一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

類似チャンネル提示機能を活用したTelegramにおける サイバー犯罪関連チャンネルの発見と分析

青砥　陸† 　インミンパパ†† 　吉岡　克成†††,††

† 横浜国立大学大学院環境情報学府
†† 横浜国立大学先端科学高等研究院
††† 横浜国立大学環境情報研究院
E-mail: †aoto-riku-hx@ynu.jp, ††{yinminn-papa-jp,yoshioka}@ynu.ac.jp

あらまし　Telegram は近年，サイバー犯罪者にとって主要な活動の場となっている．本研究では，膨大な数のチャンネルから，サイバー犯罪関連チャンネルを効率的に発見・分類するシステムを提案する．本システムは，Explore，Filter，Categorize の 3 つのモジュールで構成される．Explore モジュールでは，Telegram の「Similar Channels」機能を用いて既知のサイバー犯罪関連チャンネルに類似するチャンネルを特定する．しかし，この機能はチャンネルの購読者の重複に基づいて類似性を判定するため，無関係なチャンネルも多く含まれる．そこで Filter モジュールでは，大規模言語モデル（LLM）を用いて各チャンネルの内容を分析し，無関係なチャンネルを除外する．Categorize モジュールでは，LLM を用いてやり取りされているサイバー犯罪のトピックに基づいてチャンネルを分類する．本研究では 272 個の既知チャンネルから「Similar Channels」機能を用いて 2,273 個のチャンネルを発見し，フィルタリングで 791 個のサイバー犯罪関連チャンネルを抽出した．分類の結果，悪意のあるツールや窃取したとされるデータの取引に関するチャンネル，DDoS 攻撃などを通じてシステムの正常な機能を妨害する活動に関するチャンネルが多く観測された．

キーワード　Telegram，サイバー犯罪，大規模言語モデル，チャンネル分類

Discovery and Analysis of Cybercrime-Related Telegram Channels using the Similar Channel Recommendation Feature

Riku AOTO†, Yin Minn Pa Pa††, and Katsunari YOSHIOKA†††,††

† Graduate School of Environment and Information Sciences, Yokohama National University
†† Institute of Advanced Sciences, Yokohama National University
††† Faculty of Environment and Information Sciences, Yokohama National University
E-mail: †aoto-riku-hx@ynu.jp, ††{yinminn-papa-jp,yoshioka}@ynu.ac.jp

Abstract　Telegram has become a major hub for cybercriminals in recent years. This paper proposes a system to efficiently discover and categorize cybercrime-related channels among the vast number of channels on Telegram. The system comprises three modules: Explore, Filter, and Categorize. The Explore module utilizes Telegram's "Similar Channels" feature to identify channels similar to known cybercrime-related channels. However, because this feature determines similarity based on how many subscribers are common between channels, it often includes many irrelevant channels. Therefore, the Filter module employs a large language model (LLM) to analyze each channel's content and filter out irrelevant channels. The Categorize module uses an LLM to classify channels based on their cybercrime-related discussion topics. Through this approach, we identified 2,273 channels using the "Similar Channels" feature starting from 272 known channels, and extracted 791 cybercrime-related channels through filtering. The categorization revealed a prevalence of channels related to trading malicious tools and allegedly stolen data, as well as channels focused on activities aimed at disrupting system functionality through DDoS attacks and other means.

Key words　Telegram, Cybercrime, Large Language Models, Channel Classification

1. はじめに

近年，インターネット上でのサイバー犯罪活動は，ダークウェブから一般の SNS プラットフォームへとその活動領域を拡大している．特に Telegram [1] は，エンドツーエンド暗号化による高度な匿名性と，特別なソフトウェアを必要としない優れたアクセス性を備えており，サイバー犯罪者にとって極めて魅力的なプラットフォームとなっている．実際に，DDoS 攻撃ツールの販売，個人情報の取引，組織的なサイバー攻撃の計画に関する情報など，多岐にわたる違法活動に関するメッセージが Telegram 上で確認されている．

このような状況に対し，法執行機関やセキュリティ研究者らは，Telegram 上での犯罪活動の実態解明に取り組んでいるものの，Telegram には膨大な数のチャンネルが存在し，その作成と削除が頻繁に繰り返されるため，犯罪に関連するチャンネルを継続的に発見し分析することは，依然として重要な課題となっている．

本研究では，この課題に対し，Telegram が提供する「Similar Channels」機能と大規模言語モデル（LLM）を組み合わせた，効率的なチャンネル発見・分析手法を提案する．犯罪関連チャンネルの発見手法としては，TGStat [2] などのディレクトリサービスの活用，Telegram のグローバル検索機能，メッセージのフォワーディング関係の追跡，チャンネル招待 URL の収集など，複数のアプローチが存在する．本研究では，これらの手法の中から，比較的安定した数のチャンネルを継続的に発見できる「Similar Channels」機能に着目した．しかしながら，この機能はチャンネルの購読者の重複度に基づいて類似性を判定する [3] ため，スポーツや娯楽など，無関係なチャンネルも結果に一定数含まれる傾向がある．そこで本研究では，チャンネルの説明文や投稿内容を LLM で分析することで，監視対象として適切なチャンネルを自動的に判別する．これらのチャンネルを新たなシードとして探索を繰り返すことで，無関係なチャンネルを増やすことなく，サイバー犯罪関連チャンネルを効率的に発見することが可能となる．

チャンネル情報の収集にあたっては，Telegram の API クライアントライブラリである Telethon [4] を活用し，チャンネルのメタデータや投稿メッセージを効率的に収集するシステムを構築した．具体的には，サイバー犯罪に関連する可能性が高いと手動で判断された 272 個のチャンネルをシードデータとし，Telegram の「Similar Channels」機能を活用して 2,273 個の類似チャンネルを発見した．これらのチャンネル群の中には，サイバー犯罪とは関係ないものも含まれているため，オンプレミスで構築したローカル LLM（Llama 3.3 [5]）を用いて関係のないものを除外するフィルタリングを行い，最終的にサイバー犯罪に関係する可能性の高い 791 個のチャンネルを抽出した．この収集されたチャンネルに対して，ローカル LLM を用いた分類・分析を行うことで，Telegram 上で流通するサイバー犯罪コンテンツの実態把握を試みた．具体的には，事前に定義したサイバー犯罪のカテゴリ（不正データ取得，システム妨害，デバイスの悪用など）に基づいてチャンネルを分類し，各カテゴリの特徴を明らかにした．

本研究の主な貢献は以下の 2 点である．第一に，「Similar Channels」機能と LLM を組み合わせた効率的なチャンネル発見手法を提案し，その有効性を実証した点である．第二に，大規模なメッセージ収集・分析システムを構築し，Telegram 上でのサイバー犯罪活動の種別や特徴を明らかにした点である．

2. 関連研究

Telegram は匿名性の高さと優れたアクセス性から，サイバー犯罪を含む多様な分野での情報共有プラットフォームとして利用されている．近年，同プラットフォームは研究対象としても注目を集めており，情報拡散，コミュニティ分析，ヘイトスピーチ，陰謀論など，多岐にわたる研究が展開されている．

Curley et al. [6] は，COVID-19 パンデミック下のロックダウン反対運動における極右勢力の関与について，アイルランドの Telegram グループを対象とした分析を実施した．スノーボールサンプリングとコンテンツ分析を通じて，極右的な言説のメッセージの存在を明らかにした．Zihiri et al. [7] は，QAnon，極右，極左コミュニティにおける 350 万件以上の Telegram メッセージを分析し，各コミュニティのメッセージにおける言説の特徴と相互関連性を体系的に比較検討した．

また，Vergani et al. [8] は，COVID-19 パンデミック初期のイタリアの陰謀論チャンネルにおけるヘイトスピーチの動態を分析し，パンデミックの進行に伴うヘイトスピーチ対象の変遷過程を明らかにした．

既存研究は，特定のトピックや社会現象における Telegram の利用実態を明らかにしてきた．しかしながら，サイバー犯罪という広範なカテゴリを対象とし，効率的なチャンネル発見・分析手法を提案した研究は限定的である．本研究は，この研究領域の空白に着目し，サイバー犯罪関連チャンネルの効率的な発見と分析を可能にする新たな方法論を提案する．

3. 背景と用語

3.1 サイバー犯罪のカテゴリ

本研究では，Tsakalidis ら [9] が提案したフレームワークを基準として，サイバー犯罪の分類体系を構築する．この分類体系は，ブダペスト条約 [10] の基本原則に基づきながら，近年の技術進展と新たな脅威を反映するよう拡張を加えている．Tsakalidis らは詳細な分類を提示しているが，本研究ではブダペスト条約を補完的に参照することで，分類の妥当性と国際的な法的整合性を確保している．

本分類体系は，主として二つの目的で使用する．第一に，LLM の性能を評価する際の正解データを作成するために使用する．具体的には，研究者がチャンネルの内容を確認し，この分類体系に基づいて各チャンネルを分類することで，LLM の出力と比較するための正解データを作成する．第二に，LLM にチャンネルを分類させる際のプロンプトを作成するために使用する．LLM がチャンネルを適切に分類できるよう，この分類体系の定義や基準を参照しながらプロンプトを構築する．

サイバー犯罪は，5 つの主要カテゴリとそれぞれに付随する

サブカテゴリから構成される．以下に，各カテゴリの概要を示す．詳細な定義については，Tsakalidis ら [9] とブダペスト条約 [10] の原典を参照されたい．

カテゴリ A：コンピュータデータおよびシステムの機密性，完全性，および可用性に対する脅威．このカテゴリは，6 つのサブカテゴリを含む．違法アクセス（*A1*）：システムへの不正アクセスに関与する行為．違法データ取得（*A2*）：機密データの窃取を対象とする行為．違法傍受（*A3*）：通信の不正傍受を指す行為．データ妨害（*A4*）：データの改ざんまたは破壊に関与する行為．システム妨害（*A5*）：システムの機能を妨害する行為．デバイスの不正使用（*A6*）：サイバー犯罪のためのツールの作成，所持，または配布を指す行為．

カテゴリ B：コンピュータを介した犯罪．このカテゴリは，3 つのサブカテゴリを含む．コンピュータ関連の偽造（*B1*）：デジタル記録の偽造に関与する行為．コンピュータ関連の詐欺（*B2*）：金銭的利益を目的とした詐欺行為を対象とする行為．*ID* 関連の犯罪（*B3*）：ID の窃取と不正使用を包含する行為．

カテゴリ C：コンテンツに関連する違法行為．このカテゴリは，7 つのサブカテゴリを含む．ポルノコンテンツ（*C1*）：違法な露骨なコンテンツの作成または配布に関する行為．児童ポルノ（*C2*）：未成年者を搾取する素材に関与する行為．宗教的犯罪（*C3*）：宗教的信念を標的とする行為．サイバーいじめ（*C4*）：デジタルプラットフォームを介したハラスメントに関与する行為．違法ギャンブルおよびオンラインゲーム（*C5*）：規制されていないゲーム活動を促進する行為．スパムおよび関連する脅威（*C6*）：大量の迷惑な通信に関与する行為．人種差別およびヘイトスピーチ（*C7*）：差別的または憎悪的なコンテンツを拡散する行為．

カテゴリ D：知的財産権の侵害．このカテゴリは，2 つのサブカテゴリを含む．著作権関連の犯罪（*D1*）：著作権で保護された素材の不正使用または配布を指す行為．商標関連の犯罪（*D2*）：偽造のために商標を不正使用する行為．

カテゴリ E：複合的な違法行為．このカテゴリは，4 つのサブカテゴリを含む．フィッシング（*E1*）：機密情報を窃取するための欺瞞的な行為．サイバーロンダリング（*E2*）：不正な収益を処理するためにデジタルツールを使用する行為．サイバー戦争（*E3*）：国家安全保障または重要インフラストラクチャを標的とする行為．インターネットのテロリスト利用（*E4*）：デジタルプラットフォームを通じてテロリズムを促進または助長する行為．

3.2 Similar Channels 機能

Telegram は，ユーザーが新たなチャンネルを発見するための機能として，Similar Channels と呼ばれるレコメンデーション機能を提供している．この機能は，チャンネル間の購読者の重複度に基づいて類似性を判定し，購読者層が共通しているチャンネルほど類似度が高いと判断される．ユーザーが Telegram アプリケーション上でチャンネルを閲覧する際に，類似チャンネルが自動的に提示される．また，この機能は Telegram API からも利用可能であり，特定のチャンネル ID を指定することで，そのチャンネルに類似したチャンネルのリストを取得できる．そのため，サイバー犯罪関連のチャンネルをシードとして入力することで，類似するサイバー犯罪関連のチャンネルを探索できると考え，本研究ではこの機能を活用することとした．

しかしながら，Similar Channels 機能は購読者の重複のみに基づいて類似性を判定するため，必ずしもコンテンツの類似性を反映しているとは限らない．例えば，サイバー犯罪とは無関係なスポーツや娯楽など，全く異なるジャンルのチャンネルが類似チャンネルとして提示される可能性がある．そのため，本研究では，Similar Channels 機能で探索したチャンネルを，後述する Filter モジュールでフィルタリングし，サイバー犯罪と関連性の高いチャンネルのみを抽出する．

4. 提 案 手 法

4.1 システムの全体像

本研究では，Telegram 上のサイバー犯罪関連チャンネルを効率的に発見し，分析するためのシステムを提案する．本システムは，Explore，Filter，Categorize の 3 つの主要なモジュールで構成される．

図 1 に示すように，本システムは以下の手順で動作する．まず，Explore モジュールで類似チャンネルを探索し，その結果を Filter モジュールでフィルタリングしてサイバー犯罪関連チャンネルを抽出する．最後に，抽出されたチャンネルを Categorize モジュールでカテゴリ分類し，分析を行う．これらのモジュールが連携して動作することで，効率的なサイバー犯罪関連チャンネルの発見・分析を実現する．

4.2 Explore モジュール

Explore モジュールは，Telegram の Similar Channels 機能を用いて，既知のサイバー犯罪関連チャンネルに類似するチャンネルを探索する．このモジュールでは，手動で特定したサイバー犯罪に関連する可能性の高いチャンネルをシードとして Similar Channels 機能を実行し，類似チャンネルを探索する．

Similar Channels 機能を繰り返し活用することで，サイバー犯罪に関係する可能性が高いチャンネルを効率的に探索できる．本論文では，サイバー犯罪に関係するチャンネルを「悪性」，サイバー犯罪に関係しないチャンネルを「良性」と表現する．なお，3.2 節で述べたように，Similar Channels は購読者の重複のみに基づいて類似性を判定するため，結果には良性なチャンネルが一定数含まれる．そのため，2 回目以降の Similar Channels への入力は，悪性と判断されたチャンネルのみに限定する．具体的には，シードチャンネルに対して Similar Channels 機能を適用して得られたチャンネル群に対して，後述する Filter モジュールでフィルタリングし，悪性レベルが Malicious であるチャンネルのみを抽出する．その後，抽出された Malicious チャンネルを次の Similar Channels 機能への入力として使用することで，より効率的に悪性度の高いチャンネルを探索する．この手法により，サイバー犯罪関連チャンネルを効率的に発見し，監視対象を効果的に拡大することができる．

4.3 Filter モジュール

Filter モジュールは，Explore モジュールで発見されたチャン

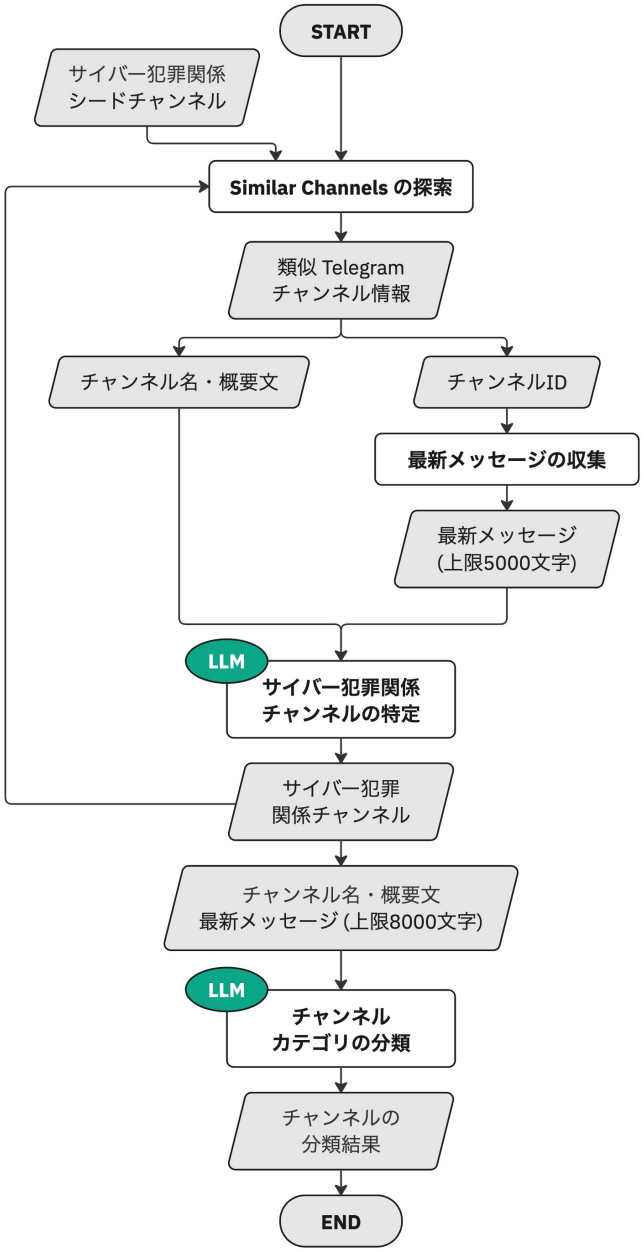

図 1: 提案システムのフローチャート

ネルの中から，サイバー犯罪に関係のないチャンネルを除外するために，大規模言語モデル（LLM）を使用する．各チャンネルのタイトル，概要文，収集したメッセージを LLM に入力し，そのチャンネルがサイバー犯罪に関連する度合いを評価する．LLM は，与えられたチャンネル情報に対して，以下の 3 段階の悪性レベルを出力する．

- **Benign**: メッセージは通常の会話を含み，サイバー犯罪関連のキーワード，ツール，エクスプロイト，または技術に関する言及がない．また，窃取データ，認証情報，機密情報の取引や共有，不正なサービスや活動の助長も示唆していない．
- **Suspicious**: メッセージにはサイバー犯罪関連のキーワードが含まれているが，文脈が曖昧で不明瞭である．または，ハッキングツールや技術に関する議論が含まれている可能性があるが，サイバー犯罪を実行する明確な意図は示されていない．
- **Malicious**: メッセージは，サイバー攻撃の積極的な調整や計画，窃取データの取引や共有，不正サービスの提供など，明確なサイバー犯罪に関連する情報が含まれている．

LLM はサイバー犯罪アナリストとして，与えられたチャンネル情報を分析し，最も適切な悪性レベルを判断する．悪性レベルが Malicious のチャンネルを，次の Categorize モジュールへの入力として使用する．

4.4 Categorize モジュール

Categorize モジュールは，Filter モジュールで抽出されたチャンネルを，前述の「サイバー犯罪のカテゴリ」に基づき，LLM を用いて分類する．LLM には，各チャンネルのタイトル，概要文，収集されたメッセージをプロンプトとして入力し，3.1 節で定義したカテゴリの中から，最も関連性の高いカテゴリを 1 つ選択させ，その理由を分析結果として出力させる．

この Categorize モジュールにより，発見されたチャンネルを体系的に分類し，サイバー犯罪活動の種類や特徴を明らかにする．

5. データ収集と分析

本章では，提案システムを用いて実施したデータ収集と分析の手順，およびその結果について説明する．

5.1 データ収集

データ収集は，2025 年 1 月 15 日から 2025 年 1 月 22 日にかけて実施した．データ収集の手順は以下の通りである．

（1） シードチャンネルの準備：二人の研究者により，サイバー犯罪に関連する可能性が高いと判断された 272 個の Telegram チャンネルをシードチャンネルとして準備した．

（2） 類似チャンネルの探索：Explore モジュールを用いて，シードチャンネルに対して Similar Channels 機能を適用し，類似チャンネルを探索した．表 1 に Similar Channels 機能の適用結果を示す．1 回目の適用では 1,360 個，2 回目の適用では 1,730 個のチャンネルを発見した．2 回目の適用では，1 回目の適用で Malicious と判定されたチャンネルから重複を除いた 502 個のチャンネルをシードとして使用した．その結果，1 回目と 2 回目の適用で，重複を除いて合計 2,273 個のチャンネルを発見した．

表 1: Similar Channels 機能適用結果

適用回数	シードチャンネル数	発見されたチャンネル数
1 回目	272	1360
2 回目	502	1730
合計	774	3090
合計 (unique)	774	2273

（3） チャンネルのフィルタリング：発見された類似チャンネル

に対して，Filter モジュールを用いてローカル LLM (Llama 3.3 [5]) を用いたフィルタリングを行い，Malicious と判定されたチャンネルを抽出した．表 2 は，Similar Channels 機能の各適用で発見されたチャンネルの悪性レベル別内訳を示したものである．表に示すように，1 回目の適用で発見されたチャンネルのうち 531 個，2 回目の適用では 282 個が Malicious と判定された．1 回目，2 回目ともに，Malicious と判定されたチャンネルが最も多く，全体の 3〜4 割程度を占めている．

表 2: 発見されたチャンネルの悪性レベル別内訳

適用回数	Malicious	Suspicious	Benign	None
1 回目	531 (39.0%)	238 (17.5%)	450 (33.1%)	141 (10.4%)
2 回目	282 (30.9%)	222 (24.3%)	260 (28.5%)	149 (16.3%)
合計 (unique)	791 (34.8%)	460 (20.2%)	732 (32.2%)	290 (12.8%)

（4） メッセージの収集：Filter モジュールで Malicious と判定されたチャンネルから，チャンネル開設時からデータ収集時点までのすべてのメッセージを収集した．その結果，合計 1,307,630 件のメッセージを収集した．

5.2 データ分析

データ収集で得られたメッセージに対して，Categorize モジュールを用いて，サイバー犯罪のカテゴリ分類を行った．

表 3 は，カテゴリ分類の結果を示している．最も多く検出されたカテゴリは「デバイスの悪用 (A6)」であり，全体の 63.0% を占めている．このカテゴリには，ゲームチートツールやアカウント売買，RAT（Remote Access Trojan）などの悪意のあるソフトウェアの取引に関する情報が含まれていた．次いで，「不正データ取得 (A2)」が全体の 11.1% を占めている．このカテゴリには，クレジットカード情報，金融機関の口座情報といった機密データの取引情報や，盗難アカウントの販売情報など，金銭的な利益を目的とした活動に関する情報が含まれる．これらの結果から，Telegram が金銭的な利益を目的とする攻撃者らの情報共有プラットフォームとして利用されている実態が示唆された．

また，システム妨害 (A5) が 5.9% と比較的高い割合を占めていることも特筆すべき点である．このカテゴリには，DDoS 攻撃の実行や，サーバーの乗っ取り，マルウェア感染など，システムの正常な動作を妨害する活動に関する情報が含まれており，Telegram がこれらの攻撃を調整・実行するための情報共有基盤として利用されている可能性が示唆された．

6. LLM の評価

本研究では，サイバー犯罪関連チャンネルの発見と分析にローカル LLM (Llama 3.3) を活用している．LLM の性能はシステムの精度に大きく影響するため，ここでは LLM の評価方法と結果について説明する．

6.1 評価方法

LLM の評価は，Filter モジュールにおけるチャンネルのフィルタリング性能と，Categorize モジュールにおけるチャンネルのカテゴリ分類性能の 2 つの観点で実施した．

表 3: カテゴリ別分類結果

カテゴリ	チャンネル数	割合
不正アクセス (A1)	2	0.3%
不正データ取得 (A2)	88	11.1%
不正傍受 (A3)	1	0.1%
データ妨害 (A4)	0	0.0%
システム妨害 (A5)	47	5.9%
デバイスの悪用 (A6)	498	63.0%
コンピュータ関連偽造 (B1)	0	0.0%
コンピュータ関連詐欺 (B2)	43	5.4%
個人情報関連犯罪 (B3)	13	1.6%
ポルノコンテンツ (C1)	3	0.4%
児童ポルノ (C2)	1	0.1%
宗教的犯罪 (C3)	3	0.4%
サイバーいじめ (C4)	9	1.1%
違法ギャンブル・オンラインゲーム (C5)	6	0.8%
スパムおよび関連する脅威 (C6)	3	0.4%
人種差別・ヘイトスピーチ (C7)	17	2.1%
著作権関連の犯罪 (D1)	10	1.3%
商標関連の犯罪 (D2)	0	0.0%
フィッシング (E1)	9	1.1%
サイバーロンダリング (E2)	9	1.1%
サイバー戦争 (E3)	21	2.7%
インターネットのテロリスト利用 (E4)	8	1.0%
合計	791	100.0%

6.1.1 Filter モジュールの評価

Filter モジュールの評価では，事前に人手で悪性度を判定したチャンネルをデータセットとして用意し，LLM によるフィルタリング結果と比較することで性能を評価した．具体的には，Benign（サイバー犯罪の兆候なし），Suspicious（サイバー犯罪の疑わしい兆候），Malicious（明確なサイバー犯罪コンテンツの存在）の 3 つのレベルごとに，30 チャンネルずつ，合計 90 チャンネルを評価用データセットとして準備した．

データセットの準備手順は以下の通りである．まず，サイバー犯罪に関連する可能性が高いと事前に判断されたシードチャンネルを基に，Similar Channels 機能を 1 回適用して得られたチャンネルから十分な数をランダムにサンプリングした．次に，一人の研究者が Telegram アプリケーションで各チャンネルのタイトル，概要文，投稿内容を確認し，4.3 節で述べた定義に基づき，各チャンネルの悪性レベルを判断して Benign, Suspicious，Malicious のいずれかに分類した．この分類によって，各悪性レベルのチャンネルを 30 個ずつ，合計 90 個を抽出して評価用データセットとした．準備した 90 個のチャンネルを Filter モジュールの LLM に入力し，出力された悪性レベルと正解データを比較して，各レベルにおける精度を算出した．

6.1.2 Categorize モジュールの評価

Categorize モジュールの評価では，事前に人手でカテゴリを

判定したチャンネルをデータセットとして用意し，LLM によるカテゴリ分類結果と比較することで性能を評価した．

データセットの準備手順は以下の通りである．まず，Filter モジュールで Malicious と判定されたチャンネルをランダムにサンプリングし，合計 100 チャンネルを抽出した．そして，Telegram アプリケーションで各チャンネルのタイトル，概要文，メッセージ内容を確認し，3.1 節で定義した「サイバー犯罪のカテゴリ」を参考に，一人の研究者が手動で各チャンネルのカテゴリを判断し，それを正解データとした．

準備した 100 個のチャンネルを Categorize モジュールの LLM に入力し，出力されたカテゴリと正解データを比較して性能を評価した．

6.2 評価結果

6.2.1 Filter モジュールの評価

表 4: Filter モジュールの性能指標 (レベル別)

レベル	Precision	Recall	F1 Score	Accuracy	Support
Benign	0.95	0.63	0.76	0.63	30
Suspicious	0.44	0.27	0.33	0.27	30
Malicious	0.56	0.97	0.71	0.97	30
macro avg	0.65	0.62	0.60	0.62	90

表 4 は，フィルタリングの性能指標をまとめたものである．Malicious と判断されたチャンネルの Recall は 0.97 と高い一方で，Precision は 0.56 と低いことがわかる．また，Suspicious と判断されたチャンネルの性能指標は全体的に低い．これは，LLM がサイバー犯罪の基準を人間よりも緩く設定している可能性を示している．一方で，Recall の値が 0.97 と高いことから，サイバー犯罪関連コンテンツを含むチャンネルの見逃しは少ないことがわかる．

6.2.2 Categorize モジュールの評価

表 5 は，LLM によるカテゴリ分類の性能評価結果である．今回の実験結果では，サイバー犯罪カテゴリ全体の性能は，F1 スコアに基づいて以下の 3 つのグループに分類できる．

- **高パフォーマンスグループ** $F1 \geq 0.75$**:** ポルノコンテンツ (C1)，違法ギャンブル・オンラインゲーム (C5)，スパムおよび関連する脅威 (C6)，著作権関連の犯罪 (D1) が含まれる．これらのカテゴリは，LLM が比較的高い精度で分類できている．特に，ポルノコンテンツ (C1) は Precision と Recall が 1.00 であり，評価した 4 つのサンプル全てを正しく判定できている．
- **中程度のパフォーマンスグループ** $0.35 \leq F1 < 0.75$**:** 不正データ取得 (A2)，コンピュータ関連詐欺 (B2)，個人情報関連犯罪 (B3)，宗教的犯罪 (C3)，サイバーいじめ (C4)，人種差別・ヘイトスピーチ (C7)，フィッシング (E1)，サイバーロンダリング (E2)，サイバー戦争 (E3)，インターネットのテロリスト利用 (E4) が含まれる．これらのカテゴリは，LLM の性能がカテゴリによって異なり，改善の余地がある．
- **低パフォーマンスグループ** $F1 \leq 0.35$**:** 不正アクセス (A1)，システム妨害 (A5)，デバイスの悪用 (A6)，コンピュータ関連偽造 (B1) が含まれる．これらのカテゴリは，LLM の分類精度が低く，改善が必要である．特に，不正アクセス (A1) とコンピュータ関連偽造 (B1) は，F1 スコアが 0.00 であり，全く分類できていない．

表 5: サイバー犯罪カテゴリ別性能評価結果

Category	**Precision**	**Recall**	**F1 Score**	**Support**
不正アクセス (A1)	0.00	0.00	0.00	2
不正データ取得 (A2)	0.50	0.55	0.52	11
不正傍受 (A3)	-	-	-	0
データ妨害 (A4)	-	-	-	0
システム妨害 (A5)	0.22	0.67	0.33	3
デバイスの悪用 (A6)	1.00	0.19	0.32	16
コンピュータ関連偽造 (B1)	0.00	0.00	0.00	1
コンピュータ関連詐欺 (B2)	0.29	0.67	0.40	3
個人情報関連犯罪 (B3)	0.44	0.50	0.47	8
ポルノコンテンツ (C1)	1.00	1.00	1.00	4
児童ポルノ (C2)	-	-	-	0
宗教的犯罪 (C3)	0.33	1.00	0.50	1
サイバーいじめ (C4)	0.50	1.00	0.67	3
違法ギャンブル・オンラインゲーム (C5)	0.83	1.00	0.91	5
スパムおよび関連する脅威 (C6)	0.67	1.00	0.80	2
人種差別・ヘイトスピーチ (C7)	1.00	0.47	0.64	15
著作権関連の犯罪 (D1)	0.88	1.00	0.93	7
商標関連の犯罪 (D2)	-	-	-	0
フィッシング (E1)	0.25	1.00	0.40	1
サイバーロンダリング (E2)	0.33	0.50	0.40	4
サイバー戦争 (E3)	0.67	0.40	0.50	5
インターネットのテロリスト利用 (E4)	0.80	0.44	0.57	9
accuracy			0.55	100
macro avg	0.44	0.52	0.48	100
weighted avg	0.71	0.55	0.62	100

ただし，コンピュータ関連偽造 (B1) やフィッシング（E1）などのカテゴリでは，評価データセットのサンプル数が非常に少ないことに注意が必要である．そのため，これらのカテゴリの F1 スコアは，モデルの性能を正確に反映していない可能性がある．今回の評価では，不正傍受（A3），データ妨害（A4），児童ポルノ (C2)，商標関連の犯罪（D2）に該当するチャンネルは発見できなかったため，評価データセットには含まれていない．

これらの結果から，以下の特徴が明らかになった．ポルノコンテンツ（C1），違法ギャンブル・オンラインゲーム (C5)，スパムおよび関連する脅威 (C6)，著作権関連の犯罪 (D1) の 4 つのカテゴリにおいて高い性能を示している．これは，カテゴリが特徴的な用語や表現パターンを持っていることによっ

て，LLM がより正確な分類を行えたためと考えられる．一方で，不正データ取得 (A2) やデバイスの悪用（A6），個人情報関連犯罪 (B3)，人種差別・ヘイトスピーチ (C7)，インターネットのテロリスト利用 (E4) など，技術的な文脈を必要とするカテゴリや，定義が曖昧で判断が難しいカテゴリでは，低いまたは中程度の性能となっている．

これらの分析結果から，現状の LLM による分類システムは，ポルノコンテンツや違法ギャンブルなど，特徴的な用語や表現を含むカテゴリについては信頼性の高い分類が可能である一方，技術的な専門性を要するカテゴリや，複雑な文脈理解が必要なカテゴリについては，さらなる改善が必要であることが示唆された．

7. ケーススタディ

本章では，前章までの分析結果を踏まえ，特に観測頻度が高く，特徴的な事例が見られたカテゴリに焦点を当て，具体的な事例を分析する．紙面の都合上，全てのカテゴリを詳細に分析することは困難であるため，本研究では 5.2 節の分析において 10 件以上のチャンネルが観測されたカテゴリに絞ってケーススタディを行う．

7.1 カテゴリ A：データとシステムの機密性，完全性，および可用性に対する犯罪

不正データ取得 (A2)：このカテゴリでは，窃取したとされるデータの販売に関する投稿が多く観測された．例えば，「パキスタンの企業のデータベースの 1 億 3000 万件の記録を販売」や「アメリカのショッピングサイトへの不正アクセスで得たデータを数千ドルで販売」といった投稿では，具体的なデータ内容や価格を提示することで，取引を誘引しようとする意図が見られた．また，ハクティビストグループを自称するアカウントによる投稿も確認され，「アゼルバイジャンの政府機関からのデータリーク」や「イスラエルの情報機関の機密データリーク」といった投稿は，政治的な動機に基づいた情報漏洩活動を主張するものであった．さらに，「インドネシア警察のデータベースに対する SQL 脆弱性」のような脆弱性を悪用したデータ窃取の手法に関する情報共有も確認された．

システム妨害 (A5)：このカテゴリでは，DDoS 攻撃を通じてシステムの正常な機能を妨害する活動に関する投稿が観測された．無料の DDoS 攻撃ツールを提供し，その使用方法を解説するメッセージや，特定の政党を批判し，そのウェブサイトを攻撃する意図を表明するメッセージが見られた．これらの投稿は，攻撃ツールや攻撃対象の情報を提示することで，サイバー攻撃を扇動する意図を持つものと考えられる．

デバイスの悪用 (A6)：このカテゴリでは，DDoS 攻撃ツール，ゲームチート，ボットネットなど，サイバー犯罪に利用される可能性のあるツールやサービスの取引に関する投稿が観測された．例えば，特定のファイアウォールサービスをバイパスする方法，Rainbow Six Siege などの人気ゲームのチートツール，DDoS 攻撃に利用可能とされるボットネットの販売などに関する情報が確認された．これらの投稿では，ツールの機能や価格が具体的に示され，購買意欲を促す内容となっていた．

7.2 カテゴリ B：コンピュータ関連の犯罪

コンピュータ関連の詐欺 (B2)：このカテゴリでは，仮想通貨に関連した詐欺的な投稿が多く観測された．例えば，「PUMP NOW! BUY NOW!」という文言で特定の仮想通貨の購入を促し，価格操作を試みる投稿や，「Invest ₹537 - ₹22 Per/Day」のように，少額投資で毎日の利益を保証すると謳う投稿が確認された．これらの投稿は，高額な利益や配当を強調することで，ユーザーを誘引する意図を持つものと考えられる．

個人情報関連犯罪 (B3)：このカテゴリでは，個人情報の不正取得・利用や ID 偽造に関連する情報の取引に関する投稿が観測された．例として，「ウクライナの退役軍人省の職員のパソコンから個人情報が流出」という主張や，「アメリカの医師のデータベース」や「インドの運転免許証情報」といった情報の取引に関する投稿が確認された．これらの情報は，個人情報の不正利用や ID 偽造に悪用される可能性が懸念される．

7.3 カテゴリ C：コンテンツ関連の犯罪

人種差別・ヘイトスピーチ (C7)：このカテゴリでは，特定の民族，人種，宗教，性的指向などに対する差別的な発言や，憎悪を煽る投稿が観測された．例えば，特定の民族を指す侮辱的な中傷や，ナチスに関連する表現，また LGBTQ+ などの性的マイノリティを標的にした敵対的な発言が確認された．これらの投稿は，排他的な感情を助長し，特定のグループに対する差別や憎悪を煽る目的を持つものと考えられる．

7.4 カテゴリ D：知的財産権侵害

著作権関連の犯罪 (D1)：このカテゴリでは，著作権で保護されたコンテンツの無許可配布に関する投稿が観測された．例えば，映画の違法アップロードを促す投稿や，「Adobe Photoshop CC 2025」のような高額ソフトウェアを無料でダウンロードできると宣伝する投稿が見られた．また，著作権侵害された書籍を利用するためのファイル変換ツールと見られる投稿も確認された．これらの投稿は，著作権侵害行為を促進する意図を持つものと考えられる．

7.5 カテゴリ E：複合的な犯罪

サイバー戦争 (E3)：このカテゴリでは，国家や組織を標的としたサイバー攻撃の計画や実行，攻撃ツールの共有，ハクティビスト活動への参加呼びかけなど，サイバー空間での紛争行為に関する情報交換が観測された．

国家間の紛争に関連する投稿として，イスラエルとイラン間のサイバー攻撃に関する情報が多く見られた．”The road is closed”というメッセージとともに，イスラエルの軍事基地を攻撃したと主張する動画が共有されており，サイバー攻撃が物理的な攻撃と連動している可能性が示唆された．また，ハクティビストグループを名乗るアカウントが特定の国のネットワークに侵入し，機密情報を窃取したと主張する投稿も観察された．

8. 結　　論

本研究では，Telegram におけるサイバー犯罪関連チャンネルを効率的に発見・分析するためのシステムを提案した．Similar Channels 機能で発見されたチャンネルを LLM でフィルタリン

グして新たなシードとして利用することで，無関係なチャンネルを増やすことなく効率的に監視対象チャンネルを拡大できることが示された．この結果は，本手法が少数のシードから効率的にサイバー犯罪関連チャンネルを発見できる可能性を示唆している．一方，LLM を用いたカテゴリ分類に関しては，特徴的な表現パターンを持つカテゴリでは高い性能を示したものの，技術的な内容の理解や定義が曖昧なカテゴリの分類には課題が残ることが判明した．

収集されたチャンネルの分析からは，Telegram 上でのサイバー犯罪に関連する情報共有の実態が明らかになった．特に，デバイスの悪用や不正データ取得に関連するチャンネルが多数を占め，これらのチャンネルでは具体的な価格や手法が提示されながら取引が呼びかけられていた．また，国家間の紛争に関連したサイバー攻撃の計画や実行に関する情報共有も確認され，Telegram が現実世界の紛争とサイバー空間での攻撃を結びつける場として機能している可能性が示唆された．

本研究は，Telegram 上でのサイバー犯罪の検出と分析における新たなアプローチを提示した．提案システムは，サイバー犯罪の兆候を早期に捉え，効果的な対策を講じる上で重要な役割を果たす可能性がある．

今後の課題として，以下の点が挙げられる．第一に，LLM を用いたフィルタリングおよびカテゴリ分類の精度向上である．特に，技術的な内容を含むカテゴリの分類精度を改善する必要がある．第二に，データセットの拡充である．一部のカテゴリでは評価用サンプル数が少なく，より多くのデータを収集して評価を行う必要がある．第三に，異なる LLM の比較検討である．本研究では単一の LLM のみを使用したが，複数の LLM を比較することで，より適切なモデルの選択が可能になると考えられる．

また，本論文ではチャンネルの探索を 2 回のみ実施したが，フィルタリングされたチャンネルを新たなシードとして探索を繰り返すことで，より多くのサイバー犯罪関連チャンネルを発見できる可能性がある．探索回数と発見されるチャンネル数の関係性を分析することで，本手法の有効性をより詳細に評価できると考えられる．

本研究で提案した手法は，他のソーシャルメディアプラットフォームにも応用できる可能性がある．今後は，より広範な適用を通じて，サイバー犯罪対策のための包括的なモニタリングシステムの開発を目指す．

9. 倫理的考慮

本研究は，Telegram 上のサイバー犯罪関連チャンネルを対象としているため，特にデータ収集と LLM の運用において，倫理的な側面について慎重な検討を行った．本章では，これらの側面に焦点を当て，倫理的な課題と対策について述べる．

まず，データ収集に関しては，プライバシー保護が最も重要な課題である．本研究では，公開されている Telegram チャンネルのデータのみを対象としているが，サイバー犯罪関連チャンネルでは，流出したとされる個人情報が画像や zip ファイルなどの形式で共有されることが多い．これらのファイルを不用意にダウンロードすることは，プライバシー侵害のリスクを高めるだけでなく，研究者自身が法的問題に巻き込まれる可能性もある．そのため，本研究では，メッセージのテキスト情報のみを収集対象とし，画像や zip ファイルなどのダウンロードは行わないこととした．この方針により，個人情報を含む可能性のあるデータの収集を最小限に抑え，プライバシー侵害のリスクを低減している．

次に，LLM の運用においては，情報漏洩のリスクについて考慮する必要がある．ChatGPT のようなクラウドベースの LLM サービスを利用する場合，入力データが外部サーバーに送信されるため，機密情報が漏洩するリスクがある．特に，本研究で扱うデータには，サイバー犯罪に関連する情報が含まれているため，情報漏洩は重大な問題につながる可能性がある．このリスクを回避するため，本研究ではローカルで動作する LLM である Llama 3.3 を採用した．ローカル LLM を使用することで，データを外部に送信することなく分析を行うことができ，情報漏洩のリスクを最小限に抑えている．これらの対策を通じて，本研究が倫理的に責任ある形で実施され，社会に貢献することを目指す．

謝辞：この成果の一部は，国立研究開発法人新エネルギー・産業技術総合開発機構（ＮＥＤＯ）の委託業務（JPNP24003）の結果得られたものです．

文　献

[1] Telegram. Telegram messenger. https://telegram.org/.
[2] TGStat. Telegram channels and groups catalog. https://tgstat.com/.
[3] Telegram. Similar channels. https://core.telegram.org/api/recommend.
[4] Lonami. Telethon’s documentation. https://docs.telethon.dev/en/stable/.
[5] Ollama. llama3.3. https://ollama.com/library/llama3.3.
[6] Christina Curley, Eugenia Siapera, and John Carthy. Covid-19 protesters and the far right on telegram: Co-conspirators or accidental bedfellows? *Social Media + Society*, Vol. 8, No. 4, 2022. Article number: 20563051221134523.
[7] Saifeldeen Zihiri, Gabriel Lima, Jiyoung Han, Jiyoung Han, Meeyoung Cha, and Wonjae Lee. Qanon shifts into the mainstream, remains a far-right ally. *Heliyon*, Vol. 8, No. 2, p. e08764, 2022.
[8] Matteo Vergani, Aylet Martinez Arranz, Ryan Scrivens, and Laura Orellana. Hate speech in a telegram conspiracy channel during the first year of the covid-19 pandemic. *Social Media + Society*, Vol. 8, No. 4, 2022.
[9] George Tsakalidis and Kostas Vergidis. A systematic approach toward description and classification of cybercrime incidents. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, Vol. 49, No. 4, pp. 730–742, 2019.
[10] Council of Europe. Convention on cybercrime, 2001. ETS No. 185.