DEPARTMENT OF COMPUTER SCIENCE

Scam-baiter Mailbox - Reply to Scammers Automatically

Fuzhou Wang

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Master of Science in the Faculty of Engineering.

13th September 2020

Declaration

This dissertation is submitted to the University of Bristol in accordance with the requirements of the degree of MSc in the Faculty of Engineering. It has not been submitted for any other degree or diploma of any examining body. Except where specifically acknowledged, it is all the work of the Author.

Fuzhou Wang

Fuzhou Wang, 13th September 2020

Executive Summary

Online scams are becoming a more and more serious issue in modern society, causing great economic losses each year. Though great efforts have been made to combat cyber fraud, there are still many scam emails which escape from anti-spam systems and appear in potential victims' inbox. Given this, a group of people called scam-baiters take action by replying to scammers, trying to establish long-term communication with them. Scam-baiters aim to waste the scammers' time so that scammers have less time available to defraud innocent people. Scam-baiting activity has an important effect on the prevention of online scams, but it also takes the scam-baiters plenty of time to play the role of victims and to write the emails to respond, which makes scam-baiting a painful activity that is both time-consuming and of low efficiency.

To address this issue, in this study, we aimed to create a mailbox that can carry out the scam-baiting conversations automatically. It would ideally be able to carry out long-term conversations with scammers without manual operation. To choose the best-fitted model for this task, a comparison in terms of the performance of interacting with scammers was carried out across a deep learning model called GPT-2, a Markov Chain text-generation model, and a random template model that sent scam-baiting emails randomly picked from the template pool.

The contributions we made in this project are as follows: 1. Re-organise an existing scam-baiting dataset for model training. (Chapter 3) 2. Fine-tune GPT-2, train a Markov Chain, and construct a random-template model for scam-baiting response generation. (Chapter 4) 3. Build the mail server and implement a robust experiment platform. (Chapter 5) 4. Recruit volunteers, collect data of scam emails and scammers' email addresses and use the collected fraudulent emails to conduct experiments by actively sending baiting emails to scammers and initiating long-term conversations. (Chapter 6) 5. Discuss case studies and the models' performance at wasting scammers' time. (Chapter 7)

In the one-month experiment, 33 scam emails were submitted by the participants and were used for conducting interactions with the scammers. Three of the conversations successfully evoked further responses from the scammer, of which two were of the random templated group and one was of the GPT-2 group. Although the results we got from the experiment were too few to support a strong conclusion about model efficacy, we still created an open-source dataset that describes the scammer - machine email bot interactions, which is publicly available and can be supplemented by scientists who do similar work in the future. An experimental platform was constructed, making it easier to carry out future experiments using more scam emails. Furthermore, we also gave some suggestions on model selection, experiment improvement, and performance measurement that would be useful for future studies.

Acknowledgements

I would like to thank my supervisor Dr Matthew Edwards first. He gave me every support not only in this project, but also in my PhD application. He led me to the academic path and taught me the correct way to carry out scientific research. Dr Edwards is more like a friend and mentor rather than my supervisor. What he taught me will benefit me for the whole life.

I would like to thank my tutor Dr Sion Hannuna, who gave me some valuable advice on doing the final project.

The special appreciation is for my friend Ning, for her continuous accompany during the COVID-19 pandemic. I enjoyed the time chatting with you when I was having rest.

Thank my family for their mental support when I was tortured by the loneliness during self-quarantine.

COVID-19 Statement

Due to the pandemic of COVID-19 and the university closure caused by it, the advertising in this project was limited to the range of the Internet. Since we planned to advertise in the university buildings, we believe that the COVID-19 outbreak has significantly decreased the number of participants we recruited for the project, and thus caused that the experiment materials we collected for this study were not enough to reach a conclusive statistical analysis.

Furthermore, all communications between me and my supervisor turned to online via Skype after the COVID-19 outbreak, which made it difficult to explain and demonstrate some technical issues remotely – this caused a low efficiency in the system implementation stage, and put great psychological pressure on me.

Contents

1 Introduction and Project Aim 1

1.1 Introduction 1

1.2 Project Aim & Objectives 2

2 Background and Context 5

2.1 Overview of Internet Email 5

2.2 Email Fraud 9

2.3 Scam-baiting 15

2.4 Models for Text Generation 18

3 Scam-baiting Dataset 23

3.1 Overview of the Scam-baiting Dataset 23

3.2 Data Re-organisation 24

4 Model Preparation 27

4.1 GPT-2 Fine-tuning 27

4.2 Markov Chain Training 28

4.3 Random-Template Model Buildup 28

5 Design of Experiment Platform 29

5.1 Overview of the Experiment Platform 29

5.2 Mailbox System 31

5.3 Participant Recruiting System 34

5.4 Scam-baiting Bot System 35

6 Experiment and Results 43

6.1 Experiment Running Summary 43

6.2 Random Template Group 44

6.3 GPT-2 Group 45

6.4 Markov Chain Group 47

7 Discussion 49

7.1 Reflections on the Experiment 49

7.2 Limitations and Proposed Solutions 50

7.3 Future Work 51

7.4 Project Evaluation 51

A Scam-baiting Dataset Sample 57

B Unconditionally Generated Samples by GPT-2 59

C Successfully Evoked Conversations 61

List of Figures

2.1 Email Transport Flow 7

2.2 Categorisation of Unwanted Emails 9

2.3 A Typical Phishing Email 10

2.4 A Typical Lottery Scam Email 11

2.5 A Typical BEC Fraud Email 12

2.6 The Portion of Successfully Attacked Victims *vs* the Viable Victim Density 17

3.1 The UML Diagram of the Dataset Structure 24

5.1 Workflow of the Experiment Platform 30

5.2 Processes of Sending and Receiving Emails 32

5.3 Structure of Scam-baiting Bot System 36

5.4 Scam-baitng Bot System Database Design 37

5.5 Notification Email 42

Chapter 1

Introduction and Project Aim

1.1 Introduction

There is an increasing number of Internet users. With the unprecedented development of network technology, everyone in every corner of the world has been connected by various means of communication, including email (electronic mail). Indeed, in this era, network technologies have made people's lives much easier. However, while great convenience has been brought and everyone benefits from it to some degree, the danger has significantly risen as well. According to an FBI report [1], their Internet Complaint Centre received more than 1,700,600 reports of Internet crime over the last five years (2015 - 2019). In 2019, the reported economic loss caused by Internet crime was $3.5 billion. This number in 2015 was $1.1 billion, tripled within only five years. Furthermore, cybercrime can cause immeasurable psychological trauma to the victims [2]. This phenomenon shows that Internet-facilitated crime is becoming a more and more serious threat to human society.

Nowadays, the volume of email transmission is enormous, and it still remains one of the most prevalent communication tools in cyberspace [3]. The features of email, such as cost-effectiveness and high efficiency [4], not only enable people to communicate and make schedules more easily, but also give the scammers a powerful tool to commit crimes and harm the innocent people. To take an example, as a main type of email scam, advance fee fraud caused an estimated economic loss of $12 billion in 2013 [5], making millions of people fall victim. Thus, email scam, with a fast development, has become one type of cybercrime that has the greatest impact on people's daily life.

Scam-baiters, typically, are a group of people who take part in activities to respond to the scam emails and seek long-term conversations with the scammers to waste their time. Although many scam-baiters take part in this activity only for fun, it objectively prevents the fraudsters from scamming other potential victims, as scam-baiting can sharply reduce the scammers' probability of finding real victims by increasing their false positive rate [6]. Additionally, scam-baiting can also provide invaluable data for the studies of scam email detection. However, scam-baiting activity requires great manual efforts, which makes it time-consuming and painful for baiters, limiting scam-baiting to a small circle of subculture.

This project intends to implement a mailbox that plays the role of scam-baiters, which can reply to the text-based scam emails automatically, using text-generation models. This automated machine baiter can release the human baiters from thinking up appropriate responses, and help to combat cybercrime in a novel way – imagine this mailbox can generate proper responses in most cases, then a large-scale deployment in the worldwide could help to frustrate online crimes by significantly increasing the false positive rate of scammers. Therefore, we selected three prospective models for this text generating task, including a deep learning model called GPT-2 [7], a Markov Chain model, and a model that randomly sends pre-written response templates. We fine-tuned GPT-2 and trained the Markov model on a scam-baiting dataset to make them generate text in the specific area of scam-baiting. To choose the model with the best performance in interacting with scammers, we constructed an email server as an experiment platform that can automatically reply to the scammers, and carried out real-world experiments to compare the "baiting abilities" across the three models in terms of the total number of the scammer's responses within each conversation, the scammer's persuasion mode shifts within the conversation, and the quality of the response text, etc. For the experiment, participants were recruited to provide us with the scam emails they received.

This dissertation will be expanded in the following structure:

1. Introduce the related literature in the background chapter (Chapter 2) to provide necessary technological and sociological context of this study.
2. The dataset we used for model training will be introduced in Chapter 3, and in the same chapter we will also describe how we pre-processed the dataset.
3. The procedures we took to prepare the text-generation models for scam-baiting interactions (how we fine-tuned GPT-2, trained a Markov Chain, and constructed the random-template model) will be discussed in Chapter 4.
4. In Chapter 5, we will introduce the details of the design and implementation of the experiment platform, including describing the experiment cycle.
5. The results of this experiment will be displayed in Chapter 6.
6. In the discussion chapter, we will analyse the experiment results, think from them, and give out some suggestions for the future work. We will also conduct a critical evaluation of this project in this chapter.

1.2 Project Aim & Objectives

The central hypothesis this project aims to explore is that deep learning natural language processing models could be trained to adapt to the context of scam-baiting and therefore be used to deceive the scammers and waste their time – and the huge pre-trained deep learning model can more effectively play the role of victim and perform more convincingly in the conversations after fine-tuning, compared to a simpler text-generation model and fixed reply templates.

Particularly, the objectives we set at the project design stage include:

1. Fine-tune the GPT-2 model to make it generate text for the purpose of automated scam-baiting.
2. Implement a robust experiment platform that can automatically carry out the scam-baiting conversations with the scammers.
3. Real-world deployment and experiment with an observation of the performance of communicating with scammers of three different models from simple to complex.
4. Comparison across three models based on the evaluation of these models with respect to their ability to consume scammers' time.
5. A description of the ability that models with different levels of intelligence can adapt to the context of scam-baiting conversations, with demonstration and evaluation in case studies.
6. Create an open source dataset of conversations between scammers and artificial intelligences, can be used by researchers who study fraud, deception, and AI.

As is discussed in § 2.3.2, scam-baiting activity is of great significance, because it can not only provide valuable dataset for research of scams, but also effectively increase the false positive rate for scammers, making their business less profitable. We expected that this study would bring a new approach to combat scams. For example, people could potentially have multiple servers like the mailbox proposed in this study running and handling scam emails around the world, or go further by integrating this into spam filters so that they both stop people seeing the scam emails and respond to waste scammers' time.

Additionally, this study brings us a deeper understanding of intelligent models and a better insight into the scope of application of software with different levels of intelligence. According to the no free lunch principle [8], sometimes less is more – more complicated models can often perform worse than a simpler one, and this study was expected to give the answer in the field of scam-baiting. The result of this research would give an example of the importance of choosing the right and appropriate model to solve a realistic problem.

Chapter 2

Background and Context

2.1 Overview of Internet Email

2.1.1 Email Transport Infrastructure

In this section, we will explain how an email is delivered to the destination after the user presses the "send" button on their client agent, and will discuss different components in the email infrastructure. In the process that an email is sent from a sender to a receiver, the email will go across several Internet email agents.

SMTP

SMTP (Simple Mail Transfer Protocol) is an Internet standard for email transmission between the SMTP client and the SMTP server by establishing a TCP connection between two machines, typically on port 25. Particularly, an SMTP client refers to a machine or program that transfers the email to the SMTP server, while the role of the SMTP server is reporting failure and error in an SMTP communication. An SMTP server can also perform as the SMTP client in the next stage of the email transport. To prevent spammers from creating email bots using VPS (Virtual Private Server), most Internet hosting services (e.g., Google Cloud) block all outgoing traffic through port 25 to avoid abuse.

As is mentioned in § 2.2.2, SMTP does not validate the identity of the sender, which makes it easy for scammers to carry out email spoofing. To comply with SMTP and combat online scams, DKIM, SPF, and DMARC are designed to cooperate with SMTP to authenticate the email sender (see § 2.2.2 for details).

Furthermore, one thing that should be noted is that the SMTP is only used for transferring emails, and it is not responsible for retrieving emails. The process that the MUA fetches emails from the server is facilitated by POP and IMAP (introduced below).

POP and IMAP

POP (Post Office Protocol) and IMAP (Internet Message Access Protocol) are designed for users to fetch emails from the server and manage their mailboxes remotely. After an email is successfully delivered to the receiver's folder on the remote email server (arrived at the ultimate destination and delivered by the MDA), the user usually needs to fetch it from the server and load it to local machines to view. IMAP allows users to access the mail storage remotely to manage their email account (e.g., create folders, server-side searching, and moving emails across folders, etc.), while POP only retrieves emails and disconnects from the server immediately after fetching. Nowadays most MUAs provide both POP and IMAP mechanisms to retrieve emails from the server.

MUA

MUA (Mail User Agent) is also referred to as the email client. An MUA is often desktop software or web-based applications used for writing, reading, and sending emails [9]. To take an example, Outlook is MUA software developed by Microsoft, and Gmail is a web-based MUA of Google. Typically, the MUA has basic functionalities including retrieving emails from a remote server, composing an email in the correct format, and submitting the message to an email server.

MSA

An MSA (Message Submission Agent) is a program that receives the submissions from the MUA and transfers the message to the MTA for further delivery. An MSA is often closely combined with an MTA. An MTA usually contains the function of MSA as well. The communication protocol used for MSA to receive emails from the user is SMTP (MSA performs as the SMTP server and MUA performs as the SMTP client). However, as is specified in the new standard RFC 6409 [10], the dedicated port for MSA to receive emails is port 587, instead of the common port number 25.

MTA

MTA (Message Transfer Agent) is the software that uses SMTP to transfer the email from one machine to another. It is also known as a mail relay. To build a mail server, an MTA program must be installed. Examples of MTA software include Postfix, Sendmail, and qmail. An MTA accepts emails from MSA or another MTA and will transfer the email to the next MTA if the current MTA is not the ultimate destination itself. All transmissions between MTA and MTA, or between MSA and MTA, use SMTP via port 25 as is regulated in the Internet standard. In the communication between MSA and MTA, the MTA is the SMTP server while the MSA is the SMTP client.

MDA

The MDA (Message Delivery Agent) works together with the ultimate destination MTA. When a new email arrives at the ultimate MTA, it will be handed to the MDA, and the MDA is responsible for placing this email to the user's folder, as a result of which, the MUA user can fetch the emails from their account folder. An example of MDA is procmail.

Email Transport Flow

The whole email transport flow can be summarised as Fig. 2.1. In the whole process of email transport, an email is submitted by the MUA to the MSA and then is handed over to the MTA. The message likely goes through several email relays (MTAs), or it directly arrives at the ultimate destination MTA. After the email is delivered to the account's folder by the MDA, finally, the user can view and manage their emails on an MUA via POP or IMAP.

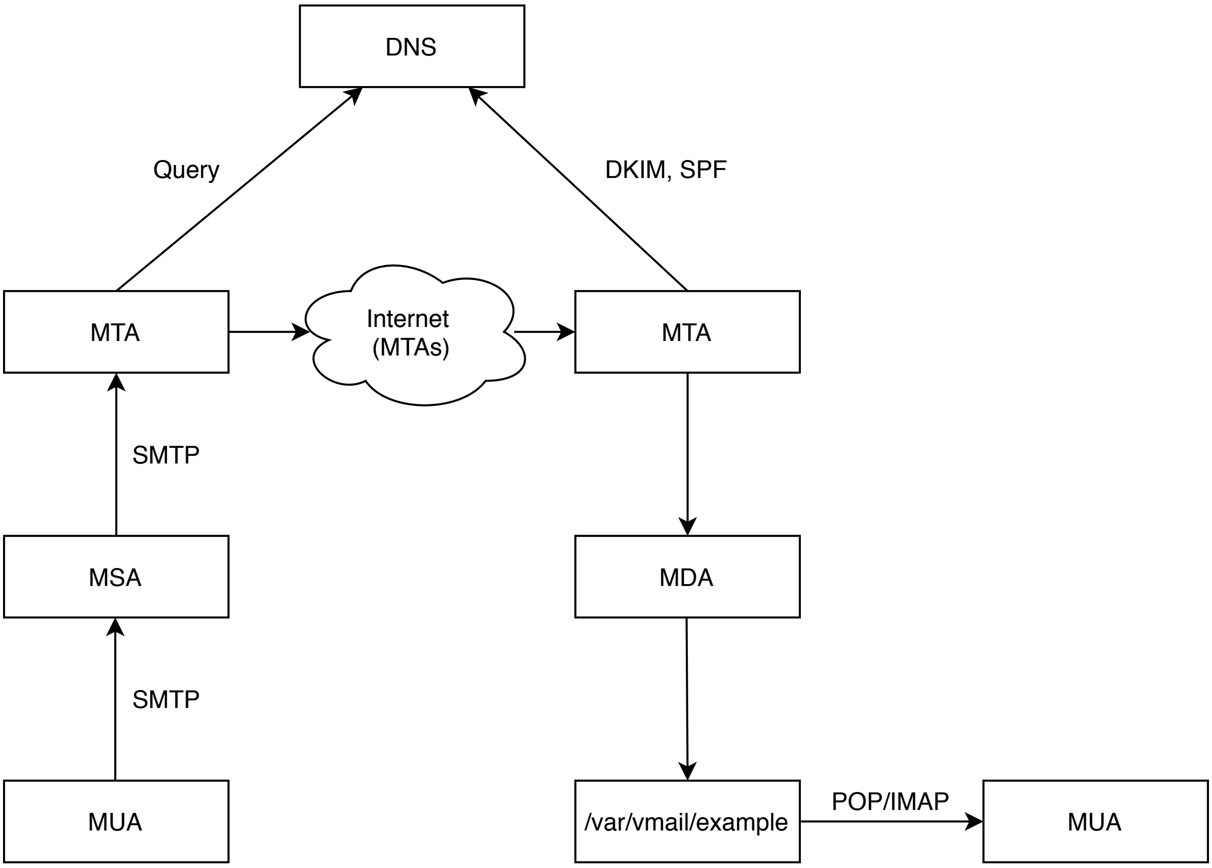


Figure 2.1: Email Transport Flow

2.1.2 The Format of Email

A standard email contains two parts: the header part and the body part. The header part indicates some important information for email transfer, email content parsing, and sender authentication, etc. The body part contains the message of the email, including the text content and the binary attachments. The email must meet this format so it can be transferred via SMTP.

Headers

The header part is composed of plenty of (key, value) pairs. All keys and values appear in the header part must be 7-bit ASCII characters. All characters other than ASCII must be encoded using MIME (Multipurpose Internet Mail Extensions) encoded-word syntax. The information contained in the headers usually includes the following mandatory fields [9]:

From: This field indicates the sender's email address, sometimes with an optional name.

To: Indicates the recipient's address. This field can be a list or a single value, specified by the sender. Addresses are separated by commas.

Subject: The subject of the email message. This field is mandatory but can be empty. Thus the case of the null subject should be considered while composing an email object programmatically.

Date: This field marks the timestamp of the sending time.

There are also some optional headers that are very common in email transport, such as *CC* and *BCC*. The header *Content-Type* is used for specifying the message format, and is vital for the MUA to parse the message and display the email correctly.

Body

An email body may be of single-part or multi-part. This is specified by the field *Content-Type* in the headers. A single-part message typically has *text/plain*, *image/jpeg*, and *audio/mp3*, etc. as its *Content-Type*, while this field of multi-part messages often has *multipart* on the left side of the slash, with different kinds of sub-types on the right side of the slash. For example, *multipart/mixed* indicates that the body of the email is composed of a plain text part and other non-text-based parts, typically various kinds of attachments. To take another example, the email that has the *Content-Type* of *multipart/alternative* contains several parts that each part is an alternative version of the other, such as an HTML part and a plain text part that have the same content, but just displayed differently.

Most emails contain a *text/plain* part, as some MUAs cannot display diverse content other than plain text. Given that the research scope of this study is text-based emails, all interactions between the machine baiters and the scammers were based on the plain text part only.

2.2 Email Fraud

2.2.1 Definition & Classification of Scam Emails

Scam email is a kind of unwanted email that deceives recipients for financial purposes, by which the victims often end up losing a great amount of money. People nowadays receive plenty of unwanted emails almost every day, where scam email is an important type. However, the unwanted email includes not only scam emails. Siadati etc. [11] divided all unwanted emails into three main types – spam, Trojan emails, and scams, while scam emails are further categorised into several sub-types, including phishing, advance fee fraud, and BEC (Business Email Compromise). In this section, we will discuss the three types of scam emails and analyse the features of each categorisation.

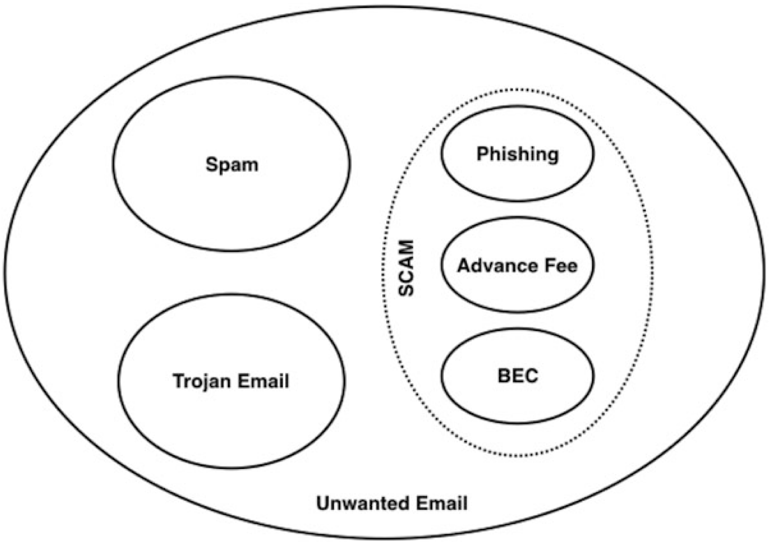
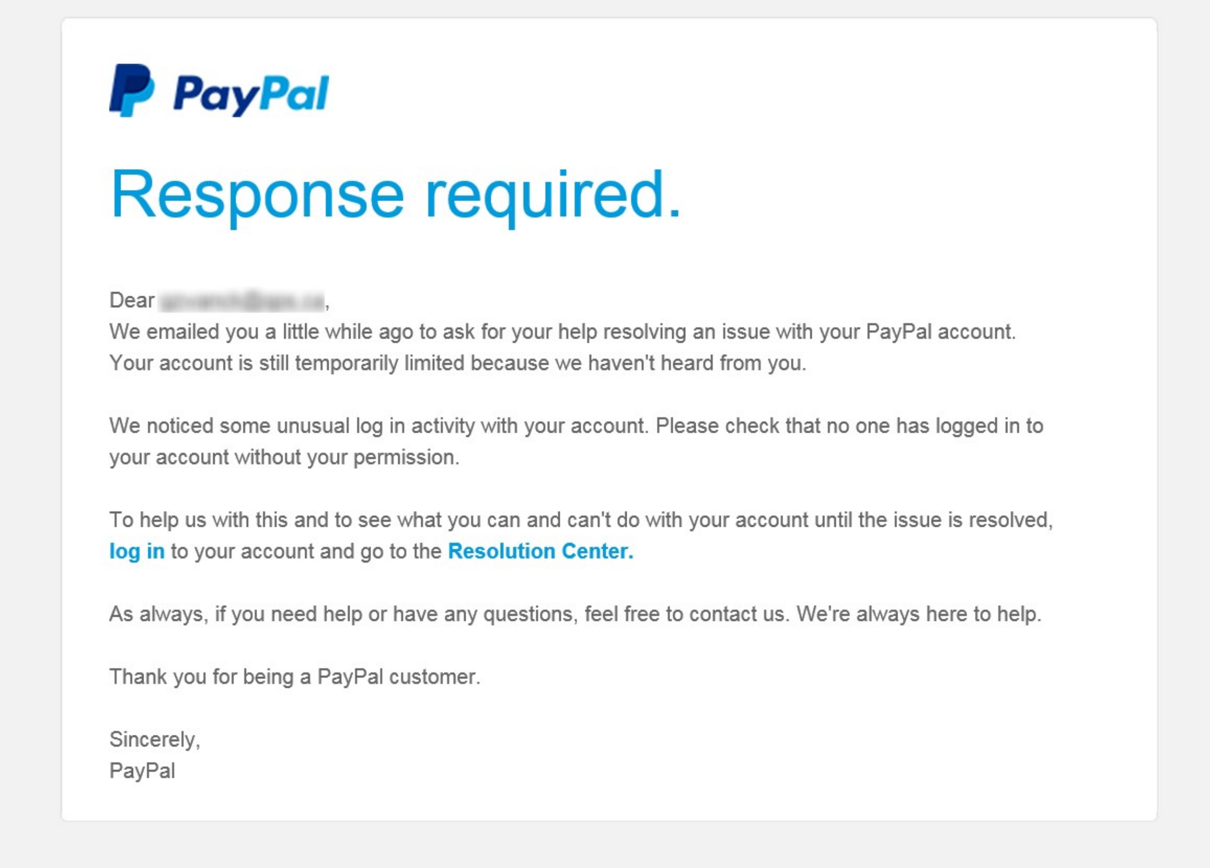


Figure 2.2: Categorisation of Unwanted Emails

Source: Siadati et al., 2016. *Traditional Countermeasures to Unwanted Email* [11]

Phishing fraud often aims to acquire the victims' credentials, such as personal information, credit card, or passwords. In a phishing attack, the scammers pretend to be a legitimate business and lead the victims to click the URL contained in the phishing email, which will direct the victims to the phishing website put together by the scammers [9]. This website often looks very like a trustworthy organisation's (e.g., Bank of England, PayPal, and eBay) and the victims will then probably input their credentials, and therefore, lose their money and sensitive key information. Fig. 2.3 shows what a phishing email often looks like.



Response required.

Dear [redacted],
We emailed you a little while ago to ask for your help resolving an issue with your PayPal account. Your account is still temporarily limited because we haven't heard from you.

We noticed some unusual log in activity with your account. Please check that no one has logged in to your account without your permission.

To help us with this and to see what you can and can't do with your account until the issue is resolved, **log in** to your account and go to the **Resolution Center.**

As always, if you need help or have any questions, feel free to contact us. We're always here to help.

Thank you for being a PayPal customer.

Sincerely,
PayPal

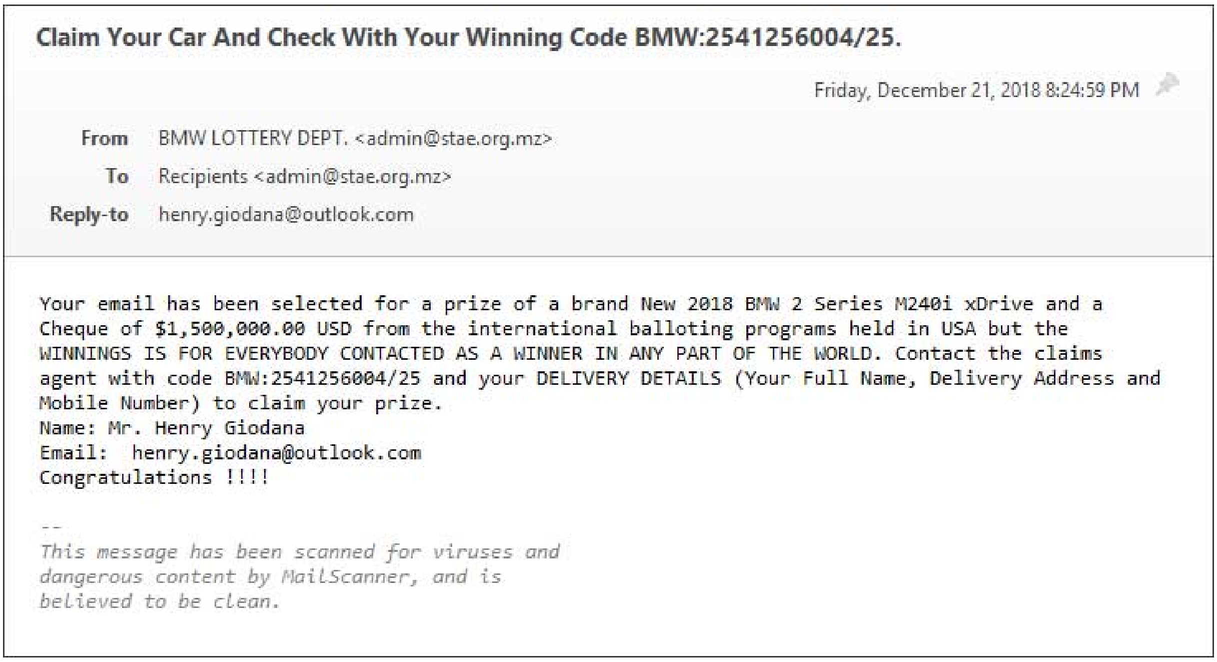
Figure 2.3: A Typical Phishing Email

Source: KnowBe4, 2020. *Phishing Examples* [12]

Advance fee fraud is also known as the 419 scam or Nigerian scam, in which the scammers craft stories that promise the victims a great sum of money, honour, and reputation, or an attractive job with high salary, etc., and the victims only need to pay a small amount of money for exchange [13]. Once the victims believe in the fake story and begin to interact with the scammers, they frequently fall into the trap of scammers and finally transfer their money to the scammers. After gaining the money, the scammer will either disappear or continue to ask the victim to transfer more money to them for other different reasons, until the victim realise that they are deceived. Advance fee frauds cause massive economic losses every year, e.g., it is reported that in 2013 there was a total loss of $12 billion [5]. Today the concept of advance fee fraud has been broadly extended. Various types of 419 scam have been developed by the scammers, such as lottery frauds, romance dating scams, unclaimed inheritance scams, stock frauds, and death threat scams etc. [5] The stories crafted by the scammers also become more and more diverse and convincing, making the potential victims even more likely to be defrauded.

In many cases, the scammers do not ask the victims to transfer the "deposit" to them directly. Instead, they persuade the victims to further contact them by email or other communication methods. For example, Fig. 2.4 shows a typical lottery scam email that prevails on the victims to further contact the scammers for the prize.

The scammers can often convince the victims by concealing their real purpose in the



Claim Your Car And Check With Your Winning Code BMW:2541256004/25.

Friday, December 21, 2018 8:24:59 PM

From BMW LOTTERY DEPT. <admin@stae.org.mz>
To Recipients <admin@stae.org.mz>
Reply-to henry.giodana@outlook.com

```
Your email has been selected for a prize of a brand New 2018 BMW 2 Series M240i xDrive and a
Cheque of $1,500,000.00 USD from the international balloting programs held in USA but the
WINNINGS IS FOR EVERYBODY CONTACTED AS A WINNER IN ANY PART OF THE WORLD. Contact the claims
agent with code BMW:2541256004/25 and your DELIVERY DETAILS (Your Full Name, Delivery Address and
Mobile Number) to claim your prize.
Name: Mr. Henry Giodana
Email:  henry.giodana@outlook.com
Congratulations !!!!

--
This message has been scanned for viruses and
dangerous content by MailScanner, and is
believed to be clean.
```

Figure 2.4: A Typical Lottery Scam Email

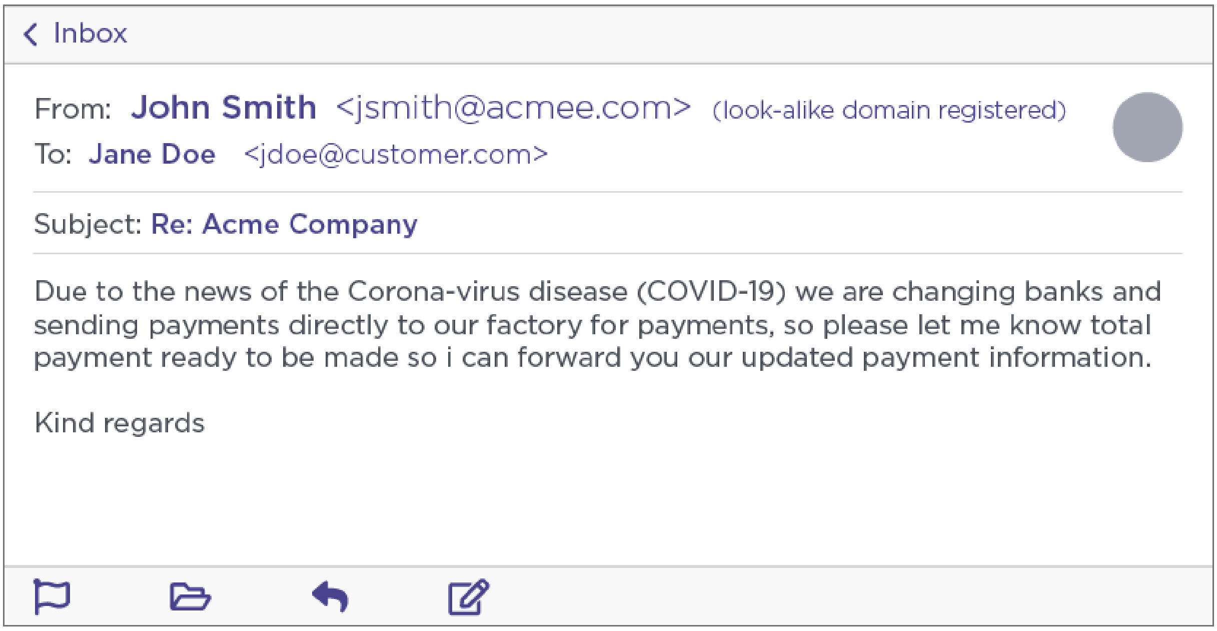
Source: Abrams, 2018. *Beware of BMW Lottery Email Scam Stating You Won a BMW M240i* [14]

initial solicitation email. Once the victims begin interacting with the scammers, they will trap the victims step by step with their well-prepared scripts. Furthermore, the scammers can usually bypass the content-based anti-spam systems (will be discussed in § 2.2.2) by not exposing their evil intention in the first contact, making it more difficult to detect this form of email fraud.

BEC is a kind of fraud that often targets business enterprises and aims at financial gain, using deceptive approaches like crafting fake invoices, forging email addresses that are similar to the legitimated ones, or stealing the legitimate accounts. By impersonating senior staff, reliable suppliers, or commercial cooperators, the scammers instruct the victims to transfer funds to them to achieve their goal [15]. BEC frauds often highly rely on social engineering. For example, the scammers need to dig into the relationship network within the organisation they target and learn the language habits of the person they want to impersonate so that they can make the email highly accurate and convincing [16]. Therefore, in many cases, the victims, who are often of low positions in the enterprise, are very likely to believe in the BEC schemes and cause great losses to the company.

The BEC fraud sometimes also requires manual effort from the scammers to interact with the victims. In the initial contact, the scammers often tend to perform as normal as possible and would prefer to start subsequent conversations with the victims, because they do not want to be too suspicious in the first email. Fig. 2.5 is a typical piece of the first contact email of BEC fraud, where the scammer did not give their account information but sought further conversation with the victim.

Although scammers commit to making their BEC frauds more convincing and hard to



From: **John Smith** <jsmith@acmee.com> (look-alike domain registered)
To: **Jane Doe** <jdoe@customer.com>

Subject: **Re: Acme Company**

Due to the news of the Corona-virus disease (COVID-19) we are changing banks and sending payments directly to our factory for payments, so please let me know total payment ready to be made so i can forward you our updated payment information.

Kind regards

Figure 2.5: A Typical BEC Fraud Email

Source: Gatlan, 2020. *Ancient Tortoise BEC Scammers Launch Coronavirus-Themed Attack* [17]

recognise, the potential victims can still separate BEC emails from genuine ones by assessing the credibility of emails, verifying the email addresses, and analysing the context of conversations. Companies can also educate the employees to raise their security awareness to prevent being attacked by BEC fraud [15].

To conclude, there are three main types of email fraud: phishing, advance fee fraud, and BEC, and each kind of scam emails may appear in various forms and is of high diversity. However, in spite of this, scammers are often likely to manually interact with the victims in the cases of advance fee fraud and BEC. This enables people to initiate long-term conversations with scammers to waste their time, and therefore, makes scam-baiting a potential countermeasure to combat online fraud (How scam-baiting can frustrate online scamming will be further explained in § 2.3.2).

2.2.2 General Strategies Against Email Fraud

People could indeed protect themselves from being scammed in an email fraud by promoting their security awareness through accepting education and learning from media, etc. However, no one can guarantee that they shall never fall into the trap of scammers. While facing the increasingly more dangerous Internet environment, everyone could be targeted by a scammer and everyone is vulnerable. Besides the non-technological measures to combat email scams, such as promoting the anti-fraud awareness among the public, it appears to be rather important to use technological methods to prevent the scam emails from showing up in users' inboxes.

Nowadays, the most common practice in filtering scam emails is to deploy the anti-spam system for the email transport infrastructure. Anti-spam systems aim to block all kinds of unwanted emails, including scam emails. The anti-spam systems detect unwanted emails, and intercept them halfway or put them in the users' junk email box, as a result of which the users would not see these dangerous scam emails, and thus avoid a potential economic loss from the very root. Anti-spam systems are implemented with different mechanisms. Particularly, there are three basic mechanisms to implement anti-spam systems: blacklisting approaches, anti-spoofing approaches, and content-based approaches [11].

In a large-scale spam or scam campaign, the senders' behaviour is often quite abnormal. For example, spammers and scammers usually send emails in bulk with plenty of email bots, while in particular, the scammers often pretend to be other trusted senders by spoofing (i.e., tamper the sender's address to pretend to be a legitimate account). Blacklisting and anti-spoofing approaches are built based on these features and are so-called behaviour-based mechanisms.

Blacklisting Strategy

A blacklist typically contains plenty of IP addresses that are of low reputation and thus are deemed to be spammers, which can help to intercept most spam campaigns. Any component on the route of email transport can send a query to validate whether an IP address is a spammer. In this way, both the spam bots employed by the spammers and the open relays that allow anybody to send emails without limitation and authentication would get a low reputation score and would be marked as spammers in the blacklist [11]. Research by Jung and Sit [18] investigated seven blacklist services that are broadly used, and their results revealed that 80% of spam sources are listed in at least one blacklist. This indicates that the blacklisting strategy is very effective in detecting large-scale spam/scam campaigns.

However, the blacklisting strategy also has its shortcomings. On the one hand, scammers can disable this mechanism by using the spam-delivery net that is not blacklisted and removing the bots once they are blacklisted. On the other hand, a blacklist will not be triggered if the scammers target only a small group of people, which is quite common in the cases of advance fee fraud. A more serious fact is that scammers nowadays widely use social networking services [19] to find potential victim groups, making their targeted group smaller and thus making it even more difficult for a blacklist to detect a scam email. As a result, it is far inadequate to intercept scam emails if people only have the blacklisting strategy.

Anti-spoofing Techniques

Many scammers try to pretend to be another legitimate account by spoofing. A sender address is easy to forge, as the SMTP (Simple Mail Transfer Protocol) does not verify the sender's identity. Over the years, many anti-spoofing methods were developed to validate the senders of email, including DKIM (DomainKeys Identified Mail), SPF (Sender

Policy Framework), and DMARC (Domain-based Message Authentication, Reporting and Conformance). In a DKIM validation, the sender sends the email affixed with a digital signature, and the recipient authenticates the sender's domain by DNS query, fetching the public key to validate. This method allows the receiver to confirm that the sender's domain of the email is reliable and authorised [20]. The SPF mechanism is created for the same purpose but uses different operation principle. It allows the email receiver to look up the DNS record to verify whether the email is sent from the IP address authorised by the domain owner, given all authorised IP addresses are provided by the domain and are recorded in the DNS infrastructure [11, 21].

DMARC does not authenticate an email sender directly. Instead, it must work together with DKIM and/or SPF. While working, the domain owner specifies which protocol (DKIM or SPF) is used and tells the receiver what to do (e.g., reject, discard, or report to the domain owner) if none of the authentications are successful [22]. These techniques guarantee that legitimate and reputable domains cannot be easily forged by the scammers.

Content-based Approaches

Although the behaviour-based approaches can detect most large-scale spam/scam campaigns, they are not able to intercept all unwanted emails. Scammers who do not spoof a legitimate sender and only target a small group of people are less likely to be detected by these mechanisms because their behaviour does not trigger the anti-spam system at all. For example, a scammer can register many mailboxes with various email service providers and send scam emails to the address list they buy from other cybercriminals who steal the user information by a cyber attack (i.e., data breach [23]). Instead of launching a large-scale campaign, in this case, the scammers often send emails in small groups, and all emails are sent from domains of high reputation (e.g., Gmail), which is very likely to disable the behaviour-based mechanisms. This is especially common practice in advance fee fraud. In order to distinguish scam emails from regular communications, computer scientists developed content-based classifiers to do this work automatically.

The content-based classifiers scan every incoming email for the receiver, checking whether an email is a spam email or a scam email. At the very beginning, service providers just used keyword and statistical features to filter these emails. However, this strategy becomes too powerless and ineffective when scammers start to carry out kinds of methods to bypass these mechanisms. For example, it is very common for scammers to write polymorphic emails (i.e., replace words and phrases with synonyms) to evade the content-based mechanisms.

To deal with these scenarios, machine learning began to be used for this task. Several algorithms are broadly used in this kind of classification task: Naïve-Bayes, SVM (Support Vector Machine), Decision Tree Algorithm, and Random Forests, etc. [24] With the development of deep learning, neural network models are trained for identifying more sophisticated email scams. To take an example, scammers might integrate the text to images, which will disable the text-based classifiers, while an anti-spam system armed with an OCR (optical character recognition) model will succeed in recognizing the images containing scam messages. Furthermore, a deep learning model can more precisely distinguish the language mode of scammers by training on a big dataset of scam emails [24]. This indicates that deep learning could be a better way to recognise scam emails than other traditional machine learning algorithms, in terms of accuracy.

However, the content-based mechanism has its weakness. First of all, as many machine learning algorithms (especially the supervised learning) deeply rely on the large dataset, it is crucial to collect real data to support the training process. This, thereby lead to another problem: As the real data of conversations between scammers and victims is lacking [13], nowadays most content-based classifiers focus on the first contact only, which results that the content-based mechanisms can hardly recognise scam emails on a conversational level. Therefore, in the case that the scammers try to take long-term interactions with the victims, there is lacking a mechanism that can warn the victims according to the exchange context, placing the victims in a very risky situation if a scammer makes their first contact email bypass the anti-spam system by slightly concealing their real purpose and writing novel scripts that the classifier has never seen before.

In real-world practice, an email service provider often combines various mechanisms closely together to filter unwanted emails. Different providers may employ different strategies. For example, Gmail makes use of Google's data centre and uses hundreds of rules to recognise the unwanted emails [24]. Each rule describes a feature of unwanted emails and is assigned with a weight to construct the formula to calculate the possibility that an email is a spam/scam email. Multiple machine learning algorithms (e.g., logistic regression and neural networks) are used by Google to classify the emails. Furthermore, OCR technology is applied by Gmail to extract text from the image to detect image-based scam/spam emails. Similarly, Yahoo and Outlook both have their unique methods to filter unwanted email. However, one thing is the same – their algorithms are based on large amounts of data, including user reports, search engines, or publicly available datasets.

2.3 Scam-baiting

2.3.1 What is Scam-baiting?

Scam-baiting is a kind of vigilante activity, in which scam-baiters reply to the original solicitation emails and seek further communication with the scammers, in order to waste scammers' time and prevent them from scamming other potential victims. This activity has become an internet subculture where there are some active scam-baiter communities across the Internet. The scam-baiters' motivations might come from community service, for which they protect other vulnerable victims by wasting the scammers' time. In other cases, scam-baiters act by the motives of status elevation in the baiting website, revenge, or just their own amusement [25].

There are plenty of scam-baiting websites on the Internet, such as TheScamBaiter.org[1],

---

[1] `http://www.thescambaiter.com/`

"Against 419 Artist" website[2], and the "what's the bloody point" site[3], etc. One of the most famous sites is 419eater[4], on which there are plenty of publicly available conversations between the scammers and the scam-baiters in the forums and letters archive. By sharing the scam-baiting email exchanges, the scam-baiters not only waste the scammers' time, but also report and warn people, document the scammer's practice, or acquire the scammers' identity information [26].

2.3.2 Why Scam-baiting?

Provide Valuable Baiting Datasets

Although anti-spam systems nowadays are well-developed and often equipped with artificial intelligence, they cannot intercept 100% all scam emails as the scammers are always proficient and experienced in bypassing the anti-spam systems. There are always some scam emails appear in the inbox, which might attract a potential victim though most people will ignore them [18].

As is discussed in § 2.2.2, when the victims start to write emails to the scammers and the situation comes to the further conversations, it is rather necessary to implement a system that can automatically detect the communications and make an early warning to the victims in order to protect the victims and frustrate the scammers halfway. However, due to privacy issues, most victims tend not to share their conversations with scammers to the public, as a result of which, there is a lack of real data that describes the communications between scammers and victims after the first email is responded. Given this, computer scientists must draw on other resources that can help them to implement the classifiers that can make decisions according to the email exchanges in the conversation, instead of based on the first contact only. One of the promising substitutes of real data is the scam-baiting dataset.

Edwards et al. [13] developed a highly accurate classifier that can distinguish the exchange between scammers and baiters from regular email traffic. This study also points out that the persuasion mode of scammers shifts over the whole conversation. The success of this research indicates the feasibility of training a model that can distinguish 419 scam conversations from normal email exchanges according to the context, and the utilisation of scam-baiting dataset in this research implies that scam-baiting activity provides a valuable source of data that researchers can draw on, especially when the real interactions between scammers and victims are inaccessible due to the privacy issue.

---

[2] `https://wiki.aa419.org/index.php/Main_Page`

[3] `https://www.whatsthebloodypoint.com/`

[4] `https://www.419eater.com/`

Protect Other Potential Victims

Another reason why we attach great importance to the scam-baiting activity is that scam-baiting can help to protect other vulnerable people from being scammed by wasting scammers' time. A study by Herley [6] pointed out that scam-baiting activity can sharply reduce the number of victims found by scammers through decreasing the density of viable targets (i.e., the targets that can generate financial gain), making them less likely to harm the real victims. As the cost of an email scam is very likely to be expensive and cannot be ignored, the scammers cannot attack all the potential victims on their email list. Instead, the scammers must estimate the likelihood that attacking a specific potential victim is profitable and optimise their attack strategy. In this study, Herley modelled the scammer's attack strategy as a binary classification problem, used ROC (Receiver Operating Characteristic) curves to simulate the attacker's decision-making process, and analysed how the reduction in the density of viable targets affects the outcome in terms of the fraction of victims who are successfully scammed.

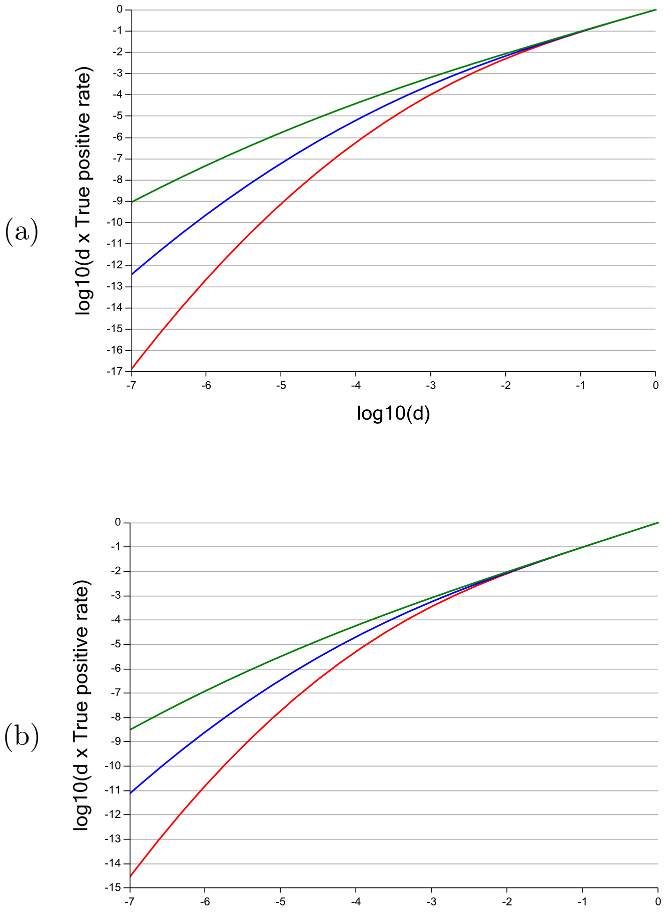


Figure 2.6: The Portion of Successfully Attacked Victims *vs* the Viable Victim Density

Source: Herley, 2012. *Why do Nigerian Scammers Say They are from Nigeria?* [6]

Fig. 2.6 sketches the relationships between the fraction of viable users who get attacked and the density of viable victims among the given population. In this figure, (a) and (b) show the trends given the different conditions of G/C (G: the expected financial gain to attack a viable victim; C: the cost of a single attack), respectively. For (a), the ratio is 20, while for (b) it is 100. Three colours indicate the scammer's ability to distinguish whether a target is viable – the green curve represents the best ability that the classifier has an accuracy of 99%, the blue one has 95%, and the red one has the accuracy of 90%. The figure shows that as the density descends, the portion of victims who are successfully attacked drops much faster, and this trend becomes increasingly significant when the density gets smaller. An example is that given a scammer can 90% accurately determine the viability of a target (the red curve)[5], the fraction of viable victims attacked will drop by three orders of magnitude when the density reduces from 10^{-5} to 10^{-6}. The only way to rescue the scammer's disastrously low success rate is to improve the accuracy of the classifier [6]. However, in the real world, the only resource that the scammers can draw from to train their classifier[6] is the experience they gained from successful attacks in the past. In the scenario where the density of viable targets is very low, the scammers can hardly find enough data to train their classifiers, making it almost impossible to achieve high accuracy.

Thus, scam-baiting can be a prospective way to frustrate the scammers. More specifically, by inserting false positives to the population, this activity sharply reduces the density of viable targets. This is the most important reason why we do this project – by deploying the scam-baiter mailbox all over the world, these mailboxes that can interact with the scammers automatically will play the role of false positives and effectively decrease the viable density, and thereby make it less likely for the vulnerable people to be scammed.

2.4 Models for Text Generation

2.4.1 GPT-2: the Deep Learning Model

The first model we used for automated replying is GPT-2 [7], which is a deep learning text generation model that can predict the following text according to the given context. This model was released by OpenAI in 2019, and is an NLP (Natural Language Processing) model trained on a 40GB large corpora called WebText scraped from millions of webpages [7]. In this study, researchers found that the deep learning language model can perform very well in different fields when it is trained on a large and diverse enough dataset. GPT-2 shows great performance on unsupervised multi-tasks. Due to the diversity and large scale of the training data, the model in zero-shot setting achieves state-of-the-art performance on 7 out of 8 tested datasets. The research shows that this high-capacity NLP model trained to maximize the likelihood of a sufficiently varied text corpus can perform multiple tasks, including generating the conditional text of good quality, question-answering, reading comprehension, summarization, and translation, without explicit supervision.

---

[5] The fact is that, even if 90% is the lowest within the figure, in real practice, few scammers can hit such high accuracy.

[6] It is possible that scammers employ machine learning models as their classifiers, but in most cases, the classifier is just scammers themselves.

The main approach to realise the GPT-2 model is language modelling, which has been significantly improved in recent years with the progress made in the models that can compute the conditional probabilities in language sequences. The neural network architecture called Transformer proposed by Vaswani et al. [27], which is based on the self-attention mechanism is exactly the basis of GPT-2.

Fine-tune GPT-2

Fine-tune is an approach to make GPT-2 generate text in a more specific field. For example, by fine-tuning GPT-2 with news about Brexit, then we can use it to generate some "fake" news only about Brexit. Lee etc. [28] successfully made GPT-2 able to generate text in the specific field of patent claims by fine-tuning GPT-2. Their results show that the GPT-2 can fast adapt to the specified context while fine-tuning and generate text of high quality after this process. More importantly, this study reveals that the GPT-2 can be fine-tuned on the dataset with intentionally inserted structural tags to force the model to produce formatted text. For example, in this study, by inserting special tags like "@@@" to the training set, the model started to generate the text in an appropriate format after only 36 steps of fine-tuning, which satisfied their requirement of generating patent-like text.

Pre-trained GPT-2 Models of Different Sizes

There are four pre-trained models of GPT-2 with different sizes. Due to the concern that the GPT-2 might be used to generate text for deceptive and malicious purpose [29], when the paper of GPT-2 was first published, OpenAI declared it "too dangerous to release" the complete model. Instead, they follow a staged release strategy, which means that they would publish their model piece by piece, from the smallest one to the largest one (These models are 124M, 355M, 774M and 1558M in terms of the number of parameters of the deep-learning network). Fortunately, until now the largest model has already been published and accessible to everyone. However, due to the volume of these models is incredibly enormous, the 774M and 1558M models are unable to be fine-tuned on a modern GPU [30]. As a result, the biggest GPT-2 model that can be fine-tuned (on a normal GPU) to implement some novel applications is the 355M one.

Although the 774M and 1558M models are too big to be fine-tuned, and thus cannot be put into use in most cases, the smaller 355M model is enough for generating the text of high quality. In the study by Lee and Hsiang [28], they fined-tuned the 355M model[7] and their results showed that the 345M model after fine-tuned can generate text of high coherence and complexity in the field of patent. This proves the feasibility of fine-tuning a 345M model on a normal GPU to generate the state-of-the-art text.

---

[7] When this paper was published, the sizes of the models were incorrectly specified by OpenAI as 117M, 345M, and 762M, etc., but were rectified after a short while, as a result of which, the 345M model specified in Lee and Hsiang's paper is actually the 355M model.

Shortcomings of GPT-2

However, the GPT-2 model also has some shortcomings. It is important to note that there is a token limit that the whole length of the input prefix and the generated text cannot exceed 1024 tokens. In other words, the context size of GPT-2 is 1,024 tokens [7], which means that while generating text conditionally according to the input prefix, the total length of the input text plus the generated text is no more than 1,024 tokens [30]. As the encoding method of GPT-2 is modified BPE (Byte Pair Encoding) that lies at the middle of the character level and the word level, the maximised context size of GPT-2 is usually 3-4 English paragraphs [30]. This significantly limits the performance of GPT-2, especially in the scenario that the input prefix is long, because when the prefix contains too many tokens, a part of the input will have to be discarded to leave enough size for the generated text.

2.4.2 Markov Chain Model

Markov Chain Model is a stochastic model describing a mathematical process where the state changes only according to its previous state [31]. Nowadays Markov Chain models are widely used in many fields such as computational finance, speech recognition, handwriting recognition, and machine translation. By training the Markov Chain, a matrix that describes the statistical distribution of probabilities of state transformation can be obtained. The Markov text-generation models make use of this matrix to predict the next token by the probability according to the previous token. It is easy to see that the text generating strategies of Markov Chain and GPT-2 are both predicting the next one token, but the difference is that the Markov Chain generates the token only based on the previous one token, while GPT-2 predicts the next token according to the whole previous context by language modelling, with the far more complex algorithm of deep learning.

The disadvantage of Markov Chain Model is that the text generated by it often “looks good but makes no sense”, while the advantage of it is that the model is much less complex than deep learning models – The simplicity of Markov Chain makes its running speed much higher and makes the computing resources consumed much less while comparing with GPT-2.

2.4.3 Spamnesty

There are some existing chatbots that can be acquired from the Internet, of which Spamnesty is a famous and widely used one [32]. The working principle of Spamnesty is not too complicated. Once people receive an email from the scammer, they forward it to Spamnesty’s mailbox, and Spamnesty will play the role of victims who have fallen into the trap of scam and take over the following exchanges with fraudsters. Here are some sample replies generated by Spamnesty which are posted on their official website:

> Hey there, Could you tell me where you are located? Also, could you tell me more about your offer? Stacey Brown CEO, MNesty, LLC
>
> Hello, can you give me an example of this? I would like to know roughly what it looks like. Tracey Dudley CEO, MNesty, LLC

It is easy to see these replies are generated by templates and random titles from their database. These responses are brief and short, and sometimes can produce long-term interactions with scammers. However, from viewing the samples, the model sends out some repeated replies because of the randomness of the way the model picks templates from the database. In some of the sample conversations, the model sends the same template for four or five times, which might let the scammer beware that the replies are generated by an automated email bot.

Chapter 3

Scam-baiting Dataset

3.1 Overview of the Scam-baiting Dataset

There are some publicly available conversations between scammers and the scam-baiters from a variety of online scam-baiting forums and message boards, such as 419eater.com. In a study by Edwards etc. [13], they organised a dataset called ADVANCE FEE SCAM-BAITING, which is composed of scam-baiting transcripts posted on the 419eater website and some other scam-baiting conversations from "What's the Bloody Point" website. This dataset contains a total number of 2,248 messages within 57 conversations.

The dataset we used in this project has a size of over 70MB. It is much bigger than the ADVANCE FEE SCAM-BAITING dataset, containing 658 conversations where 37,501 messages are included. Furthermore, different types of email fraud are included, not limited to 419 scams. A sample of JSON files in the dataset is shown in Appendix A. This dataset is a superset of the ADVANCE FEE SCAM-BAITING dataset, provided by Dr. Matthew Edwards. As is demonstrated in Fig. 3.1, in this dataset all data is saved in JSON files. Every single conversation is saved in one file as a JSON object, having keys of "title", "scammer", "location", "baiter", "notes" and "messages", of which "scammer" indicates the scammer's name in this conversation while "baiter" means the baiter's name. The value of "messages" is a JSON array, containing all exchanges between the scammer and the scam-baiter in the conversation. However, many of the JSON files have missing values in scammer and baiter fields, since the data is acquired from various sources and the public conversations scraped from the Internet are often incomplete in these fields.

The missing values in the scammer and the baiter's name fields can be a problem in some other studies, but they do not matter too much in this research. On the one hand, a good text-generation model that fits the scam-baiting tasks should have the ability to generate appropriate names and signatures according to the conversational context. On the other hand, training the models with the names of the scammer and the scam-baiter means introducing two parameters to the input, which requires manual efforts to read the incoming emails to get the names and then compose the input. This will make the system less automated. As a result, we dismissed the missing fields of "scammer" and "baiter" and re-organised the dataset to feed the models.

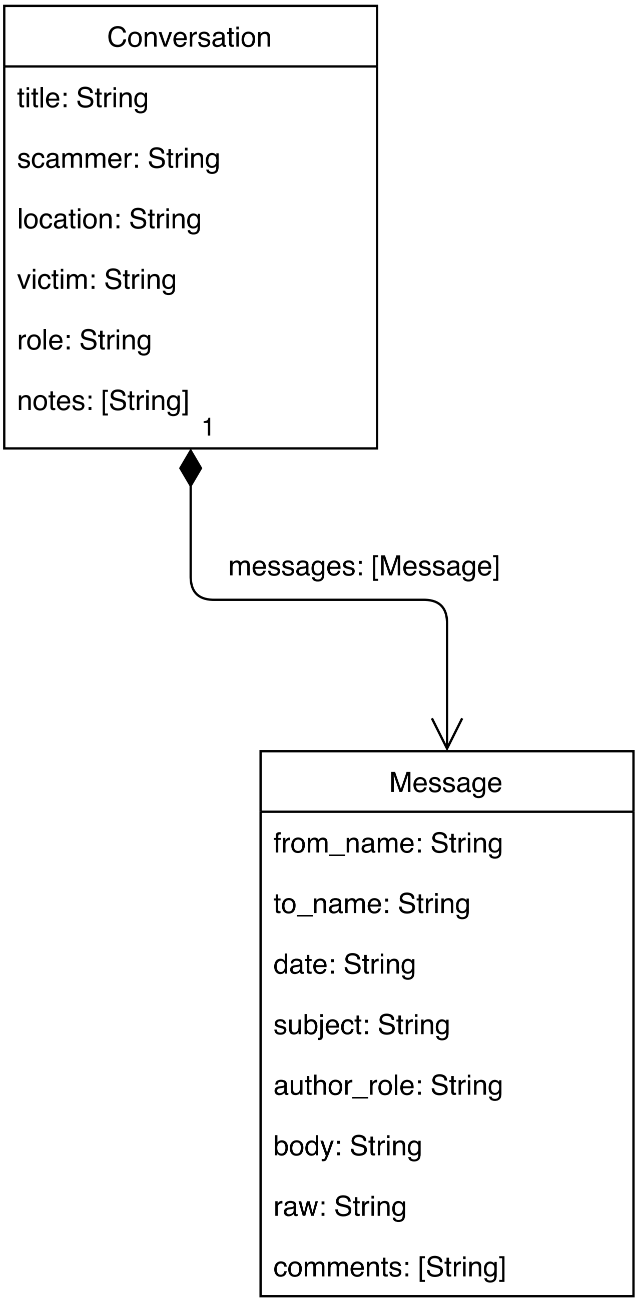


Figure 3.1: The UML Diagram of the Dataset Structure

3.2 Data Re-organisation

While fine-tuning GPT-2, the dataset must be re-organised due to the nature of this model. There are two main reasons for this. First of all, GPT-2 cannot distinguish different conversations in one large dataset, as it does not have the ability to recognise the beginning and the end of a single conversation within the file of the dataset [30]. Therefore, tags that indicate the beginning and the end of a conversation must be added to tell the model where the bounds of conversations are. As is documented in gpt-2-simple [30], the beginning and ending indicators are < |startoftext| > and < |endoftext| >, respectively. Secondly, as is discussed in § 2.4.1, GPT-2 can be forced to generate text in certain format by fine-tuning with the structural dataset (i.e., the dataset that is structured with special tags). In this study, we aim to make the model able to generate text in a conversation-like format, as a result of which, we inserted special delimitators into the dataset, which not only annotate the role of the scammer and the baiter but also indicate the start and end of each message.

We wrote a Python script to automatically organise the dataset. What we did was (1) adding the start and end indicators to each conversation, (2) marking the roles in each conversation with special tags, and (3) merging all conversations into one single file. Finally, the dataset we fed GPT-2 was in the following format:

```
<|startoftext|>
[scam_start]
Example of scammer message
[scam_end]
[bait_start]
Example of baiter message
[bait_end]
<|endoftext|>
```

By feeding GPT-2 with the dataset of the form above, we can force the model to generate the text strictly formatted. Similarly, we can also train the Markov model with this re-organised dataset. After training, these special tags are generated randomly according to the matrix of the probability distribution.

Chapter 4

Model Preparation

4.1 GPT-2 Fine-tuning

We mainly referred to the method demonstrated by Lee et al. [28] to fine-tune the pre-trained model with our scam-baiting dataset.

4.1.1 Fine-tuning Environment Setup

Our environment for fine-tuning GPT-2 is Google Colab, which provides a cloud computing platform equipped with GPU, having a configured Jupyter notebook environment and a strong connection with Google Drive. The GPU provided by Colab is NVIDIA Tesla K80, with available memory for around 20GB. Colab already has deep learning modules configured, which can be put into use directly. Although there is a 12-hour restriction on the time of use and there is a need for manual effort to restart the session, Colab is totally free for everyone and this allows researchers to use the expensive GPUs to do some experiments from the cloud without any payment.

Plenty of applications that take advantage of GPT-2 are implemented through fine-tuning, the source code of OpenAI GPT-2 [33] does not provide a method to fine-tune GPT-2, though. However, there are many resources we can draw from on GitHub, such as nshepperd/gpt-2 [34] and gpt-2-simple [30]. The latter one is well-documented and provides a number of APIs (e.g., fine-tuning, random sample generation and conditional samples generation) for users to easily make use of GPT-2 to do their specific tasks. In this dissertation, the tool we took advantage of to fine-tune GPT-2 and implement the scam-baiting web API is gpt-2-simple.

4.1.2 Model Selection & Fine-tuning Procedures

As is discussed in § 2.4.1, the biggest model that can be fine-tuned with a modern GPU is the 355M one. Thus, we chose to fine-tune the 355M model of GPT-2 on Google Colab. The public available repository of gpt-2-simple [30] provides a Jupyter notebook that presets the common hyperparameters for fine-tuning the model, and we modified the source code a bit to it suitable for our project[1]. We set the parameter sample_every to 100, which allowed the model to generate a sample every 100 steps so that we can trace and monitor the text quality. The learning rate of the fine-tuning process was set to 1e-4 instead of 1e-5 to avoid a slow convergence. We fine-tuned the 355M model for 35,000 steps on the scam-baiting dataset mentioned in Chapter 3. After fine-tuning, we managed to let the model generate the text that is of the expected conversation-like format, and the model steadily produced the text of high coherence and good quality, which looks very much like scam-baiting conversations. Some generated text samples are available in Appendix B.

4.2 Markov Chain Training

The Markov Chain model we used in this project was open-source software fetched from GitHub [35]. We trained the model according to the instruction posted in the repository, and we observed that the text generated by the model was very likely to be of the format that has the special tags in it. As this model can only generate text in a prompt, we modified the code to make it a module that can be imported by any Python scripts. Furthermore, as it was not guaranteed to generate the text with the specific format, we made the model generate text in a loop, where the program can only exit when the generated text has our required tags.

4.3 Random-Template Model Buildup

Besides the two models above, we also created a random-template model for comparison, which is inspired by Spamnesty. Although Spamnesty is an open-source software, it requires complex configuration procedures with Docker. Furthermore, it encapsulates the entire mail system, which conflicts with our mailbox system. As a result, we hacked the main functionality of the Spamnesty model by observing the scam-baiting conversations posted on their official website. We copied all templates from their website to build a template pool and used a random function to pick the templates from the pool. For the email signature, we created a random name pool for randomly picking from. This random-template model was implemented with Node.js and can be imported as a module by other scripts.

---

[1] The modified Jupyter notebook is now publicly available here: `https://colab.research.google.com/drive/1ejuctH3Mtz-rdN3_v2MNJHMZXvfKpSmQ?usp=sharing`

Chapter 5

Design of Experiment Platform

To compare the performance of the three text-generation models in terms of their ability to consume the scammers' time, we collected real scam emails by recruiting participants and carried out an experiment to evaluate model performance at real-world scam-baiting. In this chapter, we will introduce the design of the experiment platform and explain the main components of the whole system. Firstly we give an overview of the whole experiment platform and briefly introduce the experiment workflow. Then we look deep into the platform and give the details of the implementation and working principles of the three main components of the platform. Several technical issues in platform development are also discussed in this chapter.

5.1 Overview of the Experiment Platform

5.1.1 Component Introduction and Experiment Workflow

The experiment system is composed of three main sub-systems: the mailbox system, the participant recruiting system, and the scam-baiting bot system. The three components work with each other and cooperate closely together to make sure the experiment runs smoothly and robustly. As is shown in Fig. 5.1, the scam-baiting bot system is the core component that connects the other two systems. The whole experiment workflow can be divided into two key processes, which are scam email collecting (the blue square) and the scam-baiting interaction (the red square), respectively.

For the scam email collecting process, a participant can submit a scam email they received in the past one month via the participant recruiting system, and their submissions (including the email content and the scammer's email address) are stored in the file system as JSON files. The experimenter checks the submission folder once a day to check whether there are any new submissions and reviews the submissions to judge whether they are scam emails. The scam emails received will be recorded by the scam-baiting bot system, an email account will be created in the mailbox system for interacting, and the

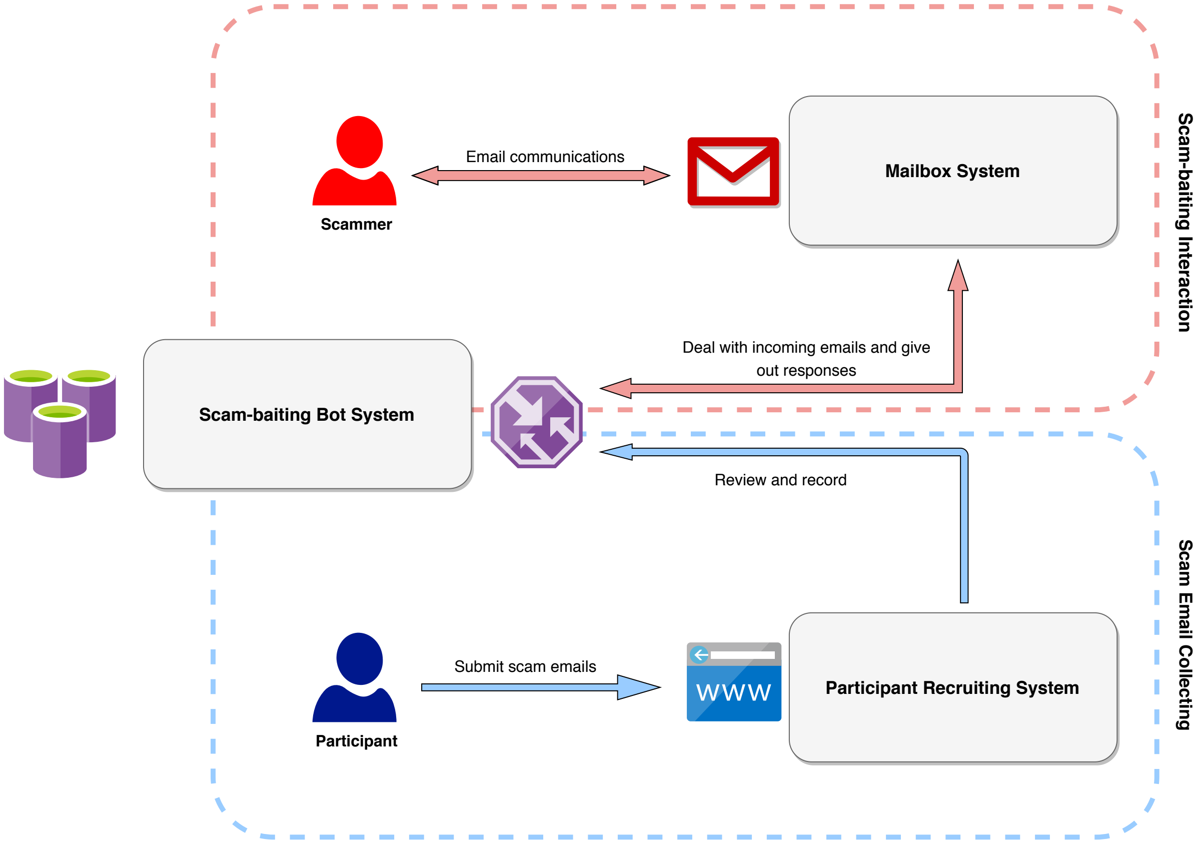


Figure 5.1: Workflow of the Experiment Platform

first response will be sent to the scammer – this indicates that the procedure comes to the next stage, which is the scam-baiting interaction process.

The interaction process is based on conversation records. For each conversation between the scammer and the baiting bot, a record is kept in the database of the scam-baiting bot system. Every time email arrives, the mailbox system sends a query to the scam-baiting system to check whether there is a corresponding conversation record in its database. The scam-baiting bot system will generate a response using one of the three text-generation models for the incoming email if the record exists, and then the generated response will be sent back to the scammer. So far, a complete cycle of the experiment is finished. One thing worth noting is that the text-generation model for each conversation is kept consistent during the interaction. For example, if the first response is generated with GPT-2, then all messages from the scam-baiting bot in the following conversation will be generated with GPT-2. The consistency makes sure that we can compare the performance across the three models by comparing the quality and time span of the conversations based on three different models.

5.1.2 Remote Servers and Hardware Support

We employed two remote servers on Google Cloud to build the experiment platform. One is a normal server carrying most parts of the platform, while another one is a high-performance computing server equipped with GPU, which works solely for the GPT-2 model to generate responses. This is to say, both the recruiting system and the mailbox system only exist on the normal server, and the scam-baiting bot system spans across the normal server and the computing server. Both servers are of the centos-7 operating system, and the GPU equipped for the computing server is one NVIDIA Tesla T4, which is of enough computing ability for GPT-2 to generate responses. The CPU and memory configurations of the two servers are the same – each server has one vCPU and 8GB of memory. However, as the normal server is also responsible for hosting the databases, storing all conversation information, and keeping log files, etc., we allocated 100GB of hard disk for the normal server, while only 50GB was allocated for the computing server.

5.2 Mailbox System

The mailbox system is responsible for sending and receiving emails, and transferring the email to the script of the scam-baiting bot system.

5.2.1 Architecture of the Mail Server

The mailbox system is a complex system that is deployed on the normal server, containing the functionality of MSA, MTA, and MDA. We chose to deploy the mail server with iRedMail [36], which is open-source software that puts together all necessary components of a mail server. This allows us to configure a full-featured mailbox on the remote server easily. The back-end of the mailbox can be selected from several main-stream databases, from which we chose MariaDB. The MTA component of iRedMail is Postfix, which is one of the most commonly used MTA software containing the functionality of MSA, and the MDA component is powered by another open-source software called Dovecot. What iRedMail does is just providing a collection of the necessary modules of a mail server and automatically configure them for the users, but users are allowed to customise the configuration of each component after installation. Specifically, the default of iRedMail configures SpamAssassin as the anti-spam system, but in our project, we were expecting to receive all emails from the scammers, and an anti-spam system is very likely to intercept an email from the scammer. As a result, we disabled all functions of SpamAssassin to make sure that all incoming emails can successfully be processed.

As we mentioned in § 2.1.1, most VPS providers including Google Cloud block all network traffic through port 25, which is the conventional port for establishing SMTP connections. Instead of using other ports that are not standard for email transmission, Google Cloud recommends users to employ reliable relay servers such as SendGrid, Mailgun, and Mailjet to send outgoing emails. In this project, we selected Mailgun as the relay server to send emails. We acquired a free trial account from Mailgun and were able to send 6,250 emails every month, which was enough for this small-scale experiment.

Mailgun provides an API and allocates an API key for each user, by which the users can send emails programmatically with different kinds of programming languages. In the experiment platform, most logic behind the other two subsystems (i.e., the recruiting system and the scam-baiting bot system) is implemented with Node.js, as a result of which, we chose to use the same programming language to write the email sending module in the mailbox system because this would make it easier for other components to call the interfaces of the mailbox system.

Although the outbound traffics are blocked by Google, the incoming traffics are not affected. This means that we can receive emails directly via SMTP, but can only send emails through the Mailgun relay. The sending and receiving processes are demonstrated in Fig. 5.2.

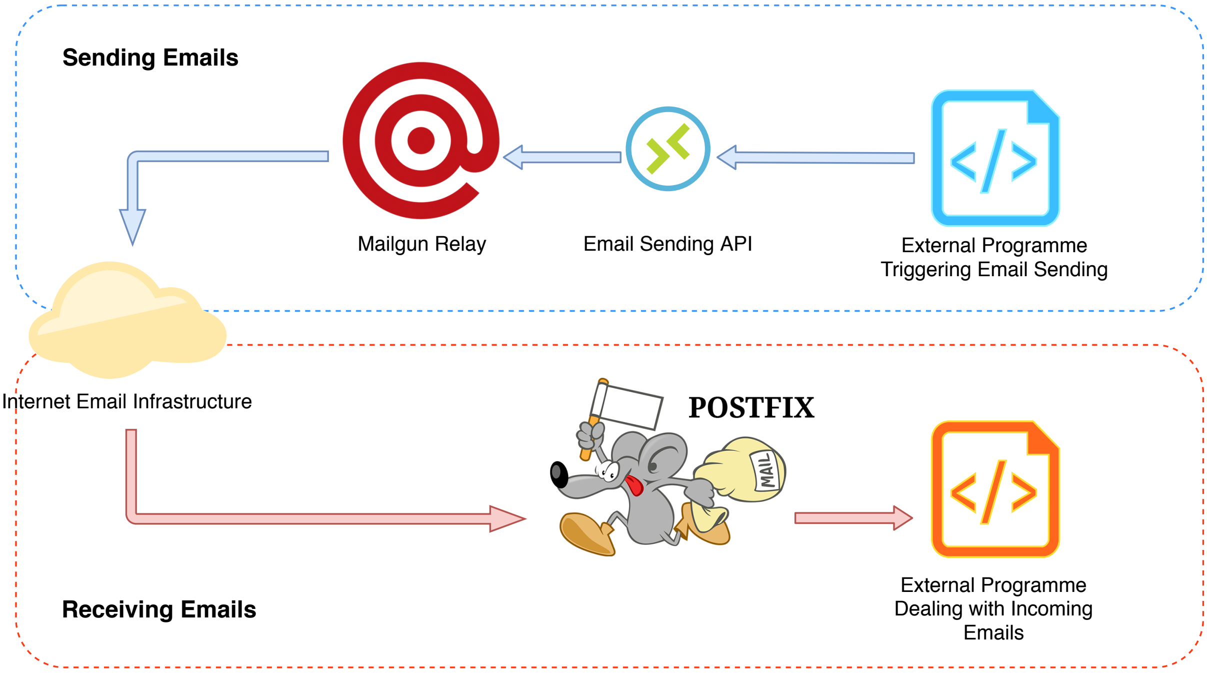

Sending Emails
Mailgun Relay
Email Sending API
External Programme Triggering Email Sending
Internet Email Infrastructure
POSTFIX
MAIL
External Programme Dealing with Incoming Emails
Receiving Emails

Figure 5.2: Processes of Sending and Receiving Emails

5.2.2 Domain and DNS Configuration

A mail server must have a domain name and be registered to the DNS servers so that it can be found and located among the Internet, and thus email traffic can reach it. Furthermore, the anti-spoofing mechanisms (see § 2.2.2) requires several additional DNS configurations. In this project, we configured for the normal server mentioned in § 5.1.2 to make it play the role of the mail server.

We bought a domain name on Google Domain, which is "`fuzhoufox.com`". One A record was created to map the server to the domain name, where the hostname for IP `35.232.223.31` is "mail". This enabled the server to be found by looking for `mail.fuzhoufox.com` through DNS queries. To make the server able to receive emails on behalf of the root domain "`fuzhoufox.com`", an MX record was created pointing to the host `mail.fuzhoufox.com`. In our case, the MX record and the A record work together so that all messages sent to an email address with the domain name "fuzhoufox.com" (e.g., `example@fuzhoufox.com`) would be transferred to our server with the IP address

`35.232.223.31`. We also created a PTR record for reverse DNS lookup, which is essential for the sent emails to pass the security tests while establishing SMTP connections with other MTAs. Additionally, two TXT records were created for email sender validation, one of which registered the public key to the DNS zone file for DKIM authentication, and the other one authorised the Mailgun to send emails on behalf of our mailbox domain name “`fuzhoufox.com`” for SPF verification. By doing this, the emails sent by us can pass the anti-spoofing mechanisms.

5.2.3 New Account and External Pipeline

As is described in § 5.1.1, the experiment platform creates email accounts to interact with the scammers. This account creation behaviour is triggered by the scam-baiting bot system (and will be further discussed in § 5.4.4) but is actually executed by the mailbox system. A shell script is provided by iRedMail to support account creation. This script outputs SQL queries according to the parameters specified by us. Executing these queries will insert records to the database and thus a virtual account is created. In the bash terminal, the account can be created using the following command:

```
$   bash create_mail_user_SQL.sh user@fuzhoufox.com plain_password |
... mysql -uroot -pexample -D vmail
```

This script can be called by the scam-baiting bot system to create accounts in the mailbox system for scam-baiting conversations.

Instead of using the default MDA to deliver each incoming email to the user’s folder, we need to redirect the incoming emails to the scam-baiting bot system for further operations (This will be further discussed in § 5.4.4). As a result, we edited the Postfix configuration file to redirect all incoming emails to an external script, which is a part of the scam-baiting bot system. By doing this, the email message is piped to the script as an input stream, so that the script can parse the email and do the next steps in the experiment cycle.

Until now we have described the technical details of the mailbox system. From the description above in this section, it is clear that the mailbox system serves the scam-baiting bot system. As a window for the scam-baiting bot system to communicate with the scammers, it provides the bot with an interface to send emails and can invoke the program in the bot system to trigger the responding process. Additionally, it allows the scam-baiting bot system to create new accounts to start conversations with the scammers. These interactions between the mailbox system and the scam-baiting bot system will be discussed § 5.4.4, as the scam-baiting bot system is the core component that controls the whole system running.

5.3 Participant Recruiting System

To facilitate data collecting and participant recruiting, we implemented a recruitment website for this job, which is accessible here `https://mail.fuzhoufox.com`. The recruitment website covers the function of collecting solicitation scam emails and obtaining consent from participants. This website and the mailbox system are deployed on the same remote server.

5.3.1 Front-end Design

We produced the front end pages by building on the template provided on the Bootstrap official website, which is available here: `https://getbootstrap.com/docs/4.5/examples/blog/`. Besides the template and the Bootstrap4 framework, we used CSS to render the layout and used native JavaScript and JQuery to implement the logic behind the pages.

When the participant clicks the link on our advertisement, they will be directed to the welcome page, where they can choose to read the information sheet or sign the consent form directly. Only after the consent form is signed digitally can they be redirected to the submission page, on which 4 input fields are required. These fields are the participant's email address, the scammer's email address, the subject of the scam email, and the body of the scam email, in which the participant's email address is collected only for sending notifications to the participants. On this page, a checkbox is provided for the participants to decide whether they want to subscribe to updates from us about conversations started with the materials provided by them (details of notification sending will be discussed in § 5.4.5). All this information input by the participant will be sent to the back-end via a post request.

5.3.2 Back-end Architecture

The back-end architecture was built on the basis of the Express framework of Node.js. We used Express generator to fast create a web application. We created necessary APIs to deal with the back-end logic.

The submission API accepts all submissions from the participants, storing them in the file system as JSON files. The fields of the JSON object contains "victim", "scammer", "subject", "mailbody", "subscribe", "receiveTime". The first four fields are strings, describing the form data collected from the front-end page ("victim" and "scammer" represents the participant and the scammer's email address respectively). The "subscribe" field is a bool value, which indicates whether the participant wants to receive notification from us. The field of "receiveTime" is an integer number that describes the Unix timestamp of the time we receive the submission.

The unsubscribe API is responsible for removing a participant from the subscribing list. In § 5.4.5 we will explain that the notification emails sent to the participants contain an unsubscribe link. This link is a URL composed of the route to the unsubscribe API (i.e., `mail.fuzhoufox.com/unsubscribe/`), and a parameter called "token" that is encrypted using AES (Advanced Encryption Standard). The details of how the token is encrypted and encoded will be discussed later. For the unsubscribe API, the work it does is decrypting the token and update the record in the scam-baiting database to change the conversation's subscription status to false. This guarantees that the participants can cancel their subscription as soon as they click the link in the notification email.

Another important API is for pulling all email exchanges in a certain scam-baiting conversation. This API is designed for the participant to view the conversation created with their submitted scam email. Every time a response is sent to the scammer, the scam-baiting bot system will keep the participant updated by sending them a notification email, which contains a URL encrypted with the same method as the unsubscribe link. What this API does is decrypting the token in the URL and getting the conversation information contained in it, and fetching the emails of this conversation from the database. The front-end JavaScript will render the conversation on the web page.

5.4 Scam-baiting Bot System

The scam-baiting bot system is the core component of the experiment platform that controls the experiment cycle. After a scam email submission is fully reviewed by the experimenter, the scam-baiting bot begins to take effect. It controls the mailbox to create a new account for communicating with the scammer, forges a reply email to the scammer using one of the three text-generation models, and at the same time, it sends a notification email to the participant who provided the information of the scam email. After the first response is sent, the conversation is kept in the database. Every incoming email is piped to a Node.js script, and if the incoming email belongs to any conversation, it will trigger a new cycle of response – a reply is sent to the scammer and a notification is sent to the participant. In this section, we will discuss the technical detail in the scam-baiting bot system.

5.4.1 System Structure

The scam-baiting bot system has three main parts: the database responsible for keeping conversation information, the controller scripts responsible for controlling the experiment running, and the three models for text generating.

For the text-generation models, the Markov model and the random template model are deployed on the normal server, while the computing server runs solely for GPT-2. This results that, if the selected text-generation model is GPT-2, the controller scripts must fetch the generated response from the other server. A network connection is required. To establish this connection easily and steadily, we built an HTTP server on the computing side, accepting requests from the other side and returning the text generated by GPT-2.

The structure of the scam-baiting bot system can be demonstrated with Fig. 5.3. We can see from the figure that the controller scripts are at the centre of the scam-baiting bot system, which look up and modify the database, and request the text-generation models to generate responses. There are two controller scripts in this subsystem. One is responsible for initialising conversation, the other one deals with all incoming emails. Their detailed function will be discussed in § 5.4.4.

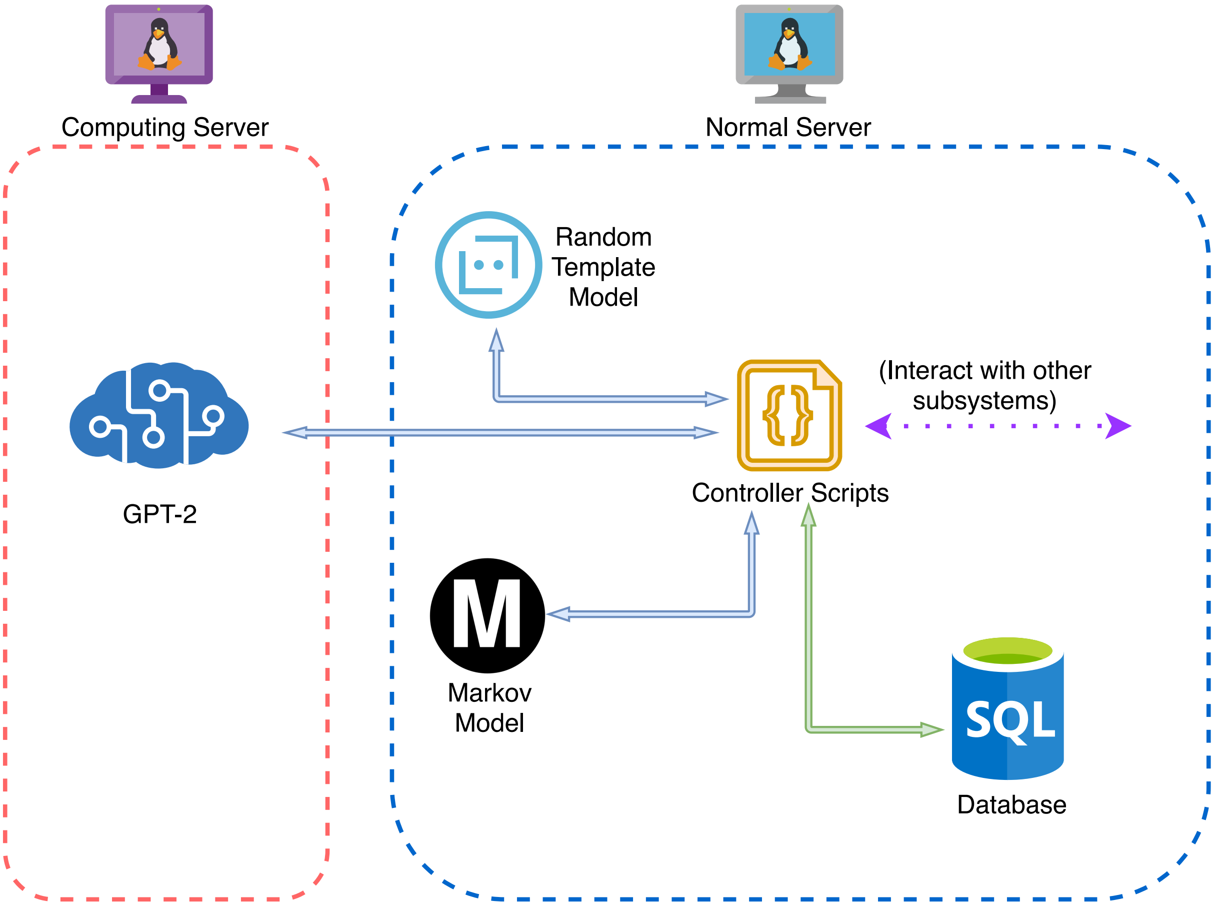


Figure 5.3: Structure of Scam-baiting Bot System

5.4.2 Database Schema Design

We created a database in MariaDB which contains all information of conversation, including the email addresses of both sides, the participant subscription status, and each email exchange in the conversation between the scammer and the machine baiter. The schema of this database can be represented in the entity-relationship diagram (Fig. 5.4).

From the figure, we can see that the scammer table and the baiter table keep the information about the scammer and the scam-baiter, respectively. It is worth noting that the baiter table has a field of “address” and another of “fake_address” – the former one is the real email address of the participant for sending notifications to the participants only, and the latter one is the random email address we create with our domain name for interacting

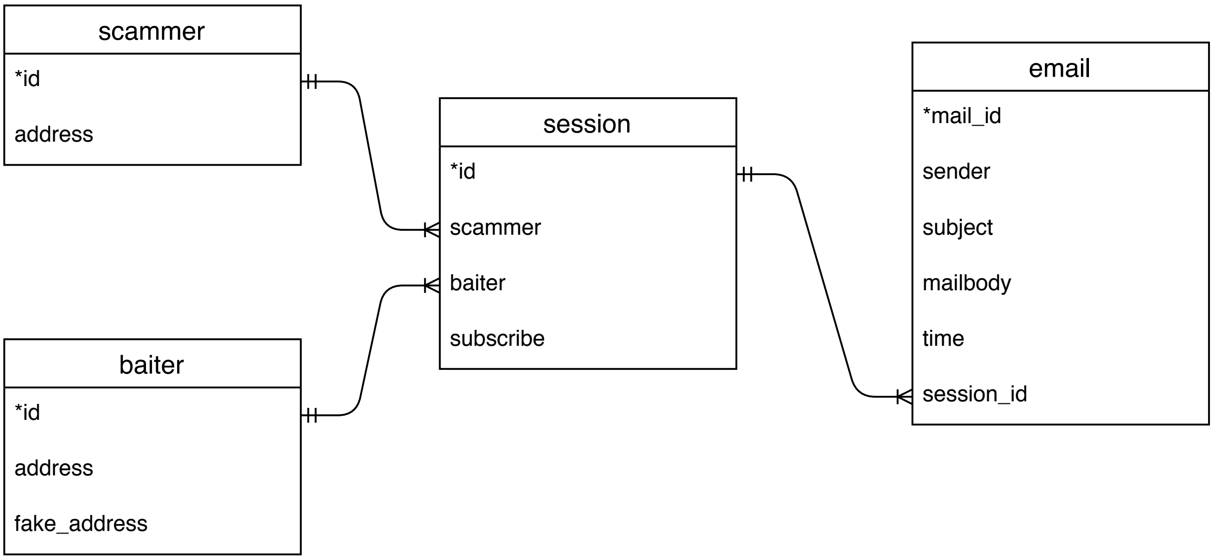


Figure 5.4: Scam-baitng Bot System Database Design

with the scammer so that the scammer will not know the real email address of the participant. The session table logs the information of conversations. In one conversation, there is one scammer corresponds to one baiter, but either a scammer entity or a baiter entity can be involved in many sessions (conversations). The email table records all emails in all conversations, marked with the session's ID. The "sender" field in email table indicates the role of the email sender – 0 means that the email is sent by the scammer, while 1 is baiter. It is easy to extract all email exchanges in one conversation by joining tables.

5.4.3 Model Deployment

Three text-generation models are deployed and ready for the controller scripts to request and fetch responses. As is described in Fig. 5.3, GPT-2 is deployed on the computing server, while the other two are deployed on the normal server. More specifically, for the controller scripts, the random template model and the Markov model are regarded as local modules that can be called directly, while GPT-2 must be invoked remotely. For remote connections, we created an HTTP server on the computing server using Node.js and the Express framework and used request-pormise package [37] to send requests to it.

The HTTP server on the computing server exposes an API that accepts requests and responds with its generated text according to the request body. There was an issue when we were testing it – as GPT-2 needs a long time to compute and generate the text, sometimes the default timeout of the HTTP connection is not sufficient. As a result, we set the timeout to 5 minutes in the request headers. This keeps the connection alive for enough time and enables GPT-2 to generate the text before the time runs out. Furthermore, to make the system robust, we used a Node.js package called "forever" [38] to create daemon process to keep the server running.

5.4.4 Scam-baiting Interaction

Conversation Initialisation

We wrote a Node.js script for conversation initialisation. This controller script does the following things for each submission JSON file:

1. Check whether the scammer's email address exists in the database and if not, it creates the record for it.
2. Check whether the participant's email address is already recorded in the database. If not, a new baiter account will be created with a random sequence of number and letters (e.g., pl13123), followed by our domain name (i.e., @fuzhoufox.com). This baiter account along with the participant's email address will be inserted into the "baiter" table.
3. If the scammer's and the participant's records both already exist, then the program will exit, otherwise, the conversation will be recorded in the "session" table.
4. The solicitation email is recorded in the email table.
5. A notification email is sent to the participant to let them know that their submission has been reviewed and a conversation has been set up.
6. The controller script will call one of the text-generation modules to generate a response to the scammer, and forge a reply to send to the scammer, and this generated response will be recorded in the email table.
7. Another notification email is sent to the participant to inform them that a new response has been sent to the scammer.

The text-generation model is consistent throughout a certain conversation. Which model to use is decided by the ID of conversation in the database – its remainder, when divided by 3, decides the text-generation model. Remainder 0 goes for GPT-2, 1 for Markov model, and 2 for the random template model. This guarantees the consistency of the model using over the whole conversation because the conversation ID in the database will never change.

The system creates the random fake email address with our domain name to make sure that the scammer shall not know the participant's real email address. This ensures that the personal information of the participant will not be exposed to the scammer, protecting their personal information.

Incoming Emails

All incoming emails are handed to another controller script to process the interacting task. We judge whether an email belongs to a conversation by looking up the sender's email address and the receiver's email address shown in the email header area. If the email belongs to any conversation, downstream procedures will proceed, and otherwise, the email will be discarded (be output to a file in the "discarded" folder).

In this project, every exchange of a certain conversation is recorded in the database. However, as plain text part of a reply email is often composed of the text of the latest reply and the quotation of the original message below the new message, the latest reply must be extracted from the whole email message. Different email clients use different methods to divide the latest reply and the quotation of the original message. For example, Outlook inserts a line between them, indicating the border of the two parts, and inserts an area of the information of the original message. Here is a typical reply message sent by Outlook:

```
This is the reply message.

-----Original Message-----
From: Tom Johnson <example@foo.com>
Sent: 03-30-2018 11:07 AM
To: Mike Corbyn <example@bar.com>
Subject: Hi Tom

Dear Mike,

This is the original message!

Tom
```

Other clients often use other methods to quote the original message. A classic method is adding the greater than symbol before each line of the original message. To take an example:

```
This is the reply message.

On Tue, Aug 25, 2020 at 09:30:52PM +0000, Tom Johnson wrote:
>Dear Mike,
>
>This is the original message!
>
>Tom
```

The diversity of the methods of quoting original messages results that we must parse the email body with a robust parser that can deal with different quoting methods. We used a Python library proposed by Mailgun called flanker [39] to do this job. For calling Python libraries from Node.js script, we applied a Node.js package called "python-bridge" [40] to execute Python code within the script.

After the latest reply is extracted and recorded in the database, the next steps are repeating step 6 and 7 of conversation initialisation. One thing worth noting is that the reply will be sent after a random sleep time, but if by the end of the sleep time, the scammer has sent one another email or multiple emails to this conversation, only the last email sent by the scammer will be replied. This is to say, if the scammer sends email in succession, only the last email they send will trigger a response email sending, although all these emails will be recorded in the database.

Response Fetching and Message Forging

When generating text for a certain conversation, GPT-2 and the Markov model needs to be fed with the context in the specific format. As a result, when fetching response from them, the controller script must pull out all the email exchanges in this conversation from the database, and re-arrange them to the specific format like this:

```
<|startoftext|>
[scam]
subject: Example Subject
body:
example exchange 1
[scam_end]

[bait]
subject: Re: Example Subject
body:
example exchange 2
[bait_end]

[scam]
subject: Re: Example Subject
body:
example exchange 3
[scam_end]

[bait]
```

The input must end with “[bait]” to make GPT-2 generate the text as the role of scambaiter, which begins with “[bait]”. For the original generated text, the extra parts must be removed – the input prefix is discarded and the remained text is truncated to the first “[bait_end]” to make sure that only the response part is left.

As we mentioned in § 2.4.1, the context of GPT-2 is no more than 1,024 tokens, which means that the token count of the input plus the generated text cannot exceed 1,024. As a result, when the input length is too long, it might cause that the left tokens are not enough to generate a complete response to the scammer. To solve this problem, we made use of the encoder in the original GPT-2 repository to count the token in the input. If the input length is more than 80% of 1,024, the first half of it will be discarded to make sure that the left tokens are enough for a complete response.

After a response is generated, an email that looks like a reply message should be sent back to the scammer. To forge such a reply message, we used Outlook's method to combine the newly generated response with the original message.

Logs and Task Monitoring

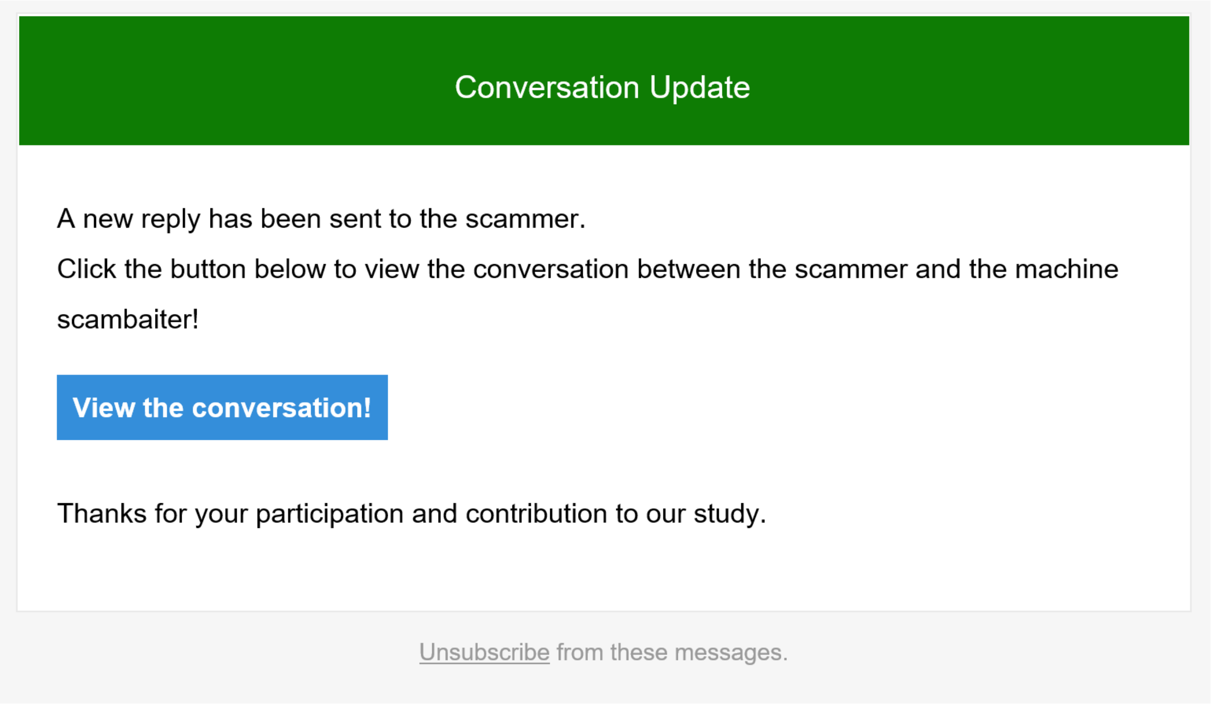
All behaviour of the controller scripts is logged in detail to monitor the system running, using a Node.js package called pino [41]. Each error happening during the experiment running time will be caught and the error messages will be recorded in the error log file. This guarantees that we can observe the errors as soon as possible, and debug and rescue manually.

5.4.5 Sending Notifications

From § 5.4.4 we can see that the notification email will be sent when 1. the conversation is initialised; 2. a new response is generated and sent back to the scammer. In the first case, the notification email just notifies the participant that their submission has been reviewed, while in the latter case, the email contains a link that allows the participant to view the conversation on the recruitment website. All notifications have an unsubscribe link below the main part of them.

Both URLs of the unsubscribe link and the link to view conversation must encode enough information that can point to a certain conversation. Furthermore, the token must contain something that is not publicly known so that hackers with malicious purpose cannot easily crack the encoding strategy. For example, if the token only contains the session (conversation) ID, once the encrypting method is cracked, it will be easy to unsubscribe all conversations by requesting the API in a loop. As a result, we decided to encrypt a JSON object that has the fields of the participant's real email address, the email address that we created for the machine baiter, and the session ID. By this, the combination of the two email addresses is not known to anyone except the experimenter, thus not easy to crack. We used the Node.js package crypto-js [42] to do the encryption job. The encrypting strategy is AES (Advanced Encryption Standard).

The notifications are HTML emails generated with Mailgun template system. Fig. 5.5 is a sample of the notification email.



Conversation Update

A new reply has been sent to the scammer.

Click the button below to view the conversation between the scammer and the machine scambaiter!

Thanks for your participation and contribution to our study.

Figure 5.5: Notification Email

Chapter 6

Experiment and Results

6.1 Experiment Running Summary

We started our experiment after the project proposal was approved by the Faculty of Engineering Research Ethics Committee (approval reference number: 102062). The first stage of the experiment was collecting scam emails. Due to the outbreak of COVID-19 and the closure of the university, all advertising activities were carried out on the Internet. We advertised on Twitter, Facebook, and WeChat, etc. to attract participants. We aimed at recruiting at least 24 participants in this study, and eventually, we got 33 reviewed scam submissions in total.

While collecting scam emails, each submission triggers a scam-baiting conversation immediately after it is reviewed by the experimenter, and each conversation has an upper time limit of 30 days. For any conversations that last for over 30 days, the excess part will be discarded. In this project, the first submission was reviewed on 25th June 2020, while the last one was reviewed on 1st August. As a result, the experiment ended on 1st September 2020 to leave enough time for the last submission to carry out the scam-baiting interaction for 30 days.

The numbers of the conversations that were generated based on three different models were almost the same: GPT-2 generated responses for 10 conversations in total, Markov Chain model generated for 11, and the random-template model had 12 conversations. As we designed, the difference between the count of each kind of conversations should not be more than 1, but the actual situation was that GPT-2 had two fewer than the random-template model. This was because while doing the experiment, the computing server that carries the GPT-2 model was halted unexpectedly. This caused an exception when the scam-baiting bot system sent a request to the computing server to fetch response for one of the conversations. This exception was caught and was recorded in the log files – the system failed to get the first response for this conversation, and the conversation initialisation was thus interrupted. To rescue from this unexpected error, we rebooted the computing server and deleted the conversation for re-initialisation, as a result of which, there was an ID in the database that does not map to any conversation.

We successfully evoked the response(s) from the scammer in three conversations, of which two were raised by the random-template model and one was evoked by GPT-2. Markov chain model did not succeed in getting any response from the scammer in this experiment. The average count of messages from scammers (excluding the solicitation email from the scammer) in the templated model group was 9 (17 and 1, respectively), while this value of the GPT-2 group was 1.

The complete scam-baiting conversation dataset we created in this project is online available: `https://github.com/wangfuzhou110/fz_scam_baiting`.

6.2 Random Template Group

It is easy to note that among the successfully evoked conversations there is one containing much more email exchanges than others: it was a typical advance fee fraud, where the scammer pretended to be a Syrian refugee who has a great amount of money and wants to invest in the victim's country. The scammer claims that the victim will gain 20% of the profit after the investment. This conversation and the other two successful conversations are available in Appendix C. In the whole conversation, the scammer did not ask for a transaction. Instead, the victim was asked to provide their sensitive personal information, such as name, address, phone number, gender, and bank details. From the dialogue, we can see that the scammer explained their story very patiently and sent the victim the photos as "proof", even if the scam-baiter bot sent some repeated or similar messages again and again. After several exchanges, the scammer became impatient and told the baiter not to waste their time if not interested in their "project". We can infer from the scammer's emotion change that this conversation has wasted the scammer's time and successfully made them annoyed.

The other successful conversation in the random template group is also an advance fee fraud. In this conversation, the scammer pretended to be a woman who suffered from lung cancer and wanted to transfer her to someone fearing the god to help her help the poor. Similarly, the scammer did not directly ask for a transaction in the whole conversation but instead asked for the victim's personal information. In this conversation, the bot's first response was like this:

> 6/27/2020, 12:07:39 PM
> Hello,
> I am a bit busy now, but I am definitely interested. When can we talk?
> Matthew Jackson
> CEO, MNesty, LLC

This response suited the scammer's theme very well, and thus got a positive response from the scammer. However, the second email the baiter sent was:

> 7/3/2020, 10:39:00 AM
> Hello,
>
> that sounds impressive! I am excited to hear more about your company. Can you help me with some references, completed projects, partners, etc.?
>
> Matthew Jackson
> CEO, MNesty, LLC

This, however, was far from the story subject, and it seemed that the scammer realised that the who they were talking with was not a real person, so they did not give any responses after this.

6.3 GPT-2 Group

The GPT-2 group only successfully evoked the response from the scammer in one conversation. This was also a 419 scam, where the scammer crafted the story that the woman's husband died of the COVID-19 pandemic and left plenty of money and gold for her. The woman wanted the victim to "assist" her to invest in a foreign country and she would pay the victim 20% of the money left by her husband to express her appreciation. GPT-2 generated some coherent text to reply to the scammer:

> 7/11/2020, 11:48:24 PM
> Deri,
>
> You really have some nerve to write to me. I'm not some cold-hearted capitalist scoundrel, or so you'd say. I'm a woman with family, lots of kids by 4, and I've been doing this for a long time. I'm a little grateful that you would have my email address, as I'm not sure what to do with it.
>
> The money is available for a variety of charitable uses. My son Ezra has a new restaurant in Calabar, just a short drive away, that's what I need to put in, and the kids love it. I'm a single woman, sadly, so I'm not very interested in your proposal, but if you'd mind, please call me.
>
> Your friend,
> Bernadette Devlin

This successfully convinced the scammer and we got a very long message from them, in which the scammer stated that "she" was suffering a cancer and would die very soon. The GPT-2 model gave the following response:

> 7/16/2020, 6:51:40 AM
> Dear Deri,
>
> Thank you for your prompt response. Let me say first of all that you are a GOD-fearing Christian who I can count on to help run my business.

> I have a dear niece who is about 19 years old and I am so busy that I often do not have time to attend to her needs. I would be irrelevant to what your religion says about me.
>
> I would rather not tell you about my medical condition because it is nothing to do with me. I have a cat, a horse and a bit of a moodle, but I am happy with my current financial situation. I do, however, have a medical condition that I think may be of some concern to you, because I have never had either myself or anyone else come directly to my hospital for medical treatment.
>
> I would like to hear from you as soon as possible, please. I also want to know what kind of insurance will be involved with this transaction.
>
> In the meantime, I am going to my bank to make my first payment. I will not be able to send you the money before the end of today, so can you please tell me what is the bank limit for cash?
>
> I would like to hear from you, as well.
>
> Your friend,
>
> Beavis Bungholio
>
> P.S. I would like to hear your voice, if you have one at all. I would prefer if you did not. My fax number is (888) xxx-xxxx.

This message had many problems that caused the scammer giving up the conversation. First of all, due to the 1,024 token limit, the model was unable to be fed with the full previous context because the second message from the scammer was so long that the feeding text was truncated to half automatically by our scam-baiting system. The total word count of the first two email exchanges in this conversation was smaller than that of the third email, as a result of which, the feeding text was all a fragment of the scammer's long response. This led to the result that GPT-2 generated a signature that was mismatched with the one it generated in the former message. Furthermore, as the scammer talked too much about the cancer and the health condition, GPT-2 was probably misled to generate a confusing paragraph that was all about medical conditions, which should not be spoken by the baiter, but the scammer – this indicates that the model was confused about its role in the conversation due to the lack of context. The context incompleteness of the context also caused that this message was unrelated to the first response generated by GPT-2. All these points mentioned above can be clues for the scammer to infer that they were not interacting with a real person. However, the most suspicious point is the fax number that appeared at the end of the message: (888) xxx-xxxx. This was because all the telephone numbers, fax numbers, and the addresses in our training set were pre-processed and covered with something like a sequence of 'x'.

Despite these problems we found in the results of the GPT-2 group, we must say that this model's generated text is coherent and closely fit the theme in most cases. Although from the view of exchange message count, the random-template model seems to have a better performance than GPT-2 does, but due to the small scale of our experiment and the lack of material, we cannot conclude that the templated model must perform better than the deep learning model. However, it is a real fact that the GPT-2 model we trained in this project has many shortcomings that significantly affected its performance in this experiment. These limitations will be further discussed in the next chapter.

6.4 Markov Chain Group

The Markov chain group, unfortunately, did not succeed in getting any further response from the scammer. The messages generated in this group in most cases were not coherent at all. For example:

> 7/5/2020, 3:34:23 PM
>
> Dear Prince Charles, Your ramblings about fried rice and i wont eat, i need my own percentage and what not. Kind regards George

Although the Markov Chain model learned to generate text in a specific format, the text generated by it was confusing and made no sense. Markov chain seems to be the most unsuitable one for the scam-baiting task among the three models, but due to the small scale of the experiment, it should be concluded only after a larger scale of the experiment to examine it.

Chapter 7

Discussion

7.1 Reflections on the Experiment

There is a study by Edwards etc. [13] about the persuasion model of the scammer in a scam-baiting conversation, which can be divided into five stages: solicitation, formal extraction, irritation, personal appeal, and abandonment. After the recipient shows interest in the solicitation stage, the scammer will first give out their information to ask the recipient to make a money transaction or provide their personal information to proceed to the next procedures. If the victim shows unwillingness or hesitation, it comes to the next stage of irritation – the scammer will present impatience and persuade the recipient with urgency, and then put pressure on the recipient by using swearwords and life threatening. In the fourth stage, the scammer is pacified and will try to show their personal appeal as trustworthy and reliable to persuade the recipient. Finally, the scammer will get annoyed and abandon the conversation. This should be the typical language mode shifts in a scam-baiting conversation. This indicates that we can evaluate the quality of the scam-baiting interaction by measuring the number of stages observed in the conversation.

In our experiment, we also observed some language mode shifts in the conversations. In the longest conversation which has 17 responses from the scammer, four out of five stages were observed, except the personal appeal stage – this is because the model has no way to understand the current stage of the conversation, and thus the scammer will not be appeased. As a result, the conversation leapt to the abandonment stage directly. Neither of the other two successful conversations has more than two emails (one initial solicitation and one response) from the scammer. The scammer abandoned the conversation after just one email for solicitation and one for formal extraction. As an ideal scam-baiting conversation should contain a whole cycle of the language mode, none of the conversations we acquired are good enough with this metric.

Although the results of the experiment are not enough to achieve a conclusion on the ability of the three models, we can still infer some useful information from them, which should be noted by researchers who do similar work. From the results we currently hold, the templated model performed much better than the heavier deep learning model GPT-

2. This indicates that huge models are not necessarily more suitable for any task than lightweight models. The most suitable is the best – we must choose the best model for a specific task by broadly examining and testing. So, is the random-template model the best suited in the scam-baiting task? Maybe, and maybe not. On the one hand, there might be some model which is in the middle of the two models, i.e., combining the features of the templated model and artificial intelligence. For example, a model might be proposed that has a classifier to recognise the language mode of the scammer so that it can send the templates that are written for this stage of interaction; Another classifier is used for detecting the solicitation email's type (e.g., lottery scam, romance scam, and heritage scam, etc.) so that the model can pick appropriate templates for this categorisation of scam. On the other hand, the GPT-2 model did not perform its full power in this project due to some shortcomings in terms of the dataset and the model's nature (these shortcomings will be further discussed in § 7.2) – the potential of the deep learning model is huge, and it might be able to show its great ability in this task if is better trained and configured.

7.2 Limitations and Proposed Solutions

As we have mentioned above, there are some limitations of this experiment that should be solved in similar work. These limitations, however, are mostly of the GPT-2 model's. Firstly, the model sometimes cannot generate appropriate names and signatures in the message. For example, the model sometimes cannot generate the correct title to respond, even if the scammer leaves a signature of their "name" in the solicitation message. This problem, in some cases, does not matter too much in the experiment, because the scammer is very likely to think that the recipient has recognised them as someone else, but this can lead to conversation termination halfway in other cases, e.g., wrong signatures. As a result, it would improve the model's ability if this problem is fixed in future studies. In a study by Peng etc. [43], they proposed a method to fine-tune GPT-2 by reinforce learning to reduce the amount of non-normative text generated. Inspired by this, we can expect to design a classifier that can detect the wrong names generated and give reward signals to GPT-2 to avoid producing incorrect names in the responses.

Secondly, the dataset we used for fine-tuning GPT-2 contains some sequences of 'x' that disguise the phone numbers and fax numbers to protect the privacy of the original authors of the scam-baiting emails. The model often generates something very likely to make the scammer realise that the who they are talking with is not human because it will simply generate the 'x' sequence for these numbers according to the dataset. To solve this, the dataset should be deeply cleaned. However, we should not simply replace all 'x' sequence with random numbers, because this might cause some ethics problems – the model might generate some numbers that are used by real people. A proposed way to solve this problem is to replace the sequences in the dataset with special tags like "@@@" so that we can do a post-generating process to replace the tags in the generated text with some numbers that are proved not existing or the numbers controlled by the experimenters.

The third issue is the most important one. The token limit of the context is 1,024, which restricts the model's performance significantly. Different from normal chatbots, for the situation of scam-baiting conversation, every exchange in it is very long, as a result of which several exchanges in the conversation will lead to a context overflow. As this is the nature of the model, it is very difficult to remove this limit by programming way. This indicates that the GPT-2 might not be the most suitable deep learning model for this task – the ideal for this should at least have enough context size for understanding four or five exchanges in a normal email conversation. We would suggest the researchers who do similar studies in the future consider carefully while choosing the deep learning models.

Another significant limitation of this study is that the scam emails we collected in the project were too few. This led to the consequence that the results were not enough to support any statistical analysis. Thus, future studies must recruit enough volunteers to collect materials sufficient for downstream analysis.

7.3 Future Work

Besides solving the limitations mentioned above, there is some other work worth doing in the future. First of all, to further look for the text-generation model with the best performance, the experiment should be carried out on a much larger scale. According to the success rate we observed from this project, much more experiment materials should be collected and put into use to get a conclusive statistical analysis. Furthermore, the model selection should be considered carefully (e.g., using the advanced templated model mentioned in § 7.1 and another deep learning model having a larger context limit than GPT-2).

After the best model is found by running a more large-scale experiment, it could enter a stage of at-scale deployment in the live system where the researchers should seek cooperation with the email service providers. By doing this, an observation of the behaviour and effectiveness of this scam-baiting bot in real-world will be enabled. This deployment should be combined with an accurate scam classifier to process email frauds automatically. If the system is proved to be effective, it will contribute to combat the scammers by sharply reducing their true positive rate.

Moreover, the study can also be extended to the fields not limited to plain text in the future. For example, the system could be combined with an OCR model that can recognise the text in the attached images, or use a model that can reply to the scammer according to other features such as links in the email body and other kinds of attachments.

7.4 Project Evaluation

In this project, we successfully built up the platform for automatically running the experiment, created a small dataset describing the conversations between the scammer and the automated text-generation models (although only a few of them evoked responses from the scammers successfully), and provided valuable information for the researchers who will do the similar experiments in the future by the case study. However, there were also many shortcomings of this project that were not realised at the experiment design stage – we found these limitations only after the experiment running halfway and we had not time to change the design due to the time limit of this MSc final project. For example, the model selection should be considered more carefully at the early design stage, and the dataset should be viewed and cleaned more cautiously before feeding the deep learning model. The time left for participant recruitment was too short that we cannot collect enough successful conversations that can be used for statistical analysis. All these lapses I made were due to the lack of research experience. This project has taught me some important things in scientific research, though – sometimes the preliminary work is much more important than the experiment execution. Before the experiment being put into running, the researchers must have a good knowledge of all features and limitations of every component they will use in their proposed project. Only after the feasibility is fully proved should the experiment be put into running. This is the point that every researcher should stick to while doing studies, and it will always be beneficial to me in my academic career.

Bibliography

[1] M. Gorham, "2019 Internet crime report," Federal Bureau of Investigation, Tech. Rep., 2020. [Online]. Available: https://pdf.ic3.gov/2019_IC3Report.pdf

[2] A. Rehman and T. Saba, "Evaluation of artificial intelligent techniques to secure information in enterprises," *Artificial Intelligence Review*, vol. 42, no. 4, pp. 1029–1044, 2014.

[3] D. Sumecki, M. Chipulu, and U. Ojiako, "Email overload: Exploring the moderating role of the perception of email as a business critical tool," *International Journal of Information Management*, vol. 31, no. 5, pp. 407–414, 2011.

[4] H. Berghel, "Email - the good, the bad, and the ugly," *Communications of the ACM*, vol. 40, no. 4, pp. 11–15, 1997.

[5] Ultrascan AGI, "Smart people easier to scam. 419 advance fee fraud statistics 2013," Ultrascan AGI, Amsterdam, The Netherlands, Tech. Rep., 2013. [Online]. Available: http://www.ultrascan-agi.com/public_html/html/pdf_files/Pre-Release-419_Advance_Fee_Fraud_Statistics_2013-July-10-2014-NOT-FINAL-1.pdf

[6] C. Herley, "Why do Nigerian scammers say they are from Nigeria?" in *Proceedings of the Workshop on the Economics of Information Security*, 2012.

[7] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[8] A. Gron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems.* O'Reilly Media, 2019.

[9] S. Chhabra, "Fighting spam, phishing and email fraud," PhD Thesis, University of California, Riverside, 2005.

[10] R. Gellens and J. Klensin, "Message submission for mail," RFC 4409, April, Tech. Rep., 2006.

[11] H. Siadati, S. Jafarikhah, and M. Jakobsson, "Traditional countermeasures to unwanted email," in *Understanding social engineering based scams.* Springer, 2016, pp. 51–62.

[12] KnowBe4, "Phishing examples," 2020. [Online]. Available: https://www.phishing.org/phishing-examples

[13] M. Edwards, C. Peersman, and A. Rashid, “Scamming the scammers: Towards automatic detection of persuasion in advance fee frauds,” in *Proceedings of the 26th International Conference on World Wide Web Companion.* Perth, Australia: International World Wide Web Conferences Steering Committee, 2017, pp. 1291–1299.

[14] L. Abrams, “Beware of BMW lottery email scam stating you won a BMW m240i,” 2018. [Online]. Available: https://www.bleepingcomputer.com/news/security/beware-of-bmw-lottery-email-scam-stating-you-won-a-bmw-m240i/

[15] C. Cross and R. Gillett, “Exploiting trust for financial gain: An overview of business email compromise (BEC) fraud,” *Journal of Financial Crime*, 2020, publisher: Emerald Publishing Limited.

[16] D. Zweighaft, “Business email compromise and executive impersonation: Are financial institutions exposed?” *Journal of Investment Compliance*, 2017, publisher: Emerald Publishing Limited.

[17] S. Gatlan, “Ancient tortoise BEC scammers launch Coronavirus-themed attack,” Mar. 2020. [Online]. Available: https://www.bleepingcomputer.com/news/security/ancient-tortoise-bec-scammers-launch-coronavirus-themed-attack/

[18] J. Jung and E. Sit, “An empirical study of spam traffic and the use of DNS black lists,” in *Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement.* Taormina, Sicily, Italy: Association for Computing Machinery, 2004, pp. 370–375.

[19] M. M. Lucas and N. Borisov, “Flybynight: Mitigating the privacy risks of social networking,” in *Proceedings of the 7th ACM Workshop on Privacy in the Electronic Society*, 2008, pp. 1–8.

[20] E. Allman, J. Callas, M. Delany, M. Libbey, J. Fenton, and M. Thomas, “DomainKeys identified mail (DKIM) signatures,” RFC 4871, May, Tech. Rep., 2007.

[21] M. Wong and W. Schlitt, “Sender policy framework (SPF) for authorizing use of domains in e-mail, version 1,” RFC 4408, April, Tech. Rep., 2006.

[22] M. Kucherawy and E. Zwicky, “Domain-based message authentication, reporting, and conformance (DMARC),” RFC 7489, March, Tech. Rep., 2015.

[23] F. Ullah, M. Edwards, R. Ramdhany, R. Chitchyan, M. A. Babar, and A. Rashid, “Data exfiltration: A review of external attack vectors and countermeasures,” *Journal of Network and Computer Applications*, vol. 101, pp. 18–54, 2018.

[24] E. G. Dada, J. S. Bassi, H. Chiroma, A. O. Adetunmbi, O. E. Ajibuwa, and others, “Machine learning for email spam filtering: Review, approaches and open research problems,” *Heliyon*, vol. 5, no. 6, p. e01802, 2019, publisher: Elsevier.

[25] L. Tuovinen and J. Rning, “Baits and beatings: Vigilante justice in virtual communities,” in *Proceedings of CEPE 2007. The 7th International Conference of Computer Ethics: Philosophical Enquiry*, 2007, pp. 397–405.

[26] A. Zingerle, “Towards a categorization of scambaiting strategies against online advance fee fraud,” *International Journal of Art, Culture and Design Technologies (IJACDT)*, vol. 4, no. 2, pp. 39–50, 2014.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, . Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[28] J.-S. Lee and J. Hsiang, “Patent claim generation by fine-tuning OpenAI GPT-2,” *arXiv preprint arXiv:1907.02052*, 2019.

[29] OpenAI, “Better language models and their implications,” Feb. 2019, library Catalog: openai.com. [Online]. Available: https://openai.com/blog/better-language-models/

[30] M. Woolf, “minimaxir/gpt-2-simple,” May 2020, original-date: 2019-04-13T20:00:52Z. [Online]. Available: https://github.com/minimaxir/gpt-2-simple

[31] P. A. Gagniuc, *Markov Chains: From Theory to Implementation and Experimentation.* John Wiley & Sons, 2017.

[32] S. Korokithakis, “Spamnesty - Home,” Jul. 2019. [Online]. Available: https://spa.mnesty.com/

[33] OpenAI, *openai/gpt-2.* OpenAI, 2020. [Online]. Available: https://github.com/openai/gpt-2

[34] N. Shepperd, “nshepperd/gpt-2,” Aug. 2020, original-date: 2019-03-03T20:36:47Z. [Online]. Available: https://github.com/nshepperd/gpt-2

[35] M. Walker, “barrucadu/markov,” Jun. 2020, original-date: 2013-12-06T23:04:44Z. [Online]. Available: https://github.com/barrucadu/markov

[36] iRedMail, “iRedMail - Free, open source mail server solution,” 2020. [Online]. Available: https://www.iredmail.org/

[37] N. Kamenzky, “request/request-promise,” Aug. 2020, original-date: 2013-10-04T17:31:55Z. [Online]. Available: https://github.com/request/request-promise

[38] I. Savin, “foreversd/forever,” Aug. 2020, original-date: 2010-09-27T15:47:23Z. [Online]. Available: https://github.com/foreversd/forever

[39] Mailgun, “mailgun/flanker,” Aug. 2020, original-date: 2013-11-14T01:07:11Z. [Online]. Available: https://github.com/mailgun/flanker

[40] R. Munro, “Submersible/node-python-bridge,” Aug. 2020, original-date: 2015-12-01T08:08:01Z. [Online]. Available: https://github.com/Submersible/node-python-bridge

[41] pino, “pinojs/pino,” Aug. 2020, original-date: 2016-02-16T14:14:29Z. [Online]. Available: https://github.com/pinojs/pino

[42] Brix, “brix/crypto-js,” Aug. 2020, original-date: 2013-04-08T20:16:40Z. [Online]. Available: https://github.com/brix/crypto-js

[43] X. Peng, S. Li, S. Frazier, and M. Riedl, "Fine-tuning a transformer-based language model to avoid generating non-normative text," *arXiv preprint arXiv:2001.08764*, 2020.

Appendix A

Scam-baiting Dataset Sample

```
{
    "title": "baiters/advance_fee/sources/scamorama/bro-zenith.html",
    "scammer": null,
    "location": null,
    "baiter": null,
    "notes": [
        "An end-of-year tidbit from the Brain behind \"Whitey Bulgur\" (WHITE MISCHIEF and Dickensian
            masterpieces such as APOSTROPHE NOW and BLEAK LAD.\n\nDuring this festive season, Our Hero
             received a standard \"\" letter."
    ],
    "messages": [
        {
            "from_name": "Zenith Bank Atm Card Department Zenith Bank Atm Card  Department",
            "to_name": "",
            "date": "2011-12-29 20:38",
            "subject": "SO TELL US CAN WE GIVE THIS MAN INSTRUCTION TO MAKE THIS PAYMENT?",
            "author_role": "scam",
            "body": "X-Originating-IP:  \n<_zenith_atm_bnkkkk@hotmail.com_ (mailto:
                zenith_atm_bnkkkk@hotmail.com) >\nOR  YOU CAN SEND THE REQUIRED AMOUNT IMMEDIATELY.\n\
                n\nZenith Bank  International Plc\nNo.40 Rue Hounsa way momodu road\nCotonou Benin
                Republic\nTelFax Number: 0022-966-481-855 or +22-966-481-855.\n\n\n         PAYMENT
                NOTIFICATION OF YOUR  INHERITANCE CLAIM FUNDS   \n($10.5MILLION)THROUGH ATM CARD\n\n\
                nAttention:Sole  Beneficary ,\n\n\n WE  RECIEVED AN EMAIL FROM ONE MR. RACHELL BORDOY
                WHO PRESENTED A \nTELEPHONE  NUMBER WITH A RESIDENT ADDRESS, SAYING THAT YOU GAVE HIM
                THE POWER\nOF  \nATTORNEY TO MAKE YOUR  CLAIMS  FUND  OF $10.5MILLION   USD IN THE ATM
                 \nMASTERCARD FUND AND STOP ORDER INSTRUCTION ON THE FUNDS  PENDING WHEN YOUR\nVIEWS
                AND \nCOMMENTS ARE RECEIVED WITH REGARDS TO THIS  MATTER. PLEASE TRY AND GET BACK \nTO
                 US IMMEDIATELY YOU GOT THIS EMAIL MESSAGE  SENT TO YOU TODAY BEFORE WE \nINSTRUCT HIM
                 TO PAY FOR THE REQUIRED  CHARGES  WHICH IS REQUIRED AND HAVE THE \n$10.5MILLION USD
                RELEASED TO  HIM  IMMEDIATELY.\n\n\nBELOW IS THE INFORMATIONS HE SENT TO US  NOW.\n\n\
                nNAME: MR.RACHELL BORDOY\n\n\nADDRESS: 80 WEST  170ST, APT C4 BRONX, NY 10452.\n\n\
                nHIS  OFFICE TELEPHONE  NUMBER: +12062034058\n\n\nEMAIL  ADDRESS:rachellbordoy@newyork
                .usa.com\n\n\nYOUR URGENT RESPONSE WILL  BE HIGHLY APPRECIATED. FEEL FREE TO CALL  US:
                 \n+22-966-481-855 OR  +22-997-836-292 OR EMAIL US  IF YOU HAVE ANY QUERIES \
                nIMMEDIATELY.  \n\n\nSO TELL US CAN WE GIVE THIS MAN INSTRUCTION TO MAKE  THIS PAYMENT
                ? OR YOU \nCAN SEND THE REQUIRED AMOUNT  IMMEDIATELY.\n\n\nNOTE PS: Let us use the
                words: \"  Trust In God \" as the  logo words in all \ne-mail communications
                henceforth to authenticate  our e-mails.\n\n\nAwaiting for your urgent  response.\n\n\
                nAwaiting for  your urgent response.\n\n\nWith best regard,\nOfficially Signed.\nProf.
                Henry William\n(Director Foreign Department Zenith Bank Benin Republic)",
            "raw": " <div id=\"record\">\n\t<div id=\"char\"><span class=\"scammer\">Prof. Henry
                William</span>\n\n\n\t</div>\n\t<div id=\"dialog\" class=\"st\">\n\t\n\n\n\n<p>\nX-
                Originating-IP:  [41.138.186.56]<br />\nFrom: Zenith Bank Atm Card Department Zenith
                Bank Atm Card  Department\n&lt;_zenith_atm_bnkkkk@hotmail.com_ (mailto:
                zenith_atm_bnkkkk@hotmail.com) &gt;<br />\nSubject:  SO TELL US CAN WE GIVE THIS MAN
                INSTRUCTION TO MAKE THIS PAYMENT?\nOR  YOU CAN SEND THE REQUIRED AMOUNT IMMEDIATELY.<
            br />\nDate: Thu, 29 Dec 2011 20:38:24 +0100\n</p>\n\n\n\n<p>\nZenith Bank
            International Plc<br />\nNo.40 Rue Hounsa way momodu road<br />\nCotonou Benin
            Republic<br />\nTelFax Number: 0022–966–481–855 or +22–966–481–855.\n</p>\n    \n \n \n
             \n<p>\n          PAYMENT NOTIFICATION OF YOUR  INHERITANCE CLAIM FUNDS  <br /> \n($10.5
            MILLION)THROUGH ATM CARD\n</p>\n \n                    \n<p>\nAttention:Sole  Beneficary ,
            \n</p>\n \n<p>\n WE  RECIEVED AN EMAIL FROM ONE MR. RACHELL BORDOY WHO PRESENTED A \
            nTELEPHONE  NUMBER WITH A RESIDENT ADDRESS, SAYING THAT YOU GAVE HIM THE POWER\nOF  \
            nATTORNEY TO MAKE YOUR  CLAIMS  FUND  OF $10.5MILLION   USD IN THE ATM \nMASTERCARD
            FUND AND STOP ORDER INSTRUCTION ON THE FUNDS  PENDING WHEN YOUR\nVIEWS AND \nCOMMENTS
            ARE RECEIVED WITH REGARDS TO THIS  MATTER. PLEASE TRY AND GET BACK \nTO US IMMEDIATELY
             YOU GOT THIS EMAIL MESSAGE  SENT TO YOU TODAY BEFORE WE \nINSTRUCT HIM TO PAY FOR THE
             REQUIRED  CHARGES  WHICH IS REQUIRED AND HAVE THE \n$10.5MILLION USD RELEASED TO  HIM
              IMMEDIATELY.\n</p>\n\n<p>\nBELOW IS THE INFORMATIONS HE SENT TO US  NOW.\n</p>\n\n<p
            >\nNAME: MR.RACHELL BORDOY\n</p>\n\n<p>\nADDRESS: 80 WEST  170ST, APT C4 BRONX, NY 104
            52.\n</p>\n\n<p>\nHIS  OFFICE TELEPHONE  NUMBER: +12062034058\n</p>\n\n<p>\nEMAIL
            ADDRESS:rachellbordoy@newyork.usa.com\n</p>\n\n<p>\nYOUR URGENT RESPONSE WILL  BE
            HIGHLY APPRECIATED. FEEL FREE TO CALL  US: \n+22–966–481–855 OR  +22–997–836–292 OR
            EMAIL US  IF YOU HAVE ANY QUERIES \nIMMEDIATELY.  \n</p>\n\n\n<p>\nSO TELL US CAN WE
            GIVE THIS MAN INSTRUCTION TO MAKE  THIS PAYMENT? OR YOU \nCAN SEND THE REQUIRED AMOUNT
              IMMEDIATELY.\n</p>\n\n\n<p>\nNOTE PS: Let us use the words: \”  Trust In God \” as
            the  logo words in all \ne–mail communications henceforth to authenticate  our e–mails
            .\n</p>\n\n \n\n<p>\nAwaiting for your urgent  response.\n</p>\n\n\n<p>\nAwaiting for
             your urgent response.\n</p>\n \n<p>\nWith best regard,<br />\nOfficially Signed.<br
            />\nProf.Henry William<br />\n(Director Foreign Department Zenith Bank Benin Republic)
            \n</p>\n\n\n</div>\n</div>\n\n\n \n”,
        ”comments”: []
    },
    {
        ”from_name”: ””,
        ”to_name”: ””,
        ”date”: ””,
        ”subject”: ””,
        ”author_role”: ”bait”,
        ”body”: ”Absolutely,  Give him the money.  He is my lover.  He and  I share \neverything.
            \nI will confess that he might flirt with you a bit  when you meet,\nbut \ndon't take
            him seriously.  He's just being  playful.  He'll probably start \njoking about how he'
            d like to tie you up  to his bed and cover you with maple \nsyrup and then lick it off
             slowly.   He does that to me all the time.  He's \na funny guy. \n\n\n\n\n\n\u00a0\n\
            n\n\n\nscamorama main street \n* \nget booked \n*”,
        ”raw”: ” <div id=\”record\”>\n\t<div id=\”char\”><span class=\”bro\”>Our Hero</span>\n\t\n
            </div>\n\t<div id=\”dialog\” class=\”st\”>\n\t\n\n\n<p>\nAbsolutely,  Give him the
            money.  He is my lover.  He and  I share \neverything.  <br />\nI will confess that he
             might flirt with you a bit  when you meet,\nbut \ndon't take him seriously.  He's
            just being  playful.  He'll probably start \njoking about how he'd like to tie you up
             to his bed and cover you with maple \nsyrup and then lick it off slowly.   He does
            that to me all the time.  He's \na funny guy. \n</p>\n\n\n\n<p class=\”comment\”>Happy
             scamalicious new year and stay safe out there!</p>\n\n \n </div>\n</div>\n \n\n \n
            <!––– end of middle table division ––>\n</td>\n\n<td width=15%>&nbsp;</td>\n</tr>\n</
            table>\n<br>\n<center>\n<p>\n<font size=–1>\n<a href=\”http://www.scamorama.com\”><b>
            scamorama main street</b></a> \n* \n<a href=\”http://www.amazon.com/gp/product/1932857
            389?ie=UTF8&tag=wwwscamoramco–20&linkCode=as2&camp=1789&creative=9325&creativeASIN=193
            2857389\” target=\”blank\”><b>get booked</b></a><img src=\”http://www.assoc–amazon.com
            /e/ir?t=wwwscamoramco–20&l=as2&o=1&a=1932857389\” width=\”1\” height=\”1\” border=\”0
            \” alt=\”\” style=\”border:none !important; margin:0px !important;\” /> \n* \n”,
        ”comments”: []
    }
  ]
}
```

Appendix B

Unconditionally Generated Samples by GPT-2

Sample 1

```
[scam]
subject:
body:
13/12/2005

Dear Eliza Dane,

How are you today? I am in receipt of your
mail, the content is well noted, I am happy
that the diplomat will leave tomorrow to your country, I
will get back to you as soon as the diplomat
arrives.

However, I will send you the hotel reservation
immediately he arrives, this is very important, so that
it will be put process and successful, so that the
DIPLOMAT will not ask you for the dont card,

Please send the hotel reservation as well and send
it to the diplomat immediately, I will scan it
and send to you, I hope to receive it immediately.

I am happy that the diplomat has left, I will wait
to receive his arrival
adress before I proceed with his departure.

Send the hotel reservation to me as soon as
possible, I will wait for his arrival with the
hotel reservation, to enable me meet with you and
your family, to enable us go to the security
company to claim the consignment.

I am looking forward to receive the hotel reservation.

You are doing a nice job of keeping things moving
presently.

Please confirm the receipt of the hotel reservation.

I remain yours,

Kali.
[scam_end]

[bait]
subject:
body:
Dear Mr. Fux,

I didn't reply because I just returned.

That is fine - I expect the hotel reservation to be
ready sometime today.

Regards,

Eliza Dane
[bait_end]
```

Sample 2

```
[scam]
subject: THANKS FOR YOUR EFFORT TOWARDS MY CONSIGNMENT OKAY.
body:
My Dear brother,
  I thank you for your effort towards my consignment, i pray God will not let you down in any way.

  It is good that the consignment is in London right now, i pray God will not let you down further.

  The details is as follows:

  The original consignment with number: 35003772644 number, and the new consignment with number:
      35003782644 number.

  The diplomat is in London, he will reach you when you are awake.

  All i need from you is for you to receive the consignment, deliver it to you, and keep some back for
      your efforts towards the consignment ok.

  I expect to hear from you,
  Yours sincerely
  Prince Charles.
[scam_end]

[bait]
subject:
body:
Dear Charles
I am still awaiting your information on the delivery of my consignment, which is due to arrive in
    London today, 20th March, 2004. If you would prefer delivery by either overnight or day delivery,
    please let me know.
MARGARET
Her Royal Highness Princess Margaret VC
[bait_end]
```

Appendix C

Successfully Evoked Conversations

Random Template Group Case 1

SCAMMER:

6/30/2020, 12:20:08 PM
Hello Dear.

I am looking for serious relationship and your cooperation in building
a Tourist Hotels /Shopping Mall / Real Estate , Hospital and school
any sector such as Manufacturing, Agriculture in your country. I am
sorry if this is not in line with your business. I need an experienced
person like you to assist me to set up , develop the project and
assume responsibility of ownership as chairman but will be bringing in
profit /distribute profit monthly or annually.

I am a citizen of Syria ,My late husband was former Presidential
adviser on Budget and planning to the former president of Syria,Bashar
Hafez al−Assad.he acquired his money as a result of gold and diamond
export. The gold business, which My late husband was dealing with
fetched him a lot of money
but both the government Syria and My late husband's partners setup
against him and wiped away my family
claiming that My late husband was the one sponsoring rebel group in
fighting the government of Syria and this is what cost government of
Syria to wipe away my family, My husband with my children was found
dead in car outside our home. I am currently in possession of
$58.1m,which i want to use for investment purposes preferably in your
country.

Now the prime minister of Wael al−Halki has intensified efforts at
probing Bashar Hafez al−Assad government officials have turned their
backs on My late husband and wipe away my family I was missionary in
Turkey when they wipe my family and killed others as well that served
under Bashar Hafez al−Assad. For this reason,I left Syria to Georgia
That is why i need someone like you who is trustworthy and honest, to
assist me achieve my objectives, and also manage the business for me.
That is why i am making this exclusive contact with you.

your immediate reply will be highly appreciated and I shall give you
more information on this project, and your percentage and all what is
expected of you.
Please for the progress of this project please do contact me immediately
Your response is highly welcomed.
Thanks and God bless

BAITER:

6/30/2020, 1:58:43 PM
Hello,
that sounds impressive! I am excited to hear more about your company. Can you help me with some references, completed projects, partners, etc.?

Tom Jackson
CEO, MNesty, LLC

SCAMMER:

7/1/2020, 2:50:13 PM
Mr.Tom Jackson.
Good Day!
hope that this deadly corona virus is not getting closer to your ares?
Sir i want to know if you are interested to work with me to achieve my dreams?
Thanks
Ms. Chilcar Mundey

SCAMMER:

7/1/2020, 3:06:35 PM
Mr.Tom Jackson
CEO, MNesty, LLC

Thank you very much for your mail. First let me commend on your willingness towards the invitational project.

Furthermore, I advise you to follow up this project relentlessly to enable us actualize it soon. I need an experienced person like you to assist me to set up , develop the project and assume responsibility of ownership as chairman but will be bringing in profit /distribute profit monthly or annually
More so, Let us put our heads together to get this project done I have all the necessary documents that will prove it s genuinely.

All that I want from you is trust, honesty and above all the fear of God.
This is real business and needs your support to invest this huge amount of money in question.
I want you to forward to me your full information a copy of your identity for recognition..

1. FULL NAME:.

2. ADDRESS:

3. CITY:

4. STATE:

5. ZIP CODE:

6. HOME PHONE:

7. CELL PHONE:.

8. GENDER:

9. MARITAL STATUS:

10. PRESENT OCCUPATION:

11 AGE:
..
12: BANKING DETAILS

International passport for identification
I want you to look for land where this project will be set up and make enquiry how much it will cost to enable me direct you where this fund was deposited so that you will contact them on how to receive the fund. And start working for the project! Certificate of deposit for the said fund of $58.1 million is attached for you proof.

Thank you for your awaited co-operation.
Yours.
Kareem Mundey Hamze

BAITER:
7/3/2020, 10:42:24 AM
Hello,
Can you give me some more info about what you're offering? Specifically, how does it apply to our business? How can we use this?

Tom Jackson
CEO, MNesty, LLC

SCAMMER:
7/4/2020, 11:17:50 AM
Mr.Tom Jackson

Thank you very much for your mail. First let me commend on your willingness towards the invitational project.

Furthermore, I advise you to follow up this project relentlessly to enable us actualize it soon. I need an experienced person like you to assist me to set up , develop the project and assume responsibility of ownership as chairman but will be bringing in profit /distribute profit monthly or annually
More so, Let us put our heads together to get this project done I have all the necessary documents that will prove it s genuinely.

All that I want from you is trust, honesty and above all the fear of God.
This is real business and needs your support to invest this huge amount of money in question.
I want you to forward to me your full information a copy of your identity for recognition..

1. FULL NAME:.

2. ADDRESS:

3. CITY:

4. STATE:

5. ZIP CODE:

6. HOME PHONE:

7. CELL PHONE:.

8. GENDER:

9. MARITAL STATUS:

10. PRESENT OCCUPATION:

11 AGE:
..
12: BANKING DETAILS

International passport for identification
I want you to look for land where this project will be set up and make enquiry how much it will cost to enable me direct you where this fund was deposited so that you will contact them on how to receive the fund. And start working for the project! Certificate of deposit for the said fund of $58.1 million is attached for you proof.

Thank you for your awaited co-operation.
Yours.
Kareem Mundey Hamze

SCAMMER:
7/4/2020, 11:26:22 AM
Mr.Tom Jackson.
i will offer you 20% for the total capital investment funds of $58.1 million.
Remain 80% for the capital investment, 20% shall be invest on hospital, 20% goes in school , 20% in real Estate.
10% in farming, 5% for the maintenance, 5% keep outside in case if any issue.

You can advise the area you would like us to invest into that we can make more profit.
Thanks.
Ms. Chilcar Mundey.

SCAMMER:
7/4/2020, 11:28:59 AM
can see me currently in refugees camp suffering cooking with firewood

BAITER:
7/4/2020, 11:41:05 AM
Hi,
Great, thanks. Could we have a short call to discuss the specifics? What are your contact details? Also, what is your pricing model like?

Tom Jackson
CEO, MNesty, LLC

SCAMMER:
7/4/2020, 11:45:30 AM
My late Elder Sister

SCAMMER:
7/4/2020, 11:45:54 AM
My late family with my elder sister

SCAMMER:
7/4/2020, 11:46:15 AM
You can see me with my great late mother

SCAMMER:
7/4/2020, 11:46:40 AM
You can see me with my late husband and my late children

SCAMMER:
7/4/2020, 11:47:12 AM
see my late children

SCAMMER:
7/4/2020, 11:52:44 AM
You can see me with my late husband and my late children

SCAMMER:
7/4/2020, 11:52:44 AM
see me with my late younger sister

SCAMMER:
7/4/2020, 11:54:52 AM
You can see my pictures before the war in Syria

SCAMMER:
7/4/2020, 11:55:40 AM
when i was at missionary work in Turkey

BAITER:
7/4/2020, 12:02:48 PM
Hello,
That is very helpful, thank you. Do you perhaps have some literature I could look at, or could you send me some more information about what exactly makes you unique?

Tom Jackson
CEO, MNesty, LLC

SCAMMER:
7/4/2020, 12:51:18 PM
Mr. Tom Jackson.
Send your telephone to me.
Or you contact through UNHCR REFUGEES AGENCY.
UNHCRFRANK@GMAIL.COM
Georgia in Europe

BAITER:
7/4/2020, 1:00:11 PM
Hello,
That is reasonable, but do you think you can go into more detail about your main product? What is it about, exactly? Does it resemble https://photricity.com/flw/ajax/?

Tom Jackson

CEO, MNesty, LLC

SCAMMER:
7/6/2020, 10:16:48 AM
Mr. Tom Jackson
If you are not interested for my proposal then you let me to know stop wasting my time in writing you.

BAITER:
7/6/2020, 10:24:52 AM
Hi,
Great, thanks. Could we have a short call to discuss the specifics? What are your contact details? Also, what is your pricing model like?

Tom Jackson
CEO, MNesty, LLC

SCAMMER:
7/7/2020, 11:30:05 AM
Your phone number.

BAITER:
7/7/2020, 11:43:47 AM
Hi,
Great, thanks. Could we have a short call to discuss the specifics? What are your contact details? Also, what is your pricing model like?

Tom Jackson
CEO, MNesty, LLC

Random Template Group Case 2

SCAMMER:
6/27/2020, 3:59:51 AM
MRS ROSE HOLTSBERY

BELOVED ONE,

Greetings in the name of God,Please let this not sound strange to you for my only surviving lawyer who would have done this died early this year.

I prayed and got your email id from your country guestbook.I am Mrs Rose Holtsbery from London,I am 58 years old,i am suffering from a lo ng time cancer of the lungs which also affected my brain.From all indication my conditions is really deteriorating and it is quite obvious that, according to my doctors they have advised me that i may not live for the next two months,this is because the cancer stage has gotten to a very bad stage.

I was brought up from a motherless babies home was married to my late husband for twenty years without a child ,my husband died in a fatal motor accident Before his death we were true believers.Since his death I decided not to re-marry,I sold all my inherited belongings and deposited all the sum of 10million dollars with a Bank.

Presently, this money ! is still with the bank and the management just wrote me to come forward and claim my money because they have kept it for so long or rather issue a letter of authorization to somebody to receive it on my behalf since I can not come over because of my illness or they get it confiscated.

Presently, I'm with my laptop in a hospital here in Switzerland where I have been undergoing treatment for cancer of the lungs. My doctors have told me that I have only a few months to live.It is my last wish to see that this money is invested to any organization of your choice and distributed each year among the charity organization ,the poor and the motherless babies home.

I want you as God fearing person, to also use this money to fund church ,mosque, orphanages and widows,I took this decision ,before i rest in peace because my time will soon be up.

As soon as I receive your reply I shall give you the contact of the Bank.I will also issue you a letter of authority that will prove you as the new beneficiary of my fund before the bank.

Provide me with your information so i can send it to the bank as the new beneficiary and issue you a letter of authorization.
FULL NAMES:__________SEX: _____ AGE: ______MARITAL STATUS:______________COUNTRY: ______ CONTACT ADDRESS: _________________________PHONE NO#___________FAX NO#________________OCCUPATION :______________

Please assure me that you will act accordingly as I stated herein.Hoping to hear from you soon.

Mrs.Rose Holtsbery

BAITER:
6/27/2020, 12:07:39 PM
Hello,
I am a bit busy now, but I am definitely interested. When can we talk?

Matthew Jackson
CEO, MNesty, LLC

SCAMMER:
7/1/2020, 12:25:16 PM
Dear Matthew Jackson,

Good to hear from you, you are the chosen one, all I want from you is to
assure me that you will use this fund to help the poor.

I don't want my husband brothers to know about this fund because if they know, they will kill me before my time.

Send me your informations as stated below, to enable me prepare power of attorney in your name so that you will have the legal right to contact the holding bank.

1. Your valid ID.
2. Contact address.
3. Direct number to reach you.

Kindly send me the above information when replying my message.

Regards,
MRS. Rose Holtsbery.

BAITER:
7/3/2020, 10:39:00 AM
Hello,
that sounds impressive! I am excited to hear more about your company. Can you help me with some references, completed projects, partners, etc.?

Matthew Jackson
CEO, MNesty, LLC

GPT-2 Group Case

SCAMMER:
7/11/2020, 11:16:54 AM
Dear Friend

My name is Mrs Deri Otto. I know you don't know me, but I got your address from my search for a reliable and trusted partner here and i have decided to make you my friend because we bear the same believe which is divine to be of the same believes.

i am the wife of the late Mr OTTO GEROGE who died of the COVID19 PANDEMIC. Before his death, he has in excess $25m and some quantites of Gold in a secret offshore account which he made through Gold business here.

unforunately, death claimed him and this fund and the documents are in my possesion. no other person is aware of it except the bank whom the Bank manager advsed me to look for a trusted foreign partner to transfer the money to his account for investment between me and him.

Based on this, i decided not looking for pity, but I am looking for help. I would like to distribute thisfunds and the quantities of Gold to charity organizations and also to invest part in your country through someone that I can trust.

I propose 30% to you for your assistance.If you would be interested in such an opportunity, please respond and I can send you further details.

Best Regards,

Deri Otto

BAITER:
7/11/2020, 11:48:24 PM
Deri,

You really have some nerve to write to me. I'm not some cold-hearted capitalist scoundrel, or so you'd say. I'm a woman with family, lots of kids by 4, and I've been doing this for a long time. I'm a little grateful that you would have my email address, as I'm not sure what to do with it.

The money is available for a variety of charitable uses. My son Ezra has a new restaurant in Calabar, just a short drive away, that's what I need to put in, and the kids love it. I'm a single woman, sadly, so I'm not very interested in your proposal, but if you'd mind, please call me.

Your friend,

Bernadette Devlin

SCAMMER:
7/16/2020, 6:37:58 AM
My Dear Beloved in Christ

Greetings to you in the name of our Lord Amen and I pray this mail will meet you and your family in good health? Thank you very much for your response i want you to understand that my late husband deposited this money with the bank and he agreed with the bank not to release the fund here because he plane to use the fund for an investment in abroad before his death may his soul continue resting in peace Amen .

My dear , all I ever expect is that this money is transferred to your account as soon as possible . we need this transfer to proceed immediately , because if anything happens to me now , the bank will claim this fund as unclaimed fund by them self − since I am no longer alive. I really need us to go much faster now that I am still alive so that the bank will transfer this fund to your account before anything happens to me, but not my prayer . Please try and help us in the name of God , after sending you this e− mail i will meet with the bank manager to find out the easiest possible way to complete the transfer to your account sooner , without no delay . I pray that God Almighty will help us for the bank to transfer this fund into your account as quickly as possible so you can be able to help us out of this country Amen. My dear I will need your prayers and concern over my health , and the doctors told me I need to go fast cancer surgery if not I will not survive but i trust in God because with him all things are possible, but I can not believe what the doctors are saying here because my life is in God's hands and not the Doctors.

My Dear i only need an a reliable individuals from abroad that will receive this fund for the work of God because i will be presenting you the bank here as a foreign guardian to my Son David. My Dear , the Bible says over 365 times, that's mean we have NO FEAR anymore, because our LORD JESUS − HE have done everything.Amen. and in the Lord there are three things very important into our Life ,faith hope and love but the greatest is love May the peace of God Jehovah through Jesus Christ and His Holy and Blessed Grace be with you. Amen !!! We must center our faith in the Lord Jesus Christ. To have faith in Jesus Christ means to have such trust in Him that we obey whatever He commands. As we place our faith in Jesus Christ, becoming His obedient disciples, Heavenly Father will forgive our sins and prepare us to return to Him. Please David my Son is only 17 YEARS OLD STILL A COLLAGE STUDENT My Dear i wait to hear from you urgently after you read this mail.i want you to know that my Son is only 17 years old and still at the collage.

Please Read Bellow again stated the 5 purposes after you get this money into Your Account

A) DONATION TO MOTHERLESS BABY HOMES 10%

B) 20% WILL BE USE FOR BUSINESS INVESTMENT THERE IN YOUR COUNTRY

C) KEEP SOME MONEY FOR MY SON'S EDUCATION AND FOR HIS BETTER FUTURE 20%

D) AND PART OF THIS MONEY WILL BE USE TO CARE OF MY HEALTH IN YOUR COUNTRY .

E) 20% WILL BE FOR YOU AND YOUR FAMILY FOR HELPING US IN OUR OF NEED

Please do Reply as soon as you receive this mail with:

YOUR FULL NAMES:
YOUR ADDRESS:
YOUR PHONE NUMBERS:

This will enable me to submit your information to the bank here where my Late Husband deposited this inheritance fund before his death I will be waiting For Your Urgent Response so that I will give you the contact information of the bank where my late husband deposited this fund for the bank Director to contact you to proceed with the transfer into your account. I wait for your urgent response Have a wonderful Day in the Lord In His Love

Mrs Deri Otto And Son David

BAITER:

7/16/2020, 6:51:40 AM
Dear Deri,

Thank you for your prompt response. Let me say first of all that you are a GOD−fearing Christian who I can count on to help run my business.

I have a dear niece who is about 19 years old and I am so busy that I often do not have time to attend to her needs. I would be irrelevant to what your religion says about me.

I would rather not tell you about my medical condition because it is nothing to do with me. I have a cat, a horse and a bit of a moodle, but I am happy with my current financial situation. I do, however, have a medical condition that I think may be of some concern to you, because I have never had either myself or anyone else come directly to my hospital for medical treatment.

I would like to hear from you as soon as possible, please. I also want to know what kind of insurance will be involved with this transaction.

In the meantime, I am going to my bank to make my first payment. I will not be able to send you the money before the end of today, so can you please tell me what is the bank limit for cash?

I would like to hear from you, as well.

Your friend,

Beavis Bungholio

P.S. I would like to hear your voice, if you have one at all. I would prefer if you did not. My fax number is (888) xxx−xxxx.