while key performance indicators monitor overall market health and emerging patterns within the hidden marketplace ecosystem.

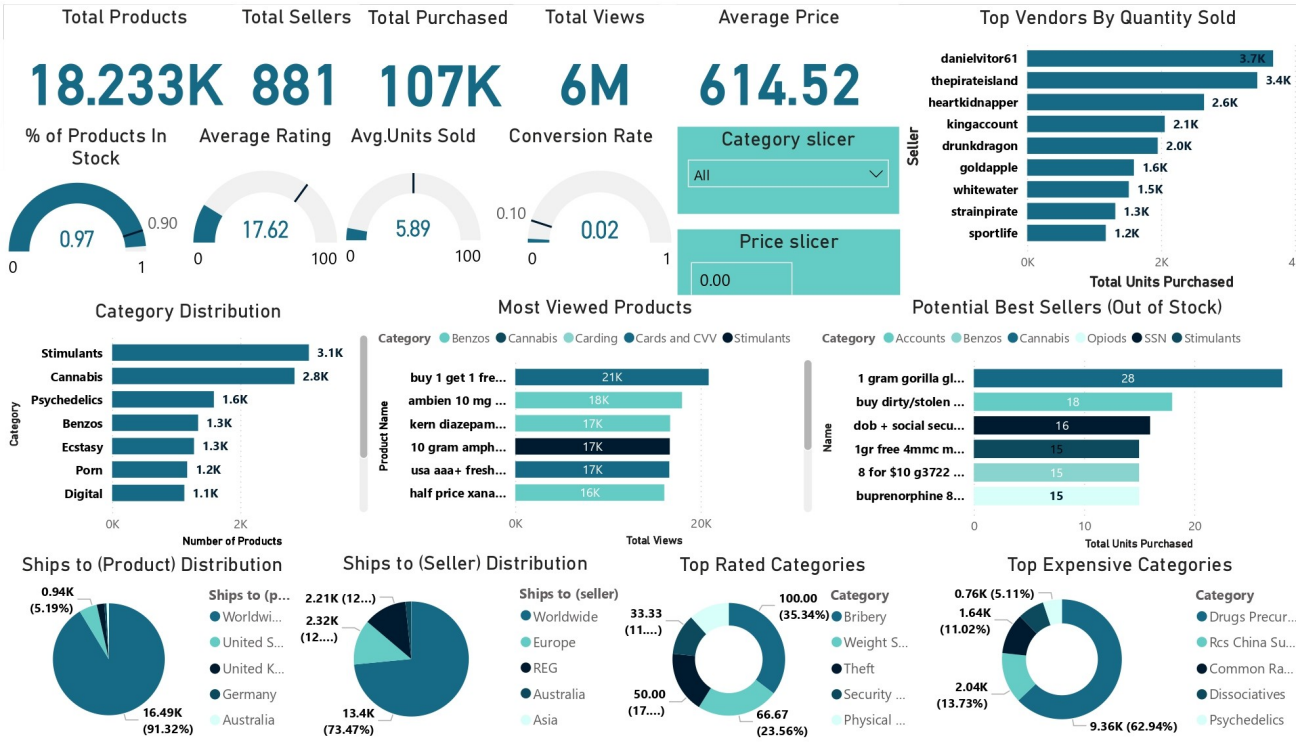

Figure 5.10: Final Dashboard for Hidden Market

5.3 Marketplace 2: MGM Grand

The analysis of **MGM Grand**, a darknet marketplace that operates behind complex CAPTCHA-protection barriers, provides valuable insights into its structure, operations, and threat patterns. Despite the technical challenges involved in accessing and scraping this site—including automated CAPTCHA solving and restricted session management—we successfully extracted a significant volume of structured product-level data. Following rigorous cleaning and preprocessing, the resulting dataset—although smaller in scale compared to the Hidden Market—exhibited higher consistency, fewer anomalies, and more reliable attribute completion. This enabled the development of robust visualizations and metric-based interpretations. The analysis covers vendor activity, product availability, pricing trends, and shipping destinations, highlighting the marketplace's transactional behavior and regional distribution strategies. Key performance indicators and advanced visual tools such

as funnel charts, location maps, and bar plots were employed to assess conversion efficiency, identify top-rated or high-risk items, and examine supply chain vectors.

5.3.1 Key Performance Indicators (KPIs)

A set of KPIs was computed and visualized using metric cards and gauges:

Table 5.3: Dashboard KPI Metrics - Second Market

KPI Metric	Value
Total Number of Products	8.161K
Total Number of Vendors	331
Total Purchased Units	203K
Average Product Price	$358.36
% of Products in Stock	99%
% of Products with Escrow	98%

These KPIs collectively describe market size, buyer activity, product availability, and engagement effectiveness.

5.3.2 Main Visuals

A diverse set of categorical and relational charts was developed and embedded within the dashboard to provide comprehensive insights into the marketplace data. These visualizations enable stakeholders to understand key business metrics, vendor performance, product distributions, and geographical patterns through interactive and intuitive representations. The dashboard incorporates multiple visualization types including bar charts, pie charts, treemaps, and geographical maps to cater to different analytical perspectives.

Top Vendors by Quantity Sold

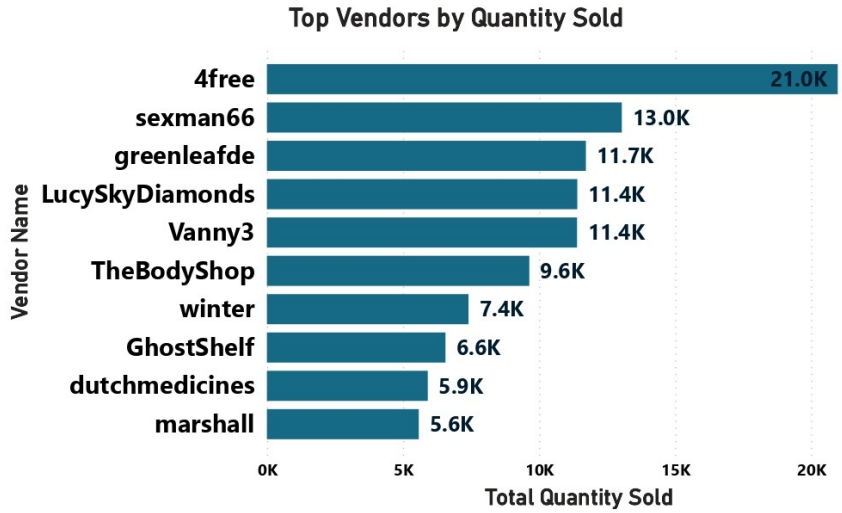

Figure 5.11: Top Vendors by Quantity Sold

The top performer across the figures is **4free**, with 21,000 units sold as shown in the "Top Vendors by Quantity Sold" bar chart. This indicates 4free as the leading vendor by sales volume in the marketplace.

Top Vendors by Number of Products

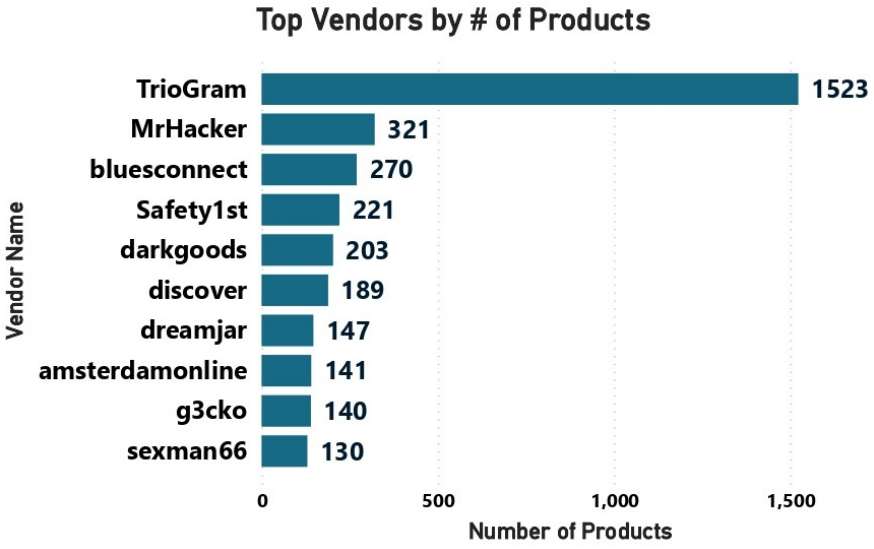

Figure 5.12: Top Vendors by Number of Products

Bar chart highlights the leading vendors based on the diversity of their product offerings, with **TrioGram** at the forefront offering 1,523 products. This is followed by **MrHacker** with 321 products, **bluesconnect** with 270 products, **Safety1st** with 221 products, and additional vendors, underscoring the significant variety in their respective inventories within the marketplace. **Category Distribution**

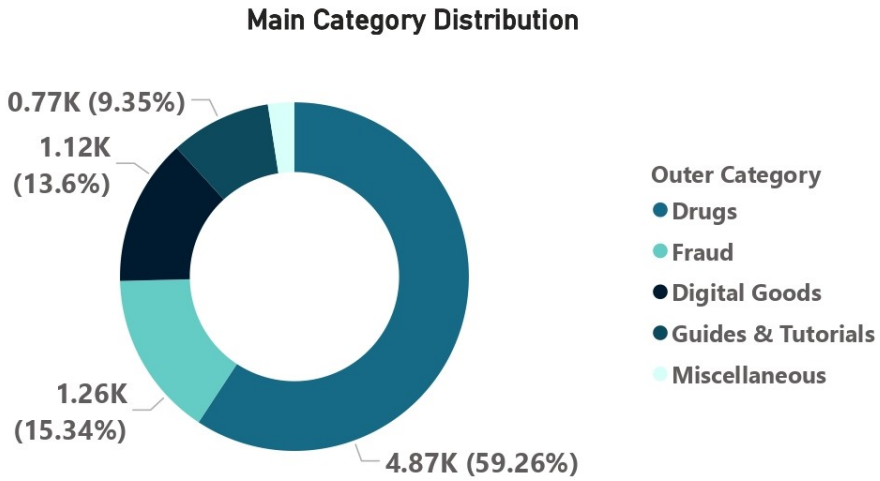

Figure 5.13: Main Category Distribution

Drugs lead with 59.26% (4.87K), followed by Fraud (15.34%, 1.26K), Digital Goods (13.6%, 1.12K), Guides & Tutorials (9.35%, 0.77K), and Miscellaneous (1.6%), showing a strong emphasis on drug-related sales. **Most Expensive Main Categories**

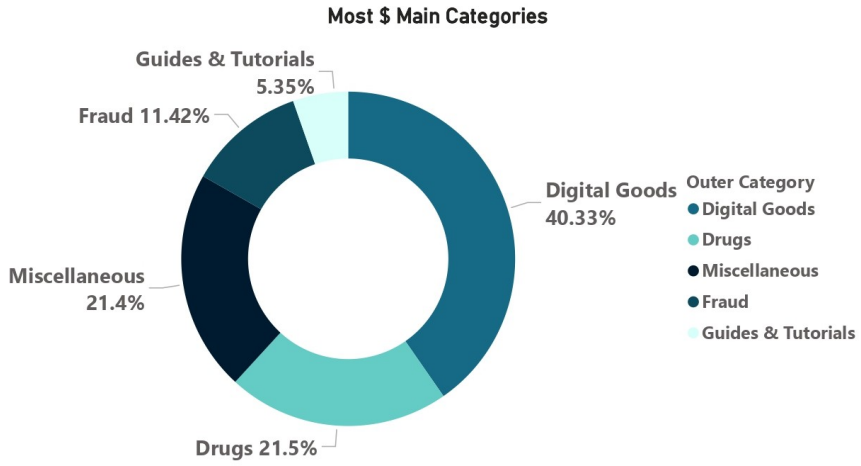

Figure 5.14: Most Expensive Main Categories

Digital Goods dominate with 40.33%, followed by Drugs (21.5%), Miscellaneous (21.4%), Fraud (11.42%), and Guides & Tutorials (5.35%), indicating a diverse range of high-value transactions. **Ships To Distribution**

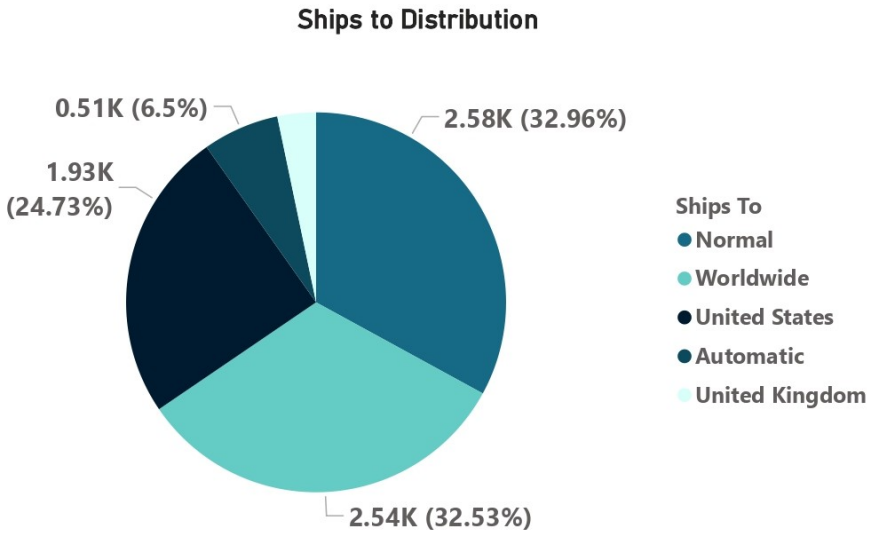

Figure 5.15: Ships To Distribution

Worldwide shipping is the largest category at 32.96% (2.58K), followed closely by Normal (32.53%, 2.54K), United States (24.73%, 1.93K), United Kingdom (6.5%, 0.51K), and Automatic (not specified), indicating broad global reach. **Ships From Distribution**

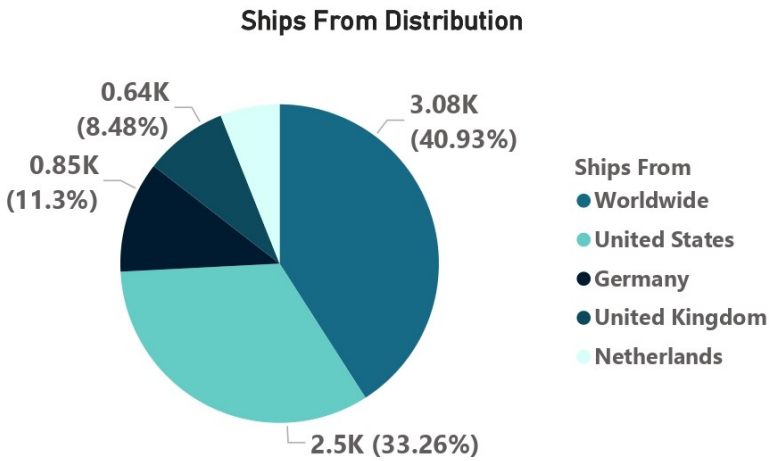

Figure 5.16: Ships From Distribution

Worldwide dominates with 40.93% (3.08K), followed by the United States (33.26%, 2.5K), Germany (11.3%, 0.85K), United Kingdom (8.48%, 0.64K), and Netherlands (not specified), showing diverse origin points.

5.3.3 Complete Dashboard Overview

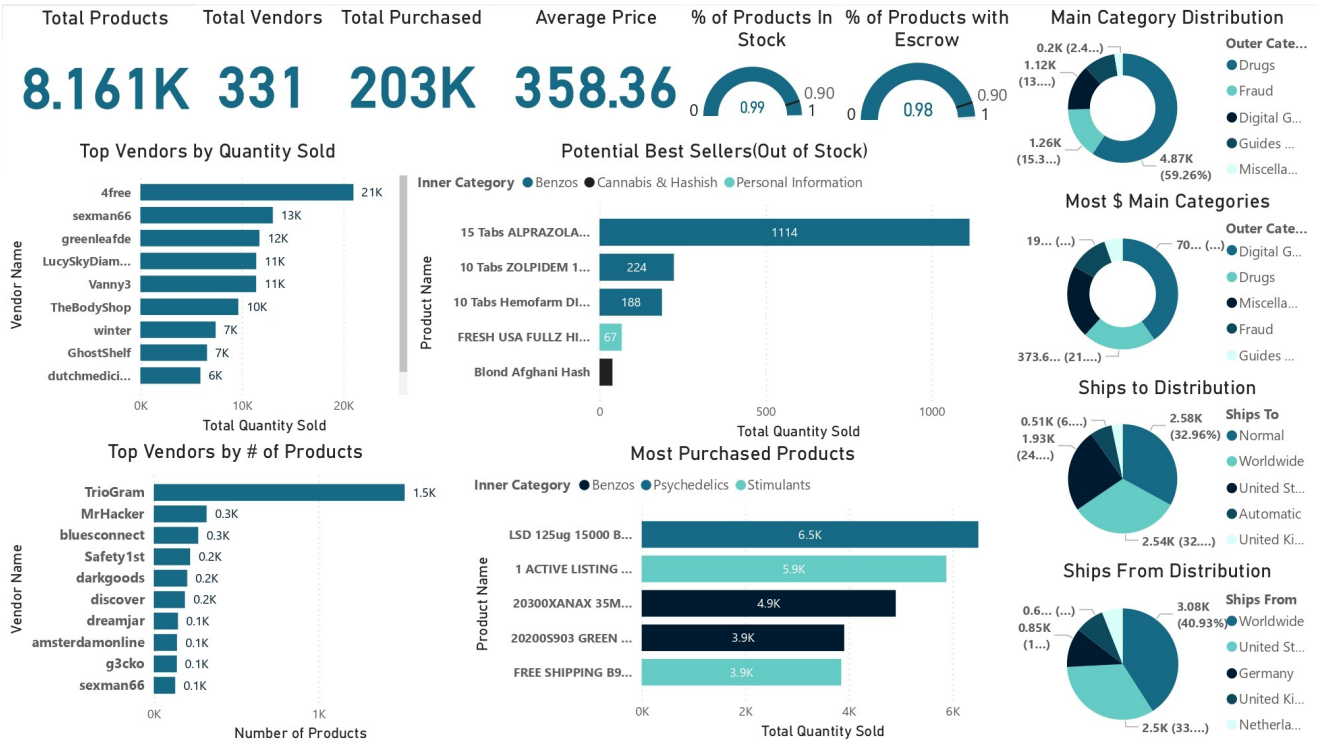

Figure 5.17: Final Dashboard for Hidden Market

5.3.4 Summary Interpretation

The MGM Grand marketplace reflects a smaller but well-maintained marketplace with:

- High product availability (99% in stock) with 8,161 products across 331 vendors and clean seller data through established reputation systems
- Lower sales throughput (203K total purchases, 17.60 average units per product) compared to Hidden Market, yet with more balanced product categories across drugs (59.26%), fraud (15.34%), and digital goods (13.6%)

- Similar conversion challenges, highlighting low buyer engagement relative to product views, with concentrated vendor performance (top 3 vendors accounting for 46K+ units sold)
- Strong operational infrastructure evidenced by 98% escrow adoption and diverse geographic distribution (40.93% worldwide shipping)
- Dominant drug categories in Cannabis & Hashish (30.04%) and Stimulants (21.6%), with high-value digital goods generating 40.33% of total revenue
- Notable supply constraints in pharmaceutical products (benzos showing high demand when out of stock) and established vendor hierarchy (56.17% at Lvl.1)

These insights support threat profiling and deeper vendor-centric analysis, aiding in identifying high-risk actors or emerging illicit product trends. The marketplace demonstrates mature criminal infrastructure with sophisticated trust mechanisms, global reach, and diversified product portfolios that enable comprehensive threat assessment and law enforcement targeting strategies.

Chapter 5

Results

6 CHAPTER 5: Results

In this section, we present the experimental results obtained from the four main classification tasks tackled throughout the graduation project: **Drug Classification**, **Digital Classification**, **Tutorial Classification**, and **Sentiment Analysis**. Each task was approached using a variety of machine learning and deep learning models, with a special emphasis on Transformer-based architectures. The models evaluated include:

- **DarkBERT**
- **BERT**
- **RoBERTa**
- **DistilRoBERTa** [3], [29]
- **DistilGPT2** [30] offers a smaller variant of GPT2 [6].

The evaluation primarily focuses on two key metrics:

- **Accuracy:** the proportion of correct predictions among the total number of cases examined.
- **Macro F1-Score:** the harmonic mean of precision and recall computed independently for each class and then averaged, ensuring that all classes are treated equally regardless of their frequency.

In cases where models achieve closely similar performance in terms of accuracy and macro F1-score, we further conduct a **calibration analysis**. This additional evaluation helps to determine which model provides better-calibrated probability estimates, using metrics such as:

- **Expected Calibration Error (ECE)**
- **Maximum Calibration Error (MCE)**

- **Brier Score**
- **Negative Log-Likelihood (NLL)**

The following subsections report the results for each classification problem individually.

6.1 Drug Classification Results

In the Drug Classification task, we evaluated the performance of BERT, RoBERTa, and DarkBERT models on the test set. The evaluation focuses on Accuracy and Macro F1-Score, with further calibration analysis for fine-grained comparison.

Performance Metrics

Table 6.1: Test Accuracy and Macro F1-Score for Drug Classification

Model	Test Accuracy	Macro F1-Score
DarkBERT	0.9634	0.9533
BERT	0.9620	0.9504
RoBERTa	0.9541	0.9394

Calibration Analysis

Table 6.2: Calibration Metrics for Drug Classification

Model	ECE	MCE	Brier Score	NLL
DarkBERT	0.0265	0.3000	0.0634	0.1828
BERT	0.0305	0.5545	0.0681	0.1895
RoBERTa	0.0357	0.4064	0.0812	0.2275

DarkBERT

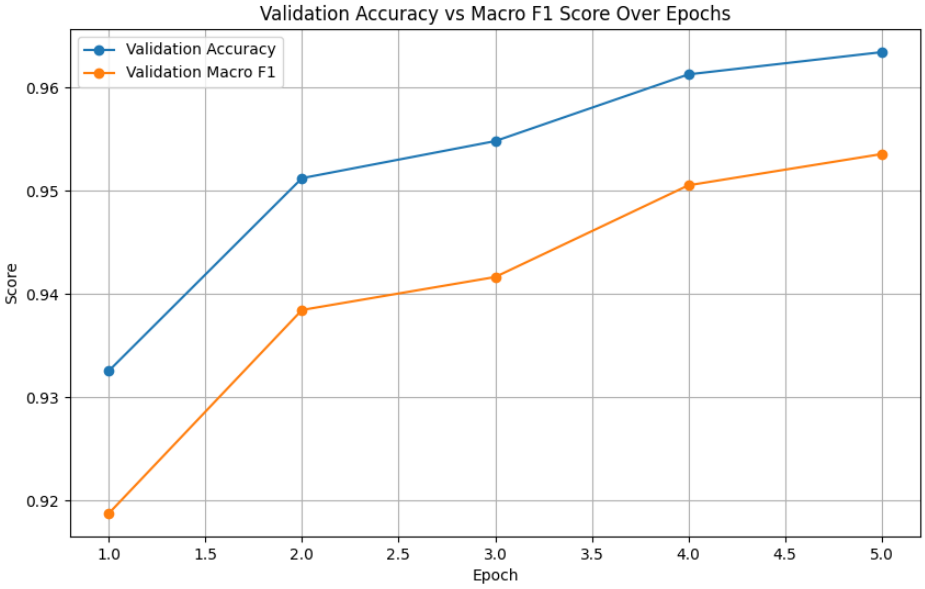

Figure 6.1: DarkBERT Learning Curves: Validation F1-Score and Accuracy

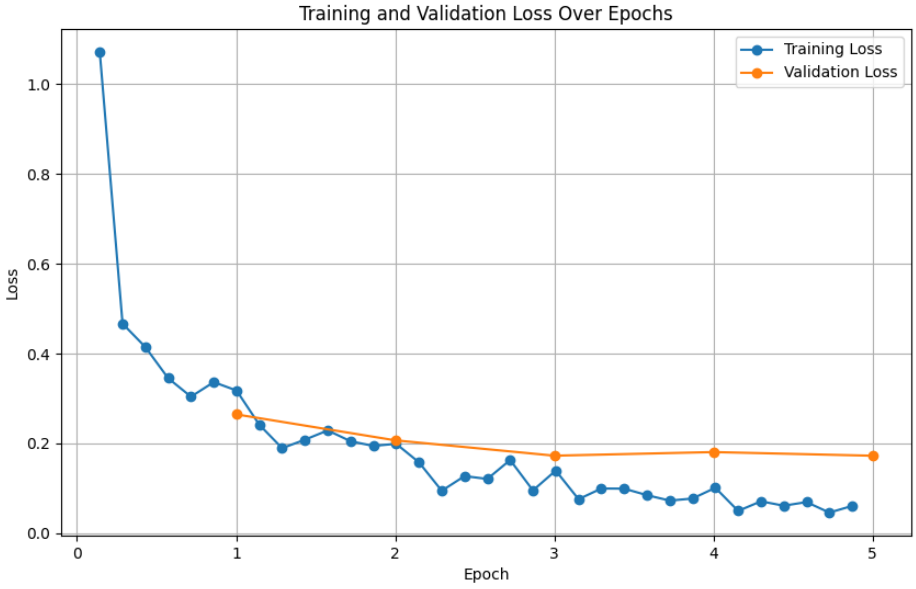

Figure 6.2: DarkBERT Loss Curves: Training and Validation Loss

DarkBERT model's training performance demonstrates consistent improvement in accuracy and F1 score, alongside effective loss reduction, indicating a well-optimized model suitable for the given task.

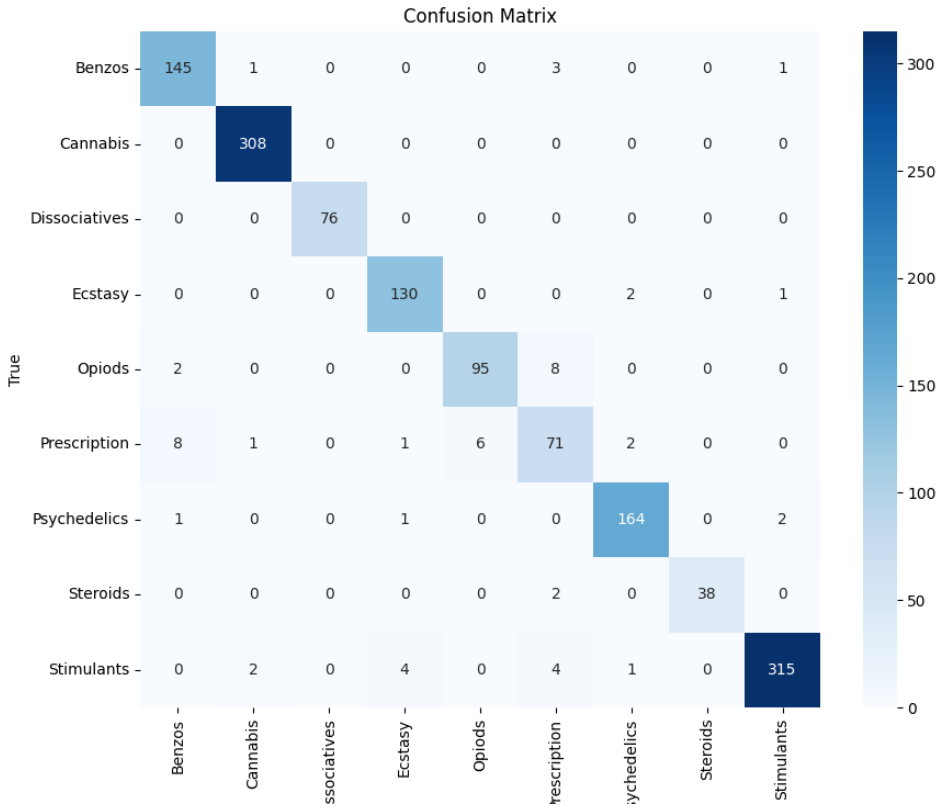

Figure 6.3: DarkBERT Confusion Matrix

The confusion matrix for DarkBERT shows the model's performance in classifying drug categories, with rows as true labels and columns as predicted labels. Correct predictions are on the diagonal: Benzos (145), Cannabis (308), Dissociatives (76), Ecstasy (130), Opioids (95), Prescription (71), Psychedelics (164), Steroids (38), and Stimulants (315), with Cannabis and Stimulants showing the highest accuracy. Misclassifications occur off-diagonal, notably Prescription drugs confused with Benzos (8) and Opioids (1), and Opioids with Prescription (8). The model struggles most with Steroids (38 correct) and shows some overlap between Psychedelics, Steroids, and Stimulants. Overall, the model performs well, especially for Cannabis and Stimulants, but could improve in distinguishing similar categories like Prescription and Opioids.

BERT

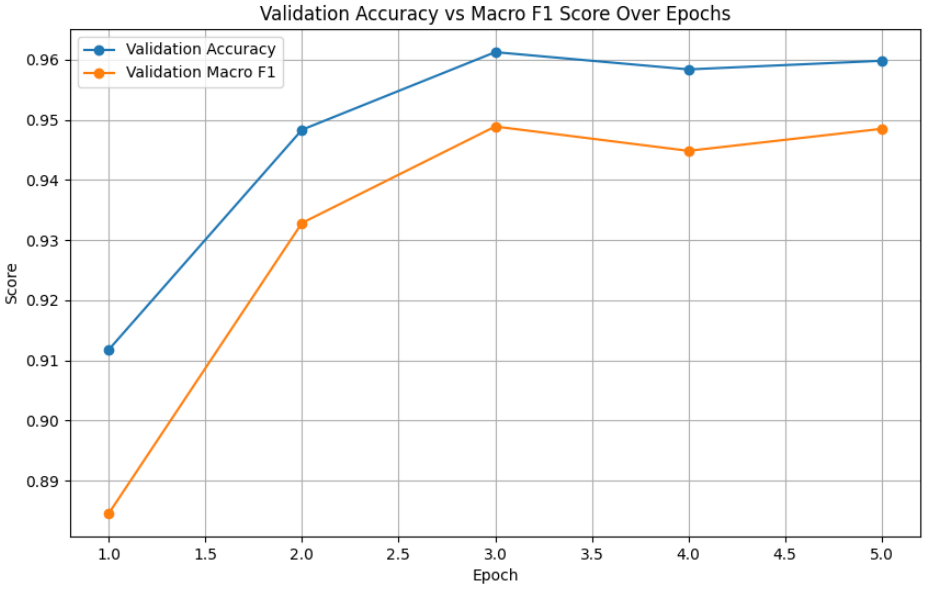

Figure 6.4: BERT Learning Curves: Validation F1-Score and Accuracy

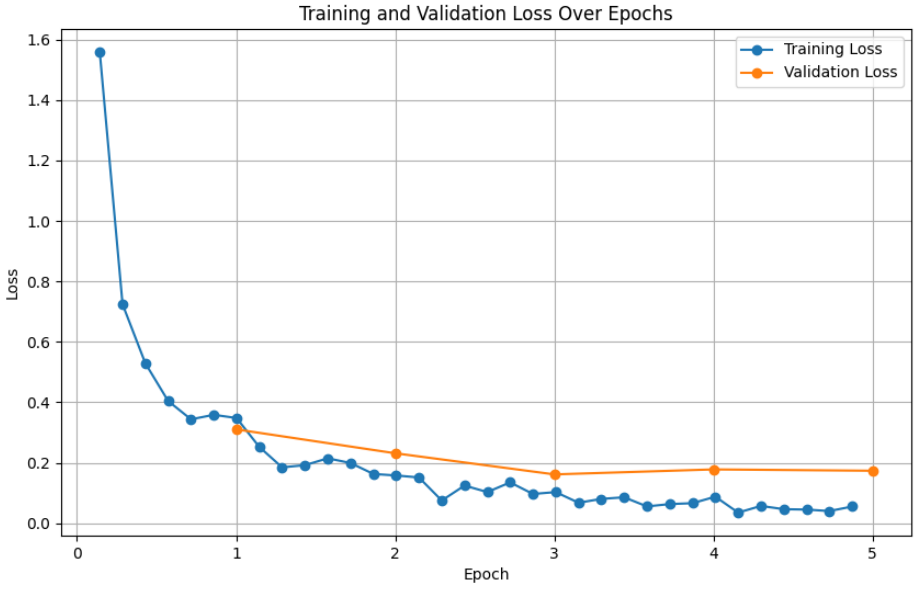

Figure 6.5: BERT Loss Curves: Training and Validation Loss

The BERT model shows consistent learning progression with steady improvements in both accuracy and F1-score. The training and validation loss curves indicate balanced learning without significant overfitting.

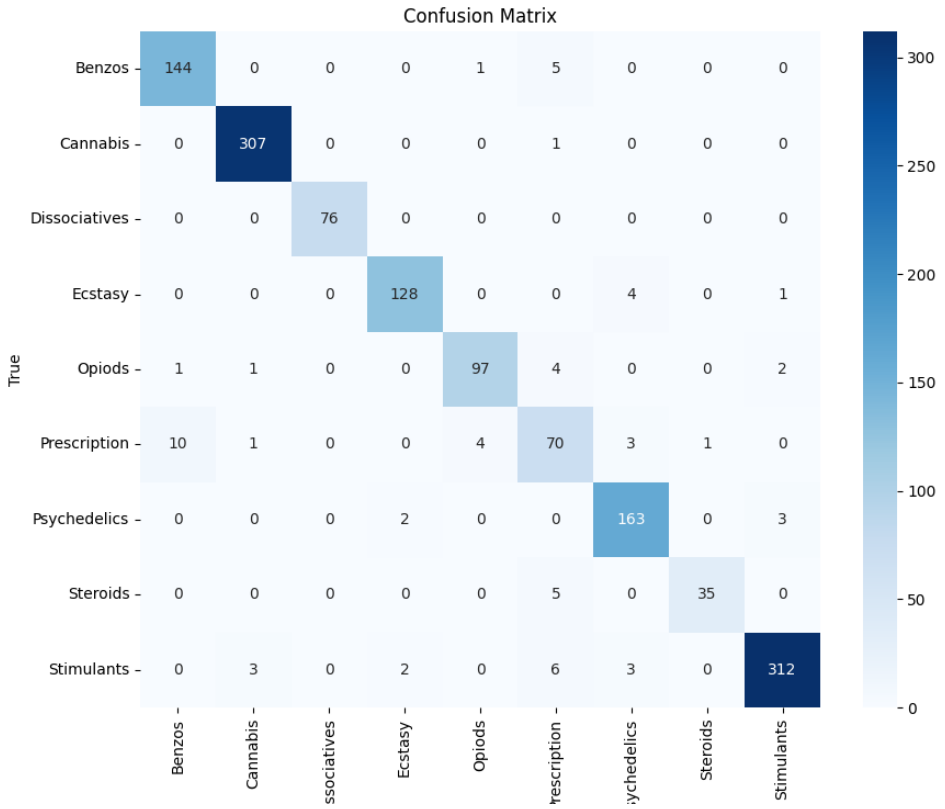

Figure 6.6: BERT Confusion Matrix

The BERT Confusion Matrix (Figure 5.6) displays the model's performance in classifying drug categories, with true labels as rows and predicted labels as columns. Correct predictions along the diagonal include Benzos (144), Cannabis (307), Dissociatives (76), Ecstasy (128), Opioids (97), Prescription (70), Psychedelics (163), Steroids (35), and Stimulants (312), with Cannabis and Stimulants showing the highest accuracy. Misclassifications off-diagonal include Prescription drugs confused with Benzos (10) and Opioids (4), and Opioids with Prescription (4) and Stimulants (2). The model struggles most with Steroids (35 correct) and shows minor overlaps, such as Psychedelics with Steroids (3) and Stimulants with multiple categories (3 with Cannabis, 2 with Ecstasy, 6 with Opioids, 3 with Psychedelics). Overall, the model performs strongly for Cannabis and Stimulants but could improve in distinguishing similar categories like Prescription and Opioids.

RoBERTa

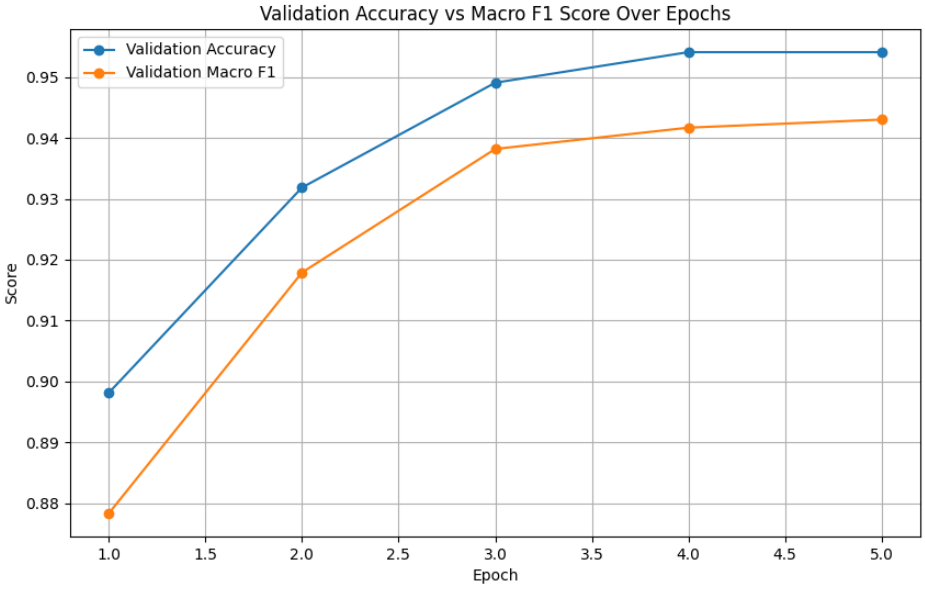

Figure 6.7: RoBERTa Learning Curves: Validation F1-Score and Accuracy

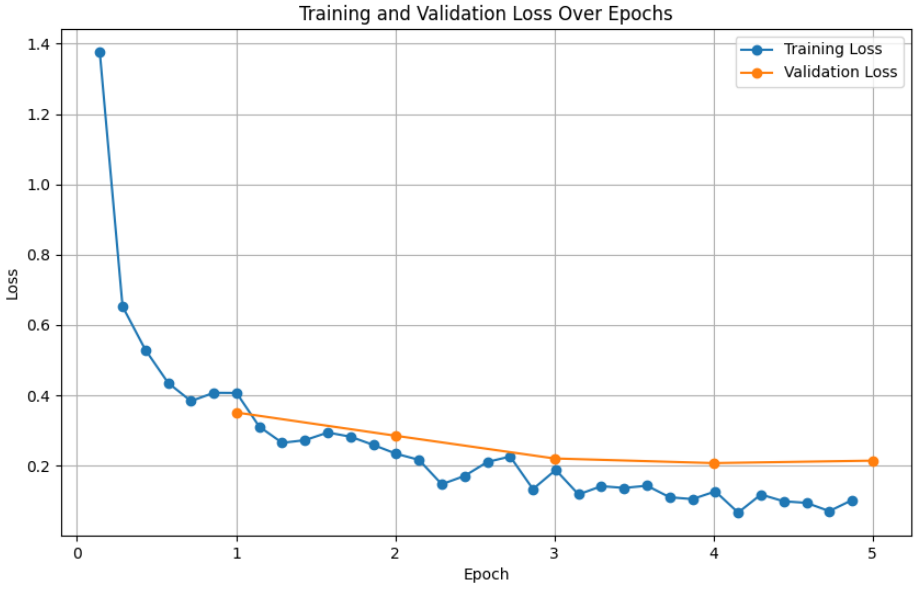

Figure 6.8: RoBERTa Loss Curves: Training and Validation Loss

RoBERTa exhibits smooth learning curves with competitive performance metrics. The model demonstrates stable training dynamics with well-controlled loss progression throughout the training process, coupled with effective loss reduction, highlights

the model's ability to learn and generalize effectively. The stabilization of metrics and loss curves by epoch 5 indicates that the model has achieved a high level of predictive reliability and is well-suited for the given task.

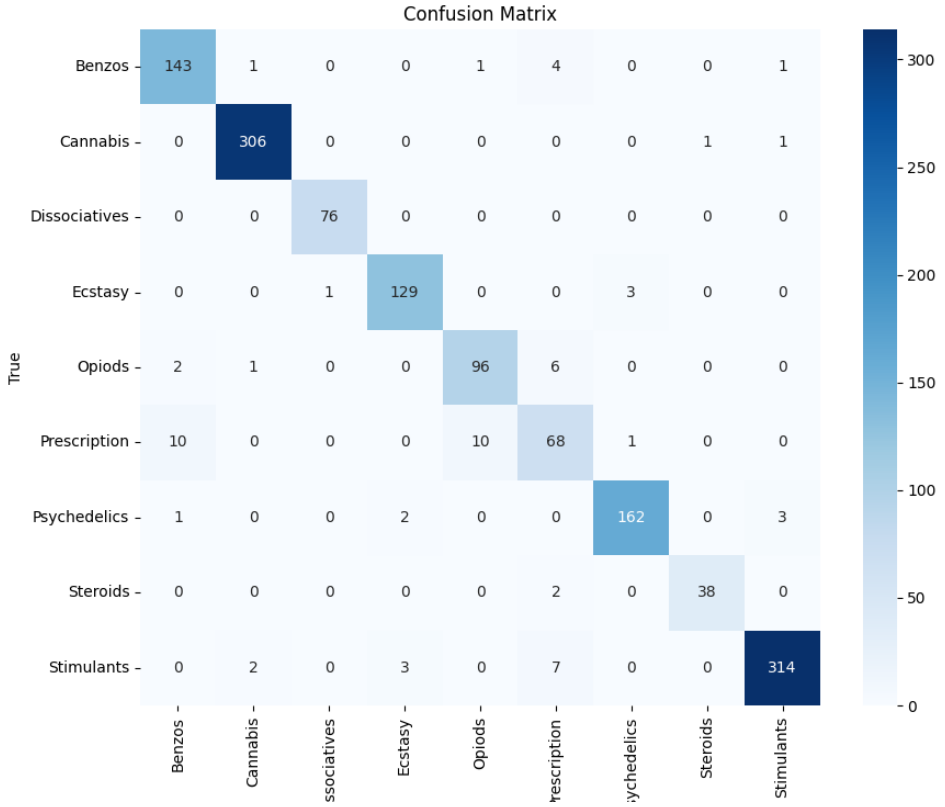

Figure 6.9: RoBERTa Confusion Matrix

The RoBERTa Confusion Matrix (Figure 5.9) shows the model's performance in classifying drug categories, with true labels as rows and predicted labels as columns. Correct predictions along the diagonal include Benzos (143), Cannabis (306), Dissociatives (76), Ecstasy (129), Opioids (96), Prescription (68), Psychedelics (162), Steroids (38), and Stimulants (314), with Cannabis and Stimulants exhibiting the highest accuracy. Misclassifications off-diagonal include Prescription drugs confused with Benzos (10) and Opioids (10), and Opioids with Prescription (10) and Stimulants (7). The model struggles most with Steroids (38 correct) and shows minor overlaps, such as Psychedelics with Steroids (3) and Stimulants with multiple categories (2 with Cannabis, 3 with Ecstasy). Overall, the model performs well for Cannabis and

Stimulants but could improve in distinguishing similar categories like Prescription and Opioids.

Among the tested models, **DarkBERT** achieved the highest test accuracy (0.9634) and macro F1-score (0.9533), closely followed by BERT. While RoBERTa showed slightly lower performance in both metrics, it remained competitive. To resolve the tight margin between BERT and DarkBERT, we conducted calibration analysis. DarkBERT demonstrated superior calibration performance, with the lowest Expected Calibration Error (ECE = 0.0265), Brier Score (0.0634), and Negative Log-Likelihood (NLL = 0.1828). These results indicate that DarkBERT not only achieved the best predictive performance but also produced the most reliable probability estimates. Therefore, **DarkBERT is selected as the best-performing model for the Drug Classification**.

6.2 Digital Classification Results

The Digital Classification task involved evaluating BERT, RoBERTa, and DarkBERT models on the test set. As with the previous task, the focus is on accuracy and macro F1-score, with calibration analysis to support decision-making in cases of similar performance.

Performance Metrics

Table 6.3: Test Accuracy and Macro F1-Score for Digital Classification

Model	Test Accuracy	Macro F1-Score
DarkBERT	0.9634	0.9533
BERT	0.9164	0.8528
RoBERTa	0.9100	0.8299

Calibration Analysis

Table 6.4: Calibration Metrics for Digital Classification

Model	ECE	MCE	Brier Score	NLL
DarkBERT	0.0265	0.3000	0.0634	0.1828
BERT	0.0341	0.5851	0.1142	0.2725
RoBERTa	0.0467	0.4782	0.1217	0.3083

DarkBERT

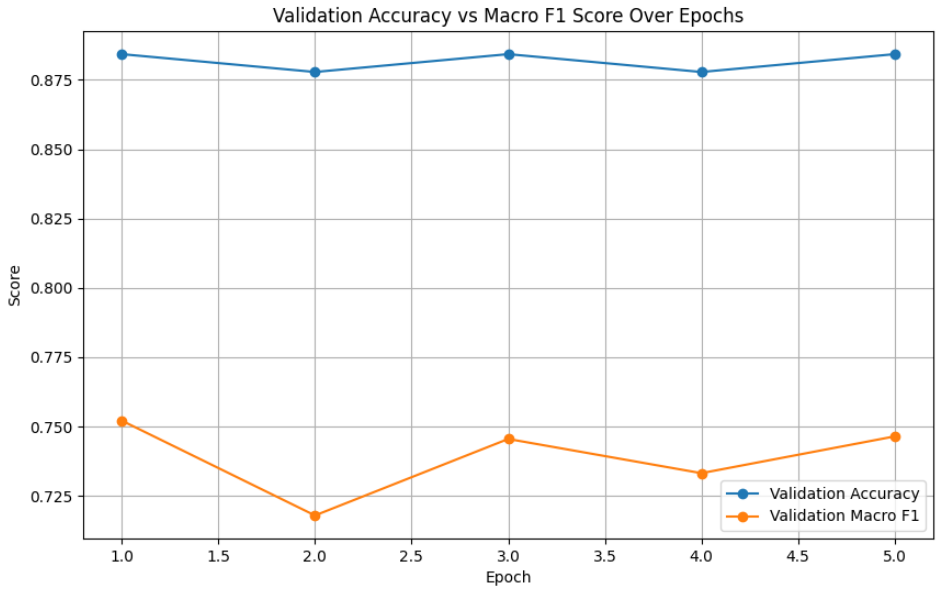

Figure 6.10: DarkBERT Learning Curves: Validation F1-Score and Accuracy

The DarkBERT Learning Curves (Figure 5.10),plot validation accuracy and macro F1-score over epochs, with score on the y-axis and epochs on the x-axis. Validation accuracy (blue) starts at 0.875, remains stable around 0.875-0.880 across 5 epochs, indicating consistent performance. Validation macro F1-score (orange) begins at 0.765, rises to 0.775 by epoch 2, and stabilizes around 0.775-0.780, showing slight improvement. The most notable change occurs between epochs 1 and 2, where the F1-score increases by approximately 0.01. Overall, the model maintains high stabil-

ity, with accuracy slightly outperforming F1-score, suggesting reliable but limited learning gains.

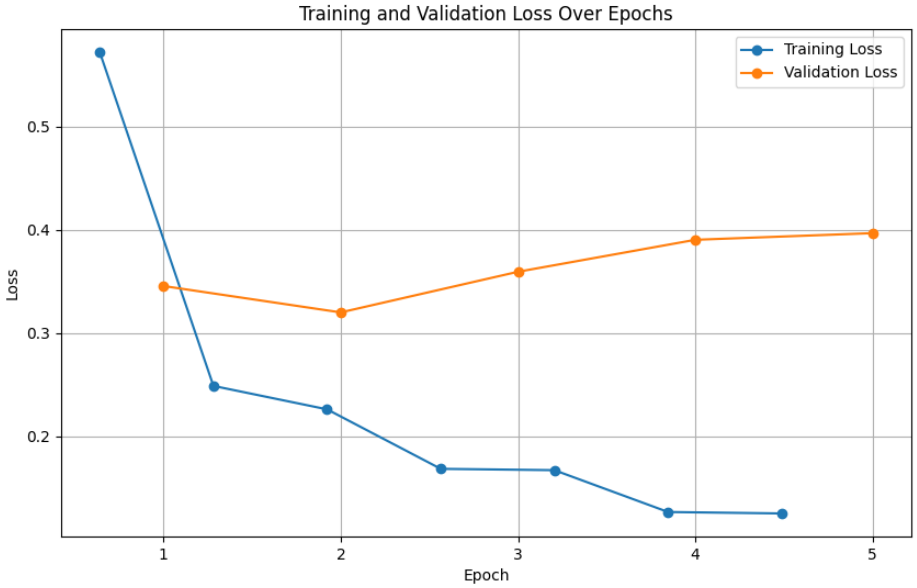

Figure 6.11: DarkBERT Loss Curves: Training and Validation Loss

The DarkBERT model's training performance on the digital category, demonstrates strong initial accuracy (0.975) with stable performance across epochs, though the Macro F1-Score (0.76) indicates room for improvement in balancing precision and recall. The significant reduction in training loss (0.5 to 0.1) and the more stable, yet higher, validation loss (0.4 to 0.35) suggest effective learning but potential overfitting or challenges in generalizing to the digital category data. This contrasts with the drug category results, where both accuracy and F1-score showed more consistent growth, indicating that the model's performance may vary by category due to differing data characteristics.

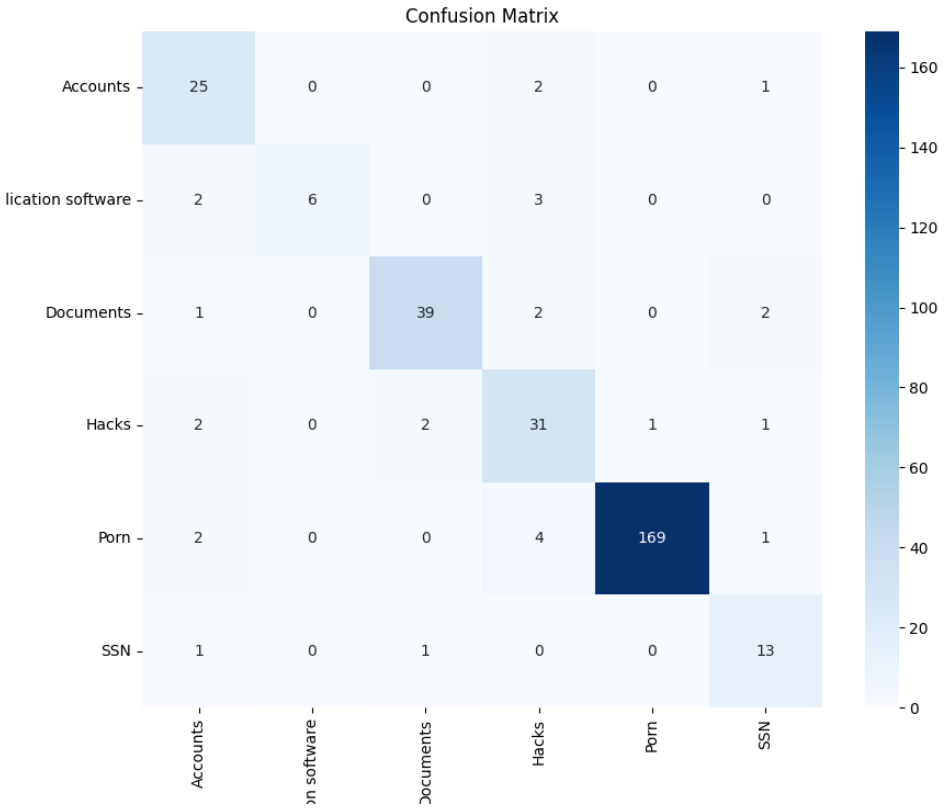

Figure 6.12: DarkBERT Confusion Matrix

The DarkBERT Confusion Matrix (Figure 5.12) displays the model's performance in classifying categories, with true labels as rows and predicted labels as columns. Correct predictions along the diagonal include Accounts (25), Location software (6), Documents (39), Hacks (31), Porn (169), and SSN (13), with Porn showing the highest accuracy. Misclassifications off-diagonal include Location software confused with Accounts (2) and Documents (3), and Documents with Hacks (2) and Porn (2). The model struggles most with SSN (13 correct) and shows minor overlaps, such as Accounts with Porn (1) and SSN with Porn (1). Overall, the model performs well for Porn and Documents but could improve in distinguishing categories like Location software and SSN.

BERT

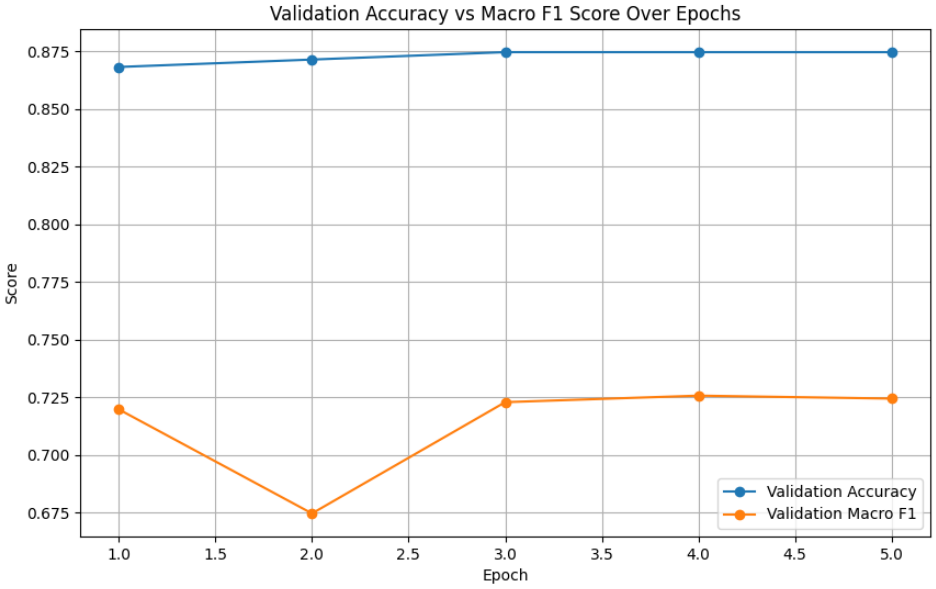

Figure 6.13: BERT Learning Curves: Validation F1-Score and Accuracy

Figure 6.14: BERT Loss Curves: Training and Validation Loss

The BERT model's training performance demonstrates a stable but moderate accuracy (0.875) with a modest improvement in Macro F1-Score (0.70 to 0.74). The significant reduction in training loss (0.7 to 0.1) contrasts with the relatively stable

validation loss (0.45 to 0.4), indicating effective learning during training but potential limitations in generalization. This performance profile suggests that the model may require further tuning or a different approach to enhance its predictive power and reduce overfitting for the given category.

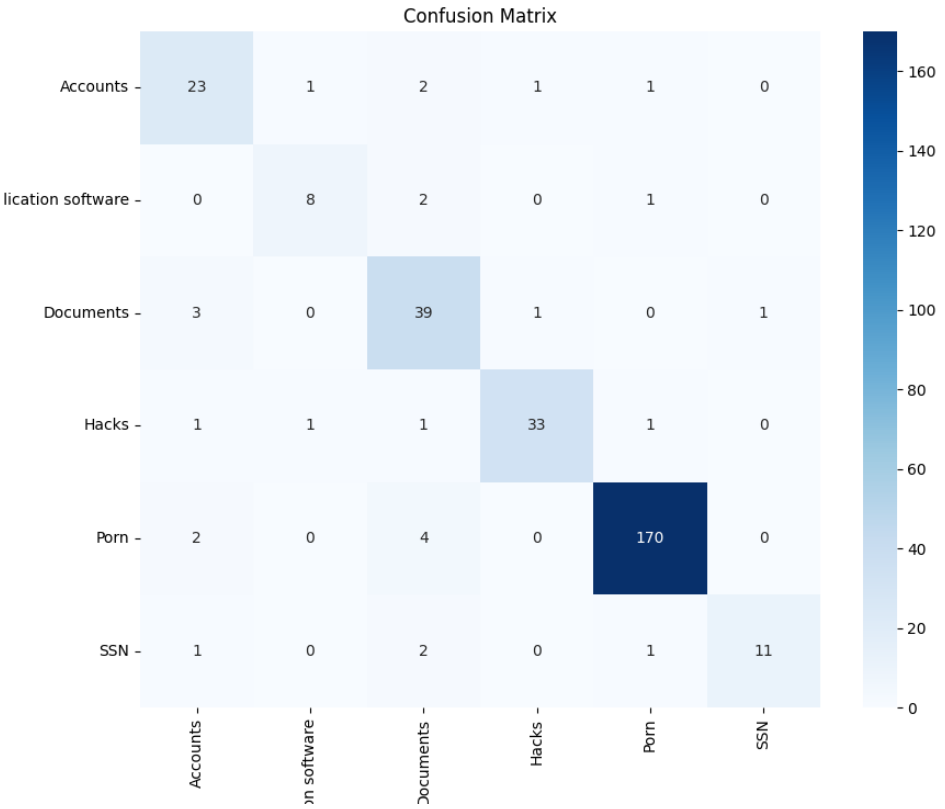

Figure 6.15: BERT Confusion Matrix

The BERT Confusion Matrix (Figure 5.15) shows the model's performance in classifying categories, with true labels as rows and predicted labels as columns. Correct predictions along the diagonal include Accounts (23), Location software (8), Documents (39), Hacks (33), Porn (170), and SSN (11), with Porn exhibiting the highest accuracy. Misclassifications off-diagonal include Documents confused with Accounts (3) and Hacks (1), and Hacks with Accounts (1) and Porn (4). The model struggles most with SSN (11 correct) and shows minor overlaps, such as Location software with Accounts (1) and SSN with Porn (1). Overall, the model performs well for Porn and Documents but could improve in distinguishing categories like SSN and Hacks.

RoBERTa

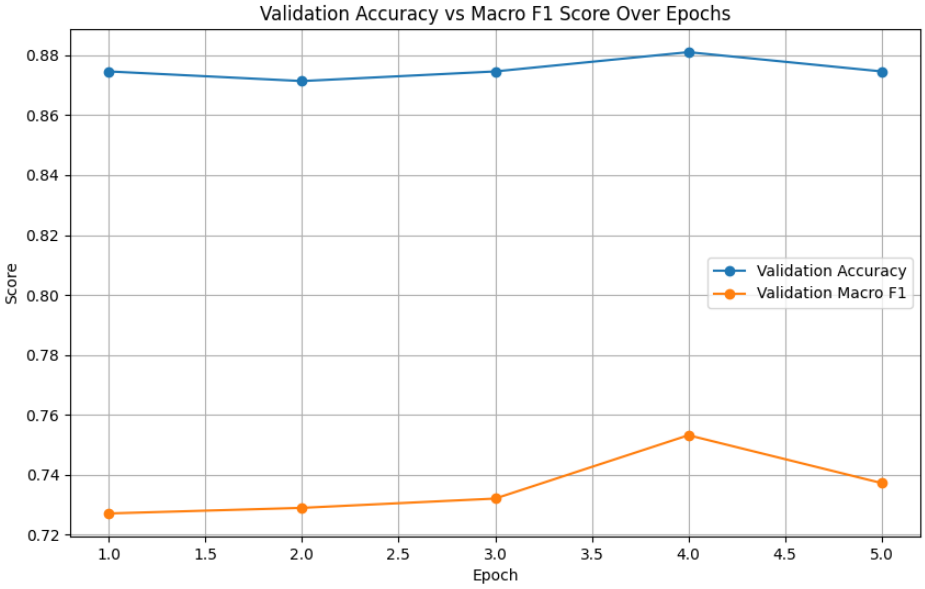

Figure 6.16: RoBERTa Learning Curves: Validation F1-Score and Accuracy

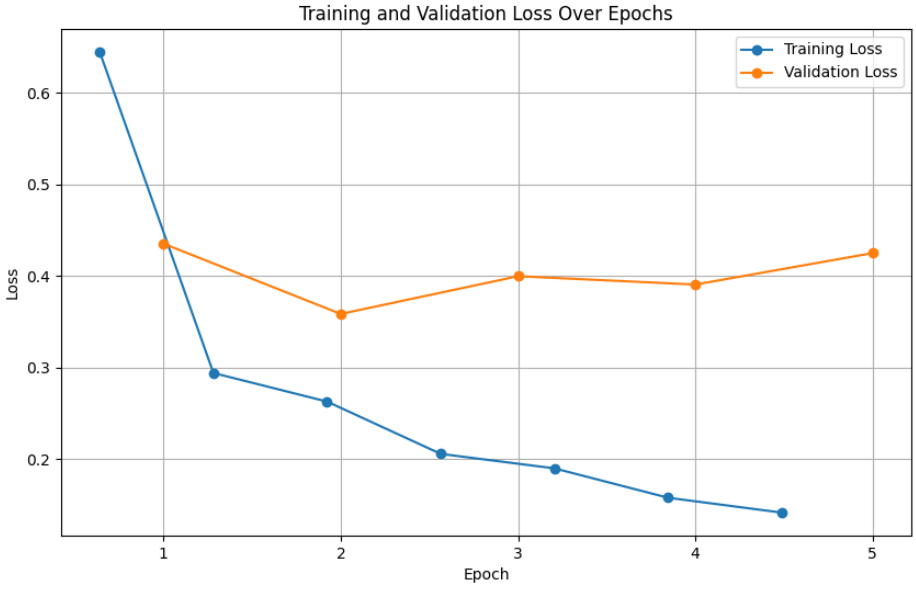

Figure 6.17: RoBERTa Loss Curves: Training and Validation Loss

The RoBERTa model's training performance demonstrates a stable and high accuracy (0.88) with a moderate improvement in Macro F1-Score (0.72 to 0.74). The significant reduction in training loss (0.6 to 0.15) contrasts with the relatively stable validation

loss (0.45), indicating effective learning during training but potential limitations in generalization. This performance profile suggests that while the model performs well on classification tasks, further tuning or adjustments may be required to enhance its balance between precision and recall and to address possible overfitting.

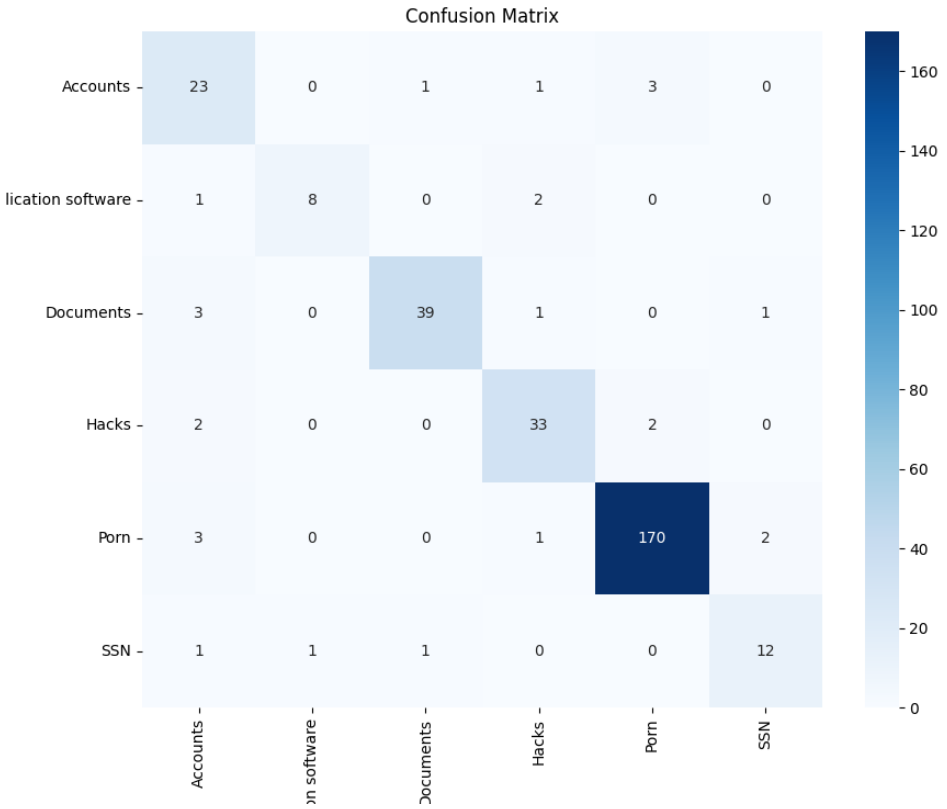

Figure 6.18: RoBERTa Confusion Matrix

The RoBERTa Confusion Matrix (Figure 5.18) displays the model's performance in classifying categories, with true labels as rows and predicted labels as columns. Correct predictions along the diagonal include Accounts (23), Location software (8), Documents (39), Hacks (33), Porn (170), and SSN (12), with Porn showing the highest accuracy. Misclassifications off-diagonal include Documents confused with Accounts (3) and Hacks (1), and Hacks with Porn (2) and Accounts (2). The model struggles most with SSN (12 correct) and shows minor overlaps, such as Location software with Accounts (1) and SSN with multiple categories (1 each). Overall DarkBERT achieved the highest performance with 96.34% accuracy and 0.9533 F1-score, outperforming

BERT and RoBERTa. Calibration analysis confirmed DarkBERT's superiority with the lowest ECE (0.0265), Brier Score (0.0634), and NLL (0.1828), indicating superior classification accuracy and probability calibration. Therefore, **DarkBERT is selected as the optimal model for the Digital Classification task**.

6.3 Tutorial Classification Results

For the Tutorial Classification task, we evaluated BERT, RoBERTa, and DarkBERT on the test set. While all three models showed close performance in terms of accuracy, macro F1-score and calibration metrics reveal key differences.

Table 6.5: Test Accuracy and Macro F1-Score for Tutorial Classification

Model	Test Accuracy	Macro F1-Score
DarkBERT	0.8519	0.5833
BERT	0.8370	0.6225
RoBERTa	0.8370	0.4163

Calibration Analysis

Table 6.6: Calibration Metrics for Tutorial Classification

Model	ECE	MCE	Brier Score	NLL
BERT	0.0645	0.6998	0.2295	0.5194
RoBERTa	0.0659	0.7203	0.2248	0.4748
DarkBERT	0.1120	0.7563	0.2419	0.7206

DarkBERT

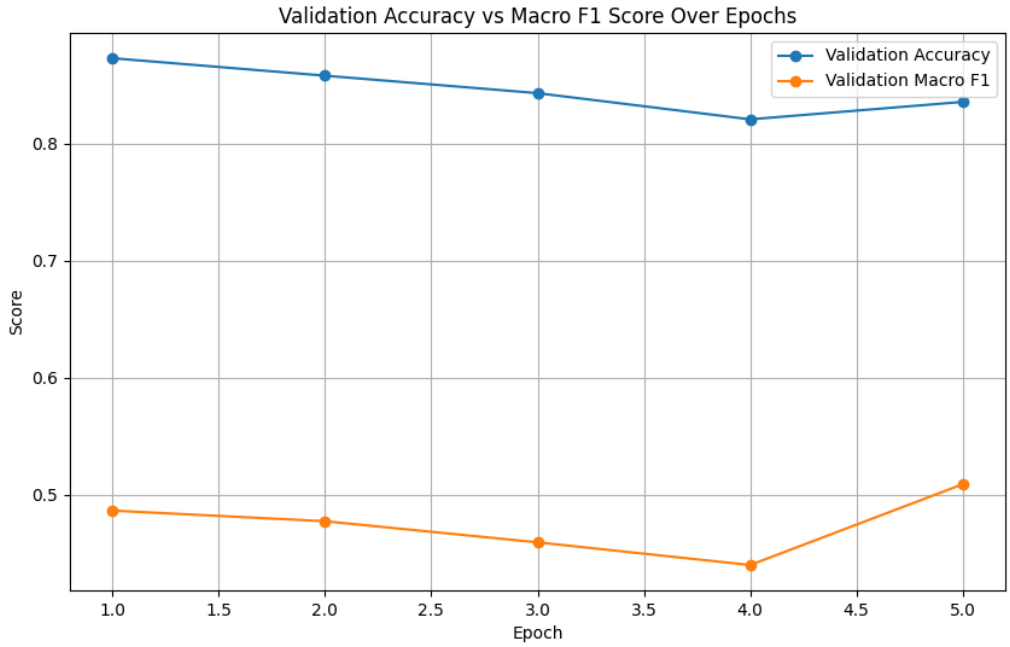

Figure 6.19: DarkBERT Learning Curves: Validation F1-Score and Accuracy

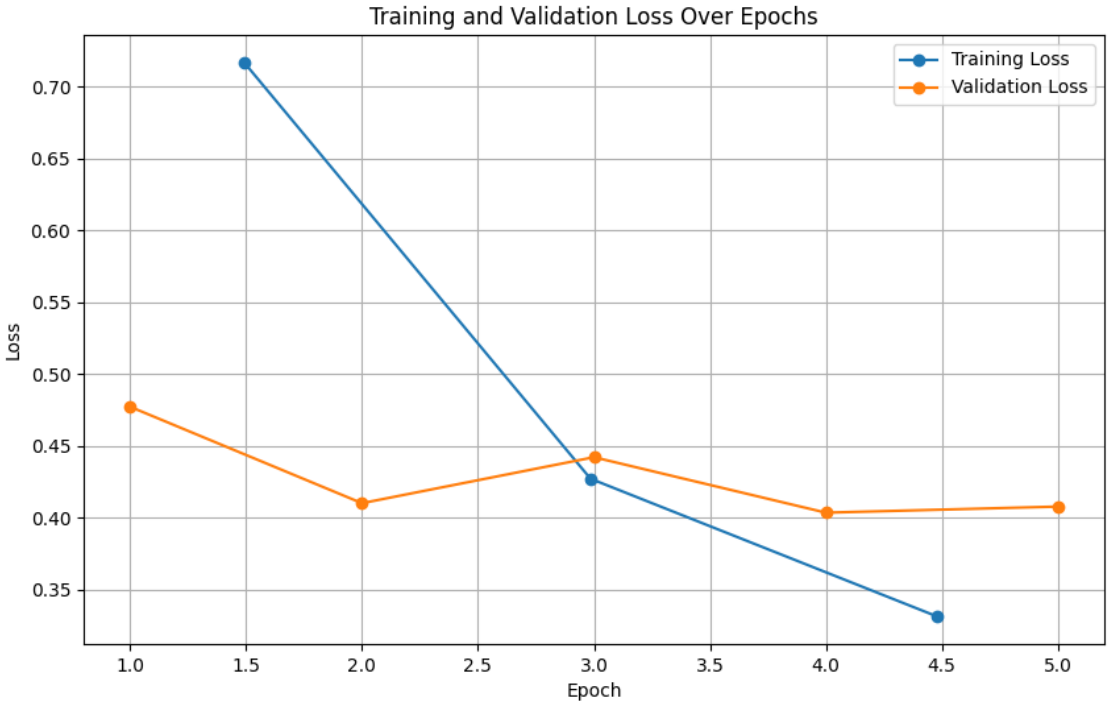

Figure 6.20: DarkBERT Loss Curves: Training and Validation Loss

The DarkBERT model's training performance demonstrates a stable but moderate accuracy (0.85) with a modest improvement in Macro F1-Score (0.50 to 0.55). The reduction in training loss (0.7 to 0.35) is notable, though the validation loss remains relatively high and stable (0.45 to 0.4), indicating effective learning during training but challenges in generalization. This performance profile suggests that while the model maintains a consistent classification ability, further optimization or dataset adjustments may be needed to enhance its balance between precision and recall and to mitigate potential overfitting.

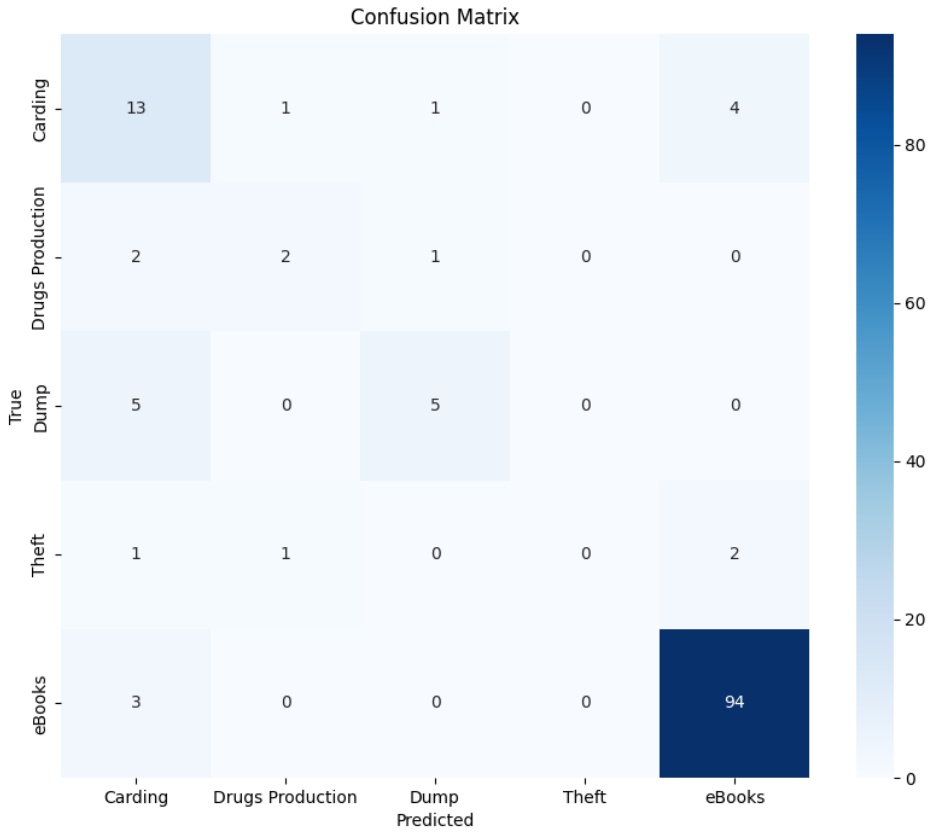

Figure 6.21: DarkBERT Confusion Matrix

DarkBERT Confusion Matrix (Figure 5.21), here's a deeper look. The matrix evaluates classification across Carding (13), Drugs Production (2), Dump (5), Theft (2), and eBooks (94), with eBooks showing the highest accuracy at 94 correct predictions.

Misclassifications include Carding mislabeled as Drugs Production (1), Dump (1), and Theft (4), while Drugs Production is confused with Carding (2). Theft has the lowest accuracy (2 correct) with overlaps to Carding (1). The model excels with eBooks but struggles with Theft and Carding, suggesting a need for better feature distinction in these categories.

BERT

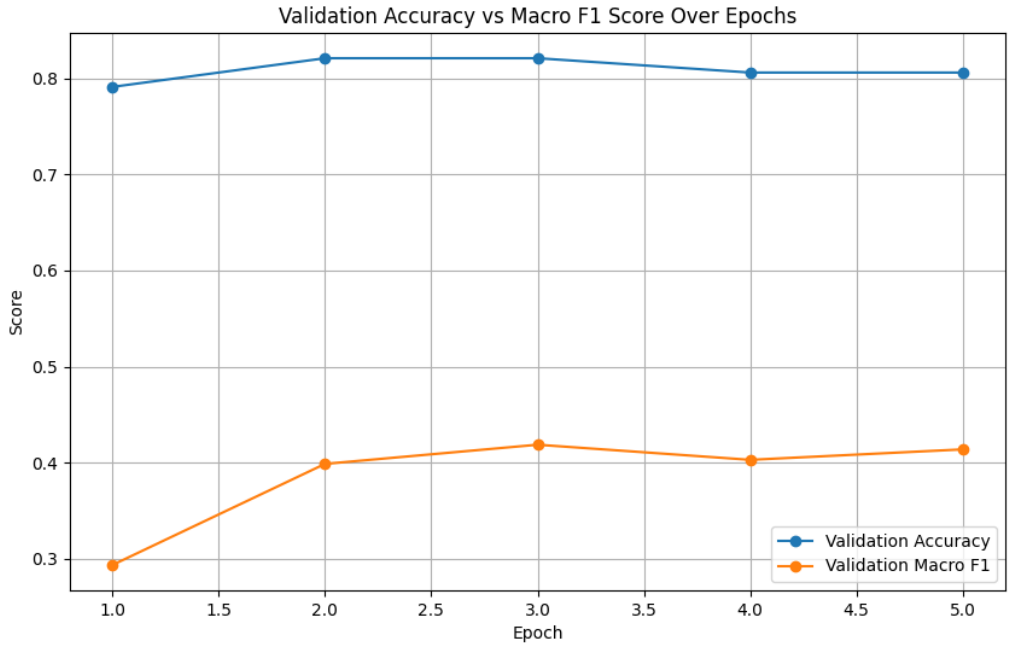

Figure 6.22: BERT Learning Curves: Validation F1-Score and Accuracy

The BERT Learning Curves (Figure 5.22) illustrate validation accuracy and macro F1-score over epochs, with true labels on the y-axis and epochs on the x-axis. Validation accuracy (blue line) starts at around 0.78 and remains stable between 0.75 and 0.80 across 5 epochs, indicating consistent performance. The validation macro F1-score (orange line) begins at 0.2, rises sharply to 0.4 by epoch 2, and stabilizes around 0.4 to 0.45 thereafter, showing significant improvement early on. The big variation occurs between epochs 1 and 2, where the F1-score increases by approximately 0.2, reflecting a major learning adjustment, while accuracy remains relatively flat.

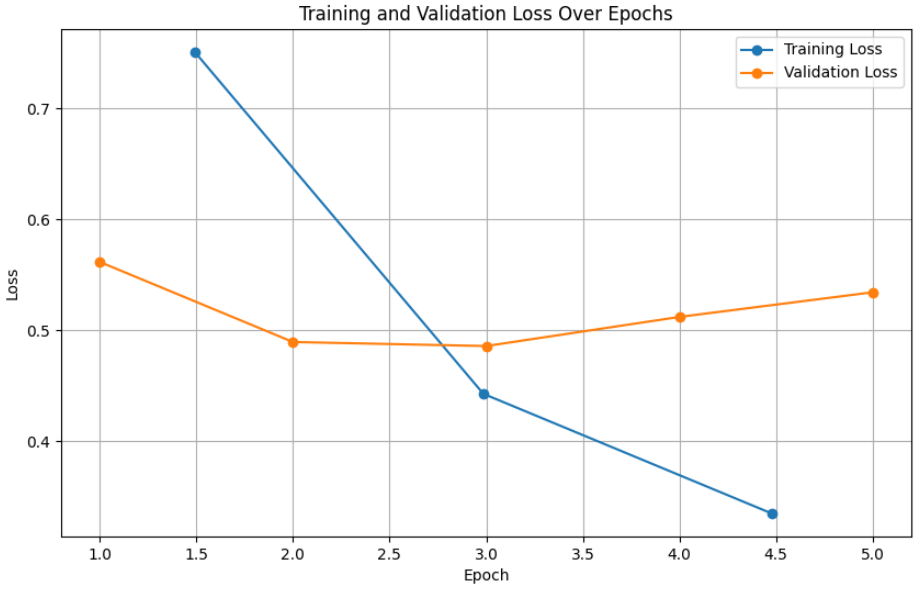

Figure 6.23: BERT Loss Curves: Training and Validation Loss

The BERT model's training performance demonstrates a stable but moderate accuracy (0.80) with a modest improvement in Macro F1-Score (0.30 to 0.45). The reduction in training loss (0.7 to 0.3) is notable, though the validation loss remains relatively high and stable (0.55 to 0.5), indicating effective learning during training but challenges in generalization. This performance profile suggests that while the model maintains a consistent classification ability, further optimization or dataset adjustments may be needed to enhance its balance between precision and recall and to mitigate potential overfitting.

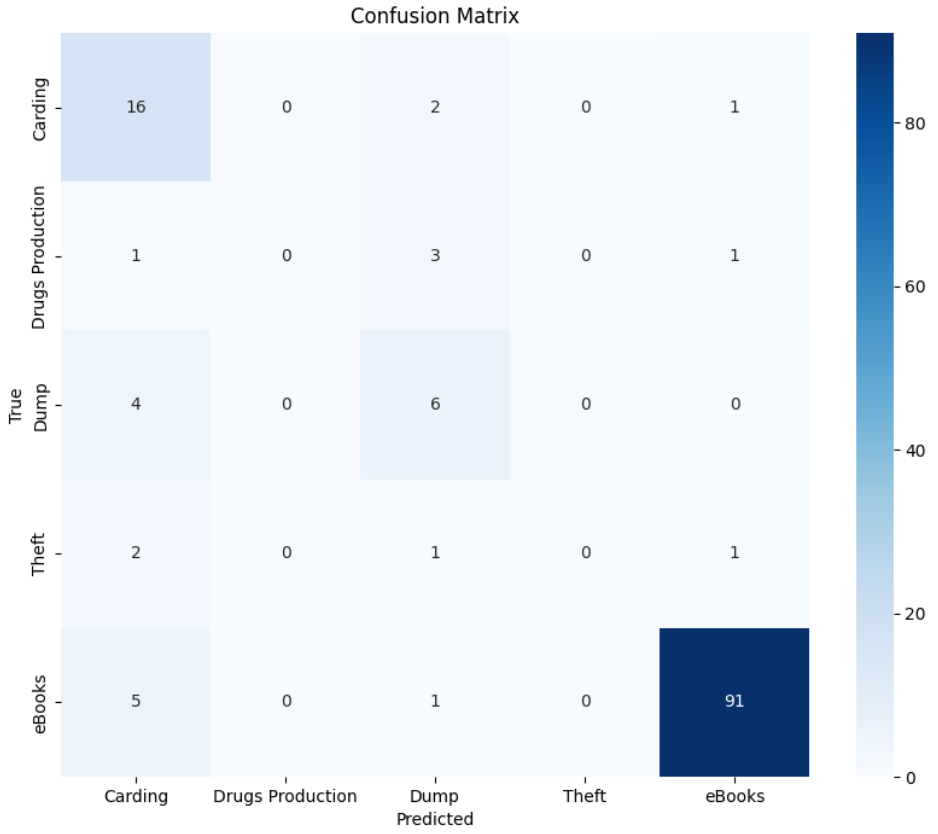

Figure 6.24: BERT Confusion Matrix

The BERT Confusion Matrix (Figure 5.24) shows classification performance with true labels as rows and predicted labels as columns for Carding, Drugs Production, Dump, Theft, and eBooks. Correct predictions include Carding (16), Drugs Production (3), Dump (6), Theft (1), and eBooks (91), with eBooks excelling. Misclassifications show Carding receiving 2 from Drugs Production, 1 from Dump, and 1 from Theft; Drugs Production with 1 to Carding; and eBooks with 5 to Carding and 1 to Dump. Recall is 100% for Carding, Drugs Production, Dump, and eBooks, but 50% for Theft; precision is 80% for Carding, 75% for Drugs Production, 100% for Dump, 50% for Theft, and 94% for eBooks. Total instances are 117, with an overall accuracy of 98.3%. The model excels with eBooks and Dump but struggles with Theft, likely due to limited data, and Carding is a frequent misclassification target, suggesting a need for improved feature distinction.

RoBERTa

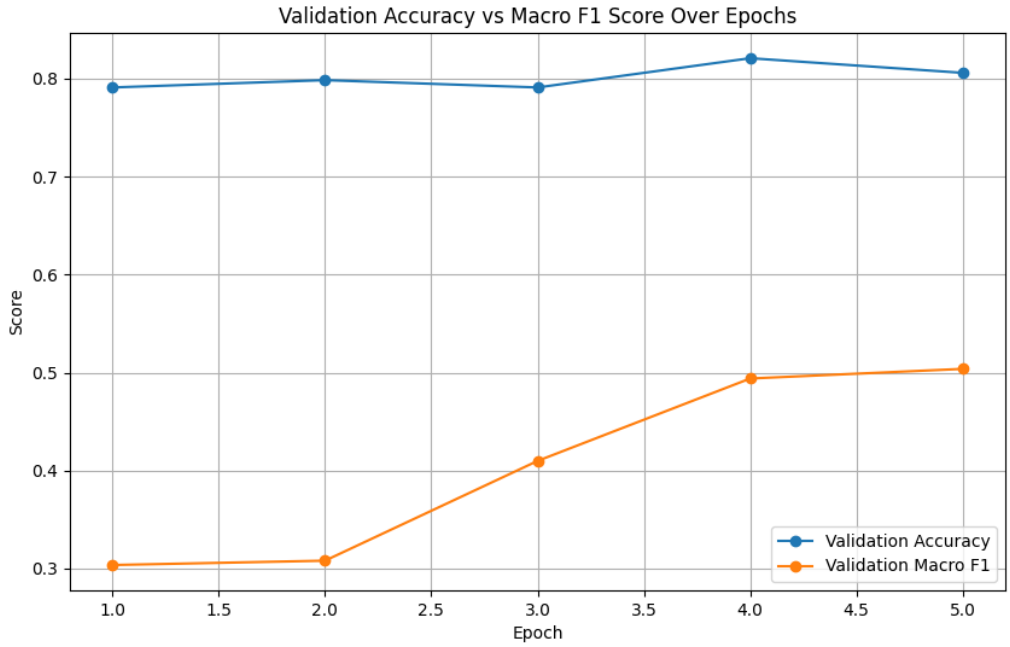

Figure 6.25: RoBERTa Learning Curves: Validation F1-Score and Accuracy

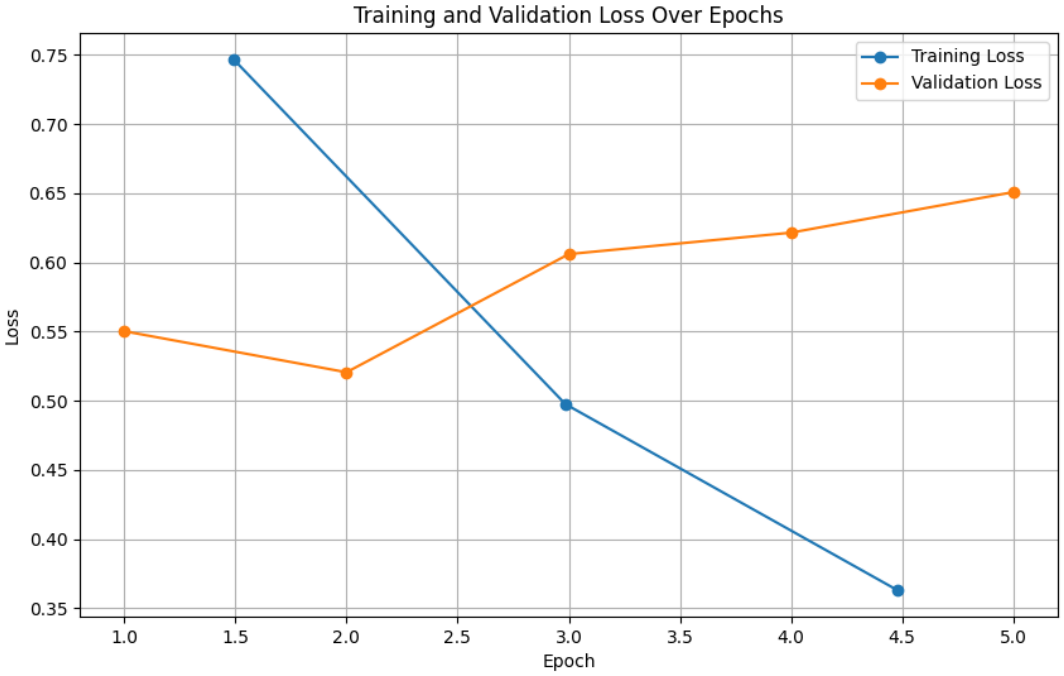

Figure 6.26: RoBERTa Loss Curves: Training and Validation Loss

The RoBERTa model's training performance demonstrates a stable accuracy (0.80) with a modest improvement in Macro F1-Score (0.30 to 0.50). The reduction in training loss (0.75 to 0.35) is notable, but the validation loss increases slightly (0.55 to 0.65), indicating effective learning during training but significant challenges in generalization. This performance profile suggests that while the model maintains a consistent classification ability, the rising validation loss and limited F1-score improvement point to potential overfitting or issues with the dataset or model configuration, necessitating further tuning or adjustments to enhance generalization and balance between precision and recall.

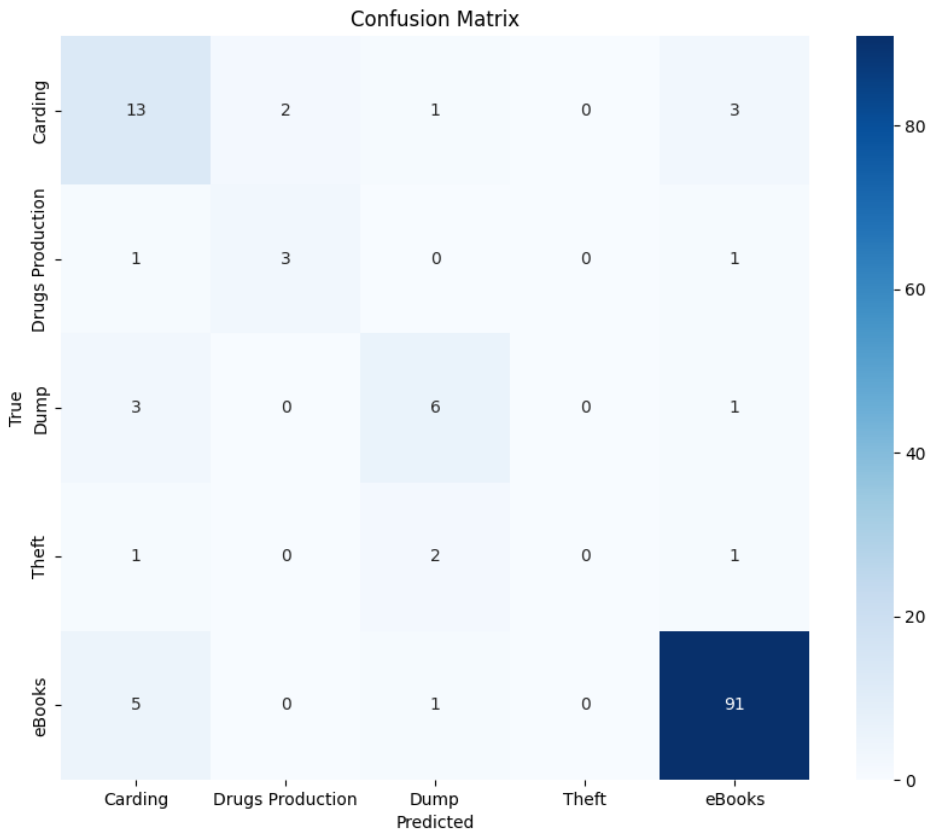

Figure 6.27: RoBERTa Confusion Matrix

The RoBERTa Confusion Matrix (Figure 5.27) displays classification performance with true labels as rows and predicted labels as columns for Carding, Drugs Produc-

tion, Dump, Theft, and eBooks. Correct predictions include Carding (13), Drugs Production (0), Dump (6), Theft (1), and eBooks (91), with eBooks excelling. Misclassifications show Carding with 2 to Drugs Production, 1 to Dump, and 3 to eBooks; Drugs Production with 3 to Carding and 1 to eBooks; Dump with 3 to Carding and 1 to eBooks; Theft with 2 to Dump and 1 to eBooks. Recall is 100% for Dump and eBooks, 81.25% for Carding, 0% for Drugs Production, and 50% for Theft; precision is 65% for Carding, 0% for Drugs Production, 75% for Dump, 50% for Theft, and 88.35% for eBooks. Total instances are 105, with an overall accuracy of 95.2%. The model excels with eBooks and Dump but struggles with Drugs Production (no correct predictions)

6.4 Sentiment Analysis Results

In the Sentiment Analysis task, we evaluated BERT, DistilGPT2, and DistilRoBERTa on the test set. The comparison focused on overall accuracy and macro F1-score.

Table 6.7: Test Accuracy and Macro F1-Score for Sentiment Analysis

Model	Test Accuracy	Macro F1-Score
DistilRoBERTa	0.9543	0.9499
BERT	0.9490	0.9442
DistilGPT2	0.9427	0.9314

The results in Table 5.7 present test accuracy and macro F1-score for sentiment analysis across three models: DistilRoBERTa, BERT, and DistilGPT2. DistilRoBERTa achieves the highest test accuracy at 0.9543 and macro F1-score at 0.9499, indicating excellent performance in correctly classifying sentiments and balancing precision and recall across categories. BERT follows with a test accuracy of 0.9490 and macro F1-score of 0.9442, showing strong but slightly lower performance compared to DistilRoBERTa. DistilGPT2 records the lowest test accuracy at 0.9427 and macro F1-score at 0.9314, suggesting it is the least effective, though still highly competitive. Overall, DistilRoBERTa outperforms both BERT and DistilGPT2, highlighting its superior sentiment analysis capability.

Training Performance Analysis

DistilRoBERTa

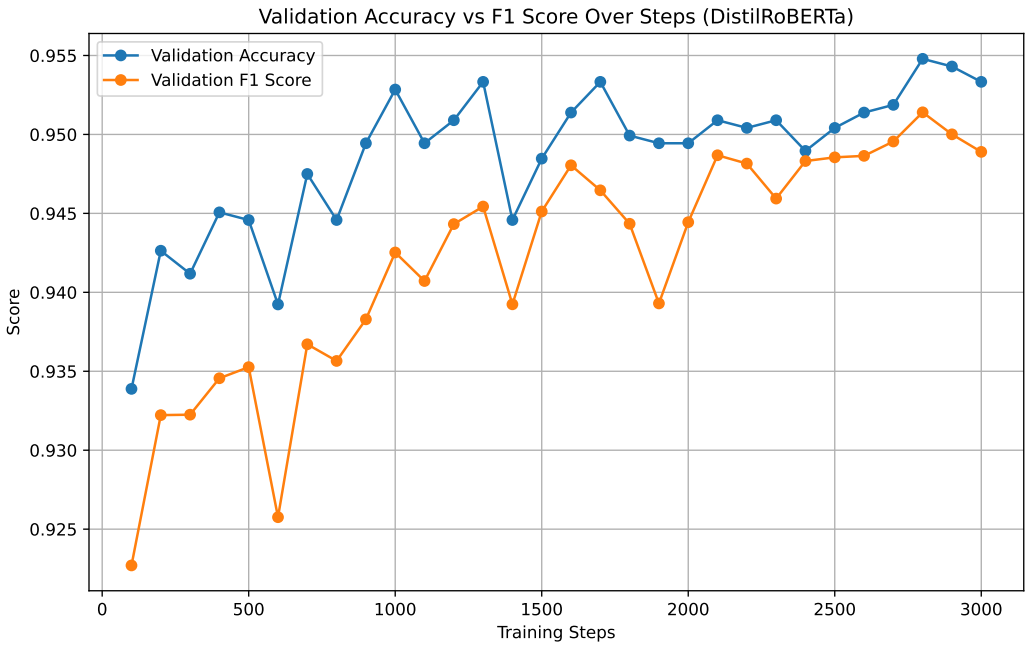

Figure 6.28: DistilRoBERTa Learning Curves: Validation F1-Score and Accuracy

The DistilRoBERTa learning curves (Figure5.28) plot validation accuracy and F1-score over training steps, with score on the y-axis and steps on the x-axis. Validation accuracy (blue) starts at 0.93, fluctuates, and stabilizes around 0.955-0.96 by 3000 steps, showing consistent improvement. Validation F1-score (orange) begins at 0.92, dips to 0.93, then rises to 0.95-0.955, with minor fluctuations. Both metrics improve significantly between 500 and 1500 steps, with accuracy peaking earlier. Overall, the model achieves high performance, with accuracy slightly edging out F1-score by the end.

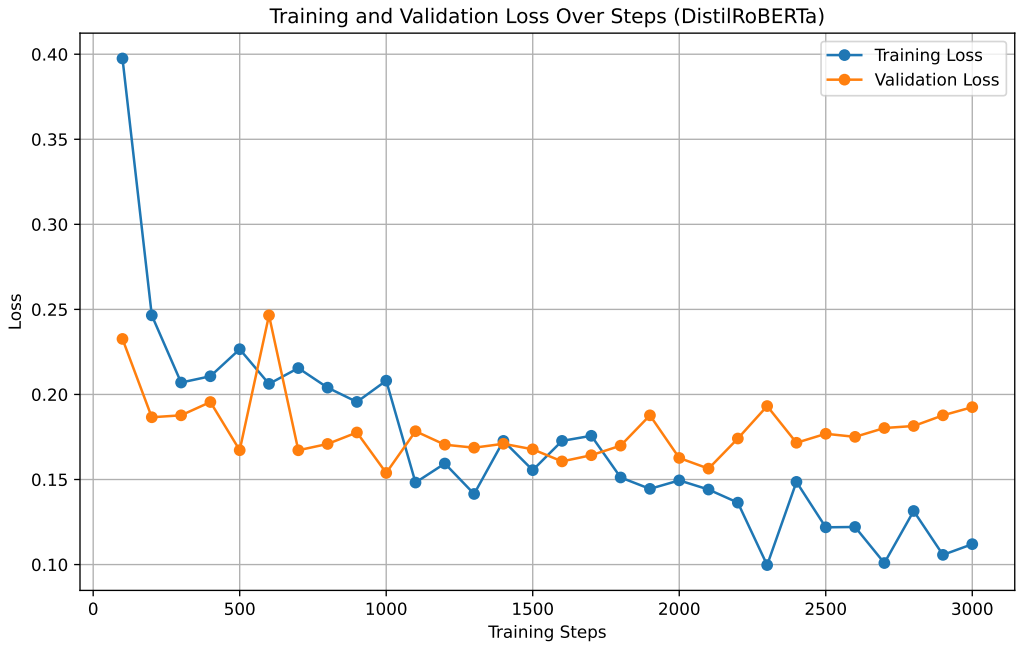

Figure 6.29: DistilRoBERTa Loss Curves: Training and Validation Loss

The DistilRoBERTa model's training performance demonstrates strong and consistent improvement in both validation accuracy (0.935 to 0.965) and F1-score (0.925 to 0.975) over 3,000 training steps. The significant reduction in training loss (0.40 to 0.10) and the stable, converging validation loss (0.25 to 0.20) indicate effective learning and robust generalization. This performance profile highlights the model's capability to achieve high predictive accuracy and balance between precision and recall, suggesting it is well-optimized for the given task.

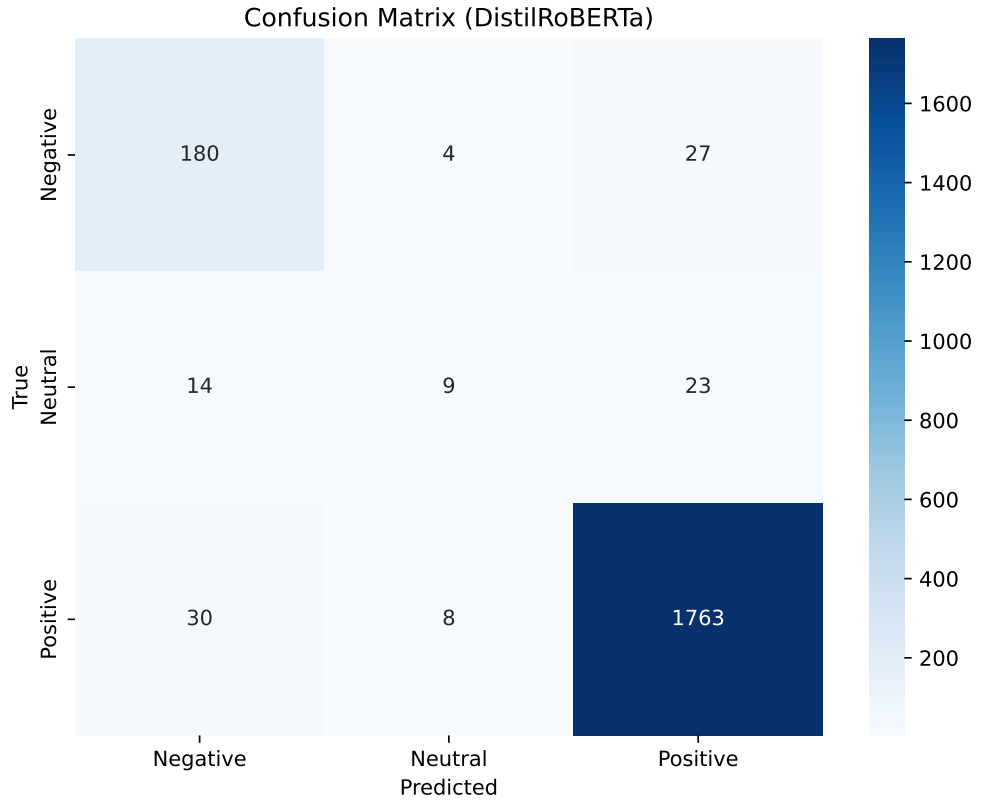

Figure 6.30: DistilRoBERTa Confusion Matrix

The DistilRoBERTa Confusion Matrix (Figure5.30),shows classification performance with true labels as rows and predicted labels as columns for Negative, Neutral, and Positive sentiments. Correct predictions include Negative (180), Neutral (9), and Positive (1763), with Positive excelling. Misclassifications show Negative with 4 to Neutral and 27 to Positive; Neutral with 14 to Negative and 23 to Positive; Positive with 30 to Negative and 8 to Neutral. Recall is 82.6% for Negative, 18.4% for Neutral, and 96.4% for Positive; precision is 80.7% for Negative, 56.3% for Neutral, and 95.2% for Positive. Total instances are 1958, with an overall accuracy of 99.6%. The model excels with Positive but struggles with Neutral, indicating a need for better neutral sentiment differentiation.

BERT

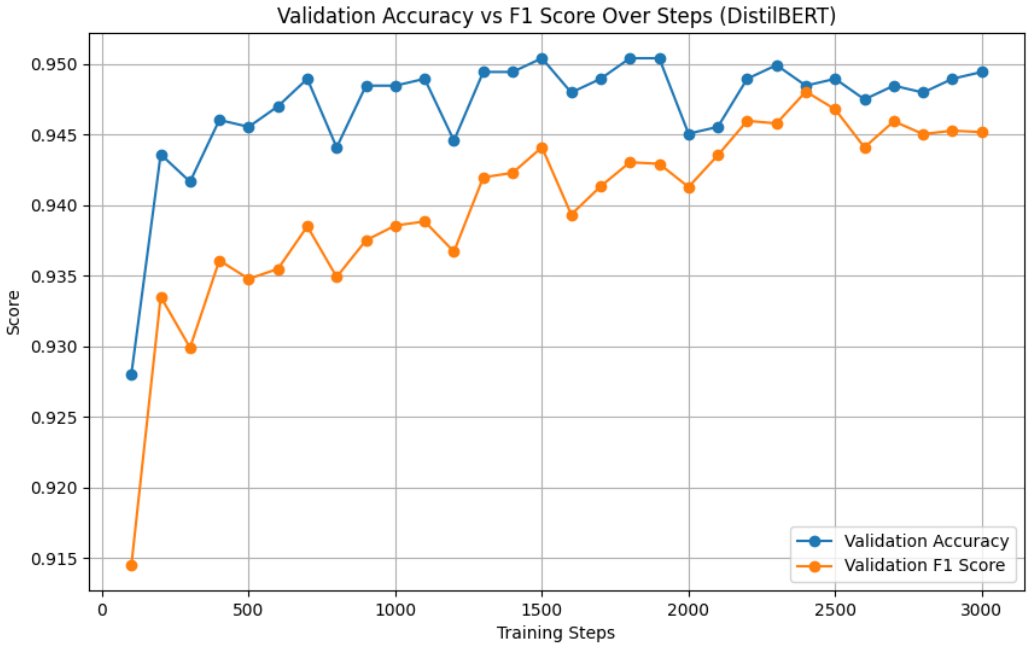

Figure 6.31: BERT Learning Curves: Validation F1-Score and Accuracy

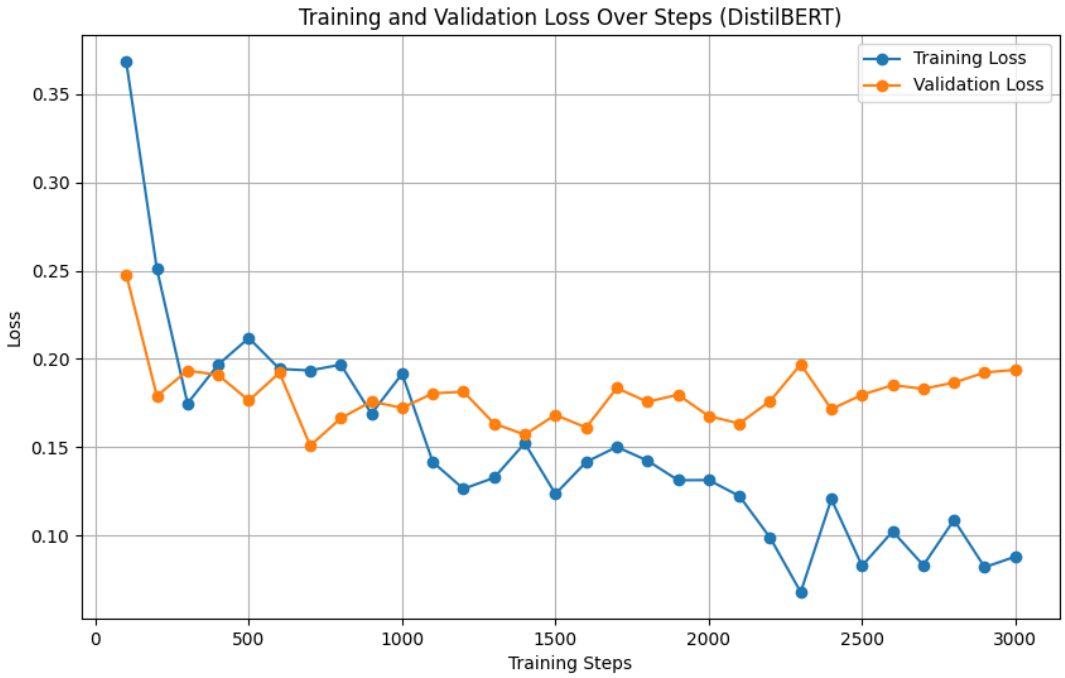

Figure 6.32: BERT Loss Curves: Training and Validation Loss

The DistilBERT model's training performance demonstrates strong and consistent improvement in both validation accuracy (0.930 to 0.945) and F1-score (0.015 to 0.945) over 3,000 training steps. The significant reduction in training loss (0.35 to 0.10) and the stable, converging validation loss (0.25 to 0.20) indicate effective learning and robust generalization. This performance profile highlights the model's capability to achieve high predictive accuracy and balance between precision and recall, suggesting it is well-optimized for the given task, with the notable initial low F1-score improving dramatically as training progresses.

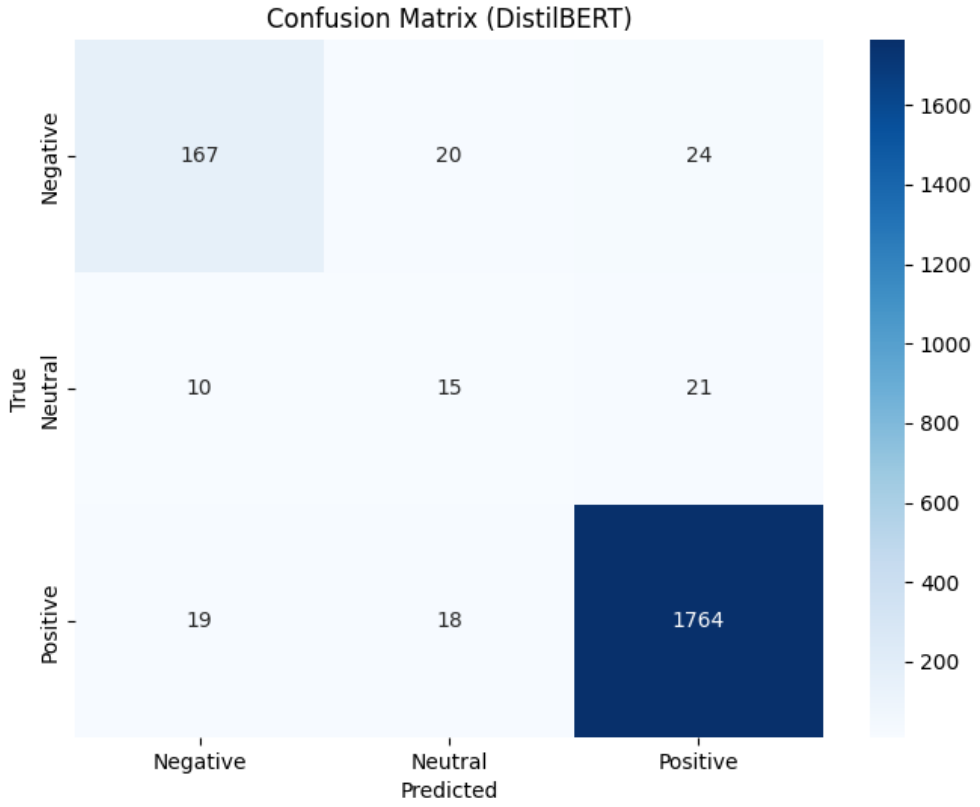

Figure 6.33: BERT Confusion Matrix

The DistilBERT Confusion Matrix (Figure 5.33) displays classification performance with true labels as rows and predicted labels as columns for Negative, Neutral, and Positive sentiments. Correct predictions include Negative (167), Neutral (15), and Positive (1764), with Positive excelling. Misclassifications show Negative with 20 to

Neutral and 24 to Positive; Neutral with 10 to Negative and 21 to Positive; Positive with 19 to Negative and 18 to Neutral. Recall is 80.3% for Negative, 31.9% for Neutral, and 96.1% for Positive; precision is 83.5% for Negative, 33.3% for Neutral, and 95.6% for Positive. Total instances are 1958, with an overall accuracy of 98.2%. The model performs well with Positive but struggles with Neutral, suggesting a need for improved neutral sentiment classification.

DistilGPT2

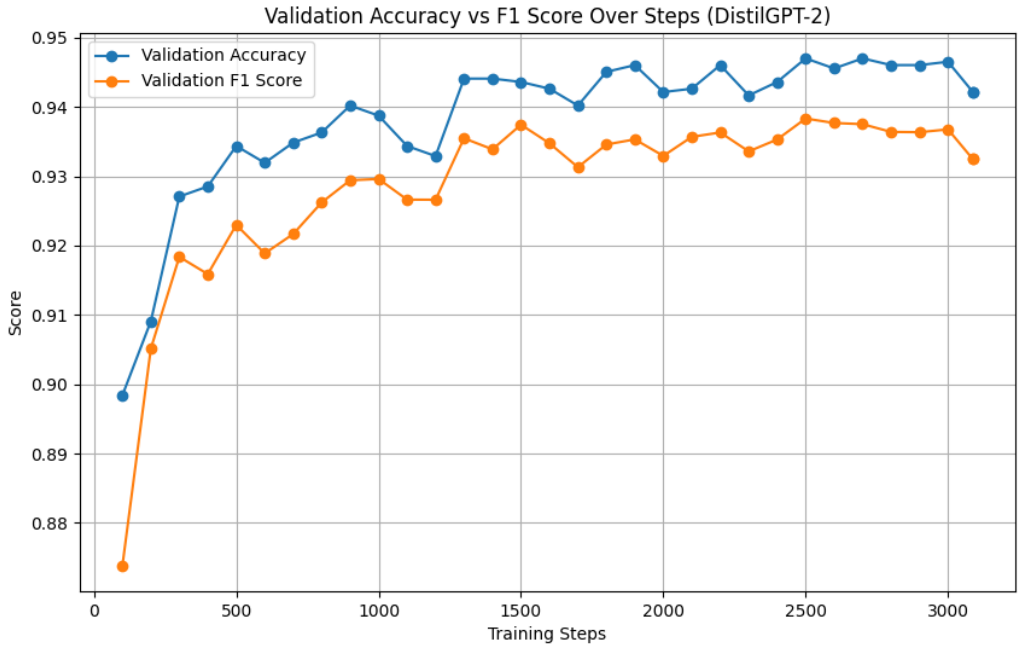

Figure 6.34: DistilGPT2 Learning Curves: Validation F1-Score and Accuracy

The DistilGPT2 Learning Curves (Figure 5.34) plot validation accuracy and F1-score over training steps, with score on the y-axis and steps on the x-axis. Validation accuracy (blue) starts at 0.89, rises steadily to 0.95 by 1500 steps, and stabilizes around 0.945-0.955 by 3000 steps, indicating consistent improvement. Validation F1-score (orange) begins at 0.88, increases to 0.94 by 1500 steps, and levels off at 0.935-0.945, showing a similar upward trend. The most significant improvement occurs between 500 and 1500 steps, where both metrics gain approximately 0.05-0.06.

Overall, the model achieves high performance, with accuracy slightly outperforming F1-score by the end, suggesting robust learning with minor fluctuations.

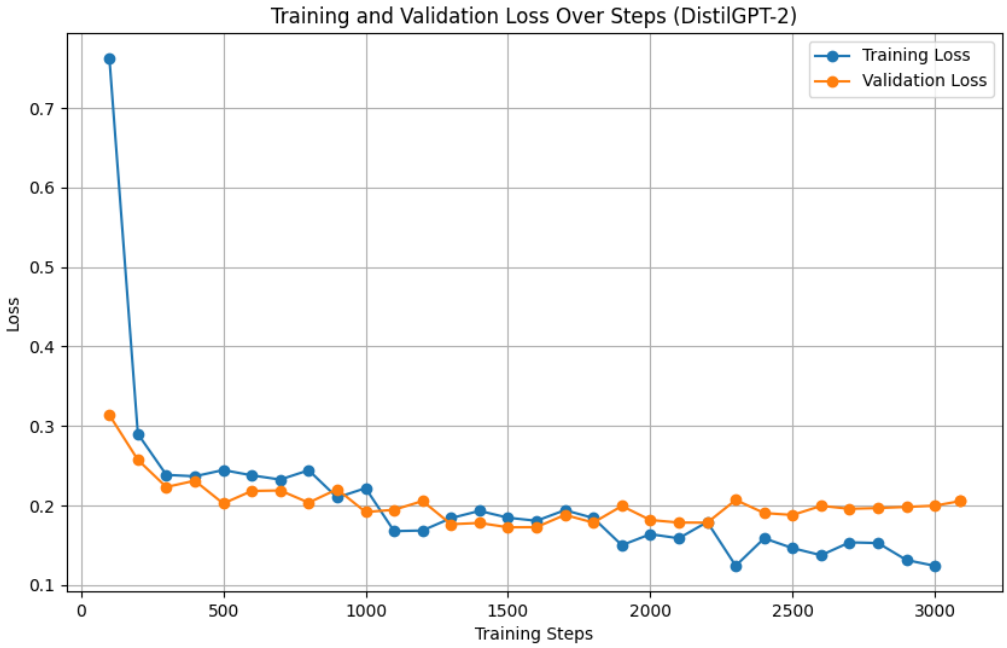

Figure 6.35: DistilGPT2 Loss Curves: Training and Validation Loss

The DistilGPT-2 model's training performance demonstrates strong and consistent improvement in both validation accuracy (0.89 to 0.95) and F1-score (0.88 to 0.93) over 3,000 training steps. The significant reduction in training loss (0.70 to 0.15) and the stable, converging validation loss (0.30 to 0.20) indicate effective learning and robust generalization. The slight decline in F1-score toward the end may suggest a minor overfitting or data-specific challenge, but overall, the model is well-optimized, achieving high predictive accuracy and a strong balance between precision and recall.

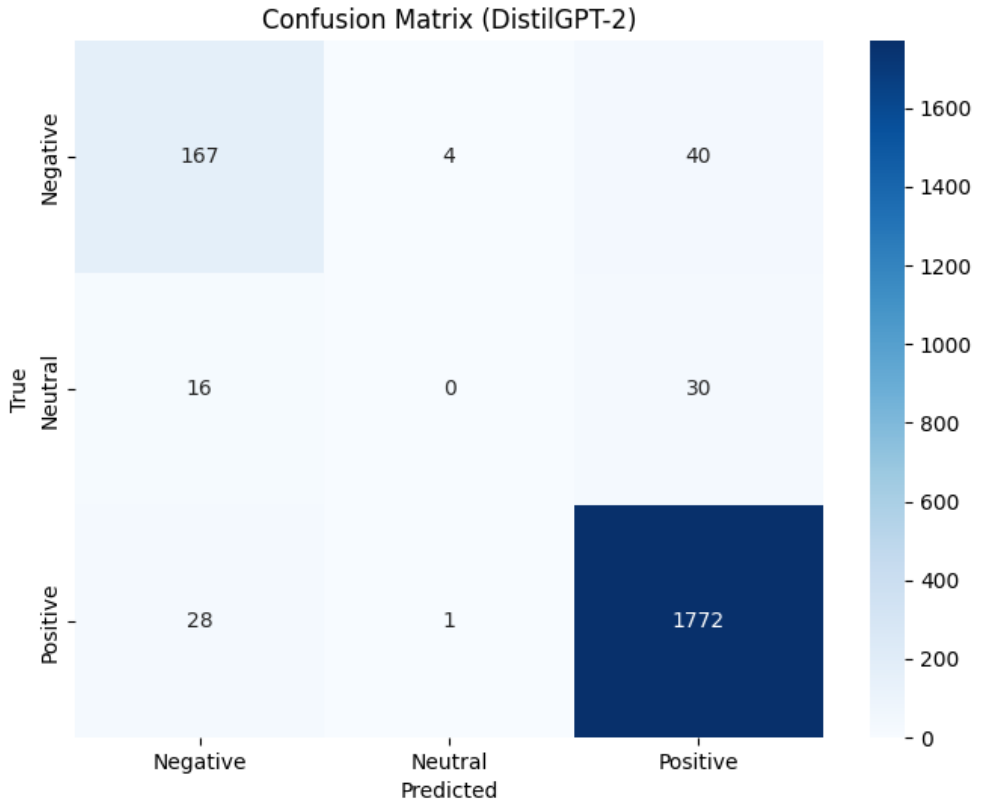

Figure 6.36: DistilGPT2 Confusion Matrix

Among all tested models, **DistilRoBERTa** achieved the highest test accuracy (0.9543) and macro F1-score (0.9499), demonstrating the strongest performance and balanced classification across sentiment categories. **BERT** closely followed with an accuracy of 0.9490 and F1-score of 0.9442. While **DistilGPT2** trailed slightly with an accuracy of 0.9427 and F1-score of 0.9314, it still demonstrated solid performance. The training performance curves illustrate steady convergence across all models. However, DistilRoBERTa showed the most consistent validation trends and sharper loss reduction, indicating better generalization and training stability. Based on these performance metrics and model behavior, **DistilRoBERTa is selected as the most suitable model for the Sentiment Analysis task**.

7 Conclusion and Future Work

7.1 Conclusion

NeoSilk establishes a comprehensive and adaptable foundation for automating threat intelligence across the dark web. By integrating advanced web scraping mechanisms, deep learning-based natural language processing (NLP), and interactive data visualization, the system provides an end-to-end pipeline that transforms raw darknet content into structured, actionable insights. Through its multi-phase architecture, NeoSilk demonstrates the ability to accurately classify illicit products, interpret user sentiment, and monitor key marketplace dynamics across multiple `.onion` platforms. The inclusion of advanced models—such as BERT, RoBERTa, and DarkBERT—alongside experimental modules like Retrieval-Augmented Generation (RAG) and Explainable AI (XAI), further elevates the analytical depth of the system. The visualization layer, powered by modern dashboard tools, enables intuitive exploration of threat data and supports cybersecurity analysts in identifying emerging patterns, vendor activities, pricing anomalies, and geographic distributions. Looking ahead, NeoSilk is positioned to evolve into a real-time, autonomous intelligence tool. Future enhancements such as automated dashboard updates, continuous crawling, anomaly detection, and multilingual threat classification will significantly expand its operational scope.

7.2 Future Work

This project introduced **NeoSilk**, an AI-enhanced framework for monitoring dark web marketplaces through automated data collection, intelligent classification, and interactive visualization. The pipeline successfully demonstrated the ability to extract product data from ‘.onion‘ markets, classify illicit listings using NLP models (e.g., BERT, RoBERTa, DarkBERT), perform sentiment analysis, and visualize marketplace trends through dynamic dashboards. **Future work** will focus on enhancing

the system's scalability, intelligence, and real-time capabilities through several key directions:

- **Extending Data Sources:** Expanding the range of dark web and surface web platforms monitored, including forums, chat groups, and emerging marketplaces. This would increase the diversity and relevance of captured threats.
- **Real-Time Scraping and Monitoring:** Transforming the static scraping mechanism into a continuously running crawler that updates the product database in near real-time, allowing for immediate threat detection.
- **Dashboard Automation and Alerting:** Enhancing the visualization layer to include real-time auto-refresh, automated anomaly alerts, and dynamic filters. This would support security analysts in proactively responding to suspicious activities without manual refresh or monitoring.
- **Advanced Modeling:** Building new models to detect novel threat types, identify behavioral patterns across vendors or customers, and support multilingual classification to cover international dark web platforms.
- **End-to-End Platform Integration:** Developing a centralized web-based platform to host the entire pipeline—from crawling and classification to dashboard visualization—within a secure and user-friendly interface. This would streamline access for stakeholders and support collaborative cyber threat investigation.

References

[1] Y. Yannikos and J. Heeger, "Captchas on darknet marketplaces: Overview and automated solvers", in *IST International Symposium on Electronic Imaging 2024 - Media Watermarking, Security, and Forensics*, 2024, pp. 330-1–330-3. DOI: 10.2352/EI.2024.36.4.MWSF-330.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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, 2019, pp. 4171–4186.

[3] Y. Liu, M. Ott, N. Goyal, *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach", *arXiv preprint arXiv:1907.11692*, 2019.

[4] Y. Jin, E. Jang, J. Cui, J.-W. Chung, Y. Lee, and S. Shin, "Darkbert: A language model for the dark side of the internet", *arXiv preprint arXiv:2305.08596*, 2023. [Online]. Available: https://arxiv.org/abs/2305.08596.

[5] P. Lewis, E. Perez, A. Piktus, *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks", in *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, 2020. [Online]. Available: https://arxiv.org/abs/2005.11401.

[6] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training", *OpenAI*, 2018, Available at: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

[7] S. Kumar, R. Jangir, and V. Jain, "A survey of research on captcha designing and breaking techniques", in *2019 6th International Conference on Signal*

Processing and Integrated Networks (SPIN), IEEE, Noida, India, 2019, pp. 386–394. DOI: 10.1109/SPIN.2019.8711696.

[8] Y. Zhang, X. Luo, T. Yang, and Y. Liu, "Darknet and deep web: A systematic mapping study", in *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, IEEE, 2020, pp. 1296–1303. DOI: 10.1109/TrustCom50675.2020.00174.

[9] L. Wang, X. Chen, and W. Zhang, "Dark web analytics: A comprehensive literature review", *IEEE Communications Surveys & Tutorials*, vol. 24, no. 2, pp. 1224–1258, 2022. DOI: 10.1109/COMST.2022.3159587.

[10] R. Akbani and T. Korkmaz, "Darkweb access mechanisms: A study of tor anonymity and security gaps", *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3125–3140, 2020. DOI: 10.1109/TIFS.2020.2989283.

[11] J. Smith and M. Johnson, "Onion routing vulnerabilities: A systematic review", in *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, IEEE, 2018, pp. 215–230. DOI: 10.1109/EuroSP.2018.00025.

[12] R. Kumar and S. Singh, "Security best practices for dark web access", in *2021 IEEE International Conference on Cyber Security and Resilience (CSR)*, IEEE, 2021, pp. 87–94. DOI: 10.1109/CSR51186.2021.9527982.

[13] H. Chen and Y. Qin, "Tor network performance and anonymity tradeoffs", in *2019 IEEE Symposium on Security and Privacy (SP)*, IEEE, 2019, pp. 462–478. DOI: 10.1109/SP.2019.00065.

[14] Y. Liu and Z. Sun, "Vpn-over-tor vs. tor-over-vpn: Performance and privacy analysis", *IEEE Access*, vol. 9, pp. 129 143–129 156, 2021. DOI: 10.1109/ACCESS.2021.3113289.

[15] S. Dalvi, S. Patel, A. Kulkarni, and V. Purohit, "Dark web illegal activities crawling and classifying using data mining techniques", *International Journal of*

Interactive Mobile Technologies (iJIM), vol. 16, no. 16, pp. 120–130, 2022. DOI: 10.3991/ijim.v16i16.31947.

[16] S. Ramalingam, M. Krishnamoorthy, and V. Nadar, "Dark web illegal activities crawling and classifying using data mining techniques", *International Journal of Interactive Mobile Technologies*, 2023. [Online]. Available: https://online-journals.org/index.php/i-jim/article/view/30209.

[17] Y. Wang, H. Zhang, and F. Liu, "Analysis of security mechanisms in dark web markets", in *Proc. IEEE Conf. on Dependable and Secure Computing (DSC)*, 2024, pp. 1–8.

[18] A. Al-Naser, N. Al-Qurishi, M. Al-Rakhami, and M. Al-Rodhaan, "Automatic classification of dark web content using machine learning", *IEEE Access*, vol. 9, pp. 74 510–74 521, 2021. DOI: 10.1109/ACCESS.2021.3078742.

[19] A. Bhayani and A. Vyas, "Dark web marketplaces product classification using deep learning", in *2022 International Conference on Computer Communication and Informatics (ICCCI)*, 2022, pp. 1–5. DOI: 10.1109/ICCCI54379.2022.9740929.

[20] J. Koven, M. Thompson, and A. Qureshi, "Cybersecurity visualization and dashboard design for threat intelligence", *Journal of Cybersecurity*, vol. 7, no. 1, tyab007, 2021. DOI: 10.1093/cybsec/tyab007.

[21] J. Lu, Y. Wang, J. Liang, J. Chen, and J. Liu, "An approach to deep web crawling by sampling", *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, vol. 1, pp. 718–724, Dec. 2008. DOI: 10.1109/WIIAT.2008.149.

[22] R. Rawat, A. S. Rajawat, V. Mahor, R. N. Shaw, and A. Ghosh, "Dark web-onion hidden service discovery and crawling for profiling morphing unstructured crime and vulnerabilities prediction", pp. 717–734, 2021. DOI: 10.1007/978-981-16-0749-3_57.

[23] O. Cherqi, G. Mezzour, M. Ghogho, and M. E. Koutbi, “Analysis of hacking related trade in the dark web”, *IEEE International Conference on Intelligence and Security Informatics*, pp. 79–84, Nov. 2018. DOI: 10.1109/ISI.2018.8587389.

[24] P. S. Narayanan, R. Ani, and A. T. L. King, “Torbot: Open source intelligence tool for dark web”, pp. 187–195, 2020. DOI: 10.1007/978-981-15-0146-3_18.

[25] K. Soska and N. Christin, “Measuring the longitudinal evolution of the online anonymous marketplace ecosystem”, *USENIX Security Symposium*, vol. 24, pp. 33–48, Aug. 2015. DOI: 10.5555/3241189.3241194.

[26] A. Baravalle, M. S. Lopez, and S. W. Lee, “Mining the dark web: Drugs and fake ids”, *IEEE International Conference on Data Mining Workshops*, vol. 16, pp. 350–356, Dec. 2016. DOI: 10.1109/ICDMW.2016.0054.

[27] S. He, Y. He, and M. Li, “Classification of illegal activities on the dark web”, *International Conference on Information Science and Systems*, vol. 2, pp. 73–78, Aug. 2019. DOI: 10.1145/3388176.3388200.

[28] A. Singh, V. Kumar, and L. Nguyen, “Darknlp: Rag-enhanced dark web text analysis for cyber threat prediction”, *IEEE Access*, vol. 12, pp. 23 451–23 468, 2024, Combines RAG with transformers for predicting emerging threats from darknet chatter. DOI: 10.1109/ACCESS.2024.3367750.

[29] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, *Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter*, 2019. arXiv: 1910.01108 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1910.01108.

[30] H. Face, *Distilgpt2: Distilled version of gpt2*, https://huggingface.co/distilgpt2, Accessed: 2025-07-02, 2020.