SCIS 2024 2024 Symposium on
Cryptography and Information Security
Nagasaki, Japan, Jan. 23 – 26, 2024
The Institute of Electronics,
Information and Communication Engineers

Discord 上のサイバー犯罪に対する ChatGPT を利用した情報収集システム
ChatGPT Assisted Information Collection System for Cybercrime on Discord

川口 大翔 * Yamato Kawaguchi
Yin Minn Pa Pa † Yin Minn Pa Pa
吉岡 克成 ‡ Katsunari Yoshioka
松本 勉 ‡ Tsutomu Matsumoto

あらまし 近年，SNS を利用した詐欺や犯罪が増加している．特に Discord というプラットフォームでは，多くのユーザと簡単に交流することができ，若い世代を中心に利用者数が増加している．しかし，認証情報や違法薬物の販売，詐欺といった犯罪が行われているケースがあり，Discord 運営による利用規約違反のサーバの削除件数も増加している．だが，招待リンクがインターネット上の様々な場所に存在することや，非公開グループへのアクセスが複雑であること，メッセージ量が多く，スラングの使用により悪質なコンテンツの特定が難しいという問題があり，Discord で行われている悪質な活動の特定が難しい．これらの問題に対処するために，本研究では Discord 上のサイバー犯罪に対する情報収集システムを提案する．このシステムは外部ソースから招待リンクを収集し，ChatGPT を利用して情報が収集対象であるかを判断し，メッセージを収集する．このシステムで収集した情報を分析することで Discord 内の犯罪の実態を調査する．

キーワード ChatGPT，Discord，サイバー犯罪

1 はじめに

近年、SNS (ソーシャル・ネットワーキング・サービス) を使った詐欺や犯罪が増加している。特に Discord のような SNS では多くのユーザーに利用され、特定のコミュニティが多く存在し、ボットなどの様々なツールが提供されているためサイバー犯罪者にとって都合の良い環境となっている。

Discord は，メッセージやビデオ通話，音声通話でコミュニケーションを取れる SNS である．Discord ではサーバというスペースを作成することができ，ゲームなどのコミュニティを無料で作成することができる．そのサーバの中にはチャンネルを複数作成することができ，その中でテキストで会話したり音声で通話したり，ダイレクトメッセージ機能で 1 対 1 で会話したりすることができる．他の SNS と比べて気軽にテキストや音声などで暗号化された匿名のコミュニケーションを取れることや，サーバの招待やアクセス権を自由にカスタマイズできること，サーバをよりカスタマイズできるボットというツールが多く開発されているという特徴がある．2021 年時点では Discord の月間アクティブユーザ数は 1 億 5000 万人であり [1]，2022 年 5 月時点では 16 歳から 24 歳のユーザが全体の 22.2%，25 歳から 34 歳までのユーザが全体の 42%以上を占めている [2].

サーバには公開サーバと非公開サーバがある．公開サーバは公式のサーバ検索サイトで自由に検索し参加できるのに対して，非公開サーバは招待リンクがないと参加できない．これらの招待リンクは非公式のサーバ検索サイトや別の SNS，Web 上に掲載されていることがあり，そこから参加することができる．招待リンクは有効期限の設定や無効化が簡単にできるため，サーバへのユーザの流入をコントロールすることができる．

このように Discord は非常に便利なコミュニケーションツールであるが，その反面，犯罪に使われるなどの問題がある．2023 年 4 月には Discord のサーバ上で米国の機密文章を流出したという事件が発生している [3]．それに対して Discord 運営者もサーバやアカウントの削除によって対策している．利用規約違反に関するサーバ削除件数は年々増加しており，2022 年には 132,067 件のサーバが運営側によって削除された．運営側に削除されたサーバに参加していたユーザに対しても警告や処分が

* 横浜国立大学大学院環境情報学府, Graduate School of Environment and Information Sciences, Yokohama National University
† 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
‡ 横浜国立大学大学院環境情報研究院/先端科学高等研究院, Faculty of Environment and Information Sciences, Yokohama National University / Institute of Advanced Sciences, Yokohama National University

行われる場合がある．だが現状では利用規約違反と思われるサーバが多く存在する．

Discord における情報収集では，いくつか難しい点がある．Discord ではサーバを無料で簡単に作成でき，statista によると 2023 年の月間アクティブサーバ数は約 1,900 万件と推定されているため [4]，多くのサーバの中から必要な情報を取得する必要がある．サーバ名やメッセージではスラングが使われていることが多いので，それらのワードを検知する必要もある．また，非公開サーバという招待制のサーバがあるため，そのサーバに繋がる招待リンクを収集しないと中のメッセージを閲覧することができない．さらにサーバに参加しても，人間であることを証明するためにメッセージへのリアクションや Bot による検証が必要な場合があり，情報収集を完全に自動化することが難しい．

それらの問題を考慮して，本研究では Discord の情報収集システムを提案する．このシステムは，ユーザが取得したい情報のカテゴリ (例: Malicious hacking) とそれに関連するカテゴリ，取得したくないカテゴリを指定することで，ユーザの目的にあった情報を収集できる．その目的に従って ChatGPT によって検索キーワードを生成する．サーバ情報は公式と非公式のサーバ検索サイトから自動的に検索して招待リンクを収集する．その招待リンクを ChatGPT のフィルタに通すことで目的に近いサーバのみを情報収集対象とし，サーバ参加処理に関するユーザの負担を減らす．また，参加したサーバから収集したメッセージから新たなサーバの招待リンクを収集し，これに参加する．最後にメッセージの情報から ChatGPT を使って目的のサーバを特定し，ユーザに分析結果を提示する．本システムはサーバ情報を収集してリストアップするものであり，入力したカテゴリに関連する Discord 上のサーバを広く見つけることを目指す．

取得したいカテゴリとして “Drug” “Malicious hacking”, “Leak of confidential files” “Market for credit card, authentication information, hacking tools and drugs”, “Scams”，それに関連するカテゴリとして “Giveaway”, “Ethical hacking including unsuspected activities” を指定し，当該システムによる情報収集を行った結果，関連するサーバ情報を 8,476 件取得できた．また，取得したサーバ情報のうちランダムに 100 件のサーバに参加し，55 件のサーバから 1,280,548 件のメッセージを取得した．収集したサーバ情報から参加するサーバ情報を選択する検出器では，正確度が GPT3.5 では 0.765，GPT4 では 0.860 であった．収集したメッセージからシステムが出力するサーバを選択する検出器では，出力すると判断した 30 件のサーバのうち，28 件に取得対象に関連するメッセージが含まれていることを確認できた．また，収集したサーバの中からクレジットカード情報の販売や盗んだ認証情報の共有といった悪意のある事例を発見することができた．

2 関連研究

Discord における詐欺や犯罪に関する研究が報告されている．Sanden ら [5] は，薬物売買のサーバについて利用者にインタビューを行い，薬物売買のサーバには lower tier と higher tier の 2 種類に分けられることを明らかにした．Nizzoli ら [6] は，Twitter に投稿された Discord と Telegram の招待リンクを収集し，それらのサーバに参加することで仮想通貨詐欺の調査を行っている．それらの招待リンクはネットワークの構造を形成していて，サーバ内のメッセージを分析することで 2 つの詐欺が行われていることを明らかにした．Heslep ら [7] は，Disboard というサーバ検索サイトからタグを使って人種差別に関連するサーバの説明文を収集し，人種差別に関するサーバが数千件あることを明らかにした．

関連研究では特定の悪意のある活動について調査が行われていたが，本研究ではより広い範囲での悪意のある活動について調査を行った．それを実現するために，汎用的に目的の Discord サーバを収集するシステムを提案する．

3 調査手法

情報収集システムは，ユーザによる入力，キーワードの生成，サーバ情報の収集，サーバへの参加メッセージの収集，メッセージの分析の 6 つのステップに分かれている．ユーザは取得対象のカテゴリ，取得対象に関連するカテゴリ，取得対象ではないカテゴリをシステムに入力し，最終的にシステムは取得対象であるサーバの名前，説明，招待リンク，取得対象のカテゴリに当てはまるメッセージのリスト，判断の理由を出力する．図 1 にシステムの全体図を示す．

3.1 ユーザの入力

ユーザはシステムに，取得対象のカテゴリと取得対象に関連するカテゴリ，取得対象ではないカテゴリを指定する．取得対象のサーバでなくても，そのサーバのメッセージに含まれる招待リンクから取得対象のサーバに参加できる場合があり，取得対象に関連するカテゴリを指定することで探索の幅を広げる．また，取得対象ではないカテゴリを指定することで，取得対象でないサーバが ChatGPT によって取得対象であると誤検知されてしまうことを抑制する．システム使用中に取得対象が変化することやより具体的になった場合は，この入力を修正することでユーザの目的に柔軟に対応できる．今回システムに入力したカテゴリを表 1 に示す．

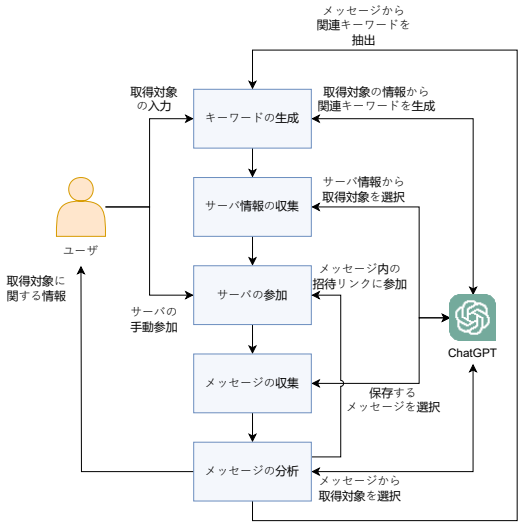

図 1: システムの全体図

表 1: 取得するサーバのカテゴリ

カテゴリ	該当するカテゴリ
対象	“Drugs”, “Malicious hacking”, “Leak of confidential files”, “Market for credit card, authentication information, hacking tools and drugs”, “Scams”
関連	“Giveaway”, “Ethical hacking including unsuspected activities”
対象外	“Porn”, “Rollplaying (sometimes called RP)”

3.2 キーワードの生成

サーバ検索時やターゲットに関連するメッセージを抽出するときにキーワードが必要になるため，ユーザの入力したカテゴリ情報を元に，カテゴリに関連するキーワードを生成する．このキーワードはサーバの検索の精度に大きく関わっているため，ChatGPT が生成した後にユーザが追加や削除などの変更を加えることができる．今回は事前調査の段階で ChatGPT を部分的に使用したキーワードリストを作成していたため，それを利用している．本研究で利用している 196 個のキーワードの一部を Listing 1 に示す．

Listing 1: キーワードリストの一部

```
0day, advanced persistent threat, adware,
amplification attack, antivirus evasion,
backdoor, bad rabbit, bait and switch,
baiting, banklog, ...
```

3.3 サーバ情報の取得

サーバ情報は，公式のサーバ検索サイト，非公式のサーバ検索サイト，Twitter や Telegram などの他の SNS，既に参加している Discord サーバのメッセージ，Web サイト，Google 検索から取得可能である．サーバ検索サイトは公式のサイトが 1 つと非公式のサイトが複数あり，他の情報源と比べて Discord のサーバを効率的に収集できるため，今回は公式のサイト 1 件と非公式のサイト 4 件からサーバ情報を取得する．取得するサーバ検索サイトは掲載数の多さや自動取得可能である点から選んでいる．サーバ検索サイトでは，キーワードの生成のステップで作成したターゲットに関連するキーワードでサーバを検索し，掲載されているサーバ名や説明，サーバへの招待リンクを取得する．また，その招待リンクが有効であるかを Discord API [8] を使って調査する．取得したサーバ情報は，ChatGPT を使ってユーザが指定した取得対象のサーバに絞る．サーバ情報には特殊な文字列を使っている場合やスラングを含んでいる場合が多く，何をしているサーバなのかを文章から読み取る必要があるため，大規模言語モデルによる処理が適していると考えられる．今回は OpenAI API [9] の ChatGPT を利用することで，大規模言語モデルを使用している．本システムでは基本的に GPT3.5 を利用していて，モデル名は gpt-3.5-turbo-1106 である．また，部分的に評価で GPT4 を利用していて，モデル名は gpt-4-1106-preview である．API 利用時は出力値が固定になるように，出力値を操作する値である temperature, top_p, seed の値をすべて 0 にしている．出力値は必ず JSON 形式になるように，response_format の type を json_object と指定している．そのプロンプトをプロンプト 1 に示す．${target_activities} には取得対象のカテゴリとそれに関連するカテゴリが，${non_target_activities} には取得対象でないカテゴリが入力され，サーバ情報は${name} と ${description} に入力される．

3.4 サーバへの参加

Discord サーバへの参加では，参加時に特定のメッセージにリアクションを押すことや，ボットによる検証，外部の Web サイトでの検証が必要な場合があり，自動化が難しいためユーザが手動で行う．

3.5 メッセージの収集

メッセージは discord.py-self [10] という Python ライブラリを使用して取得する．このライブラリでは，Discord アカウントの認証トークンを使用することで，そのアカウントが参加しているサーバ情報やサーバ内のメッセージ情報を取得できる．このシステムでは，ユーザがサーバ参加すると過去のメッセージすべて取得し，それ

プロンプト 1: サーバ検索サイトから取得したサーバ情報が取得対象であるか判断するプロンプト

```
You are a security expert. Please answer
wheter the server is related to the activity
in the example based on the name and
description of the Discord server.

### Limitations ###
The response must not contain any description
. Output should be in RC8259 JSON format with
 following keys:
- is_target: bool (whether the server is
related to the activities in the example)
- reason: str (why this server is related to
the activities in the example)

### Example of activities ###
${target_activities}

### Example of not targeted activities ###
${non_target_activities}

### Input ###
Name: ${name}
Description: ${description}
```

以降もサーバ内でメッセージが送信される度にメッセージを取得する．取得したメッセージは，メッセージ内に含まれる URL や IP アドレス，電話番号，メールアドレスなどの外部リンクを自動的に分析し，投稿されたサーバとチャンネルの情報と併せてデータベースに保存する．

参加するサーバの数が多くなるにつれ，取得するメッセージ数も多くなるため，ChatGPT によって必要なメッセージのみを保存できるようにする．これによって目的にあったメッセージのみを効率的に収集できる．

3.6 メッセージの分析

収集したサーバ内のメッセージから，そのサーバがユーザが入力した取得対象であるかを ChatGPT を使って判断する．保存されているサーバ内のメッセージの中で，キーワードの生成のステップで作成したキーワードが含まれているメッセージを取得する．ChatGPT にはトークン数の制限があるため，それらのメッセージが GPT3.5 では 16,385 トークン以内，GPT4 では 128,000 トークン以内に収まるようにメッセージ数を制限し，ChatGPT を使ってメッセージが取得対象のカテゴリに属するかとサーバが取得対象であるかを判断する．サーバ情報の取得時と同様に，出力値が固定であり，必ず JSON 形式になるように設定している．GPT4 ではより多くのトークン数を利用できるため，gpt-4-1106-preview を使用した．サーバ内のメッセージからサーバが取得対象であるかを判断するプロンプトをプロンプト 2 に示す．${target_activities} には取得対象のカテゴリが，${non_target_activities} には取得対象に関連するカテゴリと取得対象でないカテゴリが入力され，サーバ情報は${name} と${description}, ${messages} に入力される．

プロンプト 2: メッセージからサーバが取得対象であるか判断するプロンプト

```
You are a security expert. Please answer
wheter the server is related to the activity
as an example based on the messages in the
server.

### Limitations ###
The response must not contain any description
. Output should be in RC8259 JSON format with
 following keys:
- is_target: bool (whether the server is
related to the activities in the example)
- reason: str (why this server is related to
the activities in the example)
${additional_keys}

### Example of activities ###
${target_activities}

### Example of not targeted activities ###
${non_target_activities}

### Input ###
Name: ${name}
Description: ${description}
Messages:
${messages}
```

ChatGPT によりサーバが取得対象であると判断されると，システムは

- サーバ名
- サーバの説明
- サーバの招待リンク
- 取得するカテゴリに属するメッセージのリスト

- 取得対象であるという判断の理由

の情報をユーザに出力する．

メッセージに招待リンクが含まれている場合は，ChatGPT で取得対象のサーバであるか判断し，ユーザによって招待先のサーバに手動で参加される．これによって関連するサーバに参加することができ，誰でも閲覧できる場所に公開されていないようなサーバに参加できる可能性がある．

メッセージに含まれる取得対象のカテゴリに関連する単語を取得し，その単語をキーワードリストに加える．これによって，より多くのキーワードでサーバ検索やメッセージの取得を行うことができる..

4　調査結果

4.1　システムからの取得結果

Discord のサーバ情報を 2023/10/13-16 の期間に，提案システムに対して “Drug” “Malicious hacking”, “Leak of confidential files” “Market for credit card, authentication information, hacking tools and drugs”, “Scams” を対象カテゴリ，“Giveaway”, “Ethical hacking including unsuspected activities” を関連カテゴリ，“Porn”, “Rollplaying (sometimes called RP)” を対象外のカテゴリとして提示した．公式のサーバ検索サイト 1 件と非公式のサーバ検索サイト 4 件から収集した結果，サーバ情報が 8,476 件，重複を除いた招待リンクとして 7,908 件が収集できた．それらの招待リンクが使えるかどうかを 2023/12/10-11 の期間に調査したところ，有効な招待リンクは 5,718 件であり，それらのリンクが招待しているサーバ数は 5,601 件あった．そのうち GPT3.5 を使って 2023/12/9-10 の期間にサーバが取得対象であるか判断したところ，取得対象と判断されたサーバ数は 711 件あった．検索に使うキーワードは英語のみだが，別の言語が使われているサーバも取得できた．これはサーバ情報は英語で書いているが中身は別の言語が使われているといったように，別の言語と英語が併記されている場合があるためだと考えられる．

2023/12/11 に取得対象と判断されたサーバのうち 101 件に参加を試みて，100 件のサーバに参加でき，1 件のサーバでは招待リンクの有効期限切れにより参加に失敗した．参加したサーバのうち，89 件は問題なく参加でき，残りの 11 件は参加できたがサーバ内の認証に失敗し，一部のチャンネルしか閲覧することができなかった．参加時の検証は，電話番号の検証を済ませた Discord アカウントでサーバのルールに同意するものが一番多く，続いて特定のメッセージにリアクションのボタンを押す，ボットによって生成されたメッセージ内のボタンを押すものが多かった．複雑なものでは，サーバからの DM 受信設定をオフにして，ルールを読んで同意し，ボットから送信された画像内の数字を入力するという検証があった．

参加したサーバのうち 55 件において，2023/1/1-2023/12/12 の期間に送信されたメッセージ全ての収集を試みたところ，合計 1,280,548 件取得できた．メッセージには URL が合計で 17,728 件含まれており，そのうち Telegram の招待リンクが 71 件あった．

これらのメッセージを使ってサーバが取得対象であるかどうか ChatGPT を使って判断した．その結果，参加した 55 件のサーバのうち 28 件のサーバが取得対象であると判定された．それらのサーバには，クレジットカード情報の販売や盗んだ認証情報の共有という種類のものがあり，詳細はケーススタディで説明する．

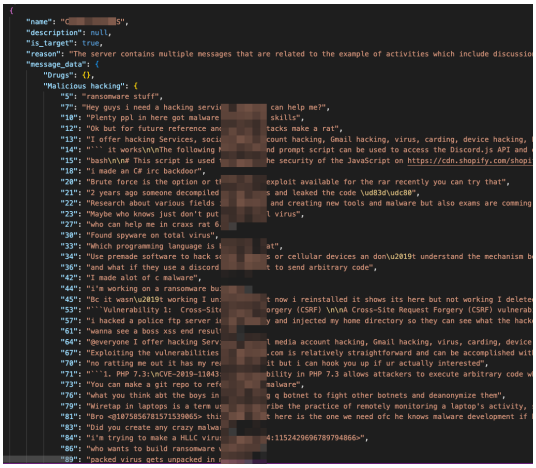

図 2: Malicious Hacking に関連するシステムの出力

4.2　Discord サーバのケーススタディ

本システムで取得されたサーバ情報やメッセージ情報から，サイバー犯罪に関連する事例について追加調査した．その結果，悪質な可能性の高いサーバをいくつか発見した．

クレジットカード情報の販売 このサーバではクレジットカードやデビットカード情報を販売していて，管理者にダイレクトメッセージを送信することでビットコインを使って購入できる．サーバ内にはサンプルとして一部が隠されたクレジットカード情報の画像載っていて，販売の信頼性を挙げている (図 3)．掲載した画像では，元から書かれている黒い線に加えて，モザイクを加えている．

盗んだ認証情報の共有 このサーバでは SilverBullet というアカウントチェッカーを利用して，漏洩したアカウント情報を使い，別のサイトからアカウント情報を盗んでいる．アカウントチェッカーではチェックするアカウント情報と Web サイトごとの設定が必要であり，これらの情報が共有や売買されている履歴があった．また，ここでは PayPal や Xbox，Azure，Outlook などのアカ

図 3: クレジットカード情報の一部を載せている投稿

ウント情報が共有されていて，特に PayPal のアカウントは 600 個以上共有されていた (図 4)．掲載画像にはモザイクを加えている．

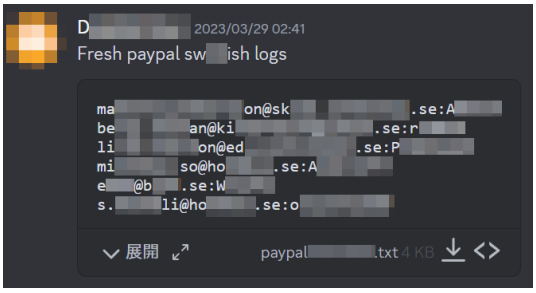

図 4: PayPal の認証情報を共有している投稿

アカウント情報の販売とハッキングサービス このサーバではゲームや SNS アカウント情報の販売と SNS アカウント等のハッキングサービスを行っている．アカウントは，勝率が高く Ban されていないゲームアカウントや，収益化されていたりフォロワーが多い SNS アカウントアカウントの販売があった．ハッキングサービスは，一般的な SNS アカウントに対してリクエスト可能であると記載されている．これらの販売やサービスは，Discord サーバ内で管理者にダイレクトメッセージを送信するか，Bot によってチケットを発行し，自動作成されたプライベートチャンネルで管理者と連絡することによって購入できる．

ゲームのハッキングツール販売 このサーバではゲームのハッキングツールの販売を行っている．1 日$3 から永続で$50 までのツールや，1 日$20 から永続で$150 までのツールなど，複数のツールと購入プランが用意されていて，これらのツールは Web サイト上で購入できる．これらのツールは定期的に更新されており，Discord サーバでサポートを受けることもできる．また，このサーバではサーバが違反により削除される可能性について話しており，仮にサーバやアカウントが削除されたとしても代替のサーバとアカウントが作成可能であるため，それほど問題ないということを発言していた．

招待リンクのみ置かれているサーバ サーバ内にはチャンネルが 1，2 個だけあり，チャンネル内には別のチャンネルの招待リンクが置かれている．これはサーバ検索サイトから直接本体のサーバに参加させずに，一度招待リンクが置かれているだけのサーバを挟むことによって，本体のサーバが削除される可能性を下げているのではないかと考えられる．また，招待リンクが置かれているサーバ自体は削除される可能性が低く，招待リンクのみを更新することで，外部からの訪問者を最新のサーバに参加させることができると考えられる．

4.3 Telegram のケーススタディ

本システムで取得されたメッセージ情報から Telegram の招待リンクを抽出し，どのような Telegram のチャンネルとグループに誘導されるのか追加調査した．

ドラッグの販売 このサーバではいくつかの薬物の販売を行っている．それに加えて，味の良さやおすすめの吸い方，接種すべきときなど商品に関するレビューが投稿されている．購入は管理者にダイレクトメッセージを送信することで行い，支払い方法は XMR, BTC, ETH, DOGE といった仮想通貨に対応しているとの記載がある．

4.4 制限

本報告で説明する実験の時点で以下の機能はまだ実装できていない．

- 取得対象のカテゴリに関連するキーワードを生成する機能
- 保存するメッセージを選択する機能
- 収集したメッセージに含まれる招待リンクから，新しいサーバに参加する機能
- 収集したメッセージに含まれる取得対象のカテゴリに関連するワードを抽出し，キーワードリストに加える機能

5 評価

ChatGPT を使った検出器についての評価を行う．評価 1 では，サーバ情報から取得対象を選択する検出器について，検出器と 2 人のサイバーセキュリティ研究者の判断を比較して評価する．評価 2 では，収集したメッセージからシステムが出力するサーバを選択する検出器について，検出器によって検出されたサーバ情報とその根拠となるメッセージ情報を，2 人の人間によって評価する．

5.1 評価 1: サーバ情報から取得対象を選択する検出器

取得対象のカテゴリとそれに関連するカテゴリとして，“Drug” “Malicious hacking”, “Leak of confidential files” “Market for credit card, authentication information, hacking tools and drugs”, “Scams”, “Giveaway”,

“Ethical hacking including unsuspected activities” を指定している．GPT3.5 を使った検出器が取得対象と判断したサーバ 1,635 件から 100 件，取得対象でないと判断したサーバ 6,840 件から 100 件をランダムに選び，それらをシャッフルして検出器の結果を削除した上で，2 人のサイバーセキュリティ研究者によってサーバが取得対象であるかを判断した．GPT のモデルによる精度の違いを検証するために，GPT3.5 と GPT4 でそれぞれ評価した．OpenAI API の rate limit の関係で，基本的には GPT3.5 を使用し，それに併せて評価では GPT4 を使用している．GPT3.5 による結果を表 2 に，GPT4 による結果を表 3 に示す．

正解値が True で予測値が True の値を TP，正解値が True で予測値が False の値を FN，正解値が False で予測値が True の値を FP，正解値が False で予測値が False の値を TN とする．そのとき，$Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$, $Presision = \frac{TP}{TP+FP}$, $Recall = \frac{TP}{TP+FN}$, $F1 = \frac{2(Precision \cdot Recall)}{Precision+Recall}$ として，小数点第 3 位は切り捨てる．それらの結果を表 4 に示す．GPT3.5 に比べて GPT4 のほうがすべての項目において精度が向上していることがわかる．

表 2: GPT3.5 によるサーバ情報が取得対象であるかの検出器の混同行列

		予測値	
		True	False
正解値	True	69	16
	False	31	84

表 3: GPT4 によるサーバ情報が取得対象であるかの検出器の混同行列

		予測値	
		True	False
正解値	True	71	14
	False	14	101

表 4: サーバ情報が取得対象であるかの検出器の精度評価値

GPT Model	GPT3.5	GPT4
Accuracy	0.765	0.860
Precision	0.690	0.835
Recall	0.811	0.835
F1	0.745	0.835

5.2 評価 2: 収集したメッセージからシステムが出力するサーバを選択する検出器

取得対象のカテゴリとして，“Drug” “Malicious hacking”, “Leak of confidential files” “Market for credit card, authentication information, hacking tools and drugs”, “Scams” を指定している．参加した取得対象の 55 件のサーバにおいて，GPT4 を使ったこの検出器が出力するサーバであると判断したサーバは 30 件あり，その根拠として挙げられたメッセージから取得対象のカテゴリと関連するメッセージについて，その正誤を 2 人の人間によって判断した．その結果，30 件中 28 件のサーバで取得対象のメッセージが見つかり，残りの 2 件で 1 つも取得対象のメッセージが見つからなかった．この検出器には複数のメッセージを入力するため，多くのトークンを入力できる GPT4 のみで評価した．また，この検出器では取得対象とは関係ないサーバの出力を防ぐ目的があるため，取得対象と判断されたサーバのうち，正しく検出できている数と誤検知している数について評価している．

1 件のサーバはスペイン語が使われているセキュリティ教育のサーバであり，関連があると判断されたメッセージのうち，取得対象である悪意のあるハッキングなどと関連のあるメッセージはなかった．もう 1 件のサーバは英語が使われている様々なトピックに関して会話を行うサーバであり，取得対象のカテゴリに関連するワードは含まれているが，そのカテゴリには属さないと考えられるメッセージが多かった．

6 考察

検出器で ChatGPT を使うことによって，少ない情報量から適切な判断ができるケースがあった．例えば，「bio tolerance」というサーバでは，説明文に「biohacking community」という記載があり，Malicious Hacking カテゴリに関連があると誤検知する可能性や，中毒や受容体など drug と明言していない表現が使われていることから見逃しがある可能性があったが，GPT3.5 と GPT4 の両方でシステムに入力した取得対象のカテゴリに属すると検知でき，検知の理由として drug に関連があると出力された．また，GPT4 でのみ適切に判断できたケースも見られた．例えば,「OnlyFans Leaks」というサーバはポルノ系であるが，サーバ名には Leaks と入っているため，背景知識がないと Leak に関連するカテゴリであると判断してしまう恐れがあった．実際，GPT3.5 では取得対象であると判断されたが，GPT4 の場合はポルノ系のコンテンツであることを認識し，取得対象ではないと判断できた．

本システムのプロンプトでは，例を与えていない，会話の流れがわからない，言語によって精度に違いが出

る，カテゴリへの関連度がわからないという問題があった．まず，プロンプトは入出力の例を与えない Zero-shot Prompting になっているが，いくつかの入出力の例を与える Few-shot Prompting を使うことで精度が向上する可能性がある．システムの取得対象を固定して，Few-shot Prompting によってプロンプトを作成したところ，精度の向上がみられた．本システムでは取得対象が可変のため，ユーザが入出力の例を与えたり，システムが自動的に例を追加することによって，判別器の精度が向上すると考えられる．次に，本システムでは取得対象のカテゴリに関連するワードを含むメッセージを ChatGPT に渡している．これでは会話の流れが分からずに，その文章のみから推測する必要があり，人間でも判断が難しい．そのため，会話の流れをそのまま ChatGPT に渡すことで，精度が向上すると考えられる．次に，フランス語やスペイン語などの英語以外の言語での誤検知が見られたため，メッセージを英語に変換してから ChatGPT で判別することによって精度が向上すると考えられる．最後に，本システムのプロンプトでは少しでもカテゴリに関連しそうなメッセージはすべて同等に扱われてしまい，カテゴリへの関連度がわからない．サーバがそれぞれのカテゴリに対しての関連度を出すことで，ユーザが見るべき情報をソートして出力することができ，よりユーザの目的に合わせた出力が可能になると考えられる．

7　まとめと今後の課題

本研究では，情報収集が難しい Discord において汎用的な情報収集システムを提案し，それを利用してサイバー犯罪に関連する情報を収集し，Discord 内で起きている悪意のある行動について調査した．その結果，8,476 件のサーバ情報を取得し，55 件のサーバに参加したところ，1,280,548 件のメッセージを取得できた．参加したサーバのメッセージ情報から，取得対象であると考えられるサーバが 28 件あるとわかり，それらの中にはクレジットカード情報の販売や盗んだ認証情報の共有といったことが行われていることがわかった．

今後は収集したメッセージに含まれる招待リンクから参加することや，メッセージに含まれる関連するワードを抽出して関連キーワードリストにフィードバックをかけるなど，システムで実装できていない部分を実装したいと考えている．また，メッセージからサーバが取得対象であるかの判別器では，会話の流れがわからない，言語によって精度が変わる，関連度がわからないという問題があり，これらを改善したいと考えている．

謝辞: 本研究成果は、国立研究開発法人情報通信研究機構 (NICT) の委託研究（JPJ012368C05201）により得られたものです。

参考文献

[1] Discord BLOG，An Update on Our Business, https://discord.com/blog/an-update-on-our-business (参照 2023.12.10).

[2] statista, Distribution of Discord.com users worldwide as of May 2022, by age group, https://www.statista.com/statistics/1327674/discord-user-age-worldwide/ (参照 2023.12.10).

[3] NHK，米機密文書流出 21 歳空軍州兵逮捕 “アクセス権限持っていた”，https://www3.nhk.or.jp/news/html/20230414/k10014038131000.html (参照 2023.12.10).

[4] statista, Number of weekly active Discord servers worldwide from 2020 to 2023, https://www.statista.com/statistics/1368309/discord-monthly-active-servers/ (参照 2023.12.13)

[5] Sanden, Robin van der, Chris Wilkins, Marta Rychert, and Monica J. Barratt. 2022. “The Use of Discord Servers to Buy and Sell Drugs.” Contemporary Drug Problems 49 (4): 453–77.

[6] Nizzoli, L., S. Tardelli, M. Avvenuti, and S. Cresci. 2020. “Charting the Landscape of Online Cryptocurrency Manipulation.” IEEE. https://ieeexplore.ieee.org/abstract/document/9120022/.

[7] Heslep, Daniel G., and P. S. Berge. 2021.“Mapping Discord’s Darkside: Distributed Hate Networks on Disboard.” New Media & Society, December, 14614448211062548.

[8] Discord Developer Portal, Invite Resource, https://discord.com/developers/docs/resources/invite#get-invite (参照 2023.12.12)

[9] OpenAI, OpenAI API, https://openai.com/blog/openai-api (参照 2023.12.12)

[10] GitHub, dolfies/discord.py-self: A fork of the popular discord.py for user accounts., https://github.com/dolfies/discord.py-self (参照 2023.12.12)