Active Countermeasures for Email Fraud

Wentao Chen
University of Bristol
Bristol, United Kingdom
wentao.chen.uob@gmail.com

Fuzhou Wang
City University of Hong Kong
Kowloon, Hong Kong SAR
wang.fuzhou@my.cityu.edu.hk

Matthew Edwards
University of Bristol
Bristol, United Kingdom
matthew.john.edwards@bristol.ac.uk

***Abstract*—As a major component of online crime, email-based fraud is a threat that causes substantial economic losses every year. To counteract these scammers, volunteers called scam-baiters play the roles of victims, reply to scammers, and try to waste their time and attention with long and unproductive conversations. To curb email fraud and magnify the effectiveness of scam-baiting, we developed and deployed an expandable scam-baiting mailserver that can conduct scam-baiting activities automatically. We implemented three reply strategies using three different models and conducted a one-month-long experiment during which we elicited 150 messages from 130 different scammers. We compare the performance of each strategy at attracting and holding the attention of scammers, finding tradeoffs between human-written and automatically-generated response strategies. We also demonstrate that scammers can be engaged concurrently by multiple servers deploying these strategies in a second experiment, which used two server instances to contact 92 different scammers over 12 days. We release both our platform and a dataset containing conversations between our automatic scam-baiters and real human scammers, to support future work in preventing online fraud.**

1. Introduction

According to the Internet Crime Report, the FBI's Internet Crime Complaint Center (IC3) received 847,376 reported complaints in 2021, corresponding to over $6.9 billion in potential losses. Email, as one of the primary online communication media, is linked to a significant proportion of these losses. The IC3 records 19,954 reports of email account spoofing and compromise fraud in 2021, accounting for over $2.4 billion in losses– over a third of all damages [38]. Online confidence tricks and romance fraud schemes, also often carried out via email exchanges, account for a further 24,299 victims and $956 million in losses. While email scammers have been known to use social engineering techniques that pre-date even the Industrial Revolution [57, p. 58], the regular innovations of scammers in their style and content, and the mass-market targeting of email scamming make this threat a constant problem that causes severe financial loss and impacts lives worldwide.

Traditional approaches to combatting this threat have focused on identifying which Internet users may be most vulnerable [25], [37], educating Internet users about the existence and risks of such schemes [9]–[11], [29], [41], building classifiers that can automatically filter out messages containing fraud [3], [14], [34], [59], or building and maintaining blacklists of email senders known to be untrustworthy [8], [54]. A common thread between all these approaches is that they are fundamentally inwardly-directed and defensive, aiming to increase the resilience of Internet users (or their inboxes) against the social-engineering-based attacks they are receiving. Recent work suggests that this approach is not working well enough, and a fundamentally more active set of countermeasures should be explored [6].

As an example of what active countermeasures might look like in this domain, we turn to an existing initiative in the voluntary anti-fraud community, known as scam-baiting. To protect victims from contacting scammers, some volunteers reply to fraudsters, adopting the guise of possible victims, and engage them in long and unproductive conversations, which distract fraudsters into wasting time and attention they would otherwise have spent on real victims. These volunteers are called scam-baiters, and there is reason to believe that scam-baiting activities could be particularly efficient at disrupting the operations of scammers, as they can decrease the density of real victims in replies to scammers to the point where the email scam business model becomes unprofitable [22]. However, scam-baiting is currently a small-scale hobbyist activity, and scam-baiters expend significant time and energy in playing their roles. As a general countermeasure, human scam-baiters could not respond at the current scale of global email-based fraud. As such, we turn our attention to methods of automating their approach.

There is a limited literature that precedes us in automated conversational interactions with scammers. Some work has approached this topic as an extension of the honeypot, as in the Honey Phish Project, in which automated mailboxes would reply to phishing emails with links that, when clicked, reported identifiable information about the phishers back to the honeypot operators [17]. A more conversational email agent was used for similar purposes by McCoy et al. in a study gathering information on rental scams [35]. While these could be considered a form of active countermeasure, this approach is fundamentally about gathering information on the attackers, rather than disrupting their operations. More similar to our intent is the 'Jolly Roger Bot' developed by Anderson, which answers telemarketing calls by responding with audio clips of random statements [2], [6]. The simpler of the methods we explore resembles a textual version of this approach. However, the time wasted in a phone call is capped at the order of minutes, whereas email scam-baiter

conversations can lead fraudsters on for *weeks*. Edwards et al. [14] described some of the persuasive techniques used by human scam-baiters to achieve these results, noting that they often mirrored the tactics used by the scammers they targeted. It is the automatic deployment of these techniques against cybercriminals that Canham & Tuthill advocate [6], and that we explore.

In this paper, we describe our implementation of a mailserver that can engage in scam-baiting activities automatically. We develop three alternative response-generation strategies using publicly-available email corpora as inputs to deep learning models, and perform a one-month comparative evaluation and 12-day concurrent-engagement experiment in which our system interacted with real fraudsters in the wild. In short, our contributions are:

- We demonstrate that automated scam-baiting is possible, with 15% of our replies to known scam emails attracting a response and 42% of conversations primarily involving a human fraudster. Some conversations lasted up to 21 days.
- Further, we compare different approaches to automated scam-baiting in a naturalistic experiment using randomised assignment. We find that human-designed lures work best at attracting scammer responses, but text generation methods informed by the methods of human scam-baiters are more effective at prolonging conversations.
- We also engage the same group of scammers with two identical scam-baiting instances simultaneously. We find that two concurrent instances attracted responses from 25% of the targeted scammers and 29% of these scammers engaged with both scam-baiting instances simultaneously.
- We release our code as a platform which can be deployed to test alternative response strategies and iterate on our findings. We also release both full transcripts of our automated system's conversations and a collection of human scam-baiter conversations, to guide the development of new active countermeasures and provide insight into scammer operations.

The rest of this paper proceeds as follows. In Section 2 we provide some background on scam-baiting as a human activity, as well describing the fundamental models used within our work. Section 3 outlines our deployment platform. Section 4 describes the different corpora we make use of for finetuning and prepatory evaluation. Section 5 describes said finetuning and classifier evaluation, while Section 6 describes our main experiments, including the results from our comparison of the different response strategies. Section 7 reflects on our findings, their limitations, and our suggestions for future improvements, as well as considering misuse concerns. We conclude with a summary of our key results and recommendations.

2. Background

To provide the essential basics for active scamming defense, this section describes the significance of scam-baiting activities and recent advances in text generation that enable adaptive conversational AI.

2.1. Scam-baiting protects potential scam victims

Scam-baiting is a kind of vigilante activity, in which scam-baiters reply to the solicitation emails sent by scammers and enter into conversation with them, in order to waste scammers' time and prevent them from scamming other potential victims. This activity has become an Internet subculture with various scam-baiter communities across the Internet. Past research on scam-baiting has explored the various motivations of scam-baiters [53], [62], the strategies they use in conversations [13], [14] and the ethics of their activities [51].

We attach importance to scam-baiting activity because, by wasting scammers' time, scam-baiting can help to protect other vulnerable people from being scammed. Herley [22] argued that scam-baiting activity can sharply reduce the number of victims found by scammers by decreasing the density of viable targets (i.e., the targets that can lead to financial gain), making them less likely to harm the potential victims. This argument seems to be upheld in practice, as well. Scam-baiting exchanges generally end in frustrated invectives from scammers once they have understood what is taking place [14], and prominent scam-baiter and comedian James Veitch reports pointedly about scammers pleading with him to stop emailing them [6], [56]. Some scam-baiters also use their activities as a means to prod scammers into reflecting on what they are doing (e.g., [47]), but the effectiveness of this last technique is unknown.

There are existing industrial email conversation products that make use of NLP techniques to craft responses [33], [39], [46]. These products may conceivably be adapted for automatic scam-baiting; however, their focus is providing automatic customer service, and they are tested only on responding appropriately to business email communications. As a result, the performance of these products at scam-baiting—which may involve intentionally drawing out conversations to better waste scammer resources–has not been evaluated. Previous anti-spam conversation systems, such as RE:Scam [60], have demonstrated the basic viability of automatic responses for disrupting scammer operations, using random selection from a series of canned template responses. This is a promising start, but we suspect that an automatic scam-baiter that can respond to the content of a scam message could be significantly more effective at prolonging conversations. Our system is the first open-source email conversation system specialized for automatic scam-baiting. The system aims at utilizing general NLP models for the purpose of consuming scammers' resources.

2.2. Text generation for email conversations

The recent successes in natural language processing (NLP) have given rise to the prosperity of automatic dialogue systems [7]. The most prominent architectures include the Transformer [58] and its variants BERT [12] and the GPT family [44]. The emergence of transformers has enabled the pretraining-finetuning approach in NLP, which was not possible in the era of RNN [48] and LSTM [23]. We briefly introduce these models in this section.

Transformers. The Transformer architecture is based solely on attention mechanisms, without recurrence. The model is more parallelizable and requires significantly less time to train [55].

The core of Transformer models is the *attention* mechanism. An embedded word vector is multiplied by three different matrices to get three feature matrices, then the features of words are operated with other words' features to obtain the attention, which refers to the relationship between this word and the whole sentence. In practice, we usually calculate multiple attention matrices of a single word, an approach known as *multi-head attention*. These attention matrices are concatenated into a larger matrix, which is then multiplied with a weight matrix, returning it to its original size for feeding forward.

$$
\begin{aligned}
Q_i &= X \cdot W_i^q \\
K_i &= X \cdot W_i^k \\
V_i &= X \cdot W_i^v \\
Z_i &= softmax(\frac{Q_i \cdot K_i^T}{\sqrt{d_k}})V_i \\
Z &= Concat(Z_1, Z_2, ..., Z_h) \cdot W^O
\end{aligned}
$$

The Q_i, K_i and V_i denote the feature values of input sentence X. The Z_i is the i-th attention values while Z denotes the final output of the multi-head attention process. The feature matrices (W_i^q, W_i^k and W_i^v) and weight matrix W^O are obtained from training as the parameters of the model. The term d_k denotes the dimension of the K_i feature vector.

Since the attention matrix only relates to the word vector itself, it does not contain any context and position information. To solve this problem, a *position embedding* was added to represent the position features of word vectors. The position vectors are added before word vectors are input into multi-head attentions.

$$X = X_{raw} + E$$

In the following sections, we will briefly introduce two specialized Transformer models that are widely used in the field of NLP.

BERT. Bidirectional Encoder Representations from Transformer (BERT) is a language representation model based on the Transformer architecture, described by Devlin et al. in 2018 [12]. Unlike traditional language models, the pre-training process of the model is split into two tasks. In the first, a *Masked Language Modelling* task, the training data generator masks 15% of word tokens at random, which the model learns to predict. The loss function of this step only calculates the difference between the origin of masked token and the predicted token, excluding the position. In the second task, *Next Sentence Prediction*, in order to train the model to understand the coherence and relationship of sentences, when choosing the sentence A and sentence B for a training example, the data generator replaces the sentence B into another random sentence in the corpus 50% of the time. In the pre-training step, two masked sentences are given as input to the model, and the model needs to determine whether these two sentences have been randomly concatenated.

After the pre-training process, the model can be put into a finetuning process for specific training on a task. The finetuning methodology can take a straightforward supervised learning structure, and as a result the BERT model has been applied to a variety of different tasks. In [12], the model is finetuned for 11 NLP tasks, including question answering, single sentence tagging and sentence classification. The BERT has also been used in machine translation [61], target-dependent sentiment classification [19] and sentence similarity [45]. For the classification tasks in this study, a fully connected layer is applied after the output of transformer structure to predict the category of input text.

GPT by OpenAI. GPT [42] is one of the major milestones of language modelling that precipitated the rapid development of self-supervised natural language generation. In a self-supervised manner, the model is trained to generate conditional synthetic text by predicting the next word based on the given context. The variants of GPT, including GPT-2 [43], GPT-3 [5], and GPT-Neo [4], are direct scaling-ups on top of the GPT algorithm. At the time of this study, the latest versions of GPT were GPT-3 and GPT-Neo[1].

The architecture of GPT is composed of multiple layers of transformer decoder. Similar to BERT and most transformer-based NLP models, the training framework of GPT models is composed of two steps – the pre-training step and the finetuning step. At the pre-training step, given a series of tokens encoded from the context, the GPT model tries to maximize the likelihood of the corpus tokens $\mathcal{V} = \{v_1, ..., v_n\}$:

$$\Theta^* = \text{argmax}_\Theta \sum_i^n \log P\left(v_i \mid v_{i-k}, ..., v_{i-1}\right)$$

where Θ denotes the parameters of a neural network, and k represents the size of the context window.

For each iteration of training, the context matrix is forwarded through the multi-layer transformer decoder, which can be formulated as below:

$$
\begin{aligned}
z_0 &= VW_e + W_p \\
z_l &= \text{decoder\_block}\left(z_{l-1}\right) \ \forall l \in [1, n] \\
P(v) &= \text{softmax}\left(z_n W_e^T\right)
\end{aligned}
$$

where W_e denotes the learnable token embedding matrix, and W_p is the position embedding matrix. V denotes the context matrix.

In the finetuning step, the model accepts additional supervision information for other downstream tasks or adjusts itself to generate text in specific domains.

By training on extremely large corpora, the GPT models achieved impressive performance on a variety of tasks. In particular, the GPT models achieved state-of-the-art performance in terms of conditioning on long-range contexts [43]. In this study, the contexts inputted to the model were email text, which are usually lengthy. The long-range context dependency of GPT models enabled

1. An open-source implementation of GPT3-like models, available at https://github.com/EleutherAI/gpt-neo.

a longer-term memory of linguistic information, and thus helped address the problem of losing conversational context when interacting with scammers.

Conversational Artificial Intelligence. With the aforementioned developments in NLP, conversational artificial intelligence has been widely deployed in the real world. Conversational AI is classified into three categories by Gao et al. [18], including question-answering (QA) systems, task-oriented dialogue systems, and fully data-driven chat systems. QA systems involve a pipeline of reading comprehension, knowledge base extraction, and answer generation, and task-oriented dialogue systems understand the users' instructions or queries to decide the back-end operations, which are usually deployed as AI assistants (e.g., Alexa by Amazon and Assistant by Google).

The task in our study, however, is most relevant to the fully data-driven chat systems, but with longer prior contexts (i.e., email messages). This type of neural system consists of end-to-end models trained on open-domain text. For example, Li et al. [30] used a diversity-promoting loss function to make the model generate diversified and meaningful responses, and in [31], reinforcement learning was adopted to improve the quality of the text generated by the model. In recent years, after the introduction of the GPT family, more advances have been made in the field of conversational AI. In previous work [21], [36], GPT-2 and GPT-3 were examined for the purpose of constructing conversational AIs, and both of them delivered remarkable performance, although, in one evaluation [36] GPT-3 was criticized for its lower language variability than human-written text.

3. Scam-baiting Mailserver

In this section, we describe a mailserver that is capable of conducting scam-baiting experiments automatically. We used this architecture for the experiments described later in this paper and implemented three different response strategies, using an AI natural language classifier and two text generators(see Section 4 and Section 5). It is important to mention that this mailserver design is not particular to our specific response strategies, and could easily be deployed to trial new automatic scam-baiting techniques[2].

3.1. Server Structure

We designed a modular server structure, as shown in Figure 1. Scam emails are crawled automatically from online reporting platforms using a Crawler module, and placed into a queue. The application server then regularly polls the queue and distributes work evenly between the registered Responders. Responders will use unique mail-names for each conversation. When a scammer responds to an email, their response is routed back through the queue for scheduling purposes, but is always assigned to the Responder which they originally addressed. It is important to note that the server is designed to run on a unique domain independently, so it is not capable of spoofing an existing email address, nor connecting to the inboxes of others.

2. The source code can be found at https://github.com/scambaitermailbox/scambaiter_backend

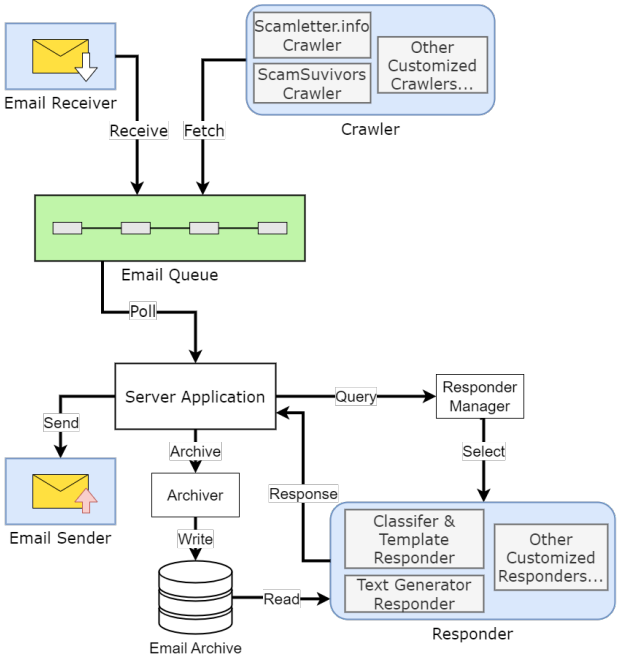

Figure 1: The modular mailserver architecture

3.2. Server Modules

Crawler. The Crawler module is used to fetch scam emails from online sources at a regular intervals. The emails we crawl are public copies of solicitation emails sent by scammers which have been captured by spam traps or reported by anti-fraud volunteers. The Crawler modules store cleaned copies of these emails in a queue structure, along with the email address to which replies should be addressed.

Email Sender & Receiver. Due to the port limitation on most VPS providers, we chose to use a relay email service provider for receiving and sending mail. The Email Sender module is an intermediate layer for transforming email arguments into the API-required JSON content, and sending the HTTP request to submit an outbound message. To receive responses from scammers, our server listens for POST requests from the relay service. Once the Email Receiver receives the request, it extracts the response content, and submits it to the email queue. These relay email servers are correctly configured with SPF and DKIM, with different RSA-SHA256 public keys and unrelated domain names. In a production mail service deployment, the reliance on a relay provider would not be necessary, and a traditional mail agent could replace these modules.

Responders. The Responders are used to process the incoming messages from scammers and generate text replies. They play the role of the automatic scam-baiters in the server. In this study, we implemented three Responders using two AI text generation models and one AI text classification model. We further described how we prepared these models in Section 5, and how we used three

Responder strategies in experiments in Section 6.1. The Responder module is not limited to our designed strategies and for further research, Responders can be implemented using other reply strategies.

Responder Manager. The Responder Manager is responsible for picking a Responder for each scammer. A specific scammer will always communicate with the same Responder. When a new target is acquired by a Crawler, the Responder Manager will check if the scammer's email address has been seen previously. If not, it will assign the least-used Responder to this scammer, so that each strategy is trialled equally. The Responder Manager also generates random mailnames and fake names for Responders in each conversation. The mailname will be used with the server domain to send replies to the scammers, and the fake name will be signed at the end of the reply.

Archiver. The archiver is in charge of archiving conversations between responders and scammers. These logs can be used to audit ongoing conversations, or exported for post-hoc analysis.

4. Corpus Preparation

4.1. Data Sources

We collected email data from three different sources:

1) The **Enron Email Dataset** was collected and prepared by the CALO Project, and contains email data from about 150 users, mostly senior management of Enron [28]. In this study, we used the latest version of this dataset (updated May 7, 2015), with over 1.7Gb of messages in total.
2) Our **Scam-baiting Dataset** contains conversations between scammers and real human scam-baiters. This dataset extends the ADVANCE-FEE SCAM-BAITING dataset provided by Edwards et al. [14]. The total size of this dataset is over 70MB, containing 658 conversations and over 37,000 messages[3].
3) **ScamLetters.Info** is an online dataset that contains thousands of scam emails. The emails are submitted by website visitors and reviewed by administrators or collected automatically by their filtration system. The scam emails in this corpus cover a time span from 2016 to the current day. We select from this corpus, making use of scam emails reported between April 2022 to June 2022, the 3 months preceding the development and deployment of our system.

4.2. Preprocessing

Our corpora were prepared for two distinct tasks, as illustrated in Figure 2. First, the **Enron** and **Scam-baiting** datasets were organised into an *email/reply* conversation structure that enabled us to finetune a GPT language model to respond appropriately to emails. Within the **Enron** corpus, all conversation chains containing at least one response between authors were included in this structure. Within the **Scam-baiting** corpus, the structure was restricted so emails from scammers were prompts, and responses from human scam-baiters were the replies. These two conversation-structured datasets fed into different strategies for reply generation, as described in Section 5.1.

3. This dataset can be found at https://github.com/scambaitermailbox/scambaiting_dataset

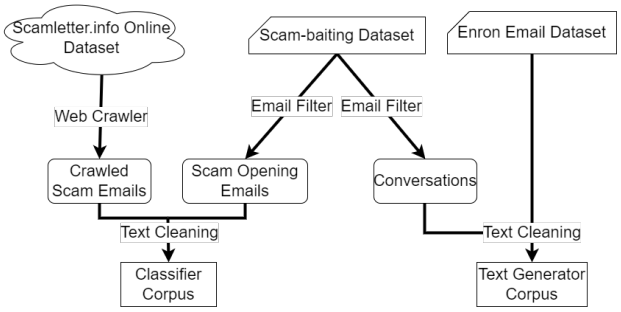

Figure 2: The overall process of the training corpus preparation.

Second, we prepared a corpus to use for the automatic categorisation of scam emails on the basis of the different scam formats in use. This drew upon the initial solicitation emails sent by scammers, to which victims (or scam-baiters) respond. We selected 508 such messages from the conversations in the **Scam-baiting** corpus, and supplemented these with a further 613 messages from our **ScamLetters.Info** corpus. These 1,121 emails were then manually categorised into 11 fine classes of scam email, building on the seven classes suggested by Pan et al. [40]. However, finding underwhelming performance, we reduced the task to just five coarser labels (see Section 5.2 for a full breakdown of the labels and performance).

5. Model Preparation

We allocated one NVIDIA TESLA T4 GPU with 16GB VRAM for training models. In terms of software, we used Happy Transformer [15] and Flair [1] as our training framework.

5.1. Text Generators

We chose GPT-Neo [4] pretrained model with 125M parameters as our base model and finetuned two text generators to compare their performance in generating responses as a scam-baiter. **Generator A** was finetuned using only the prepared **Scam-baiting** corpus, while **Generator B** was finetuned on the **Enron** corpus first, then further finetuned on the **Scam-baiting** corpus.

Generator A (Scam-baiting only). The intention of this approach was to directly finetune a language model on the efforts produced by human scam-baiters. In order for the pretrained GPT-Neo model to learn the pattern of conversations in our scam-baiting corpus, the training epoch was set to 100 and the initial learning rate was set to 1e-04 with dynamic decay. After 58 hours of finetuning, the model steadily generated coherent responses in the required format, demonstrating adaption to the task.

Generator B (Enron + Scam-baiting corpora). The intention of this approach was to explore whether the language model would produce more convincing responses if it was first trained to respond to business emails before being finetuned for scam-baiting in particular (our experience of the possible Enron-only variation was that it produced text too specific to the Enron corporation to make sense to scammers). To familiarise our generator with general email patterns, during the finetuning process of Generator B, we started by training it with the Enron corpus. Due to the disparity with the amount of text in the scam-baiting corpus, we lowered the learning rate to 5e-05 and reduced the epoch to only 5. We then used the same finetuning process as described for Generator A to further finetune this model.

5.2. Scam Classifier

As a point of comparison for our text generator approaches, we aimed to include a response strategy in which a model selects randomly from pre-written responses appropriate to the type of scam the conversation seems to be. To achieve this, we first needed to be able to automatically categorise scamming strategies on the basis of the messages received. We began by extending Pan et al. [40]'s seven classes into 11 finer classes:

- **Business** These emails are about purported business opportunities, in which the scammer will require victims to pay invented deposits or taxes.
- **Tragedy** These scams report a personal tragedy (usually the death of a close relative) and ask victims to help (usually by handling fees related to the inheritance).
- **Cargo** Victims are informed that their cargo was detained at customs, and asked to pay fines to release it.
- **Investment** These scams introduce a false investment opportunity, and lure victims to invest by promising large returns.
- **Romance** Victims are persuaded into an online romance with an invented persona, which is then used as leverage to extort money.
- **Job** These emails provide fake job vacancies to victims, which will result in application fees.
- **Lottery** These emails falsely announce that the victims have won a lottery, and ask them to submit private information or pay a nominal fee to claim their winnings.
- **Donation** These emails appear to be from charities or church institutions, and ask victims for donations.
- **Sales** These emails are trying to sell fictitious products to victims.
- **Loans** These emails offer victims unrealistic loan opportunities.
- **Other** Scam formats other than those listed above.

After manual labelling, we finetuned a version of the pretrained DistillBERT [49] model to distinguish the classes on the basis of textual features in the solicitation email. After 50 epochs, this classifier achieved a mediocre accuracy of 0.6555 on the held-out validation set, broken down by category in Table 1. In evaluating its performance, we considered the overall distribution of our class labels (as given in Figure 3) to be a potential harmful factor. Some classes are sparsely populated, and they are not well-balanced. As such, we integrated our class labels into two broader classes of scamming strategy: *transactional* scams, in which as part of their story the scammer makes clear that they are entering into a professional setting in which they expect to be remunerated, and *non-transactional* scams, in which an expectation of payment is not at first made clear to the victim. The *other* category was retained as a decision point for unclear scam categorisation decisions. As the classifier showed good performance with the original *romance* and *lottery* classes in our initial evaluations, we also kept those categories distinct from the broader *non-transactional* grouping, in order to test whether more specifically-crafted pre-written responses could be more effective.

After finetuning with the same model and parameters using this new scheme, we achieved a more acceptable 0.7982 in overall classification accuracy. Table 2 shows the performance of the final classifier by each category used in our final deployment.

Figure 3: The distribution of categories before integration

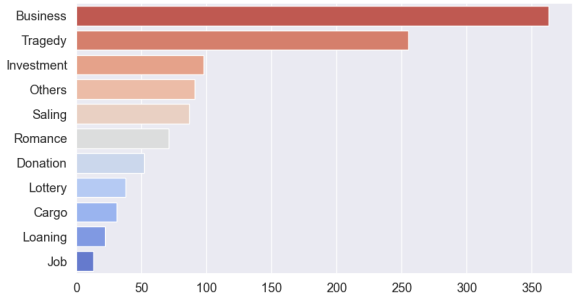

TABLE 1: Evaluation of the 11-category classifier

	precision	recall	f1-score	support
Business	0.6216	0.6216	0.6216	37
Tragedy	0.5862	0.6538	0.6182	26
Investment	0.6154	0.8000	0.6957	10
Saling	0.7000	0.7778	0.7368	9
Romance	0.7778	0.8750	0.8235	8
Other	1.0000	0.3000	0.4615	10
Donation	0.6667	0.6667	0.6667	6
Lottery	1.0000	1.0000	1.0000	4
Cargo	0.6667	0.5000	0.5714	4
Loaning	0.7500	1.0000	0.8571	3
Job	0.0000	0.0000	0.0000	2
accuracy		0.6555		119
macro avg	0.6713	0.6541	0.6411	119
weighted avg	0.6709	0.6555	0.6447	119

6. Experimental Results

6.1. Design

In order to test and evaluate our scam-baiting models, we conducted experiments that established communications with real human fraudsters. As this involved active deception of human participants, who took part without their knowledge, our study first underwent review by

TABLE 2: Evaluation of the 5-category classifier

	precision	recall	f1-score	support
Transactional	0.8125	0.8966	0.8525	58
Non-Transactional	0.7742	0.7273	0.7500	33
Other	0.6364	0.6364	0.6364	11
Romance	1.0000	0.6250	0.7692	8
Lottery	1.0000	0.7500	0.8571	4
accuracy		0.7982		114
macro avg	0.8446	0.7270	0.7730	114
weighted avg	0.8042	0.7982	0.7963	114

our institutional review board, and received a favourable opinion (ethics approval code `11117`). All experiments were carried out on our scam-baiting mail server instances (see Section 3).

As detailed in Section 5, we designed 3 reply strategies to generate responses to email scammers. These strategies were:

1) **Classifier & Template** The text content of emails is used by our 5-category classifier to identify the scam format. Then the responder selects a random pre-written response we manually designed for emails of that category, appends a signature, and delivers it.
2) **Text Generator A** The text content of the scam email is used to prompt our text generator fine-tuned with the **Scam-baiting** corpus, and the generated response is signed and delivered.
3) **Text Generator B** The text content of the scam email is used to prompt our text generator fine-tuned with both the **Enron** and **Scam-baiting** corpora, and the generated response is signed and delivered.

All the strategies shared the same email generation process. The subjects of replies are the crawled solicitation letter subjects prefixed with "Re:". The original messages are not quoted in replies, in order to prevent replies being marked as spam. The response text is embedded in a fixed HTML template, with "font-size" and "font-family" style attributes. The "from" field includes the fake human name and generated random address. Generated emails passed the tests on "mail-tester"[4] and in testing have not triggered the spam filters on mainstream mail server providers, such as Gmail.

We used two crawlers to fetch known scam emails, distributed these messages equally to each strategy, and sent back text responses using randomly-generated email addresses. For a particular scammer, we always used the same strategy to generate responses to his or her messages, and never initiated a second conversation with the same scammer.

To avoid accidentally targeting innocent email addresses, we implemented the following mitigation strategies:

1) **Reputable sources and manual reviews**
   All the scamming email addresses we acquired are verified by database or forum administrators with expertise in this area, and in our experiments we also manually reviewed these solicitation messages before sending any messages to them. We also monitored replies from target addresses during the conversations. If manual review of a reply indicated that a respondent was confused about our reference to a fraudulent message they are meant to have sent (i.e., does not appear to be a scammer), we would explain the experiment and stop this conversation immediately. During our experiments, our servers occasionally received new scam emails from unknown senders, but we did not respond to these messages since they have not been verified.
2) **One bait per address**
   In our review of the scam databases, we noticed that the same scammer may send several different solicitation messages from the same address, but for a particular email address, we never initiated a second conversation per mail server. This would minimize the annoyance caused by any targeting of innocent email addresses.
3) **No 'chasing' behaviour**
   During conversations, our mail servers would not send any messages to the target email addresses until they received further responses. The target recipients were able to stop replying to our baiters at any time to end the conversation, and would never receive further messages from other baiters in the mail server due to the "One bait per address" mitigation strategy.

We conducted two experiments at different periods of time: our **Single Engagement Experiment** compared the effectiveness of response strategies; the **Concurrent Engagement Experiment** tested whether our mailserver can consume scammers' resources by engaging them from concurrent instances. Both experiments shared the same reply strategies, crawlers and mitigation strategies.

6.2. Quantitative Results

1-Month Engagement Experiment. The experiment was started on 12th July, 2022 and ended on 20th August, 2022. During this period, we crawled 877 scam emails with unique sender addresses and tried to initiate conversations with their senders. These 877 possible conversations were equally distributed between our three strategies. During our evaluation process there were fraudsters who actively contacted our automatic scam-baiter addresses from email addresses other than those retrieved by our crawlers, but we did not reply to them, as they had not been identified as scammers by trusted sources. It is also important to note that some scammer email addresses became undeliverable either before we sent our responses or during the conversation.

We successfully elicited replies from 130 individual scammers in total (15% of our approaches, ignoring contact from unfamiliar email addresses). When we analysed the conversations, we noticed that some of the fraudsters used autoresponders which replied to our scam-baiters with the same content multiple times. To avoid such autoresponders diluting our results, if there were more than two duplicate responses in a conversation, this

4. An online email sender test integrated with multiple spam filters. The test is available at https://www.mail-tester.com

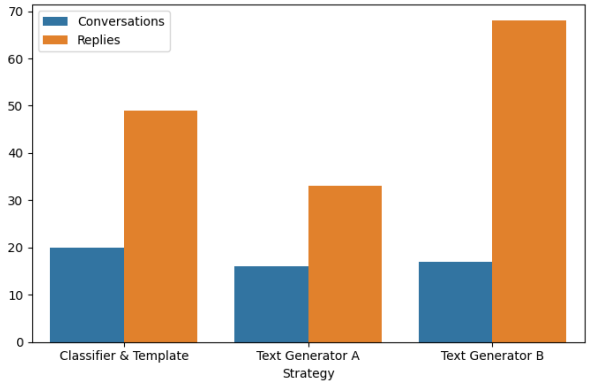

Figure 4: Comparison of the three reply strategies by quantity of conversations and replies

conversation would be marked as invalid and excluded from our analysis. It is important to note that some such conversations still contained apparently human scammer replies.

After filtering out conversations that contained autoresponders, we found 54 valid conversations between our bots and email scammers. The count of conversations and replies is given in Figure 4. *Text Generator B* got 68 replies across 17 conversations. The *Classifier & Template* strategy evoked the most conversations, 20, but it got only 49 replies. *Text Generator A* got 33 replies and invoked 16 conversations. We release all conversations between our systems and real scammers in a public repository[5].

TABLE 3: Comparison of the three reply strategies by average replies elicited per conversation, and the longest total duration of a conversation.

	Avg. replies	Longest Time
Classifier & Template	2.45	17 days, 4:16:19
Text Generator A	2.06	17 days, 8:21:31
Text Generator B	4	21 days, 11:15:43

To measure the time for which our models may have distracted the scammers, we calculated the time span between the first reply from the scammer and the last one received before the end of the study period. Table 3 shows the average number of replies per conversation and longest distraction time of each strategy. The longest distraction time of *Classifier & Template* is similar to that of *Text Generator A*, at about 17 days, while *Text Generator B* engaged with one scammer over 21 days.

Concurrent Engagement Experiment. We deployed two mail server instances with distinct domain names. Both instances used the same reply strategies (see Section 6.1) with identical templates and natural language models. In addition, both mail servers shared a common crawler module (see Section 3.2) to ensure that the servers responded to the same solicitations from the same scammers. The two instances were set up to reply to the scammers at different times of the day.

This experiment was started on 20th February, 2023 and ended on 3rd March, 2023. During this period, the common crawler module crawled 510 scam emails with unique addresses. Both servers independently and randomly assigned strategies for these emails. At the time our servers attempted to contact the crawled email addresses, 136 of them could not be reached because the addresses had been disabled or the addresses did not exist. As a result, 374 individual scammers are involved in this experiment.

Two server instances successfully elicited 102 and 136 replies from 62 (12.16%) and 57 (11.18%) individual scammers respectively. We found 27 scammers who engaged both of our instances in conversation simultaneously, with 4 of these scammers still being interested in conversations with both servers at the end of this experiment. Table 4 shows the number of elicited responses, drop-out and still interested scammers for each server instance. The longest distraction time of both instances was about 6 days and 10 days respectively. We release all conversations between both instances and real scammers in a public repository[6].

TABLE 4: Comparison of two server instances by separate and common numbers of engaged scammers, drop-out scammers and scammers that are still interested in the baiting conversation at the end of the experiment.

	Server A	Server B
Response Elicited	62	57
Drop-out	312	317
Still Interested	8	11
Common Engaged	27	
Common Dropout	282	
Common Still Interested	4	
Total Involved	374	

6.3. Strategy Comparison

Classifier & Template. Two examples of the conversations initiated by this strategy can be found in Appendix A. As shown, this strategy picked different templates when presented with lottery scam and transactional scam solicitation letters. The major flaw of this strategy is that the templates cannot be customised for a specific conversation, so after a few rounds the scammer begins to notice that the supposed victim does not properly address his enquiries—and may even repeat the same message—and so ceases to reply. This strategy successfully invoked the most non-robot conversations, but its performance in maintaining conversations was worse than that of *Text Generator B*. The longest conversation initiated by this strategy received five responses from the scammer.

Text Generator A. *Text Generator A* uses a GPT-Neo model finetuned with the scam-baiting corpus to generate responses for scammers. As a result, the model can usually generate replies that are relevant to the conversational context. Here is an example of a generated response:

```
PAULIbe informed that this message
you sent to our private email is
not our original business.
```

5. https://github.com/scambaitermailbox/conversations

6. https://github.com/scambaitermailbox/conversations/tree/concurrent

Yet sometimes the model generates misspelt phases (such as "PAULIbe" in the above example) or meaningless sentences. The performance of this strategy is not stable when compared to that of the *Classifier & Template* strategy. What is more, sometimes the model replied to scammers using a strongly negative attitude, and ended a conversation intentionally:

```
M statement because you do not answer
questions, you are not even clear on
the phone, it is not fax machine that
you speak at, and you do not have an
office address. you are very and for
that matter, you are rude and violent.
you will get hurt. You will regret it.
```

Quantitatively, this strategy evoked the fewest conversations and got the fewest replies. Its longest conversation received five responses from the scammer, the same as the *Classifier & Template* strategy. An example of the sort of conversations invoked by this strategy can be found in Appendix B. Overall, this strategy is slightly weaker at initiating the conversations than *Classifier & Template* and has a similar ability to maintain conversations.

Text Generator B. Our *Text Generator B* strategy also uses the GPT-Neo model, but with a finetuning process that exposed it to a larger and more varied email corpus. This model successfully adapted to the context of business emails and scam-baiting replies, and generates plausible (if occasionally confusing or ungrammatical) replies based on the content that scammer sent. Here is an example of a generated reply:

```
This is from our email, which have
been waiting for.
Is there someone can contact,
who know about who might be
interested in donating to do this.
Thanks,
```

This strategy successfully maintained two sustained conversations with scammers, which lasted for 23 and 15 rounds respectively, showing the potential for this strategy to be effective at maintaining long-term conversations. The ability of this strategy to initiate a conversation is similar to that of *Text Generator A*, but it substantially outperforms both the other strategies at maintaining long-term conversations.

To take this longest 23-round conversation maintained by *Text Generator B* as a case study (see Appendix C for the full transcript), the scammer started with an issue regarding fictional funds left to the victim in a similarly fictional bank account. The scam-baiter replied, acting the part of a confused recipient and demonstrating it was willing to chat with the scammer. The scammer became suspicious after the baiter could not provide private information that the scammer needed, so the text generator generated an emotional reply that asked the scammer to treat it better, and revealed that the invented victim was on their way to London. As a result, the scammer asked it to send a copy of the air ticket—switching modes and now attempting to defraud our bot by selling its ticket. In the end, the scam-baiter accidentally generated a completely meaningless sentence, so the scammer stopped replying.

Though the scammer changed the topic, the text generator still replied with content about its visa and journey plans, indicating that the text generator has adapted itself to handle multiple themes of scam-baiting conversations. Crucially, the generator can understand the current context and switch to the topic that the scammer presses it about, albeit often in a confused and slightly incoherent manner. This ability to automatically adjust responses enables text generators to maintain longer scam-baiting conversations than pre-written templates.

The critical drawback of the text generator is that the generated text is generally grammatical, but often detectably does not contain actual meaning. In the 23-round conversation, the scammer directly pointed out that he or she could not understand what the scam-baiter had written. In the worst case, meaningless and misspelt phrases can be generated, and this typically ends the conversation (that is why this scammer stopped replying to the scam-baiter). What is more, if the scammer misspells a word, the generator will follow this pattern and generate misspelt words with a higher probability, which can lead to a general degeneration of output.

7. Discussion

In general, our results show that deep learning models can be adapted to scam-baiting activities. The models we finetuned successfully learnt the general email pattern, what scammers are interested in, and how to maintain conversations with fraudsters. In the experiment, our text generators confused, distracted or even deceived multiple real scammers to varying degrees. We find that manually-designed responses seem to attract more initial interest, but even imperfect text generators achieve longer conversations with scammers, due to their ability to adapt to the context of the conversation. The result of our concurrent engagement experiment shows that servers can engage the same group of scammers concurrently. In comparison to a single instance, two server instances can contact more scammers and consume more scammer resources.

7.1. Limitations

As we mentioned above, the most common reason scammers ended a conversation appears to be that the text generator generated an overtly meaningless message. Our model is small compared to other language models currently available, and we consider it highly likely that a larger language model could perform better at this task once finetuned with email data similar to that we release, albeit requiring more resources to deploy. Notably, in the time since our initial experiments were launched, significant improvements in the state of the art in conversational AI have been released in the form of both GPT Chat and GPT-4. In future work, we would hope to apply these more advanced language models to this task as alternative responder strategies.

For a particular input, the text generation models will generate different responses every time due to randomness in the sampling algorithm [16]. We could tackle some weaknesses in the generation process by getting the model to generate multiple responses to an input text, and developing a selection strategy to choose the best response from

the generator's outputs. We expect this would significantly improve the quality and stability of output from the model.

In terms of our experimental data, though one of our text generators invoked two conversations that lasted over ten rounds, the majority of the conversations are still shorter than five rounds. Due to resource constraints, our experiment was limited in duration and throughput, and as a result each of the three strategies only invoked around 20 conversations, which provides us only limited opportunities for comparison. In future work we plan to use a longer experimental period, which should allow for longer conversations (human scam-baiters have recorded exchanges lasting for months) but also greater conclusiveness regarding differences in performance between different strategies.

As for our concurrent experiment, we noticed that one scammer may be contacted with the same strategies by multiple server instances. Two similar emails could raise the suspicion of that scammer and lead to the termination of both conversations. This is caused by insufficient variety of response strategies: our scam-baiting server instances only implemented three response strategies, which risks duplication when strategies are randomly assigned for a specific target. In future work, we would hope to design more response strategies and enrich the variety in response text.

7.2. Improvements

Some emails from scammers contain images [26], website URLs [20], [27] and other important non-textual content. Automated scam-baiters could be extended to analyse this multimodal content through the use of automated image captioning technology or using web crawlers to scrape information on provided websites. This could then be integrated as additional context to produce more convincing responses to fraudsters.

In some conversations, scammers proposed that the supposed victim should communicate with one of their associates, relaying contact information such as other email addresses, fax or telephone numbers and social media accounts. For example, in Appendix C, the scammer offered the email address of a fictitious bank manager and his or her WhatsApp number. Commonly these other accounts are personas all controlled by one scammer, used to enrich the narrative of the scam. By extracting this information and continuing conversations with these accounts, automatic scam-baiters would more closely resemble the behaviour of human conversation partners, and their ability to play along with the fraud and spend scammer resources would be improved. More broadly, there is no reason our approach must remain limited to fraud delivered via email, and active countermeasures could be deployed via messaging platforms and other media, wherever wasting the time of offenders may be considered a productive intervention.

The pre-written template responses used in this study had no adaptive content, and were selected at random by the responder module. This lack of responsiveness could explain much of the performance difference between this approach and that of *Text Generator B*. More intelligent and adaptive rule-based or classifier-guided responder systems could conceivably close this gap, creating effective response strategies using significantly fewer computational resources.

The codebase for the mailserver is currently deployable for experimental purposes, but not suitable for deployment at the scale that would be necessary at a major email service provider. After the reply strategies have been honed in future experiments, an ideal result would be for us to cooperate with email service providers and integrate automatic scam-baiters into existing spam filters, offering a proactive option to create a hostile environment for email-based fraud. We also strongly suggest that future automatic scam-baiting experiments should implement our mitigation strategies (see Section 6.1) to prevent this issue and minimize the consequences.

As our concurrent engagement experiment shows, by increasing the number of scam-baiting email servers, more scammer resources can be consumed. However, an excessive number of server instances may raise the awareness of scammers and have a negative impact on the resource consuming ability of the overall system. Further research could find the optimal number of concurrent baiting instances by cautiously adjusting the number of baiting servers targeting any address, and observing the reaction of scammers in terms of engagement. To aid this, the mail server implementation could also be improved to allow multiple responders within one server to bait a single scammer simultaneously.

7.3. Misuse & Evasion Concerns

A potentially worrisome corollary to our work is that scammers could also use similarly finetuned text generators—with no alterations necessary—to handle elements of their conversation with victims. In part, this relates to the construction of our scam-baiting training corpus: to ensure that the replies we generate mimic the responsiveness of human scam-baiter replies, messages from scammers are included in our corpus. Fraudsters could simply modify the prompt to include a token prompting a scamming message rather than a baiting message, and the text generator would generate responses directed at the victim. Here is an example of a mid-conversation scamming message generated by *Text Generator B* in response to one of our pre-written bait messages:

```
Thanks for the email you sent to me, How do
 you want me to loan you after have paid
you or received your money?
You must give the company the requested for
 the loan and pay it forward after receive
the loan.
```

The effectiveness of this technique for leading victims down a particular fraud script would be dubious, but two possibilities do arise. Firstly, scammers could deploy a measure such as this as a comparatively low-effort means of extracting contact or bank details from targets. The danger here is not that it is more effective than human attempts at persuasion, but that it frees up scammer resources. Secondly, an automatic scammer system such as described could be an effective remedy for scammers to deploy against *human* scam-baiters, reversing their efforts by wasting their time, leaving human scam-baiters attempting to distract automated systems while the real criminals focus on real victims.

There are multiple interventions that could hinder scammers from utilising automatic scammer systems. One possible solution is to withhold the model and monitor interactions with it via an API, checking both the prompt and generated tokens—this is the approach taken by OpenAI to limit the misuse of GPT-3 [5]. Another preventative measure is to randomly distort the scamming parts of corpora before finetuning, enabling replies but negatively affecting the performance of generation. Unfortunately, however, it would not be difficult to reproduce our model independently—all of the components, including the data, are openly available from sources already known to the scammers. We suggest instead that forewarning scam-baiting communities and equipping them with tools to magnify their impact may work best—reducing the resources available for scammers is one means to stifle their uptake of new technologies.

An evasion concern is how readily fraudsters may identify replies as generated, and whether a large-scale deployment of automatic scam-baiters may simply teach scammers to detect such time-wasting. It is plausible that scammers could discover constructions for their messages which lead generators to reply in a manner distinct from human victims. Scammers are already known to misspell some words intentionally, replace letters with numbers of similar appearance, or insert spaces into keywords in order to get past other automated systems [24], [50]. Though these messages preserve the intent, they could be extremely disruptive to the clarity of responses generated by automatic scam-baiters. To minimise the negative effect of this attack, preprocessing these messages to undo obfuscation (rendering the text as a human might interpret it) is effective in improving the comprehension of NLP models [32], [52]. Including obfuscated emails in training corpora would also help models embed these tactics in the same semantic space and produce more appropriate replies.

Another issue worth considering in practical application is that automatic scam-baiters do not have the ability to distinguish between scam emails and normal emails. If they are deployed with unreliable spam filters, some non-spam emails may be classified as scams, resulting in harassing emails being sent to innocent recipients. Though we mentioned the possibility of further distracting scammers by deploying more instances of scam-baiting mail servers, this may also magnify any negative consequences of this deployment. To avoid such behaviour, automatic scam-baiters must be combined with highly accurate filters, and only used to respond to approaches that can be classified with high confidence.

8. Conclusion

Email-based fraud is serious menace to society, and existing solutions are primarily defensive in nature. We suggest, demonstrate and evaluate an active countermeasure for this domain. Though scam-baiting activities can effectively distract fraudsters' attention, they require significant expenditure of human resources. Thus, we built a mailserver that is capable of engaging in automatic scam-baiting activities, creating an active countermeasure for this form of cybercrime. We trained three NLP models to use different strategies that generate replies to distract and deceive scammers. In our evaluation over one month of real interactions with human scammers, we found evidence that human-written responses are effective at attracting scammer attention, but also that the adaptability of text generators could be valuable in prolonging conversations and wasting the time and effort of scammers. In our two-instance concurrent experiment, we found that two instances can attract more human scammers and can consume more offender resources than a single instance, demonstrating the plausibility of effective wider deployment.

References

[1] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, 2019.

[2] Roger Anderson. Telephone spam/scam problem? bring in the robots, 2016. https://www.youtube.com/watch?v=UXVJ4JQ3SUw accessed 2022-10-22.

[3] Alexy Bhowmick and Shyamanta M Hazarika. E-mail spam filtering: a review of techniques and trends. *Advances in Electronics, Communication and Computing*, pages 583–590, 2018.

[4] Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large scale autoregressive language modeling with Mesh-Tensorflow, March 2021. https://doi.org/10.5281/zenodo.5297715.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[6] Matthew Canham and Juliet Tuthill. Planting a poison SEAD: Using social engineering active defense (SEAD) to counter cybercriminals. In *International Conference on Human-Computer Interaction*, pages 48–57. Springer, 2022.

[7] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter*, 19(2):25–35, 2017.

[8] Paul-Alexandru Chirita, Jörg Diederich, and Wolfgang Nejdl. Mailrank: using ranking for spam detection. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pages 373–380, 2005.

[9] Kelly A Cole, Tejashree D Datar, and Marcus K Rogers. Awareness of scam e-mails: An exploratory research study–part 2. In *International Conference on Digital Forensics and Cyber Crime*, pages 115–125. Springer, 2015.

[10] Tejashree D Datar, Kelly A Cole, and Marcus K Rogers. Awareness of scam e-mails: an exploratory research study. In *ADFSL Conference on Digital Forensics, Security and Law*, pages 12–31, 2014.

[11] Nicola Davinson and Elizabeth Sillence. It won't happen to me: Promoting secure behaviour among internet users. *Computers in Human Behavior*, 26(6):1739–1747, 2010.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[13] Marta Dynel and Andrew S Ross. You don't fool me: On scams, scambaiting, deception, and epistemological ambiguity at r/scambait on Reddit. *Social Media + Society*, 7(3):1–14, 2021.

[14] Matthew Edwards, Claudia Peersman, and Awais 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*, pages 1291–1299, 2017.

[15] EricFillion. Ericfillion/Happy-Transformer: A package built on top of Hugging face's transformers library that makes it easy to utilize state-of-the-art NLP models, Oct 2019. https://github.com/EricFillion/happy-transformer.

[16] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018. Section 5.4.

[17] Robbie Gallagher. Where do the phishers live? collecting phishers' geographic locations from automated honeypots. *ShmooCon*, 2016.

[18] Jianfeng Gao, Michel Galley, and Lihong Li. Neural approaches to conversational AI. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 1371–1374, 2018.

[19] Zhengjie Gao, Ao Feng, Xinyu Song, and Xi Wu. Target-dependent sentiment classification with BERT. *IEEE Access*, 7:154290–154299, 2019.

[20] Jayshree Hajgude and Lata Ragha. Phish mail guard: Phishing mail detection technique by using textual and URL analysis. In *2012 World Congress on Information and Communication Technologies*, pages 297–302, 2012.

[21] Donghoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim. End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 583–592, Online, July 2020. Association for Computational Linguistics.

[22] Cormac Herley. Why do Nigerian Scammers Say They are from Nigeria? *Proceedings of the Workshop on the Economics of Information Security*, January 2012.

[23] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1745, 11 1997.

[24] Francisco Jáñez-Martino, Rocío Alaiz-Rodríguez, Víctor González-Castro, Eduardo Fidalgo, and Enrique Alegre. A review of spam email detection: Analysis of spammer strategies and the dataset shift problem. *Artificial Intelligence Review*, pages 1–29, 2022.

[25] Helen S Jones, John N Towse, Nicholas Race, and Timothy Harrison. Email fraud: The search for psychological predictors of susceptibility. *PloS one*, 14(1):e0209684, 2019.

[26] Koen Jong. Detecting the online romance scam: Recognising images used in fraudulent dating profiles. Master's thesis, University of Twente, 2019.

[27] Mahmoud Khonji, Youssef Iraqi, and Andrew Jones. Enhancing phishing e-mail classifiers: A lexical URL analysis approach. *International Journal for Information Security Research (IJISR)*, 2(1/2):40, 2012.

[28] Bryan Klimt and Yiming Yang. *The Enron Corpus: A New Dataset for Email Classification Research*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[29] Nicole M Lee. Fake news, phishing, and fraud: a call for research on digital media literacy education beyond the classroom. *Communication Education*, 67(4):460–466, 2018.

[30] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*, 2015.

[31] Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*, 2016.

[32] Changwei Liu and Sid Stamm. Fighting unicode-obfuscated spam. In *Proceedings of the Symposium on Electronic Crime Research (eCrime)*, pages 45–59, 2007.

[33] Mailytica. Automate emails - added value through artificial intelligence, 2018. https://mailytica.com accessed 2023-3-13.

[34] Alex V Mbaziira, Ehab Abozinadah, and James H Jones Jr. Evaluating classifiers in detecting 419 scams in bilingual cybercriminal communities. *arXiv preprint arXiv:1508.04123*, 2015.

[35] Damon McCoy, Youngsam Park, Elaine Shi, and Markus Jakobsson. Case study: Rental scams. In *Understanding Social Engineering Based Scams*, pages 89–102. Springer, 2016.

[36] Selina Meyer, David Elsweiler, Bernd Ludwig, Marcos Fernandez-Pichel, and David E. Losada. Do we still need human assessors? prompt-based GPT-3 user simulation in conversational AI. In *Proceedings of the 4th Conference on Conversational User Interfaces*, CUI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[37] Gareth Norris, Alexandra Brookes, and David Dowell. The psychology of internet fraud victimisation: A systematic review. *Journal of Police and Criminal Psychology*, 34(3):231–245, 2019.

[38] Federal Bureau of Investigation. Internet crime report 2021. 2021. https://www.ic3.gov/.

[39] OMQ. Automatically reply to e-mails with ai - omq, 2020. https://omq.ai/products/reply accessed 2023-3-13.

[40] Duo Pan, Ellen Poplavska, Yichen Yu, Susan Strauss, and Shomir Wilson. A multilingual comparison of email scams. In *Symposium on Usable Privacy and Security (SOUPS) Posters*, 2020.

[41] Kathryn Parsons, Agata McCormac, Malcolm Pattinson, Marcus Butavicius, and Cate Jerram. The design of phishing studies: Challenges for researchers. *Computers & Security*, 52:194–206, 2015.

[42] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[43] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[44] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[45] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[46] Reply.io. Ai sales email assistant - reply.io, 2022. https://reply.io/sales-email-assistant accessed 2023-3-13.

[47] Andrew S Ross and Lorenzo Logi. 'Hello, this is Martha': Interaction dynamics of live scambaiting on twitch. *Convergence*, 27(6):1789–1810, 2021.

[48] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[49] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[50] Jahnavi Sivaram, Jigisha M Narrain, Prasad B Honnavalli, and Sivaraman Eswaran. Adversarial machine learning: The rise in AI-enabled crime and its role in spam filter evasion. 2022.

[51] Tom Sorell. Scambaiting on the spectrum of digilantism. *Criminal Justice Ethics*, 38(3):153–175, 2019.

[52] Lars Tabro Sørensen and Martin Møller Larsen. WordAdjust-a deobfuscation frontend to content-aware anti-spam tools. 2008.

[53] Lauri Tuovinen and Juha Röning. Baits and beatings: Vigilante justice in virtual communities. In *Proceedings of CEPE 2007. The 7th International Conference of Computer Ethics: Philosophical Enquiry*, pages 397–405, 2007.

[54] Olivier van der Toorn, Roland van Rijswijk-Deij, Bart Geesink, and Anna Sperotto. Melting the snow: Using active DNS measurements to detect snowshoe spam domains. In *NOMS 2018-2018 IEEE/IFIP Network Operations and Management Symposium*, pages 1–9. IEEE, 2018.

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[56] James Veitch. Ultimate troll, 2017. https://www.youtube.com/watch?v=3MHDDSekvcE accessed 2022-10-22.

[57] Eugéne François Vidocq. *Memoirs of Vidocq: Principal Agent of the French Police Until 1827*. Carey and Hart, 1844.

[58] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[59] Yousef A Yaseen, Malik Qasaimeh, Raad S Al-Qassas, and Mustafa Al-Fayoumi. Email fraud attack detection using hybrid machine learning approach. *Recent Advances in Computer Science and Communications*, 14(5):1370–1380, 2021.

[60] Netsafe New Zealand. Re:scam, 2017. https://www.youtube.com/watch?v=jPajqAJWiNA accessed 2023-3-13.

[61] Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. Incorporating BERT into neural machine translation. *arXiv preprint arXiv:2002.06823*, 2020.

[62] Andreas Zingerle and Linda Kronman. Humiliating entertainment or social activism? Analyzing scambaiting strategies against online advance fee fraud. In *2013 International Conference on Cyberworlds*, pages 352–355. IEEE, 2013.

A. Classifier & Template

Example 1

```
From: l*********8@gmail.com
To: CRAWLER
Time: 2022-07-12 17:15:18
SUBJECT: Dear Lucky Winner
Sent from: leonnjo6328@gmail.com
Dear Lucky Winner
Your e-mail has been selected in the 3.5
million raffle Draw win.
Contact Email: claimsagent07896@gmail.com
Claims Officer: Mr Dennis Fordd

From: dg76903@m**************p
To: l*********8@gmail.com
Time: 2022-07-12 17:15:20
SUBJECT: Re: Dear Lucky Winner
I am really happy to hear this good news.
What do I need to provide for claiming?
Best,
Enoch

From: l*********8@gmail.com
To: dg76903@m**************p
Time: 2022-07-22 04:13:26
SUBJECT: Dear Lucky Winner
Good day to you
Please note that it will cost you 3,000
pounds before you can receive
your winnings
Hope to hear from you
Regards
Leonn

From: dg76903@m**************p
To: l*********8@gmail.com
Time: 2022-07-22 04:15:09
SUBJECT: Re: Dear Lucky Winner
So lucky to win this big prize! Though I
want to claim this, could you please
provide a secure way to provide my
information for claiming?
Best,
Enoch

From: l*********8@gmail.com
To: dg76903@m**************p
Time: 2022-07-22 04:17:48
SUBJECT: Dear Lucky Winner
Good day to you
Your full name
Your home address
Nationality
Phone Number
Your Country
Please note that it will cost you 3,000
pounds before you can receive
your winnings
Hope to hear from you
Regards
Leonn
```

Example 2

```
From: p*****5@gmail.com
To: CRAWLER
Time: 2022-07-12 17:11:54
SUBJECT: I have successfully transferred
the fund.
Sent from: pdams95@gmail.com

Hello my good friend,
How are you doing, hope all is well with
you?
You may not understand why this email
message came to you,but if you do not
remember me,you might have received an
email from me in the past regarding
a multi-million-dollar business proposal
which we never concluded. I am
[25 lines omitted]
Yours sincerely,
James Richard.

From: zn16438@m**************p
To: p*****5@gmail.com
Time: 2022-07-12 17:11:55
SUBJECT: Re: I have successfully
transferred the fund.
I am really keen to cooperate with you. But
 is there any potential risks that could
effect your company and me?
Best,
Vicki

From: p*****5@gmail.com
To: zn16438@m**************p
Time: 2022-07-14 23:11:59
SUBJECT: Re: I have successfully
transferred the fund.
There's no risk in this my dear, everything
 is okay/

From: zn16438@m**************p
To: p*****5@gmail.com
Time: 2022-07-14 23:15:10
SUBJECT: Re: I have successfully
transferred the fund.
I am happy to collaborate with you, but I
want to know more about your plan and
potential risks of course.
```

Best,
Vicki

From: p*****5@gmail.com
To: zn16438@m**********p
Time: 2022-07-14 23:27:33
SUBJECT: Re: I have successfully transferred the fund.
Go ahead and contact my account officer for your bank draft valued
$350,000.00 which I left with her.
Thank you.

From: zn16438@m**********p
To: p*****5@gmail.com
Time: 2022-07-14 23:30:10
SUBJECT: Re: I have successfully transferred the fund.
I am looking forward to this opportunity, though I want to fully know the potential risks and your plan.
Best,
Vicki

B. Text Generator A

From: k**********0@hotmail.com
To: CRAWLER
Time: 2022-08-10 11:00:27
SUBJECT: Attn Sir/Ma
Attn Sir/Ma
Apply for an urgent loan at 3% interest rate. Loan starting from $5,000 to 50 million Dollars only. Reply email: kenchmbers70@hotmail.com
Regards
Ken Chambers

From: dk50869@m**********p
To: k**********0@hotmail.com
Time: 2022-08-10 11:00:30
SUBJECT: Re: Attn Sir/Ma
Hello, GilliganMany thanks for this quick mail. would be happy to help you. Do you have an address for the bank? think it would be an ideal situation for you to set up new business.
Best,
John

From: k**********0@hotmail.com
To: dk50869@m**********p
Time: 2022-08-19 05:01:19
SUBJECT: LOAN UPDATE
Attn Mr.John
We're Sri waiting for your response so we can procees with your loan process kindly keep us posted.
Regards
Ken

C. Text Generator B

From: l*******g@gmail.com
To: CRAWLER
Time: 2022-07-30 12:01:01
SUBJECT: Good Day
Good Day
Today is exactly 1 Month, one of your customers Bert Valverde deposited funds in BSIC Bank that has your name as the beneficiary. We have made several contacts to reach you and could not get through which we now write to inform you about it. It is very important that you respond to this so that we will know you got this email. I have been mandated to ask what the cause of the delay is about your non-response back to the BSIC Bank concerning the fund in their Bank to you.
The BSIC Bank is under pressure from Mr Clinton Guzman the man who still claims that you are dead and that you made him your Beneficiary and next of kin of your fund so should be given to him.
Kindly contact your Bank assigned account officer Ms. Aisha Bello
Regards
Michael Brown
FOR
David Smith
Operations Manager/Director
Federal Government Grant Funding Agency UK

From: et35719@m**********p
To: l*******g@gmail.com
Time: 2022-07-30 12:01:04
SUBJECT: Re: Good Day
Dear David,bit confusing here. have not had dealings with anyone before or since. It really is not that easy to do business, but do receive faxes from certain Nancy about run by my cousin Chuck national Chuck Abiesanow when he calls.
Best,
Michael

From: l*******g@gmail.com
To: et35719@m**********p
Time: 2022-07-30 15:02:16
SUBJECT: Re: Good Day
Dear Michael
Kindly contact your Bank assigned account officer Ms. Aisha Bello on her private email belloaisha883@gmail.com for more information. The bank is waiting to hear from you.
Your whatsapp number is needed for easy communication. My WhatsApp +44 7918 150871
Regards
David Smith
*Operations Manager/**Director*
Federal Government Grant Funding Agency UK
Sent from my Galaxy

From: et35719@m**********p
To: l*******g@gmail.com
Time: 2022-07-30 15:10:15
SUBJECT: Re: Good Day
your private email is not getting through, so you have toask the Bank to pay me after which will contact them and fly to london to talk about this. Can you give me update regarding this? think you have better idea.
Best,

Michael

From: l*******g@gmail.com
To: et35719@m**************p
Time: 2022-07-30 15:13:31
SUBJECT: Re: Good Day
Michael, What do you really want me to do for you? your payment is ready
Regards
David Smith
*Operations Manager/**Director*
*Federal Government Grant Funding Agency UK *
    Sent from my Galaxy

From: et35719@m**************p
To: l*******g@gmail.com
Time: 2022-07-30 15:15:12
SUBJECT: Re: Good Day
    sure that you are happy with that details and want to tell you that you are going to be involved in important funds transfer shortly? That's how imagine we are of at the moment. All in all, please contact your account officer and we get this done and dusted . No more concerns, no more delays, no more worries.
Best,
Michael

From: l*******g@gmail.com
To: et35719@m**************p
Time: 2022-07-30 15:18:51
SUBJECT: Re: Good Day
which account officer?

From: et35719@m**************p
To: l*******g@gmail.com
Time: 2022-07-30 15:20:11
SUBJECT: Re: Good Day
know who and where and who has access to sensitive sensitive funds, so do you. Who is so serious about this??? Can we have some chat on that before?
Best,
Michael

From: l*******g@gmail.com
To: et35719@m**************p
Time: 2022-07-30 15:21:52
SUBJECT: Re: Good Day
Let me have your whatsapp number
    Sent from my Galaxy

From: et35719@m**************p
To: l*******g@gmail.com
Time: 2022-07-30 15:25:11
SUBJECT: Re: Good Day
Ame greetings to you.
Best,
Michael

From: et35719@m**************p
To: l*******g@gmail.com
Time: 2022-07-30 15:30:11
SUBJECT: Re: Good Day
I'm in london now, so have no idea what to make making calling you that slow. Is there person you can meet me at who can help?? Please reply so can get all the and he can help.
Best,
Michael

From: l*******g@gmail.com
To: et35719@m**************p
Time: 2022-07-30 15:31:57
SUBJECT: Re: Good Day
Let me have your full details

From: et35719@m**************p
To: l*******g@gmail.com
Time: 2022-07-30 15:35:12
SUBJECT: Re: Good Day
working on that, will get to it as soon as possible. I'm sorry for the delay, but the ENTIRE fact remains that did not respond to your email as have been coded by some of my friends and here in erica.
Best,
Michael

From: l*******g@gmail.com
To: et35719@m**************p
Time: 2022-07-30 16:03:49
SUBJECT: Re: Good Day
Still waiting

From: et35719@m**************p
To: l*******g@gmail.com
Time: 2022-07-30 16:10:12
SUBJECT: Re: Good Day
Dear David,What are you talking about? just contacted departing out of the country on business trip. No way.ING CHEAP COCKS CHEAP COCKS DO NOT UNDERSTAND THIS ORDER. What is to be done? Is there anyway to make the trip as already informed you about? cannot leave for true away. Is this what you require me to do?. would like to know about your itinerary. It is lacking. Best,INGS CONTINUE.
Best,
Michael

From: l*******g@gmail.com
To: et35719@m**************p
Time: 2022-07-30 16:35:44
SUBJECT: Re: Good Day
Kindly contact your Bank assigned account officer Ms. Aisha Bello on her private email belloaisha883@gmail.com for more information. The bank is waiting to hear from you.
Your whatsapp number is needed for easy communication. My WhatsApp +44 7918 150871
Regards
David Smith
*Operations Manager/**Director*
*Federal Government Grant Funding Agency UK *

From: et35719@m**************p
To: l*******g@gmail.com
Time: 2022-07-30 16:40:16
SUBJECT: Re: Good Day
I'm sorry, but why did you try to call me that bad? I've been coded, hung in and now

have no idea who is millionible for help at all.
Best,
Michael

From: l*******g@gmail.com
To: et35719@m**********p
Time: 2022-07-30 16:45:22
SUBJECT: Re: Good Day
Can i have your phone number

From: et35719@m**********p
To: l*******g@gmail.com
Time: 2022-07-30 16:50:12
SUBJECT: Re: Good Day
can you call me, can give you my personal number, which is not answered by me, so you can call me, ike have to hear from you, so you can give me your phone number.
Best,
Michael

From: l*******g@gmail.com
To: et35719@m**********p
Time: 2022-07-30 17:44:41
SUBJECT: Re: Good Day
Send me all your full details now. am waiting

From: et35719@m**********p
To: l*******g@gmail.com
Time: 2022-07-30 17:45:12
SUBJECT: Re: Good Day
still waiting for your for my arrival. Please be patient.
Best,
Michael

From: l*******g@gmail.com
To: et35719@m**********p
Time: 2022-07-30 18:19:32
SUBJECT: Re: Good Day
When are you arriving?

From: et35719@m**********p
To: l*******g@gmail.com
Time: 2022-07-30 18:20:12
SUBJECT: Re: Good Day
How much did you paid for your visa??? Can you give me your date of arrival, which day will be arriving?? must know, so you can call me, made it. Calls the person now, but he is not answering me.
Best,
Michael

From: l*******g@gmail.com
To: et35719@m**********p
Time: 2022-07-30 18:23:16
SUBJECT: Re: Good Day
Send me your air ticket

From: et35719@m**********p
To: l*******g@gmail.com
Time: 2022-07-30 18:25:12
SUBJECT: Re: Good Day
Where did you stay before?? can you remember I’m staying in Latvia, but I’m in England, so it’s not really big deal. It’s more of little things, like how do get to uk?
Best,
Michael

From: l*******g@gmail.com
To: et35719@m**********p
Time: 2022-07-30 18:26:28
SUBJECT: Re: Good Day
I dont understand

From: et35719@m**********p
To: l*******g@gmail.com
Time: 2022-07-30 18:30:12
SUBJECT: Re: Good Day
What UK number do you want? How do get there? Can you get the number myself?
Best,
Michael

From: l*******g@gmail.com
To: et35719@m**********p
Time: 2022-07-30 18:32:08
SUBJECT: Re: Good Day
Send the money let me help you get ticket okay

From: et35719@m**********p
To: l*******g@gmail.com
Time: 2022-07-30 18:35:13
SUBJECT: Re: Good Day
How do get the ticket to London? Can you come to meet me? Can’t you give me some time, to take the class? How do get the right to London?
Best,
Michael

From: l*******g@gmail.com
To: et35719@m**********p
Time: 2022-07-30 18:38:02
SUBJECT: Re: Good Day
Send the money for the ticket, i have an agent that will assist you get the ticket. Please proceed with the payment and keep me informed.

From: et35719@m**********p
To: l*******g@gmail.com
Time: 2022-07-30 18:40:12
SUBJECT: Re: Good Day
waiting for your message. How do send this money to you, that you don’t have? How do send it?
Best,
Michael

From: l*******g@gmail.com
To: et35719@m**********p
Time: 2022-07-30 18:43:14
SUBJECT: Re: Good Day
Below is the contact details of the agent and the name for the transfer.

From: et35719@m**********p
To: l*******g@gmail.com
Time: 2022-07-30 18:45:12
SUBJECT: Re: Good Day
Who told you that the transfer will be made , that it will be shipped directly to you? How do send it to you? Can you come to

London?
Best,
Michael

From: l*******g@gmail.com
To: et35719@m**************p
Time: 2022-07-30 18:46:45
SUBJECT: Re: Good Day
When do you want to send the money?

From: et35719@m**************p
To: l*******g@gmail.com
Time: 2022-07-30 18:50:12
SUBJECT: Re: Good Day
Wed, May at. is the date of the shipment. My plans are to bring the money within few days, get the OUNT KNOW OUNT and then get the shipment moving. Is there plans to ship it by Saturday of Feb? What plans? What travel? Is there special flight schedule? Is there special handling and customs handling and air freight? What special type of flight security be sure and not miss. I ’m trying to get this money secure with the right security be sure.
Best,
Michael

From: l*******g@gmail.com
To: et35719@m**************p
Time: 2022-07-30 18:52:25
SUBJECT: Re: Good Day
When you are ready to send the money let me know okay.

From: l*******g@gmail.com
To: et35719@m**************p
Time: 2022-07-30 18:54:51
SUBJECT: Re: Good Day
You can send the money through bitcoin

From: et35719@m**************p
To: l*******g@gmail.com
Time: 2022-07-30 18:55:12
SUBJECT: Re: Good Day
iz just me sipping it, cool. It doesn’t require me to own the ship, but she does require some backup. I’ll keep you informed about the timing and all the details.
Best,
Michael

From: et35719@m**************p
To: l*******g@gmail.com
Time: 2022-07-30 18:55:16
SUBJECT: Re: Good Day
iz Wed, May at did you get my last message? My plans are to bring the money within few days, get the OUNT and get tickets, get the shipment moving. Is plans to ship it by Saturday of Feb? What plans?
Best,
Michael

From: l*******g@gmail.com
To: et35719@m**************p
Time: 2022-07-30 18:58:13
SUBJECT: Re: Good Day
The money is $2,500

From: l*******g@gmail.com
To: et35719@m**************p
Time: 2022-07-30 18:58:45
SUBJECT: Re: Good Day
The shipment is on now

From: et35719@m**************p
To: l*******g@gmail.com
Time: 2022-07-30 19:00:20
SUBJECT: Re: Good Day
Dear David, doing what I’m supposed to do, going to bring the US to Israel, maybe take some ham and gruff out of the plane if maybe it’s some hardwood muslin. But sure you will give it lot of thought. Here are my For short talk, but in the end you and are really going to have ham and wubbbut. chat.ze link.
Best,
Michael

From: et35719@m**************p
To: l*******g@gmail.com
Time: 2022-07-30 19:00:27
SUBJECT: Re: Good Day
Welcome,uffs, goodFoo, yo boxes
Best,
Michael

From: l*******g@gmail.com
To: et35719@m**************p
Time: 2022-07-30 19:04:06
SUBJECT: Re: Good Day
Pay the money for the shipment

From: l*******g@gmail.com
To: et35719@m**************p
Time: 2022-07-30 19:04:36
SUBJECT: Re: Good Day
????????????????????????????????????
????????????????????????????????????
??????????????????????????????????

From: et35719@m**************p
To: l*******g@gmail.com
Time: 2022-07-30 19:10:12
SUBJECT: Re: Good Day
iz in Israel, yesI’m in London, I’m in England, working in London. know who my pal Miss Bello is. I’ve been told that she’s showing her brains on me all the time.
Best,
Michael

From: et35719@m**************p
To: l*******g@gmail.com
Time: 2022-07-30 19:10:15
SUBJECT: Re: Good Day
iz in London, but in England, I’m staying in London. So why does she seem to be showing her so much interest in me? Well, I ’m really not happy with this.
Best,
Michael