DarkBERT と Llama 3.3 を用いた 初期アクセスブローカーによる投稿の分類と情報抽出

海藤 十和 * [1,a)]　伊藤 祥梧 * [1]　田辺 瑠偉[2]　インミンパパ[2]　吉岡 克成[2,3]

概要：初期アクセスブローカー（IAB）とは，企業や組織のネットワークへの不正アクセス手段（VPN や RDP 認証情報など）を売買し，攻撃者に初期侵入を提供する脅威アクターである．IAB は主にアンダーグラウンドフォーラムや SNS で活動し，特に，ランサムウェア攻撃の初期段階で重要な役割を担うため，その活動の把握は極めて重要である．IAB 活動の把握にはフォーラムや SNS の大量投稿から関連投稿の分類が必要だが，既存手法はルールベースや従来型機械学習に依存し，スケーラビリティ，文脈理解，未知表現対応に限界がある．本研究では，IAB 投稿とその関連情報を効率的かつ高精度に分類・抽出するため，ダークネット特化型モデルである DarkBERT とローカル LLM（Llama 3.3）を組み合わせた手法を提案する．本手法は，DarkBERT による高速フィルタリングとローカル LLM の精緻な分類を統合した手法に加え，販売・購入区分，アクセス種別，価格，標的組織などの情報抽出機能を備える．評価では，二段階分類が 209 件のテストデータで F1 スコア 0.944，500 件の実証実験でも適合率 0.884 を記録した．情報抽出タスクにおいても 50 件のテストデータに対し，部分一致 F1 スコア 0.858 を達成し，両タスクにおいて優れた性能を示した．さらに，全処理はローカルで完結し，認証情報や個人情報の外部流出を回避して安全性と実用性を両立する．CrimeBB データセットの 10 フォーラム（2008 年 2 月〜2024 年 5 月，約 2,518 万件）に本手法を適用し，932 個の IAB アカウントによる 2,581 件の投稿を分類・分析した．その結果，販売投稿が 82.7%を占める供給主導型の市場構造や価格分布，RDP や Shell など主要アクセス種別，政府・金融業などを標的とする IAB 活動の実態が明らかとなった．本研究は，初期アクセス売買の解明に資する基盤を提供し，将来的なサイバー攻撃への早期警戒・対策支援に応用可能である．

キーワード：Initial Access Broker 投稿, アンダーグラウンドフォーラム，DarkBERT，Llama

Detection of IAB Posts Using DarkBERT and Llama 3.3

Towa Kaidou* [1,a)]　Shogo Ito* [1]　Rui Tanabe[2]　Yin Minn Pa Pa[2]　Katsunari Yoshioka[2,3]

Abstract: Initial Access Brokers (IABs) are threat actors who sell or buy illicit access to corporate and organizational networks—such as VPN and RDP credentials—providing adversaries with initial entry points. Operating primarily on underground forums and social platforms, IABs play a pivotal role in the early stages of ransomware operations, making monitoring of their activities critically important. Tracking IAB activities requires classifying relevant posts from massive volumes of forum and social media content, but existing approaches rely on rule-based systems or conventional machine learning, which limits scalability, contextual understanding, and adaptability to novel expressions. We propose a method that combines DarkBERT, a darknet-specialized model, with a local LLM (Llama 3.3) to efficiently and accurately classify IAB posts and extract related information. Our approach integrates high-speed filtering by DarkBERT with fine-grained classification by the local LLM, and includes information extraction capabilities for sale/purchase classification, access types, prices, and target organizations. In evaluation, our two-stage classification achieved an F1 score of 0.944 on 209 test posts and precision of 0.884 in a 500-sample field study. For information extraction tasks, we achieved a partial-match F1 score of 0.858 on 50 test posts, demonstrating excellent performance in both tasks. Furthermore, all processing runs locally, avoiding external leakage of credentials and personal information while balancing security and practicality. Applying our method to ten forums in the CrimeBB dataset (February 2008 – May 2024, approximately 25.18 million posts), we classified and analyzed 2,581 posts by 932 IAB accounts. Results revealed a supply-driven market structure with 82.7% sale posts, characteristic price distributions, dominant access types such as RDP and Shell, and IAB targeting patterns focused on government and financial sectors. This research provides a foundation for elucidating the trade in initial access and can be applied to early warning and proactive defense against future cyberattacks.

Keywords: Initial Access Broker Post, Underground Forum, DarkBERT, Llama

1. はじめに

近年，企業や組織のネットワークに対する不正アクセス手段を提供する **初期アクセスブローカー（Initial Access Broker, IAB）** が注目を集めている．IAB は，仮想プライベートネットワーク（VPN）やリモートデスクトッププロトコル（RDP）などの認証情報を売買し，他の攻撃者に初期侵入の足掛かりを提供する脅威アクターであり，ランサムウェア攻撃の初期段階において重要な役割を担う．特に，IAB が提供するアクセス経路が被害組織への初動侵入手段として利用されており深刻な脅威となっている [1], [2].

組織の防御体制を強化する上で，IAB の活動を早期に検知・可視化することは重要である．このため，サイバー攻撃者の活動を把握する手法として，フォーラム内の投稿を分析してサイバー脅威インテリジェンス（Cyber Threat Intelligence, CTI）を抽出するアプローチが数多く提案されている．既存研究では，ルールベースや機械学習手法（SVM，CNN，Random Forest）を用いた投稿分類・情報抽出手法が検討され，一定の成果を挙げてきた [3], [4]．さらに近年では，ダークウェブ特有の言語的特徴に対応するため，DarkBERT などのドメイン特化型言語モデルを用いた手法も提案されている [5], [6].

しかしながら，これらの手法を IAB による投稿の分類・情報抽出に適用する際に本質的課題が存在する．第一に，従来手法の多くは限定的なデータセットや特定のフォーラムに基づく評価にとどまり，数千万件規模の実データに対するスケーラビリティや，IAB による投稿に見られる文脈の多様性への対応力について十分な検証がなされていない．第二に，サイバーセキュリティ分野では高品質なアノテーションデータの恒常的不足が構造的課題として指摘されており，教師あり学習に基づくアプローチの実運用を制約している [7], [8]．こうした課題への対応策として，近年では大規模言語モデル（LLM）の文脈内学習が注目されているが [9]，商用 LLM API の利用には訓練データを記憶・再生成するセキュリティリスクが存在し，機微な認証情報を含む CTI 分野において大きな制約となる [10], [11].

本研究では，これらの課題に対処するため，ダークネット特化型事前学習モデルである DarkBERT とローカル環境に配置された LLM を組み合わせた二段階投稿分類手法と情報抽出手法を提案する．具体的には，(1)DarkBERT [5] による高速フィルタリング（100 万件あたり約 1.5 時間）で IAB 関連候補を絞り込み，(2) ローカル LLM（Llama3.3）による文脈に基づいた精緻な分類を行う．二段階の分類処理により，ローカル LLM の高い処理コスト（100 万件あたり約 3,800 時間）と商用 API の機密情報漏洩のセキュリティリスクを回避しつつ，実用的な処理速度と高精度分類を両立する．さらに，分類された IAB による投稿に対してプロンプトベースの LLM を用いて，販売者・購入者の区別，アクセス種別，価格情報，標的組織名や業種などの属性を抽出・構造化を行う．提案手法のこれら一連の処理はすべてローカル環境内で完結するため，個人識別情報や認証情報などの機微情報を外部に送信する必要がなく，プライバシー保護の観点からも有用である．

提案手法の精度評価では，約 200 件の IAB による投稿を含むテストデータにおいて，DarkBERT による初期分類で F1 スコア 0.923，二段階アプローチで F1 スコア 0.944 を達成した．また，実際のフォーラムの投稿データにおいて，提案手法を適用して分類した結果から 500 件をランダムサンプリングしたところ適合率は 0.884 であった．さらに，提案手法の情報抽出性能を評価するために用意した約 50 件のテストデータにおいて，完全一致評価で F1 スコア 0.747，部分一致評価で F1 スコア 0.858 を達成した．

複数のアンダーグラウンドフォーラムから収集した約 2,518 万件の英語投稿を対象として，提案手法を用いて 932 個の IAB アカウントによる 2,581 件の投稿を分類・情報抽出した．さらに，抽出した IAB による投稿に関する基礎的な統計分析により，販売投稿が全体の 82.7%を占める供給主導型の構造，主要なアクセス種別，および各種機関への攻撃頻度の傾向といった，IAB が行う活動の基本的な特性が明らかとなった．本研究は，IAB による初期アクセス売買の実態解明に向けた技術的基盤を提供し，今後のランサムウェア攻撃に対する早期警戒・対策支援への応用が期待される．なお，投稿データは Cambridge Cybercrime Centre が公開する CrimeBB データセットを，研究倫理を遵守した上で使用している [12].

2. 関連研究

これまでに，フォーラムやマーケットプレイスにおける投稿内容を分析して，攻撃者の活動や脅威情報といった CTI を抽出する研究が活発に行われている．特に，アンダーグラウンド上のフォーラムにおける投稿の分類と CTI 情報の抽出に関する研究は，IAB 活動の検出に応用できる可能性がある．Portnoff らは違法取引投稿を対象に，ルールベースおよび SVM により取引関連情報を分類・抽出する手法を提案した [3]．Deliu らはハッカーフォーラム投稿の分類において，SVM，CNN，Random Forest の性能を比較して SVM が CNN と同等の精度を達成し，高速な処理によるリアルタイムへの応用可能性を示した [4]．しか

*第一著者および第二著者は同等に本研究へ寄与した.
[1] 横浜国立大学
Yokohama National University
[2] 横浜国立大学大学院先端科学高等研究院
Institute of Advanced Sciences, Yokohama National University
[3] 横浜国立大学大学院環境情報研究院
Graduate School of Environment and Information Sciences, Yokohama National University
[a)] kaidou-towa-pc@ynu.jp

しながら，これらの手法は限定的なデータセットおよび特定フォーラムに基づく評価であり，大規模データへの拡張性およびIABによる投稿特有の表現の曖昧性・文脈依存性に対する対応能力については十分に検証されていない．

このような背景から，近年，ダークウェブ特有の言語的特性に対応するドメイン特化型言語モデルの研究が進んでいる．Youngjin らはダークウェブコーパスで事前学習を行った DarkBERT を開発し，一般的な BERT モデルと比較してアンダーグラウンド特有の語彙および表現に対する理解能力の優位性を実証した [5]．加えて，Paladini らは DarkBERT を CTI 情報の抽出タスクに適用して従来手法を上回る性能を達成した [6]．これらの手法はダークウェブ言語処理において有効であるが，教師あり学習に依存するため，高品質なアノテーションデータの確保が必要である．しかしながら，訓練データ不足とアノテーションデータの作成に伴う人的コストが指摘されている [7], [8]．

そこで，これらの課題を解決するため，大規模言語モデル（LLM）を用いた文脈内学習が注目されている．Brown らは，LLM が少数の例示から文脈に基づく推論を行う能力を有しており，アノテーションが困難な専門領域への応用可能性を示した [9]．一方で，LLM の活用には新たなセキュリティリスクも指摘されている．Carlini らは LLM が訓練データを記憶し，特定のプロンプトによって機密情報が抽出されうることを示した [10]．また，Nasr らは ChatGPT のような商用 LLM に対して，訓練データの抽出が可能であることを報告している [11]．

これらの課題に対処するため，本研究ではセキュリティリスクを回避しつつ，大規模な投稿データから高速かつ高精度に IAB が行った投稿とその関連情報を分類・抽出する手法を提案することを目指す．

3. 用語定義とデータセット

3.1 IAB 投稿の定義

一般的に，IAB とは企業ネットワークへの初期侵入経路を提供する脅威アクターのことである．本研究では，IAB がアンダーグラウンドフォーラムで行った投稿のうち，**組織に対する不正アクセスの売買に関する投稿**を IAB 投稿と定義する．実験では，多様な投稿の中から IAB の活動に関連する投稿を自動分類するため，以下の 2 つの条件を同時に満たす投稿を IAB 投稿として分類する．

- **条件 1：アクセス権の取引的文脈**
  投稿がネットワークまたはシステムへのアクセス権（ログイン権限，管理者権限，リモートアクセス等）を販売，購入，提供等の取引的文脈で言及している．ただし，個人アカウントの認証情報を大量収集した combolist などは対象外とする．
- **条件 2：組織的対象**
  アクセスの対象が組織（企業，政府機関，教育機関等）であることが以下のいずれかにより確認できる．

（1） 売買対象が RDP，Shell，Webshell，Backdoor，Citrix，VDI，Active Directory のいずれかである（主に組織環境で利用されるアクセス手段）．

（2） 投稿内に組織性を示唆する内容（具体的組織名，企業ドメイン，業種，従業員規模等）が含まれる．

これらの判定基準により，大規模データセットに対する自動分類の一貫性と再現性を確保する．また，個人アカウントの売買や具体的なデータ・認証情報の販売に関する投稿は除外されるため，組織に対する初期アクセス提供のみを対象とした投稿の分類が可能となる．

表 1 CrimeBB データセット [12] の構成

フォーラム名	総投稿数	英語投稿数	開始日	終了日
BlackHatWorld	12,487,413	11,682,520	2008/02	2024/05
Nulled	9,311,477	4,667,283	2015/05	2024/02
LolzTeam	6,194,407	2,180,581	2016/08	2019/04
OGUsers	3,607,899	2,781,838	2017/06	2019/03
Cracked	2,905,208	1,844,858	2018/12	2022/10
RaidForums	1,220,894	1,065,590	2017/03	2022/01
Breached	705,922	525,724	2022/03	2023/03
BreachForums	326,407	258,301	2023/06	2024/02
XSS	300,710	30,158	2019/04	2023/03
Offensive-Community	161,476	147,861	2013/03	2018/11
合計	**37,221,813**	**25,184,714**	**2008/02**	**2024/05**

3.2 データセット

本研究では，アンダーグラウンドフォーラムにおける IAB 投稿の分類・情報抽出の性能評価および実運用規模での応用を目的として，ケンブリッジ大学が提供する **CrimeBB** データセット [12] を用いる（以降では，CrimeBB と呼ぶこととする）．CrimeBB は，アンダーグラウンドコミュニティにおけるサイバー犯罪実態の解明を目的として，複数のハッカーフォーラムに投稿されたスレッドや発言を体系的に収集・構造化したオープンなデータセットである．本研究では，2008 年 2 月から 2024 年 5 月までのデータを使用した．各フォーラムの詳細な投稿数を表 1 に示す．評価実験では，CrimeBB に含まれる 10 の主要フォーラムの投稿から，投稿本文，投稿時刻，投稿者情報を含む英語投稿を分析対象とした．具体的には，総計約 3,722 万件の投稿から約 2,518 万件の英語投稿を抽出した．

BERT 系モデルのファインチューニングには，あらかじめ用意した計 984 件のアノテーションデータを用いる．このデータセットには，IAB 投稿 136 件および非関連投稿 848 件が含まれており，5 分割交差検証により精度評価を行った．なお，データ比率（約 1:6）は，予備調査により確認された実際のフォーラムにおける IAB 投稿の稀少性を反映するとともに，モデルの学習効率を考慮した設計によるものである．IAB 投稿の収集は，RDP，Shell，VPN などの初期アクセスに関連するキーワードによる検索で抽

出された投稿からランダムサンプリングを行った．非関連投稿は，これらのキーワードに該当しない投稿から同様にランダムサンプリングを行った．アノテーション作業は，サイバーセキュリティの基礎知識を有する学生 2 名が，事前に作成したガイドラインに基づき実施した．境界事例については判定基準を明確化し，アノテータ間の一致率は，Cohen's κ 0.855 を達成した．不一致は協議により解消し，難易度の高い事例については専門家による検証を行うことで品質の確保に努めた．提案手法の分類性能評価に 209 件，情報抽出性能評価に 50 件のテストデータを用意した．

4. 提案手法

4.1 IAB 投稿分類・情報抽出の流れ

アンダーグラウンドフォーラムにおける IAB 投稿を効率的かつ高精度に分類するために，ドメイン特化型事前学習モデルである DarkBERT と，ローカル環境に配置した大規模言語モデル Llama3.3-70B [13] を組み合わせた二段階分類手法を提案する．さらに，分類した IAB 投稿に対して Llama を用いて，販売者・購入者の区別やアクセス，標的情報などを抽出する．図 1 に本研究の全体像を示す．はじめに，前処理として各投稿データに対して，改行・空白・URL・エスケープシーケンス・不正文字列等を除去し，分類および情報抽出タスクに適した形式にデータを整形する．次に，ステップ 1 で IAB 投稿の分類を，ステップ 2 で取引に関する情報の抽出・構造化を行う．最後に，さらなる分析 [14] を行う．以降では，2 つのステップを説明する．

（1） **ステップ 1: 二段階分類手法による IAB 投稿分類**

第一段階：DarkBERT による高速フィルタリング
入力：投稿本文，投稿者情報，投稿日時を含むフォーラム内の投稿データ
処理: 二値分類器として DarkBERT を用い, IAB 投稿の候補を分類
出力：IAB 投稿の候補
第二段階：Llama3.3 による高精度分類
入力：第一段階で分類された IAB 投稿の候補
処理：プロンプトに基づく IAB 投稿の分類
出力：IAB 投稿

（2） **ステップ 2: 情報抽出と構造化**

入力：ステップ 1 で分類した IAB 投稿
処理: プロンプトに基づく情報の抽出と構造化（JSON 形式）
抽出項目：販売者，購入者，初期アクセス種別，販売価格，標的組織名，連絡手段など

ステップ 1：DarkBERT による高速フィルタリング：本研究の第一段階では，ダークウェブ領域に特化して事前学習された言語モデルである DarkBERT [5] を採用する．DarkBERT は，従来の一般的な BERT 系モデルと比較して，ドメイン固有語彙に対する優れた認識性能を有しており，”RDP”，”VPN”，”creds”，”shell”などのサイバー攻撃に関連する略語・専門用語に対して高い識別精度を示すことが確認されている．本研究では，IAB 投稿の特徴に適応させるため，DarkBERT に対してファインチューニングを実施した．これらの特性により，DarkBERT は他の事前学習モデルと比較して最も高い F1 スコアを達成しており（詳細は 5 節参照），本研究におけるフィルタリング処理のモデルとして最適であると判断した．

ステップ 1：Llama3.3 による高精度分類：第二段階では，第一段階でフィルタリングされた IAB 投稿の候補に対して，Llama3.3-70B を用いたプロンプトベースでの分類を行う．利用したプロンプトは，アクセス権の取引的文脈と組織的文脈の二段階基準による階層的判定を行うように設計した．アクセス権の取引的文脈では，投稿がネットワーク・システムアクセス権を販売・提供等の文脈で言及しているかを判定し，組織的文脈では，組織環境特有のアクセス類型（RDP，Shell，Webshell 等）の検出と企業名や政府機関等の組織的指標の検出を行う．最終分類は両基準を満たす場合のみ IAB 投稿として判定する．

ステップ 2：情報抽出と構造化：第二段階で分類した IAB 投稿に対して更なる分析を可能とするために，投稿本文から取引に関する詳細情報を抽出・構造化を行う．第二段階と同様に Llama3.3-70B を活用し，事前に設計したプロンプトを用いて各投稿から必要なエンティティ情報を抽出する．具体的には，投稿内に記述されている販売者（seller）または購入者（buyer）の立場，販売されているアクセス種別（例: RDP, VPN, Shell など），価格情報，標的となる組織名やその業種，収益規模，および連絡手段を対象として抽出を行う．抽出された情報は統計分析や IAB の市場構造の可視化に資するように JSON 形式で保存する．

4.2 提案手法の評価指標

分類タスクおよび情報抽出タスクの性能を多角的に評価するため，複数の指標を採用した．はじめに，分類精度の評価では，適合率（Precision），再現率（Recall），F1 スコアに加え，不均衡データにおける少数クラス検出性能を適切に評価可能な AUC-PR を用いた．適合率 P は，モデルが IAB 関連と判定した投稿のうち正解である割合を表し，再現率 R は，実際に IAB 関連である投稿のうち正しく判定された割合を表す．F1 スコアは P と R の調和平均として定義され，以下の式で表される．ここで，TP は真陽性（True Positive），FP は偽陽性（False Positive），FN は偽陰性（False Negative）を表す．

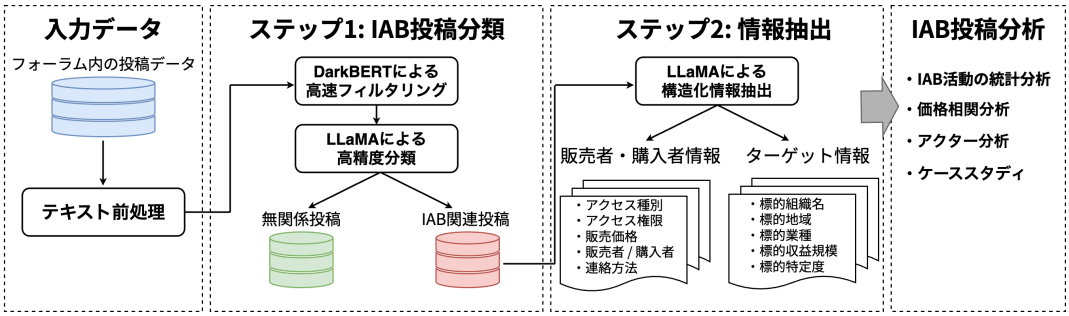

図 1　本研究の流れ

$$P = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \quad R = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \quad F1 = \frac{2PR}{P+R} \quad (1)$$

続いて，提案手法の実運用における有効性を検証するため，5 つのフォーラムを対象とした実証実験を実施した．具体的には，本システム（DarkBERT + Llama3.3 による二段階分類）が IAB 関連と判定した投稿から，フォーラム毎に 100 件の投稿をランダムサンプリングして人手による確認を行うことで適合率を検証した．ここで，適合率とは IAB 関連と判定した投稿のうち，人間の確認により実際に IAB 関連と認められた割合である．

情報抽出タスクでは，完全一致（Exact Match）と部分一致（Partial Match）の 2 種類の評価基準を用いて適合率，再現率，F1 スコアを算出した．完全一致は，抽出文字列が正解データと完全に一致した場合のみ正解と判定する．部分一致は，抽出文字列をトークン単位に分割し，正解データと比較して 50%以上のトークンが一致した場合に正解と判定する．完全一致は厳密な性能評価を可能とする一方，固有名詞や数値表記の軽微な差異によって誤判定となる場合があるため，完全一致を主指標としつつ，部分一致を補助指標として併用することで，実運用における有用性を反映した総合的評価を行った．

4.3　ベースライン手法との比較

提案手法の有効性を検証するため，ステップ 1 の分類タスクの各段階において複数のベースライン手法との比較を行う．具体的には，第一段階：BERT 系モデルの比較では，BERT [15], RoBERTa [16]，SecBERT [17]，DarkBERT [5] の比較を実施する．第二段階：大規模言語モデルの比較では，Qwen2.5-72B [18]，DeepSeek-R1-70B [19]，Llama3.3-70B [13] の比較を実施する．なお，個人情報保護の観点から Ollama[*1] フレームワークを利用し，オフライン環境で動作可能なオープンソースモデルを採用した．また，各 LLM の評価には，第一段階と同一の訓練データセットと評価指標を用いた．

[*1] https://ollama.com/

4.4　提案手法の実装

DarkBERT の設定：第一段階における DarkBERT による高速フィルタリングでは，データセットにおけるクラス不均衡の影響を軽減するため，Focal Loss [20] を採用した．IAB 投稿の希少性を考慮し，$\alpha = 7.0$，$\gamma = 3.0$ に設定した．また，分類閾値は 0.5 に設定し，学習時の正則化として weight decay 0.1，warm-up rate 0.1 を採用した．最適化手法には AdamW を用い，バッチサイズ 8，学習エポック数 10 で訓練を実施した．DarkBERT の最大入力長（512 トークン）を超える長文投稿に対しては，stride 128 の sliding window 手法を用い，各ウィンドウの予測確率の最大値を最終的な分類結果として採用した．

Llama3.3 の設定：第二段階における Llama3.3-70B を用いた高精度分類では，二段階基準による階層的判定フレームワークを設計した．基準 1（アクセス権の取引的文脈）では，投稿がネットワーク・システムアクセス権を取引的文脈で言及しているかを判定する．基準 2（組織的文脈）では，組織環境特有のアクセス類型（RDP，Shell，Webshell 等）の検出と組織的指標の検出を行う．最終分類は両基準を満たす場合のみ IAB 投稿として判定する．分類精度向上のため，段階的推論（Chain-of-Thought），少数例学習（Few-shot Learning），および JSON Schema による構造化出力制約を統合実装した．段階的推論では判定プロセスを 5 段階に分解し推論過程の追跡可能性を確保した．少数例学習では正例 2 件，負例 2 件の代表的事例を提示し，境界事例における判定精度を向上させた．

5.　評価実験

提案手法の性能を評価するために 3 つの実験を行った．具体的には，IAB 投稿の分類に用いるベースモデルの比較を行った．次に，二種類のテスト用のデータセットを用いて提案手法の分類性能と情報抽出性能の評価を行った．

5.1　ベースライン手法との比較結果

はじめに，第一段階，第二段階におけるモデルの分類性能について定量的評価を行った．第一段階における各モデルの分類性能を表 2 に示す．すべてのモデルに対し同一

表 2　BERT 系モデルの分類性能比較

モデル	正解率	適合率	再現率	F1 スコア	AUC-PR
BERT	0.955 ± 0.014	0.841 ± 0.039	0.830 ± 0.058	0.835 ± 0.018	0.863 ± 0.055
RoBERTa	0.951 ± 0.011	0.811 ± 0.054	0.853 ± 0.108	0.827 ± 0.046	**0.901 ± 0.035**
SecBERT	0.935 ± 0.015	0.766 ± 0.093	0.793 ± 0.107	0.771 ± 0.047	0.857 ± 0.033
DarkBERT	**0.964 ± 0.005**	**0.869 ± 0.056**	**0.882 ± 0.080**	**0.872 ± 0.027**	0.896 ± 0.042

表 3　大規模言語モデルの分類性能比較

モデル	正解率	適合率	再現率	F1 スコア
Qwen2.5-72B	0.972 ± 0.008	0.916 ± 0.026	0.875 ± 0.044	0.895 ± 0.029
DeepSeek-R1-70B	0.969 ± 0.004	0.891 ± 0.036	0.882 ± 0.044	0.885 ± 0.016
Llama3.3-70B(CoT なし)	0.973 ± 0.007	0.894 ± 0.036	0.912 ± 0.050	0.902 ± 0.025
Llama3.3-70B	**0.977 ± 0.008**	**0.931 ± 0.015**	**0.898 ± 0.057**	**0.913 ± 0.031**

の学習条件でファインチューニングを行い，984 件のアノテーションデータ（IAB 投稿 136 件，非関連投稿 848 件）を用いて 5 分割交差検証により評価した．この結果から，DarkBERT が正解率，適合率，再現率，および F1 スコアにおいて最も優れた性能を示すことが確認された．特に F1 スコアにおいては 0.872 を記録し，第二位であった BERT（0.835）を 3.7 ポイント上回った．一方で，AUC-PR 値については RoBERTa（0.901）が DarkBERT（0.896）をわずかに上回り，RoBERTa がより広範な分類閾値において安定した性能を発揮することが示唆された．このことから，DarkBERT の優位性は”creds”，”shells”，”RDP”といったダークウェブ特有の語彙や，暗示的な表現に対する高い理解能力に起因すると考えられる．

第二段階で使用する LLM の選定を目的として，第一段階と同様にモデルの分類性能の評価を行った．984 件のアノテーションデータセット（IAB 投稿: 136 件, 非関連: 848 件）を用いて，5 分割交差検証により各モデルの性能を評価した結果を表 3 に示す．この結果，Llama3.3-70B が Qwen2.5-72B や DeepSeek-R1-70B を上回る性能を示した．特に，Chain-of-Thought（CoT）推論を組み込んだ Llama3.3-70B が最高性能を達成し，Accuracy 0.977 ± 0.008，F1 スコア 0.913 ± 0.031 を記録した．CoT なしの Llama3.3 と比較して，Accuracy で 0.4 ポイント，F1 スコアで 1.1 ポイントの改善が確認された．また, Qwen2.5-72B は中程度の性能を示し，DeepSeek-R1-70B は相対的に低い性能となった．これは，IAB 投稿の判定に必要な段階的推論能力や，ダークウェブ特有の文脈理解において，モデル間で差異があることを示唆している．

5.2　二段階分類手法の評価結果

提案手法の二段階分類性能を評価するため，独立したテストデータセット 209 件を用いて定量的評価を実施した．各手法の分類性能を表 4 に示す．この結果から，提案手法は F1 スコア 0.944 を達成し，単体手法と比較して優れた性能を示すことが確認された．また，適合率 1.000 を記録し偽陽性を完全に排除した一方，再現率は 0.894 となった．この性能特性は，DarkBERT の高再現率（0.947）による包括的な候補抽出と，Llama3.3 の高適合率による最終判定の信頼性保証という，両モデルの特性が相補的に機能することによって実現されている．

表 4　二段階分類手法の性能比較

モデル	正解率	適合率	再現率	F1 スコア
DarkBERT	0.970	0.900	**0.947**	0.923
Llama3.3	0.976	0.946	0.921	0.933
二段階分類手法	**0.980**	**1.000**	0.894	**0.944**

実運用システムにおいて，偽陽性による誤検知の回避は極めて重要であり，本手法の適合率 1.000 という特性は高い実用性を示している．しかしながら，本評価は 209 件という比較的小規模なテストデータセットに基づくものであり，提案手法の有効性を一般化するには限界が存在する．そこで，複数の実フォーラムを対象として二段階分類手法を適用した．約 2,518 万件の英語投稿から 932 個の IAB アカウントによる 2,581 件の投稿を分類した結果を表 5 に示す．実運用環境における分類精度を検証するため，主要な 5 つのフォーラムから分類した IAB 投稿 500 件をランダムにサンプリングし，人手による精度確認を行った．その結果，提案手法の適合率は 0.884 を達成しており，アンダーグラウンドフォーラムにおいて IAB 投稿を実用的な精度で分類可能であること考えられる．

誤分類の主因としては，Nulled および BlackHatWorld において，IAB 投稿と類似する語彙を含むサーバ販売投稿の誤検出が確認された．これらのフォーラムでは「access」「RDP」「admin」といった共通語彙により，レンタルサーバ販売投稿が IAB 投稿として誤分類される事例が多く観察され，適合率の低下（Nulled: 0.74，BlackHatWorld: 0.78）の要因となった．

また，計算効率の観点からも提案手法の優位性が実証された．NVIDIA A5000 GPU 1 台環境における処理時間測定の結果，2,518 万件の全投稿処理において，Llama3.3 単体による処理時間の推定値が 87,400 時間であるのに対し，

表 5　二段階分類結果

フォーラム名	英語投稿数	IAB 投稿数	アカウント数	適合率
Nulled	4,667,283	715	333	0.74
BreachForums	258,301	424	140	0.97
BlackHatWorld	11,682,520	383	148	0.78
XSS	30,158	320	91	0.96
Breached	525,724	246	93	0.97
RaidForums	1,065,590	211	59	-
Offensive-Community	147,861	136	15	-
Cracked	1,844,858	93	21	-
OGUsers	2,781,838	41	24	-
LolzTeam	2,180,581	12	8	-
合計	**25,184,714**	**2,581**	**932**	**0.884**

表 6　構造化情報抽出の対象項目

抽出項目	説明
アクセス種別	アクセス手段の種類（RDP，VPN，WebShell 等）
アクセスレベル	アクセス権限の範囲（Admin，root 等）
価格情報	アクセス販売価格（$500，btc 0.2 等）
標的組織名	攻撃対象の組織名（企業名，政府機関等）
標的地域	標的組織の所在地域（USA，Europe，JP 等）
標的業種	標的組織の産業分野（金融業，製造業等）
標的収益規模	標的組織の収益（$5B，30M USD 等）
連絡手段	取引時の連絡方法（Telegram，Discord 等）
標的特定度	組織名の特定性（特定/不特定）

提案する二段階分類手法では 73 時間で完了し，大幅な処理時間の削減を達成した．

5.3　情報抽出性能の評価結果

ステップ 2 の情報抽出性能を評価するため，50 件のテストデータセットを用いて，IAB 投稿から抽出される 9 項目の構造化情報を評価した．評価対象となる構造化情報の一覧を表 6 に示す．評価基準としては，完全一致に加え，トークン一致率が 50%以上の場合を正と判定する部分一致を併用した．各抽出項目の定量的評価結果を表 7 に示す．この結果から，完全一致評価では Micro 平均 F1 スコア 0.747 を達成し，部分一致評価では 0.858 へと 11.1 ポイントの改善が確認された．ここで，例えば正解が「RDP」であるのに対してシステムが「RDP Access」を抽出した場合，または正解が「low price」であるのに対して「a very low price」を抽出した場合，完全一致では不一致と判定される．しかし，IAB 投稿の分析においては，抽出範囲の微細な差異よりも本質的な情報内容の一致が重要であるため，部分一致による柔軟な評価が補助指標として適切であると判断した．部分一致評価における Micro 平均 F1 スコア 0.858 は，実用的な CTI システムとして十分な性能水準であり，提案手法が IAB 投稿からの構造化情報抽出において有効であると考えられる．

6.　IAB の活動分析

提案手法を用いて分類した 932 個の IAB アカウントによる 2,581 件の投稿を対象として，IAB が行った活動の統計分析を行う．提案手法により 10 フォーラムから分類・

表 7　構造化情報抽出の性能評価

抽出項目	完全一致			部分一致		
	適合率	再現率	F1 スコア	適合率	再現率	F1 スコア
アクセス種別	0.655	0.692	0.673	0.800	0.846	0.822
アクセスレベル	0.714	0.714	0.714	0.857	0.857	0.857
標的組織名	0.737	0.824	0.778	0.842	0.941	0.889
標的地域	1.000	0.923	0.960	1.000	0.923	0.960
標的業種	0.818	0.692	0.750	0.818	0.692	0.750
標的収益規模	1.000	1.000	1.000	1.000	1.000	1.000
価格情報	0.647	0.647	0.647	0.882	0.882	0.882
連絡手段	0.727	0.800	0.762	0.864	0.950	0.905
標的特定度	0.833	0.833	0.833	0.833	0.833	0.833
Macro 平均	**0.792**	**0.792**	**0.791**	**0.877**	**0.881**	**0.878**
Micro 平均	**0.738**	**0.756**	**0.747**	**0.848**	**0.869**	**0.858**

表 8　IAB 投稿の基本統計

投稿分類			価格統計 (USD)		
項目	**件数**	**割合**	**項目**	**価格**	**-**
総 IAB 投稿	2,581	-	平均価格	15,263	-
販売投稿	2,134	82.7%	中央値	421	-
購入投稿	447	17.3%	最高価格	2,000,000	-
主要アクセス種別			**主要標的業種**		
種別	**件数**	**割合**	**業種**	**件数**	**割合**
RDP	375	14.5%	政府機関	126	13.6%
Access	102	4.0%	金融業	120	12.9%
Shell	77	3.0%	教育機関	38	4.1%
VPN	46	1.8%	医療機関	18	1.9%

情報抽出された 2,581 件の IAB 投稿の分析結果を表 8 に示す．なお，詳細な分析結果は論文 [14] にまとめることとして，本研究では基礎的な分析を行う．

分析の結果，以下の特性が明らかになった．第一に，全体の 82.7%が販売投稿で占められており，供給主導型の市場構造が確認された．第二に，アクセス種別では RDP が最も多く取引され（14.5%），次いで Access（4.0%），Shell（3.0%），VPN（1.8%）の順となっている．第三に，標的業種では政府機関が最も多く（13.6%），次いで金融業（12.9%），教育機関（4.1%），医療機関（1.9%）と続き，重要インフラ・組織を標的とする販売投稿が多い傾向が確認された．また，価格面では平均価格 15,263 USD に対して中央値 421 USD と大きな差があり，一部の高額取引（最高 2,000,000 USD）が平均値を押し上げていることが示されている．

これらの基礎的統計により，アンダーグラウンドフォーラムにおける IAB の活動は，販売投稿が圧倒的多数を占める供給主導型の市場構造を有し，特に政府機関や金融業といった重要インフラへの不正アクセス手段が活発に取引されている実態が明らかとなった．価格分布の大きな偏りは，標的組織の重要度や取得困難性から生じる価格格差に起因するものと考えられる．

7.　まとめと今後の展望

本研究では，アンダーグラウンドフォーラムにおける IAB（Initial Access Broker）投稿を効率的かつ高精度に分類・情報抽出するため，DarkBERT と ローカル環境の Llama3.3-70B を用いた手法を提案した．

二段階分類手法はテストデータで F1 スコア 0.944，実証実験で適合率 0.884 を記録し，情報抽出のタスクでは，完全一致 F1 スコアで 0.747，部分一致 F1 スコアで 0.858 を達成し，実運用可能な高精度を示した．また，CrimeBB データセットに含まれる約 2,518 万件のデータを 73 時間で分析可能にするとともに，販売投稿が 82.7%を占める供給主導型の市場や標的業種の分布など基礎的特性を明らかにした．レンタルサーバ販売投稿の誤検出や多言語対応の制約といった課題が残る一方で，本研究の成果は将来的なサイバー攻撃に対する早期警戒・対策支援に資することが期待される．

謝辞 本研究の一部は NEDO（国立研究開発法人新エネルギー・産業技術総合開発機構）の委託事業「経済安全保障重要技術育成プログラム／先進的サイバー防御機能・分析能力強化」（JPNP24003）によるものである．
また，金子翔威氏のご協力に深く感謝申し上げる．

参考文献

[1] Cyberint (a Check Point Company). Initial access brokers report 2025. Technical report, Cyberint / Check Point, 2025. Based on underground forums and dark-web marketplaces analysis over the past two and a half years.
[2] KrakenLabs / Outpost24. Demystifying initial access brokers (iabs) and links to ransomware. Technical report, Outpost24 KrakenLabs, 2024. Analyzed 152 corporate access sale offers from underground forums.
[3] Rebecca S. Portnoff, Sadia Afroz, et al. Tools for automated analysis of cybercriminal markets. In Proceedings of the 26th International Conference on World Wide Web, WWW '17, p. 657–666, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee.
[4] Isuf Deliu, Carl Leichter, et al. Extracting cyber threat intelligence from hacker forums: Support vector machines versus convolutional neural networks. In 2017 IEEE International Conference on Big Data (Big Data), pp. 3648–3656, 2017.
[5] Youngjin Jin, Eugene Jang, et al. Darkbert: A language model for the dark side of the internet, 2023.
[6] Tommaso Paladini, Lara Ferro, et al. You might have known it earlier: Analyzing the role of underground forums in threat intelligence. In Proceedings of the 27th International Symposium on Research in Attacks, Intrusions and Defenses, RAID '24, p. 368–383, New York, NY, USA, 2024. Association for Computing Machinery.
[7] Yuelin Hu, Futai Zou, et al. Llm-tikg: Threat intelligence knowledge graph construction utilizing large language model. Computers Security, Vol. 145, p. 103999, 2024.
[8] Eric Nunes, Ahmad Diab, et al. Darknet and deepnet mining for proactive cybersecurity threat intelligence. In 2016 IEEE Conference on Intelligence and Security Informatics (ISI), pp. 7–12, 2016.
[9] Tom B. Brown, Benjamin Mann, et al. Language models are few-shot learners, 2020.
[10] Nicholas Carlini, Florian Tramer, et al. Extracting training data from large language models, 2021.
[11] Milad Nasr, Nicholas Carlini, et al. Scalable extraction of training data from (production) language models, 2023.
[12] Sergio Pastrana, Daniel R. Thomas, et al. Crimebb: Enabling cybercrime research on underground forums at scale. In Proceedings of the 2018 World Wide Web Conference, WWW '18, p. 1845–1854, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.
[13] Meta AI. Llama 3.3 70b. Technical report, Meta, 2024.
[14] 伊藤祥梧, 海藤十和ほか. "アンダーグラウンドフォーラムにおける iab 活動の調査：市場動向分析とアクタープロファイリング". 情報処理学会コンピュータセキュリティシンポジウム（CSS2025）, 2025.
[15] Jacob Devlin, Ming-Wei Chang, et al. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.
[16] Yinhan Liu, Myle Ott, et al. Roberta: A robustly optimized bert pretraining approach, 2019.
[17] jackaduma. Secbert: A pretrained language model for cyber security text. `https://huggingface.co/jackaduma/SecBERT`, 2022.
[18] Qwen, :, An Yang, Baosong Yang, et al. Qwen2.5 technical report, 2025.
[19] Daya Guo, Dejian Yang, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948, 2025.
[20] Tsung-Yi Lin, Priya Goyal, et al. Focal loss for dense object detection, 2018.