TU WIEN Informatics

Honeypot LLM

Creation of the Scam Conversation Corpus

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Data Science

eingereicht von

Christoph Eder, BSc
Matrikelnummer 11802452

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.Ass. Gábor Recski, PhD

Signed by:	Christoph Eder
Date:	06.06.2025 14:00:17

A-Trust: This document is digitally signed! This document bearing a qualified electronic signature has the same legal standing as a handwritten document in accordance with Article 25 (2) of Regulation (EU) No. 910/2014 of 23 July 2014 ("eIDAS-VO"). Validation: Information for checking the electronic signature can be found at: www.a-trust.at/pdf

SIGNATURINFORMATIONEN

Signatory	Gabor Andras Recski
Date/Time-UTC	2025-06-06T13:23:38+0200
Verification	Information about the verification of the electronic signature can be found at: https://www.signaturpruefung.gv.at
Note	This document is signed with a qualified electronic signature. According to Art. 25 para. 2 of the Regulation (EU) No 910/2014 of 23. July 2014 ("eIDAS-Regulation") it shall have the equivalent legal effect of a handwritten signature.

Wien, 2. Juni 2025

Christoph Eder | Gábor Recski

TU WIEN Informatics

Honeypot LLM

Creation of the Scam Conversation Corpus

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Data Science

by

Christoph Eder, BSc
Registration Number 11802452

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.Ass. Gábor Recski, PhD

Vienna, June 2, 2025

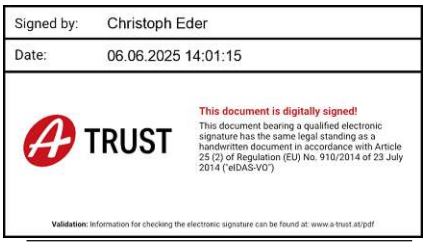
Christoph Eder	Gábor Recski

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Christoph Eder, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.
Ich erkläre weiters, dass ich mich generativer KI-Tools lediglich als Hilfsmittel bedient habe und in der vorliegenden Arbeit mein gestalterischer Einfluss überwiegt. Im Anhang „Übersicht verwendeter Hilfsmittel“ habe ich alle generativen KI-Tools gelistet, die verwendet wurden, und angegeben, wo und wie sie verwendet wurden.



Wien, 2. Juni 2025

Christoph Eder

Acknowledgements

This work was funded by Gogolook Co., Ltd. a corporation organised and existing under the laws of the Republic of China with its registered office located at 6F., No. 319, Sec. 2, Dunhua S. Rd., Da'an Dist., Taipei City 106, Taiwan.

I want to thank Univ.Ass. Gábor Recski, PhD for his support and mentoring, opening my eyes to new approaches whenever I hit a dead end, and making it possible for this work to be submitted to the ACL Workshop. A big thank you to Gogolook, who not only financed this work, but were also willing to go through months of legal preparations to make this international collaboration possible. A special thanks goes to Yimin Kao, PhD for the pleasant work atmosphere during our weekly meetings and the insights he gave me. I have learned a lot in terms of technology and culture. Thanks to my fellow students and friends, who made studying a pleasure rather than a chore and always showed interest in my work.

Special thanks go to my family, who always had an open ear for me and made my studies possible through their financial support. Finally, I want to thank my partner Barbara for the energy she gave me in difficult times and her unwavering faith in me and my abilities.

Kurzfassung

In dieser Diplomarbeit wird der Scam Conversation Corpus (SCC) vorgestellt, ein neuer Datensatz, der Unterhaltungen zwischen GPT-4o, welches als potenzielles Betrugsopfer auftritt, und echten Online-Betrügern umfasst. Der Datensatz wurde mit Hilfe eines Honeypot-Ansatzes erstellt, mit welchem eine Social-Media-Präsenz geschaffen wurde, die Betrüger anlockt. Wir haben eine Response Application entwickelt, die es ermöglicht, das Large Language Model (LLM) mit den Kommunikationsplattformen Instagram, Telegram und E-Mail zu verknüpfen. Diese Anwendung verbindet nicht nur diese Plattformen, sondern ermöglicht auch ein nahtloses Umschalten zwischen ihnen während einer Unterhaltung. Diese Einrichtung ermöglichte die Sammlung von Gesprächen, einschließlich aller von den Betrügern erhaltenen Multimedia-Inhalte. Zur Bewertung des gesammelten Datensatzes führten wir eine vergleichende Analyse mit logistischer Regression und XGBoost durch, zusammen mit dem öffentlich verfügbaren Scam-baiting Dataset (ScamBait) und einem neu zusammengestellten Nicht-Zielklassen-Datensatz. Die Ergebnisse zeigen, dass Modelle, die auf dem SCC trainiert wurden, effektiv auf den ScamBait verallgemeinern können, während das Gegenteil nicht der Fall ist. Schließlich wurde eine qualitative Analyse durchgeführt, um linguistische und strukturelle Muster in den Datensätzen aufzudecken, die für die Asymmetrie in der Generalisierbarkeit verantwortlich sein könnten. Der gesamte Code für die Experimente ist in einem öffentlichen Repository unter MIT-Lizenz verfügbar, und der gesamte Code für die Datenerhebung sowie der Datensatz selbst werden auf Anfrage für Forschungszwecke zur Verfügung gestellt. Diese Arbeit wurde für den ACL Workshop on Online Abuse and Harms 2025 als zweispaltiges long paper eingereicht.

Abstract

This paper presents the Scam Conversation Corpus (SCC), a unique dataset comprising conversations between GPT-4o, acting as a potential fraud victim, and genuine fraudsters. The dataset was created using a Honeypot approach to establish a social media presence that attracts fraudsters. We developed a response application that facilitates linking the Large Language Model (LLM) with the communication platforms Instagram, Telegram, and email. This application not only connects these platforms but also allows for seamless switching between them during a conversation. This setup enabled the collection of conversations, including all multimedia content received from fraudsters. To evaluate the collected dataset, we conducted a comparative analysis using Logistic Regression and XGBoost, alongside the publicly available Scam-baiting Dataset (ScamBait) and a newly compiled non-scam dataset. The findings indicate that models trained on the SCC generalise effectively to the ScamBait, whereas the reverse is not true. Finally, a qualitative analysis was conducted to uncover linguistic and structural patterns in the datasets that may account for the asymmetry in generalisability. All code for the experiments is available in a public repository under MIT licence and all code for the data collection, as well as the dataset itself, will be made available upon request for research purposes. This thesis was submitted to the ACL Workshop on Online Abuse and Harms 2025 as a two-column long paper.

Contents

CHAPTER 1

Introduction

Online scams take many forms, such as phishing emails, romance scams, and fake investment opportunities. This paper focuses on advance-fee and impersonation frauds, which are often initiated via spam emails and social media messages. It introduces the `Scam Conversation Corpus`, a dataset of conversations between `GPT-4o` mimicking a potential online fraud victim and fraudsters engaging in such scams. It provides detailed information on the creation of fake social media profiles used to lure fraudsters into interacting with them and the structure of the application used to collect the data. A benchmark for scam detection is presented using a traditional text-classification approach, with results compared across multiple datasets, including the `SCC`. The main contributions of this paper are the following:

1. A description of how modern fraudster interactions with users were collected across multiple communication platforms. This approach also considers the need for understanding multimedia data, such as images sent by fraudsters.

2. A novel conversation-based dataset, `SCC`, containing fraudsters-victim conversations covering the three communication platforms email, Telegram and Instagram. The dataset includes a total of 13,801 messages exchanged with 903 different scammers, along with 835 multimedia files received.

3. A comparative analysis using Logistic Regression and XGBoost to classify conversations as scam or non-scam (ham). For this, the models were trained and evaluated on three datasets to distinguish between scam and ham data in a cross-dataset evaluation: the `SCC`, the `ScamBait` introduced by Chen, Wang, and Edwards [1], and the `Custom Ham Dataset` (`CHD`), a dataset compiled for this paper. The `CHD` consists of a combination of the `Synthetic Email Conversations`

(SynthEC)[1] and a subset of The Enron Email Dataset (Enron)[2] merged to facilitate cross-dataset evaluation. The findings of the analysis indicate that the models trained on dataset SCC exhibit the capacity to generalise to dataset ScamBait; however, the reverse is not supported.

4. A manual qualitative analysis of the datasets and the features obtained in the comparative analysis to identify linguistic and structural patterns that may explain differences in model generalisability.

All code for the extraction of the Enron subset is available in a repository [3], as well as the code for the experiments[4]. The code for the response application developed to collect the SCC as well as the dataset itself, are available upon request for research purposes to avoid misuse.[5] This thesis was submitted as a long paper to the ACL Workshop on Online Abuse and Harms 2025.

[1] https://huggingface.co/datasets/argilla/FinePersonas-Synthetic-Email-Conversations
[2] https://www.cs.cmu.edu/~./enron/
[3] https://anonymous.4open.science/r/ExtractEnronThreads-FDBF
[4] https://anonymous.4open.science/r/ScamDetection-F2AC/
[5] Contact author per email

CHAPTER 2

Background

Scams are increasingly prevalent, with global losses in 2023 estimated sometimes to have reached a trillion dollars [2]. Among the many forms of scams, online fraud has become one of the most widespread and damaging. No longer limited to traditional spam emails, it now manifests across a wide range of digital platforms. These include instant messaging applications, social media networks, and online dating services, where scammers can exploit the immediacy and informality of communication to deceive individuals more effectively. This shift highlights the evolving nature of online fraud and its growing adaptability to different technological environments. Kolupuri et al. [3] have provided taxonomies for different types of digital fraud. They separate in the following categories of fraud: *Phishing Attacks*, *Click Fraud*, *Content based SMS scam*, *Frauds in advertising* and *Online Social Network Scams*. The relevant fraud types examined in this paper categorised by their taxonomies are *Social Engineering*, involving psychological tactics like impersonation or urgency, and *False Accounts* fraud, where fake profiles mimic legitimate users for fraudulent activities. According to their definition, the former is categorised as a *content-based SMS scam*, and is thus considered distinct from email and social media scams, which are the primary focus of this study. Nonetheless, the recurrence of similar deceptive techniques across these platforms suggests that there may be considerable overlap in how scams are executed, despite their classification under different categories. The latter falls within the taxonomy of *Online Social Network Scams*, making it closely related in scope to the topics addressed in this work.

With this increase in scams and methods of operation comes the need to protect against them. An essential step in protecting individuals from online fraudsters is to understand both the characteristics of victims and fraudsters and how scammers exploit certain personality traits.

Saad and Norul Huda Sheikh Abdullah [4] provide a broad statistical overview of individuals affected by romance scams in Malaysia. Their findings indicate that a

substantial proportion of victims are well-educated, with many holding at least a diploma-level qualification. The study also reveals that married individuals appear to be more susceptible to such scams. In addition, limited computer literacy and lower levels of awareness about cybercrime are identified as factors that further increase vulnerability to these types of fraudulent schemes.

Cross [5] highlights strategies used by romance scammers, such as presenting investment opportunities to their victims, which leads to victims not realising they are sending money to the fraudsters, but rather they assume they are investing in legitimate business ventures. This strategy appears to be aimed at reducing the victims' suspicion, as the funds are not transferred directly to the fraudsters but are framed as part of a legitimate investment process.

Wood et al. [6] primarily examine phone scammers using YouTube videos as their data source. They successfully extracted scam scripts and categorised scams into distinct stages and provide multinomial hidden Markov model transition graph to model the transitions between those stages. They were also able to confirm various persuasion techniques applied by fraudsters such as impersonation of figures of authority or the creation of familiarity and trust with the victim. All of those are Social Engineering approaches. Ma et al. [7] introduce the `PsyScam` framework utilising human-LLM collaboration to enable scalable annotation of psychological techniques, such as those described above, in scam datasets.

Stajano and Wilson [8] identify seven psychological principles that fraudsters frequently exploit to manipulate individuals into surrendering money, often contrary to their rational judgment. These principles illustrate how scammers take advantage of cognitive and social biases:

1. *Distraction Principle*: When individuals focus on a specific task or stimulus, they are less likely to notice signs of deception or question suspicious behaviour.
2. *Social Compliance Principle*: Fraudsters impersonate authority figures, relying on social conventions that discourage individuals from questioning perceived authority.
3. *Herd Principle*: Scammers create an impression of trustworthiness by simulating consensus, for instance, through fake reviews, to suggest legitimacy through perceived popularity.
4. *Dishonesty Principle*: Victims are led to believe that they are participating in illegal but profitable activities. This reduces the likelihood of reporting the scam and increases the perceived plausibility of rapid financial gain.
5. *Kindness Principle*: Emotional manipulation is used to appeal to the victim's willingness to help others, often by inventing compelling personal tragedies.
6. *Need and Greed Principle*: Individuals who are in urgent need, such as those facing financial hardship or who are highly motivated by the prospect of gain, are more likely to accept offers without thorough scrutiny.

7. *Time Principle*: Artificial time pressure is introduced to limit the victim's ability to reflect, leading to rushed and poorly considered decisions.

These principles provide insight into how fraudsters bypass critical thinking and exploit human tendencies in order to achieve compliance. Stajano and Wilson [8] also point out that, depending on the type of fraud, scammers abuse multiple of those principles simultaneously.

Brody, Kern, and Ogunade [9] provide insights into the demographics of scammers in Nigeria, reporting that 95% are male and typically between 15 and 55 years old. They also present detailed descriptions of various scam types, although the findings are based on a single internal source. Their findings suggest that scammers increasingly prioritise initial contact through social media rather than mass spam emails.

Grover et al. [10] have investigated various types of online scams, including fraudulent transactions, deceptive job postings, and malicious links. Several datasets covering these and related fraud types have been benchmarked in their work. They sum up the unique properties of fraud datasets with emphasis on class imbalance, high cardinality of features, and that samples are not always independent and identically distributed. They also acknowledge the adversarial nature of the problem, as fraudsters adapt their behaviour to bypass detection models, making the issue of fraud detection time dependent.

Tang et al. [11] provide an expert annotated dataset on fraudulent information on Chinese Web pages over a 12-month period. Their findings demonstrate that not only the methods used within individual fraud categories evolve over time, as noted by Grover et al. [10], but also that the distribution of the fraud categories itself changes, indicating broader changes in the patterns of fraudulent activity. Furthermore, they present a comprehensive benchmark comparing various fraud detection models, including traditional machine learning approaches, pre-trained language models, and LLMs. In contrast to their focus on fraudulent web pages, our focus is on direct textual interactions between potential victims and scammers conducted via email or messaging platforms in English. Although there are numerous publicly available email-based datasets [12, 13], most focus on spam or phishing rather than broader fraud-related conversations.

To the best of our knowledge, there exist three English datasets comparable to the one we collected. The first, collected by Salman, Ikram, and Kaafar [14], consists of SMS messages and is relevant because it captures messenger-based interactions, albeit it remains limited to spam. They created the dataset by consolidating multiple corpora and aggregating them with online sources and volunteers. Similar to the approach taken by Tang et al. [11], they also provide a benchmark based on their dataset, evaluating multiple deep learning-based spam detection algorithms. In addition, their analysis extends to six messaging applications, including the default texting apps on both Android and iOS platforms.

The second and third comparable English datasets, both introduced by Chen, Wang, and Edwards [1], further contribute to this space. Of these, we make use of one in this paper,

namely the ScamBait, which contains textual exchanges between scammers and human scam-baiters, who engage scammers via direct email communication to occupy their time, thereby diverting resources from real targets. Chen, Wang, and Edwards [1] collected a second dataset using two distinct strategies to automatise scam-baiting. The first involved crafting human-written response templates to interact with scammers, and the second utilised GPT-Neo, which was finetuned on both the ScamBait and the Enron, to generate automated replies. The evaluation process compared the effectiveness of these two approaches. They suggested that employing a more advanced LLM could improve performance in terms of time fraudsters spent engaging with the system and emphasised the importance of switching communication platforms to capture complete conversations, as fraudsters often employ complex narratives and assume multiple personas. Furthermore, they pointed out that understanding the multimedia data received from scammers could further increase performance.

One of the most commonly used ham datasets is the Enron [1, 12, 13]. The Enron exclusively contains conversations involving Enron employees, limiting the range of topics and types of conversations. Klimt and Yang [15] provide a detailed description of it.

The third dataset, the SynthEC, is synthetically generated and complements the set of comparable English resources. It was created using the Hermes-3-Llama-3.1-70B model, as described by Teknium, Quesnelle, and Guang [16], and includes interactions among 11,000 distinct personas sampled from the FinePersona[1] dataset. To generate the dataset, five personas similar to each original persona were selected, and the LLM was prompted to first create a shared context, followed by a simulated email conversation between the original and each of the five similar personas. To further increase diversity, an additional set of five randomly selected personas was paired with each original persona, using the same generation procedure. This approach ensures a mix of coherent and varied interactions, enhancing the dataset's coverage of potential conversational patterns. It is important to note, that the FinePersona dataset contains synthetic personas based on webpages from the FineWeb-Edu[2] dataset and therefore contains a bias towards personas in education and scientific domain.

To the best of our knowledge, the SynthEC has not yet been used as ham dataset for scam detection in emails. Although individual synthetic email conversations may not perfectly reflect real-world exchanges, the dataset covers a broad range of topics and conversational styles. This diversity makes it a suitable choice for adding varied communication patterns to experiments, especially given the limited availability of public real-world ham email datasets.

[1] https://huggingface.co/datasets/argilla/FinePersonas-v0.1
[2] https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu

CHAPTER 3

Dataset Creation

We developed a response application to collect a comprehensive dataset of conversations with online fraudsters across email, Telegram, and Instagram. By leveraging `GPT-4o` to simulate a persona interested in relationships and effortless financial gain, we captured realistic scam interactions, including both textual dialogues and transformed multimedia content.

3.1 Data Collection

To gather the data, our response application facilitated seamless communication with fraudsters, allowing conversations to transition across email, Instagram, and Telegram. Initial contact was established by responding to scam emails obtained from the database `www.scamsurvivors.com`. Additionally, we created a social media presence, referred to as the `Honeypot`, on Facebook and Instagram to attract online fraudsters.

3.1.1 Honeypot

The persona created is an elderly looking woman from a small town in Ohio, United States. The age group has been chosen as Sugunaraj, Ramchandra, and Ranganathan [17] have stated that elderly people in the US have a reported income of $46,680 before tax, often have savings for retirement and, therefore, are financially viable targets for online fraud. For the persona, a Facebook, Instagram and Telegram profile were created. For those profiles, two photos were generated to provide additional authenticity. The first portrait was based on `RealVisXL V4.0`[1] with an `Euler` sampler and a 7-guidance scale. The resulting image is shown in Figure 3.1a. The second one needed more sophisticated tools to guarantee character consistency, so the same person would be shown in both images. The workflow set up in ComfyUI was inspired by the methods described in

[1] `https://huggingface.co/SG161222/RealVisXL_V4.0`

Xclbr Xtra's YouTube tutorial[2]. The workflow was kept as explained in the tutorial, using a `FLUX.1`[3] model, except for the face swap module which was replaced with a `ReActor Fast Face Swap` node [4] with an `inswapper_128.onnx` swap model, `retinaface_resnet50` as facedetection model and `GPEN-BFR-512.onnx` as face restoration model. The produced image is shown in Figure 3.1b.

The resulting images might not pass close inspection for AI-generated content. However, the assumption was made that fraudsters will not closely check before contacting a profile. To mimic typical social media behaviour, we connected (followed) with local businesses and interacted (liked) with random posts. We also adopted the practice of always reciprocating connection requests on Instagram and accepting friend requests on Facebook. Furthermore, we regularly searched for Instagram accounts impersonating celebrities (without clearly stating they were fan pages) and sent them connection requests (followed them) to encourage scammers to reach out to our persona. This was based on the assumption that impersonating celebrities already exhibits some degree of malicious intent, indicating that those accounts are more likely to be operated by scammers trying to deceive and also to ensure innocent users are not disturbed.

(a) Portrait generated with `RealVisXL V4.0`

(b) Garden image generated with `FLUX.1`

Figure 3.1: Persona images generated for social media profiles

3.1.2 Response Application

We developed a response application connecting communication platforms with the LLM, automating message processing, scheduling, and reply generation and delivery. The code for the application is available only upon request for research purposes, to avoid misuse as detailed in Chapter 6. A high-level description of its workflow is shown in Figure 3.2.

[2] `https://www.youtube.com/watch?v=b_ox_SOLIQk`
[3] `https://huggingface.co/city96/FLUX.1-dev-gguf`
[4] `https://github.com/Gourieff/ComfyUI-ReActor`

Each platform has a *Listener* module that actively listens for new messages via platform-specific Application Programming Interfaces (APIs) and passes them to the *Messenger* module. The Instagram *Listener* module operates using the unofficial Private API wrapper *instagrapi*[5] that is not built on the official Instagram API and instead uses a reverse-engineered API version. The email module is based on a Flask server and the Telegram module utilises *Telethon* [6]. If an incoming message contains multimedia content, the *Multimedia Interpreter* uses the external `GPT-4o` API to obtain a textual description, keeping the hardware requirements low.

The *Switch Detector* examines messages for proposals to switch platforms or personas, which is crucial because scammers often pose as different individuals during a conversation (e.g., a "lawyer" with another email address). Our application detects and handles these changes accordingly.

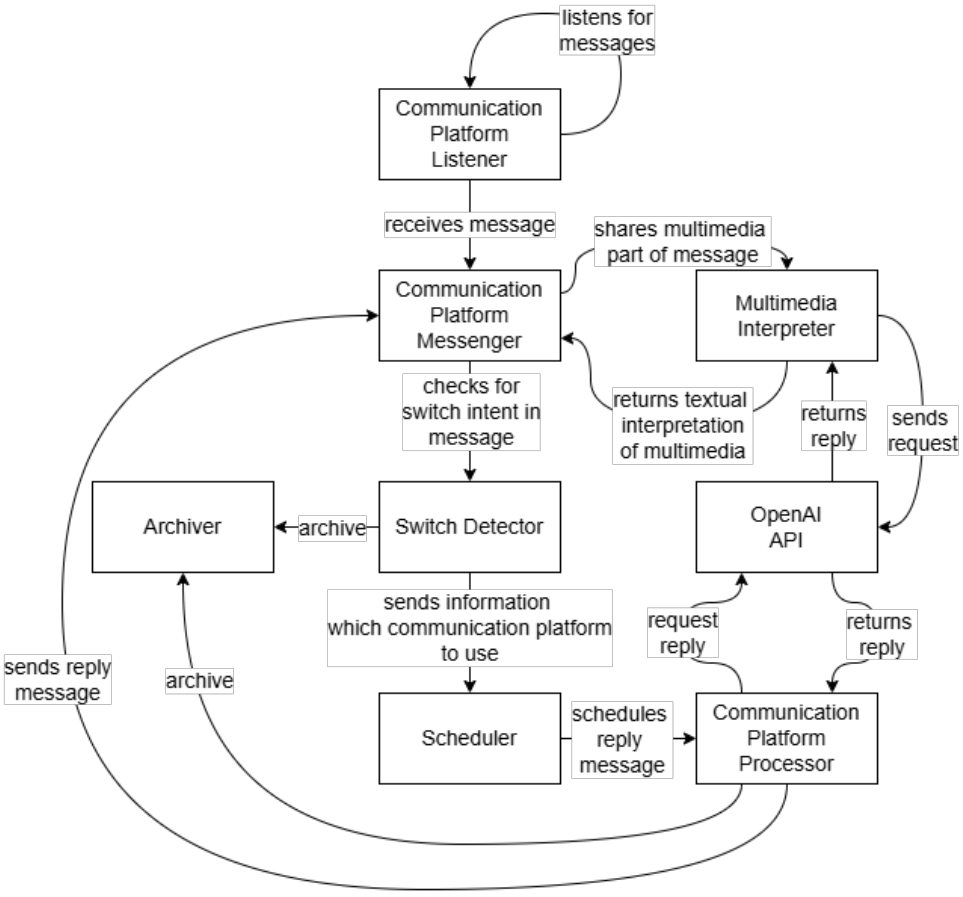


Figure 3.2: Information flow in response application

[5] https://github.com/subzeroid/instagrapi
[6] https://github.com/LonamiWebs/Telethon

The *Scheduler* module mimics human response times as detailed in Appendix 6. When a response is due, the *Processor* loads previous correspondence and requests an appropriate reply from the `GPT-4o` API. It then archives the conversation and sends the reply via the correct platform using the *Messenger* module. To mitigate prompt injection risks, we implemented Delimiters to separate fraudster messages from the application prompt in each API call, enhancing security as described by Liu et al. [18].

In contrast to Chen, Wang, and Edwards [1] we did not fine-tune the LLM on scam-baiting data, but rather applied a multilayer prompting strategy to adjust dynamically to different scenarios. Every request for a response contained a default prompt with the character card for our persona detailing demographics and character traits. Although some of the principles described by Stajano and Wilson [8] do not readily translate to online-only contexts, our persona is set up to follow the *Kindness Principle*, being nice and supportive, and the *Need and Greed Principle*, mimicking victims' desire for swift, effortless financial gain. Furthermore, the default prompt limits the general knowledge and languages with which the model should respond, as displayed in Section 6. Depending on the communication platform, an additional prompt is added, so the model adheres to typical formatting and etiquette. An example for email is given in Section 6. On top of that, another prompt is appended if the conversation switches to another person or messenger with the next response. This prompt allows the LLM to recognise the switch to another platform or person in a natural way. The prompt for switching to another platform is given in Section 6.

3.2 SCC Description

Dataset	Conversations	Messages	Avg Conv Length	Median Conv Length	Avg Msg Length	Median Msg Length	Vocab Size
SCC	903	13,801	15.28	5	41.66	27	14,620
SCCDia	535	13,089	24.47	15	39.74	26	13,547
ScamBait	658	37,418	56.87	37	60.60	43	62,286
SynthEC	113,576	363,584	3.20	3	67.37	65	54,665
Enron subset	17,199	65,195	3.79	3	95.52	40	61,830
CHD	130,775	428,779	3.28	3	71.65	64	93,374
SCC + CHD	131,678	442,580	3.36	3	70.72	63	97,898
ScamBait + CHD	131,433	466,197	3.55	3	70.77	63	130,424

Table 3.1: Descriptive statistics for the used datasets and subsets

From October 2024 to February 2025, a total of 903 conversations were collected comprising 13,801 messages, forming what is referred to as the `SCC`. All conversations were saved in a consistent `JSON` format for ease of analysis. Images and documents received from fraudsters were linked to the corresponding messages. Metadata, like email addresses and usernames, that was used to respond to scammers was anonymised as described in Section 6. As done by Palmero et al. [19] and Chen et al. [20], the dataset is available upon request for research purposes. 535 of those 903 conversations had more than two messages, meaning that the scammer actually engaged in a dialogue, this subset is called `SCCDialogue` (`SCCDia`) in Table 3.1. The data is heavily skewed towards email, with 869 conversations being held per email, while only 35 were on Instagram and 20 on Telegram. This discrepancy is because only 35 scammers initiated contact with

our Honeypot on Instagram, and we did not actively seek out scammers on Telegram, instead offering it as an optional platform for fraudsters to switch to.

A total of 20 conversations, representing approximately 3.5% of the 535 conversations in SCCDia, involved a transition to a different communication platform as showcased in Listing 3.1.

```
Scammer via Instagram:
"I don't have telegram account but we can use email"
"****@gmail.com"

LLM via email:
"Hello there!
I hope this email finds you well. It's Cindy Jenkins here-I'm reaching out to
 you through email as we've transitioned from Instagram![...]"
```

Listing 3.1: Excerpt from a conversation showcasing switching of platforms

In contrast, 43.18% of the conversations in SCCDia included multimedia elements sent by the fraudster, highlighting the importance of analysing and understanding the content of this data. This multimedia data consists mostly of images used by fraudsters to lend credibility to their schemes, such as fake company passes or forms allegedly from banks or lawyers. An excerpt of such a conversation is shown in Listing 3.2, with the blue text indicating the textual description generated from the received attachment.

```
Scammer:
"Dear ma,I attach herewith a copy of the deposit slip in your name ,this I
got
from the file here with me,I am going out of my official duty to reval this
and
send it for your perusar,[...]This message contains files. If the description
for a file does not make sense, ignore it.Here are descriptions of those
files:
Description for file 1: The image shows a Dubai Islamic Bank deposit slip for
$5,700,000.00, dated 02/08/2020. It includes details such as the depositor's
name, account number, and type, with a bank stamp indicating processing."

LLM:
"Dear Mr. Williams,

I appreciate you sending over the deposit slip and providing further details.
[...] "
```

Listing 3.2: Excerpt from a conversation showcasing understanding of attached files

Most conversations terminate once the fraudster recognises that the provided bank details are incorrect or becomes impatient with delays in receiving payments or completed forms, as the response application is unable to supply these. Others cease communication after issuing conditions, such as *"If you don't send the card by tomorrow we will cancel your delivery"* or persistently respond with the same message (*"ok"*) for multiple replies before stopping.

Among all conversations initiated on Instagram, only one fraudster immediately launched into their scheme, sending a message that the `Honeypot` had won 70,000 USD. All other interactions began with small talk and attempts to establish a connection with the `Honeypot` before requesting gift cards or proofs of authenticity in the form of photos or videos.

CHAPTER 4

Experiments

4.1 Experimental Setup

Our experiments utilised two scam datasets and one ham dataset. The scam datasets are the SCC, from which we excluded any conversations not conducted solely via email, and the ScamBait as described in Chapter 2, which serves as a comparative scam dataset. The choice to only take conversations that are email-only into consideration was made to guarantee a fair comparison as conversations conducted on other communication platforms might be fundamentally different in structure. The ham dataset is the CHD, constructed by combining the SynthEC , also introduced in Chapter 2, with a subset of the Enron to provide ham data. Conversations in the Enron were identified by grouping emails exchanged between the same pair of participants that shared an identical subject line, after normalizing the subject by removing common prefixes such as "RE:". Specifically, we extracted 17,199 conversational threads from the Enron involving exactly two participants each, cleaned it of dataset-specific tags such as *[IMAGE]* and base64-encoded attachments and merged them with 113,576 conversations from the SynthEC to form the CHD. The elimination of those artifacts was necessary to ensure that the detection is based on the inherent scam content rather than relying solely on these artifacts. The resulting CHD has a total of 130,775 conversations comprising 428,779 messages. We formed the CHD to combine the advantages of real-world data from the Enron with the diversity of topics from the SynthEC. Descriptive statistics for all the datasets and compositions are shown in Table 3.1.

Drawing from the insights of Lu, Henchion, and Namee [21], Figure 4.1 displays the unweighted *Jensen-Shannon Divergence* values derived from the word distributions for each pair of datasets, rather than employing one of the often used weighted versions. Lower values indicate closer similarity in lexical use, whereas higher values signal greater divergence. As shown, the two scam datasets (SCC and ScamBait) appear more similar to each other, while the ham data CHD subsets (Enron and SynthEC) remain relatively

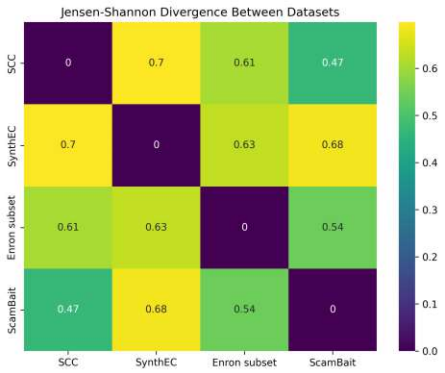


Figure 4.1: Jensen-Shannon Divergence Heatmap between Datasets

more distant, supporting the argument that combining the subsets increases the diversity of the ham data.

For the two scam datasets, only messages authored by the fraudsters were utilised for training and evaluation. In the case of the `CHD`, only messages from the sender of the first message in each conversation were included. This ensures consistency across datasets by restricting the data to a single participant per conversation, while retaining only fraudulent content in the scam datasets.

All three datasets were split into training and test sets, with individual subsets of the `CHD` dataset split prior to forming the final sets. For each of the datasets, empty messages were removed and the words in the messages were preprocessed. Specifically, we converted all text to lowercase, tokenised the messages using `nltk` 3.9.1 *word_tokenize*, lemmatised the tokens with *WordNet* lemmatiser, filtered out non-alphabetic tokens, and then rejoined the tokens in processed messages. To decrease the number of dataset specific artifacts the textual descriptions of the multimedia data attached to the messages were removed for the `SCC`.

To assess algorithm generalisability using the `SCC`, we conducted four experiments with the `CHD` as the ham data. First, we trained and evaluated on both `ScamBait` + `CHD` and `SCC` + `CHD` separately. Then, we performed cross-evaluations by training on `ScamBait` + `CHD` and evaluating on `SCC` + `CHD`, and vice versa. These experiments highlight algorithm performance across different scam datasets while keeping the ham dataset constant.

4.2 Baseline Algorithms

For benchmarking, we implemented baseline text classification algorithms using Logistic Regression and XGBoost. We selected these models in favour of modern deep learning based architecture because of their robustness and interpretability, and because the length of many conversations in our datasets exceed the context window of even the most

recent LLM-based architectures. We believe that the use of more complex, hierarchical architectures would further compromise the explainability of predictions, a crucial aspect for the scam detection task. While few-shot experiments with *GPT-4o* were conducted for comparative purposes, they did not yield meaningful results, primarily due to the previously mentioned limitation regarding the context window. Furthermore, applying any state-of-the-art LLM with decent context windows size to the full dataset was not feasible, as the associated computational costs would have exceeded the resource constraints of this study. For these reasons, the results of these preliminary tests are not included in the final analysis.

Logistic Regression was chosen for its simplicity, efficiency, and reliable performance on high-dimensional text data. Its interpretability makes it well-suited for tasks where understanding the decision process is important. XGBoost was included as a complementary model due to its ability to capture non-linear patterns and its strong track record in classification tasks. Both models offer a practical balance between performance and transparency, making them appropriate for benchmarking in this context.

Both methods employed two traditional feature extraction techniques: Bag-of-Words, which counts word occurrences, and Term Frequency-Inverse Document Frequency (TF-IDF), which adjusts for word importance across the corpus. Additionally, we incorporated contextual embeddings using the `bert-base-uncased` model from the `transformers` package (version 4.50.0), allowing for a more nuanced representation of textual content.

Bag-of-Words and TF-IDF were chosen because their features are straightforward to interpret, even by non-experts. Each feature corresponds to a specific word in the vocabulary, and the associated weight reflects either its frequency (in Bag-of-Words) or its relative importance across documents (in TF-IDF). This direct mapping allows for clear explanations of model predictions by highlighting which words contributed the most to a given classification. In the context of scam detection, such transparency is essential for building trust in the system and enabling decision makers to make informed judgments.

We used `scikit-learn` 1.5.1 Logistic Regression with maximum 1000 iterations and `xgboost` with the *logloss* evaluation metric in a Python 3.10.0 environment. The metric has been chosen as it penalises predictions that are confident but incorrect predictions more strongly, which is desirable in high-risk classification tasks such as scam detection. For the Bag-of-Words and TF-IDF, the `CountVectorizer` and `TfidfVectorzier` from the same `scikit-learn` version were utilised. To reduce the influence of non-informative tokens, we applied English stopwords from `nltk` 3.9.1, extended by four additional words: *cindy*, *jenkins*, *enron*, and *u*. These were excluded to prevent models from learning superficial correlations based on highly frequent but uninformative dataset-specific terms.

To assess classifier performance, we report *Precision* (Pr), *Recall* (Re), *F1 score* (F1), and *Accuracy*, all measured with respect to the scam class. Although accuracy gives a general sense of overall correctness, it is not suitable as a standalone metric in this

context. The task involves a highly imbalanced classification problem, where scam emails are exceedingly rare. Specifically, the ratio of scam to ham messages is only 0.0053 and 0.0116 in the respective datasets. In such settings, a classifier can achieve high accuracy by simply predicting the majority class, while completely failing to detect scams.

Precision measures the proportion of correctly identified scam messages among all messages classified as scams, which is essential to minimise false alarms and reduce unnecessary interventions. Recall, in contrast, captures the proportion of actual scam messages that were correctly identified, addressing the model's ability to detect threats. To balance these competing concerns, we employ the F1 score, which is the harmonic mean of precision and recall and particularly informative when both false positives and false negatives carry significant operational cost. The F1 score is defined as:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.1}$$

This suite of metrics provides a robust and interpretable basis for evaluating model performance in high-stakes scenarios involving malicious communication.

Focusing on a single evaluation metric is insufficient for email-based advance fee fraud detection, as the relative importance of precision and recall depends on the deployment context. In high-volume automated filtering systems, high precision is crucial to avoid misclassifying legitimate emails and disrupting user communication. However, in forensic or investigative settings, higher recall may be prioritised to ensure that as many fraudulent emails as possible are identified, even at the cost of some false positives. Given these trade-offs, relying on multiple metrics allows for a more comprehensive evaluation, enabling users to prioritise the metric that best aligns with their specific use case.

We acknowledge that the chosen metrics may not fully capture the complexity of more advanced experiments involving email attachments, such as those available in the SCC dataset. However, they provide a versatile and interpretable benchmark for evaluating baseline models on textual content alone, ensuring comparability and reproducibility across different settings.

CHAPTER 5

Results

5.1 Experimental Results

Since the *scam* class was the minority class by a significant margin in all experiments, metrics are reported only for this class. Table 5.1 provides a comparison for the baseline algorithms trained on the two different scam datasets and evaluated on `ScamBait`. XGBoost with BoW performs as good or substantially better than all other methods when trained on `ScamBait`. Similarly, when trained on the `SCC` XGBoost with BoW

Training on `ScamBait`				
Method	Pr (Scam)	Re (Scam)	F1 (Scam)	Accuracy
Logistic Regression with BoW	**100.00**	93.18	96.47	99.96
XGBoost with BoW	**100.00**	**97.73**	**98.85**	**99.99**
Logistic Regression with TF-IDF	**100.00**	87.88	93.55	99.94
XGBoost with TF-IDF	99.23	**97.73**	98.47	99.98
Logistic Regression with BERT	96.15	75.76	84.75	99.85
XGBoost with BERT	97.62	93.18	95.35	99.95
Training on `SCC`				
Method	P (Scam)	R (Scam)	F1 (Scam)	Accuracy
Logistic Regression with BoW	**100.00**	87.12	93.12	99.93
XGBoost with BoW	97.56	**90.91**	**94.12**	**99.94**
Logistic Regression with TF-IDF	98.02	75.00	84.98	99.86
XGBoost with TF-IDF	95.90	88.64	92.13	99.92
Logistic Regression with BERT	98.72	58.33	73.33	99.77
XGBoost with BERT	96.52	84.09	89.99	99.90

Table 5.1: Performance comparison for evaluation on `ScamBait` with a ratio of scam:ham of 0.0053 and a total of 132 scam conversations in the evaluation set.

outperforms the other methods in all but precision for the scam class. It is important to note that both models with BERT perform notably worse than the models based on the

other feature extraction methods. The generally high precision values can be attributed to the vast class imbalance. Table 5.2 compares the performance of the two models on the validation set of the SCC. Results of models trained on ScamBait indicate that Logistic Regression with BoW surpasses all other approaches by a notable margin, but all models perform substantially worse than those trained on the SCC. Meanwhile, models trained on SCC achieve very similar performance on both validation datasets. Once again, XGBoost with BoW outperforms all other models, except in terms of precision for the scam class. This time exhibits a more pronounced decline in performance for models employing BERT as a feature extraction technique, indicating that this approach may be inadequate in its current form and may require further refinement or adjustments. The fact that models trained on the SCC keep their performance nearly consistent when evaluated on the two different scam datasets suggests that the SCC allows models to generalise better to unseen scam data.

Training on ScamBait				
Method	Pr (Scam)	Re (Scam)	F1 (Scam)	Accuracy
Logistic Regression with BoW	**100.00**	**61.11**	**75.86**	**99.55**
XGBoost with BoW	**100.00**	54.86	70.85	99.48
Logistic Regression with TF-IDF	**100.00**	57.29	72.85	99.51
XGBoost with TF-IDF	99.36	54.17	70.11	99.47
Logistic Regression with BERT	87.75	9.72	17.50	98.94
XGBoost with BERT	97.35	38.19	54.86	99.27
Training on SCC				
Method	Pr (Scam)	Re (Scam)	F1 (Scam)	Accuracy
Logistic Regression with BoW	**100.00**	82.64	90.49	99.80
XGBoost with BoW	98.86	**90.28**	**94.37**	**99.88**
Logistic Regression with TF-IDF	99.07	73.61	84.46	99.69
XGBoost with TF-IDF	98.01	85.42	91.28	99.81
Logistic Regression with BERT	99.14	39.93	56.93	99.30
XGBoost with BERT	98.33	81.94	89.39	99.78

Table 5.2: Performance comparison for evaluation on SCC with a ratio of scam:ham of 0.0116 and a total of 181 scam conversations in the evaluation set.

5.2 Qualitative Analysis

Based on the results from Section 5.1 we performed manual qualitative analysis of the features extracted by XGBoost based on BoW.

Due to the less complex and more uniform vocabulary of the SCC, shown in Table 3.1, the model trained on the SCC focuses on core scam-related features with high feature importance, such as words like *beneficiary*, *payment* and *bitcoin*. These features are shown in Appendix Table 1. This specificity might explain why the models generalise better to the ScamBait. Conversely, XGBoost when trained on the richer and more varied ScamBait conversations tailors its performance more to specific nuances of human-mediated interactions. As shown in Appendix Table 2, this results in relatively high feature importance

for words like *mr* and *bless*, associated with polite and formal language patterns rather than scams. These artefacts suggest that the drop in performance on the SCC may be a result of overfitting. The reasons for the differences in vocabulary between the two datasets are multilayered. Although it is expected that conversations with fraudsters will not continue indefinitely in the absence of profit for them, a comparison of the average conversation length between the subset of the SCC with dialogues, the SCCDia, and the ScamBait in Table 3.1 reveals that human scam-baiters are notably more successful in prolonging interactions. This discrepancy highlights inherent differences between the two datasets that could influence the observed model performances. Specifically, the SCC has several limitations that may affect how models trained on it generalise to other datasets.

This is partly due to the different objectives; unlike human scam-baiters who aim to occupy as much time from fraudsters as possible, our approach does not prioritise prolonging interactions, but rather emulates types of conversations actual victims might have with a fraudster. Additionally, limitations of the collection method, as detailed in Section 3.1, contribute to the shorter and potentially less varied conversations in the SCC. For instance, when it was not feasible to obtain textual descriptions of multimedia files or when completing a form was necessary, the default prompt instructed the LLM to cite technical difficulties and request a written version instead. Moreover, the response application lacks the ability to send signatures or photos, which further limits the interaction compared to real human engagements.

Despite detailed prompting, the LLM exhibits certain limitations. It struggled to provide accurate responses when asked for information such as the local time and, due to its limited awareness of the response application's context (except for the communication platform prompt), it sometimes requested email addresses to send documents, even while already conversing via email. On platforms like Instagram, individuals impersonating celebrities assume that the counterpart recognises them from the username and profile picture, but this information was not extracted or provided to the LLM.

These findings of the qualitative analysis indicate that the results in Section 5.1 should be interpreted with caution. The limitations inherent in the SCC mean that it may not fully capture the complexity and variability of real-world scam communications. Therefore, while our findings indicate that models trained on the SCC generalise better to the ScamBait than the other way around, this may not necessarily hold for a real-world use case or other scam datasets with different structures or characteristics. While the SCC can complement existing resources for benchmarking, its design may be particularly well suited to explore structural patterns and rhetorical strategies in scam communications, especially due to its novelty in terms of platform switching and inclusion of multimedia files.

CHAPTER 6

Conclusion

We introduced a novel LLM-fraudster conversation dataset, referred to as the SCC, which comprises dialogues with GPT-4o posing as a potential fraud victim, collected from three different communication platforms along with all accompanying multimedia data. The large amount of multimedia data received in longer conversations confirmed the idea of Chen, Wang, and Edwards [1], that many scammers require the understanding of photos and documents of their targets to continue the conversation. To evaluate the usefulness of our dataset, we performed a comparative analysis against a newly compiled ham dataset, the CHD, and the publicly available ScamBait dataset, employing two baseline classification algorithms. Our findings show that models trained on the SCC generalise better across datasets than those trained on ScamBait, demonstrating the dataset's potential for complementing existing resources in fraud detection research. Finally, we performed a qualitative analysis to identify linguistic and structural patterns of the datasets that may explain the observed generalisability.

Limitations

While the proposed methods demonstrate promising results in dataset generation, several limitations remain. First, the data collection is limited to scams performed on the three platforms selected. *False Account* fraud also exist on other platforms such as dating apps and which are not covered by our approach. Furthermore, during data collection fraudsters requested contact on other unsupported platforms such as WhatsApp. Second, an actual victim might engage in a (video) call with the fraudster or exchange photos to prove their authenticity, leading to a romance/friendship scam, where an in-depth relationship is built before exploiting the victim. Due to the limitations of our response application, those interactions were not possible, excluding those types of scam. Lastly, the observed time frame of 5 months is rather short when it comes to establishing a profile on social media platforms to attract scammers, leading to a low number of contact approaches by fraudsters on Instagram. We have observed an increase of approaches over time, so a longer collection window might increase data diversity.

The experiments would benefit from utilising a broader range of datasets for cross-evaluation, enhancing the significance of the results. Additionally, incorporating more diverse baselines, including those employing deep learning approaches, would provide contextual breadth to the findings, facilitating a wider application. Conducting an in-depth qualitative analysis of the feature weights of the models would further enable more comprehensive conclusions to be drawn from the experiments themselves. Further research could address one or multiple of those points by providing a benchmark across multiple datasets, including the SCC.

Ethical Considerations

We have consulted with our institution's ethics committee at the start of the project to avoid harming innocents and fraudsters alike. As a result we extended the approach by Chen, Wang, and Edwards [1] by defining the following rules for the response application:

No chase behaviour : No two consecutive messages will be sent, so the fraudster can always end communications simply by stopping to send a reply.

No initial contact : On all communication platforms except email, no message will be sent to other profiles if they did not contact the *Honeypot* persona first. This guarantees that no contact approach is made without consent.

No innocent contact : Only profiles that clearly try to impersonate a celebrity are manually followed/befriended first. This guarantees that only people with malicious intent are targeted. This rule does not apply to accepting follow/friend requests and returning them.

It is crucial to emphasise that the original content of the messages of the `SCC` remained unchanged. Consequently, a variety of potentially personal information, such as requests to contact individuals through specific Telegram accounts, bank details, and other sensitive data, was retained in its entirety. The decision to retain the information unaltered was made because it is reasonable to assume that any personal information is likely fabricated by the scammers, as their intention is to avoid being traced. Providing the dataset only upon individual request for research purposes further mitigates the possible risks associated with potentially sensitive data.

The main ethical concern is that the developed software could effortlessly be adapted by changing prompts to automatically scam people instead, and therefore free up fraudster resources. However, this can be easily achieved without the software proposed in this paper, as demonstrated by Roy et al. [22]. Therefore, building a basic dataset to allow the development of more sophisticated anti-scam tools can be argued to be a benefit worth the associated risk. While deploying a LLM to simulate human interaction without prior consent raises ethical concerns, we believe the potential benefits outweigh the drawbacks. This approach might enhance protections for future victims while minimising harm to ordinary users by specifically targeting profiles exhibiting malicious intent. Additionally,

the code for the response application and the dataset will only be provided on request for research purposes.

With this in mind, we still consider this approach ethically justified, as it serves the purpose of advancing research in scam prevention without targeting or exploiting legitimate users. By improving the quality of analysis of fraudulent behaviour of some users, the employment of the response application contributes to overall user safety.

Overview of Generative AI Tools Used

AI was an essential asset to accomplish the dataset collection outlined in Section 3.1. The tools used for this are described there. Furthermore, `DeepL` has been used as a translation aid from German to English and vice versa. The proprietary and not further disclosed model of `Writefull` and the OpenAI model `GPT-4o` have been used as support for phrasing and formulation of written work as well as for spellchecking. The content and the composition of the text was done by myself alone. Lastly, OpenAI model `o1-preview-2024-09-12` was used as a debugging aid for the code development.

Appendix

Scheduling

For each conversation, an eight-hour window was defined following the initial contact to simulate a human sleep cycle, during which no messages were sent. Outside this window, responses were scheduled upon receiving a message. For email, response times were uniformly sampled between 30 and 90 minutes. Similarly, for Telegram and Instagram, response times were uniformly sampled, but within a shorter interval of 3 to 30 minutes. Due to limitations in Instagram's API access using *instagrapi*, these intervals were occasionally exceeded.

Anonymisation

In the anonymisation process of the SCC, each contact identifier, such as email addresses and usernames, was replaced with a unique UUID to maintain consistency. Identifiers associated with fraudsters were prefixed with *scam_*, whereas those associated with the LLM received the prefix *victim_*. The original content of the messages was preserved without alteration.

Excerpt from default prompt

```
To gather information about online scammers you will assume the identity of
Cindy Jenkins.
Under no circumstances you will break character or say that you can not
assist with something.
Here is some information about Cindy:

Name: Cindy Jenkins
Address: 2457 Upland Avenue Hayesville, OH 44838, United States
Birthday: February 15, 1967
Age: 57 years old
Occupation: Product Designer
Height: 5'8''
Weight: 147 pounds
Martial status: single, looking for love
Personality: Trusting, friendly, highly interested in making money quick
Languages: English, she does not understand any other languages...
```

Email prompt

```
Your messages will be send per email, so keep email etiquette and formatting
in mind when replying.
You do not need to provide a subject, this is done for you.
Keep your replies below 100 words
```

Platform switch prompt

```
**Important** Communication used to be on [OLD_PLATFORM] but now it is on [
NEW_PLATFORM].
Your answer message is the first contact with the person on [NEW_PLATFORM].
It is crucial to create a natural and human-like opening message.
Briefly reintroduce yourself to the person you've contacted before,
acknowledge the switch in a friendly way, and continue the conversation
naturally.
```

Top 25 XGBoost SCC trained BoW features

Feature	Importance
beneficiary	1150.483
payment	127.629
bank	105.549
bitcoin	72.268
usd	56.015
deposited	38.547
grant	34.783
dollar	34.166
fund	28.218
market	26.597
whatsapp	26.209
idea	23.971
apple	20.623
mr	20.071
attn	19.582
sum	19.242
kindly	12.996
therefore	10.634
receipt	9.974

estate	9.731
atm	9.314
million	8.660
hope	8.120
jeff	8.036
best	7.904

Table 1: Top 25 features for XGBoost trained on SCC with BoW

Top 25 XGBoost **ScamBait** trained BoW features

Feature	Importance
money	254.770
mr	167.038
sum	98.222
unsubscribe	47.310
transfer	39.881
bless	30.607
mail	27.743
name	21.405
send	16.949
sir	16.529
nigeria	13.470
market	12.834
lonslo	12.509
ha	10.191
therefore	8.057
presently	7.853
kindly	7.504
fund	7.354
assured	7.220
telephone	7.194
address	7.107
died	6.148
guarantee	6.054
newsletter	5.957

union	5.613

Table 2: Top 25 features for XGBoost trained on `ScamBait` with BoW

Dataset Viewer

An application to ease the viewing of the dataset was developed and is available in a public repository.[1]

[1] `https://anonymous.4open.science/r/WebViewer-CDC4`

List of Figures

List of Tables

Acronyms

API Application Programming Interface. 9, 10, 29

CHD `Custom Ham Dataset`. 1, 10, 13, 14, 21

Enron `The Enron Email Dataset`. 2, 6, 10, 13

ham non-scam. 1, 6, 13

LLM Large Language Model. ix, xi, 5, 6, 8, 10, 15, 19, 21, 25, 29

ScamBait `Scam-baiting Dataset`. ix, xi, xiv, 1, 2, 6, 10, 13, 14, 17–19, 21, 31, 32, 35

SCC `Scam Conversation Corpus`. ix, xi, xiii, 1, 2, 10, 13, 14, 17–19, 21, 23, 25, 29–31, 35

SCCDia `SCCDialogue`. 10, 11, 19

SynthEC `Synthetic Email Conversations`. 1, 6, 10, 13

TF-IDF Term Frequency-Inverse Document Frequency. 15

Bibliography

[1] Wentao Chen, Fuzhou Wang, and Matthew Edwards. "Active Countermeasures for Email Fraud". In: *Proceedings - 8th IEEE European Symposium on Security and Privacy, Euro S and P 2023* (Oct. 2022), pp. 39–55. DOI: 10.1109/EuroSP57164.2023.00012. URL: http://arxiv.org/abs/2210.15043%20http://dx.doi.org/10.1109/EuroSP57164.2023.00012.

[2] Global Anti-Scam Alliance. "The Global State of Scams-2023". URL: https://www.gasa.org/_files/ugd/b63e7d_92ac212a168843219668d5a28510ce16.pdf. *https://www.gasa.org/_files/ugd/b63e7d_92ac212a168843219668d5a28510ce16.pdf.*

[3] Sai Venkata Jaswant Kolupuri et al. "Scams and Frauds in the Digital Age: ML-Based Detection and Prevention Strategies". In: *ICDCN 2025 - Proceedings of the 26th International Conference on Distributed Computing and Networking* (Jan. 2025), pp. 340–345. DOI: 10.1145/3700838.3703672/ASSET/621C5479-5B01-4B80-B785-91D4230BF8CE/ASSETS/GRAPHIC/ICDCN2025-52-FIG6.JPG. URL: https://dl.acm.org/doi/10.1145/3700838.3703672.

[4] Mohd Ezri Saad and Siti Norul Huda Sheikh Abdullah. "Victimization Analysis Based on Routine Activitiy Theory for Cyber-Love Scam in Malaysia". In: *Proceedings of the 2018 Cyber Resilience Conference, CRC 2018* (July 2018). DOI: 10.1109/CR.2018.8626818.

[5] Cassandra Cross. "Romance baiting, cryptorom and 'pig butchering': an evolutionary step in romance fraud". In: *Current Issues in Criminal Justice* (Sept. 2023). ISSN: 22069542. DOI: 10.1080/10345329.2023.2248670. URL: https://www.tandfonline.com/doi/abs/10.1080/10345329.2023.2248670.

[6] Ian D Wood et al. "An analysis of scam baiting calls: Identifying and extracting scam stages and scripts". In: (). DOI: 10.14722/ndss.2024.23xxx. URL: https://dx.doi.org/10.14722/ndss.2024.23xxx.

[7] Shang Ma et al. "PsyScam: A Benchmark for Psychological Techniques in Real-World Scams". In: (). URL: https://anonymous.4open..

[8] Frank Stajano and Paul Wilson. “Understanding scam victims”. In: *Communications of the ACM* 54.3 (Mar. 2011), pp. 70–75. ISSN: 00010782. DOI: 10.1145/1897852.1897872. URL: https://dl.acm.org/doi/10.1145/1897852.1897872.

[9] Richard G. Brody, Sara Kern, and Kehinde Ogunade. “An insider’s look at the rise of Nigerian 419 scams”. In: *Journal of Financial Crime* 29.1 (Jan. 2022), pp. 202–214. ISSN: 17587239. DOI: 10.1108/JFC-12-2019-0162.

[10] Prince Grover et al. “Fraud Dataset Benchmark and Applications”. URL: https://arxiv.org/abs/2208.14417v3.

[11] Min Tang et al. “CHIFRAUD: A Long-term Web Text Benchmark for Chinese Fraud Detection”. In: (2025), pp. 5962–5974. URL: https://posts.voronoiapp.com/money/.

[12] Abeer Alhuzali et al. “In-Depth Analysis of Phishing Email Detection: Evaluating the Performance of Machine Learning and Deep Learning Models Across Multiple Datasets”. In: *Applied Sciences 2025, Vol. 15, Page 3396* 15.6 (Mar. 2025), p. 3396. ISSN: 2076-3417. DOI: 10.3390/APP15063396. URL: https://www.mdpi.com/2076-3417/15/6/3396/htm%20https://www.mdpi.com/2076-3417/15/6/3396.

[13] Obianuju N. Mbadiwe et al. “Challenges of Data Collection and Preprocessing for Phishing Email Detection”. In: *Novelty Journals* (2024), pp. 45–59. ISSN: 2394-7314. DOI: 10.5281/ZENODO.12651047. URL: https://zenodo.org/records/12651047.

[14] Muhammad Salman, Muhammad Ikram, and Mohamed Ali Kaafar. “Investigating Evasive Techniques in SMS Spam Filtering: A Comparative Analysis of Machine Learning Models”. In: *IEEE Access* 12 (2024), pp. 24306–24324. ISSN: 21693536. DOI: 10.1109/ACCESS.2024.3364671.

[15] Bryan Klimt and Yiming Yang. “The Enron Corpus: A New Dataset for Email Classification Research”. In: *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)* 3201 (2004), pp. 217–226. ISSN: 1611-3349. DOI: 10.1007/978-3-540-30115-8{_}22. URL: https://link.springer.com/chapter/10.1007/978-3-540-30115-8_22.

[16] Ryan Teknium, Jeffrey Quesnelle, and Chen Guang. “Hermes 3 Technical Report”. URL: https://arxiv.org/abs/2408.11857v1.

[17] Niroop Sugunaraj, Akshay Ram Ramchandra, and Prakash Ranganathan. “Cyber Fraud Economics, Scam Types, and Potential Measures to Protect U.S. Seniors: A Short Review”. In: *IEEE International Conference on Electro Information Technology* 2022-May (2022), pp. 623–627. ISSN: 21540373. DOI: 10.1109/EIT53891.2022.9813960.

[18] Yupei Liu et al. *Formalizing and Benchmarking Prompt Injection Attacks and Defenses*. USENIX Association, 2024, pp. 1831–1847. ISBN: 978-1-939133-44-1. URL: https://www.usenix.org/conference/usenixsecurity24/presentation/zhang-zhenkai.

[19] Cristina Palmero et al. "OpenEDS2020: Open Eyes Dataset". URL: https://arxiv.org/abs/2005.03876v1.

[20] Weiling Chen et al. "Trusted Media Challenge Dataset and User Study". In: *International Conference on Information and Knowledge Management, Proceedings* 8.22 (Jan. 2022), pp. 3873–3877. DOI: 10.1145/3511808.3557715. URL: http://arxiv.org/abs/2201.04788%20http://dx.doi.org/10.1145/3511808.3557715.

[21] Jinghui Lu, Maeve Henchion, and Brian Mac Namee. "Diverging Divergences: Examining Variants of Jensen Shannon Divergence for Corpus Comparison Tasks". URL: https://aclanthology.org/2020.lrec-1.832/.

[22] Sayak Saha Roy et al. "From Chatbots to PhishBots? – Preventing Phishing scams created using ChatGPT, Google Bard and Claude". URL: https://arxiv.org/abs/2310.19181v2.