Mimicking Human Responses in Spam Email Conversations with Natural Language Understanding Techniques

Gautam Mittal[1] and Melanie Tory[2]

***Abstract*— The convergence of artificially intelligent technologies and the evolving field of cybersecurity has become more apparent than it has ever been, especially with the rise of language-aware systems. Email spam has become an increasingly popular medium for attackers attempting to carry out malicious threats such as identity theft and other forms of robbery. In an attempt to mitigate this problem, language-aware systems that utilize natural language processing techniques can be used to deceive attackers and waste their time rather than having them steal from unsuspecting Internet users. Building and tweaking language-aware systems utilizing LSTM (Long Short Term Memory) neural networks, or mathematical models of large amounts of context-driven data, can be used to analyze how both attackers and users talk in written conversation, effectively allowing them to mimic human-like conversational responses. The research of this project will investigate methods for building and training a context-aware language-based network using natural language processing techniques for formulating conversational responses in context with spam emails.**

I. BACKGROUND & SIGNIFICANCE

Gleaning information from text and generating meaningful information from it has been the focus of many large technology and research firms for the past decade. Natural language processing (NLP), a branch under the field of machine learning, has been incredibly useful in solving complicated problems regarding written and spoken language. NLP techniques are omnipresent in modern day chatbots, web search, email spam detection, and mobile keyboard word suggestions.

In more recent years, the rise of neural networks, or algorithms that are modeled off of the theoretical workings of the human brain, have played a major role in machine learning and natural language processing advances. Their sophisticated algorithmic structure has allowed researchers to model more complex data and extract more detailed and hidden patterns within data of all kinds. Today, they have become a tool of quintessential importance in almost every aspect of machine learning and natural language processing.

With more sophisticated algorithms comes more sophisticated problems that can be solved. One of the earliest applications of machine learning was classifying email as spam or not spam. This was one of the first fully machine-automated tasks that previously only humans could accomplish.

In a TED talk delivered in February 2016, James Veitch, a British writer most well-known for his work regarding the understanding of spam feedback loops, explained his take on replying to spam emails. He found that it was possible to keep a conversation with a spammer through broken, looping, conversations. With the help of machine learning, automating conversation responses would be trivial to automate. Currently the most common approach is to use deep (multilayer) long-short term memory (LSTM) neural networks, as they are especially flexible with modeling sequential data found in most conversations. Veitch managed to find that by maintaining conversations with spammers, he was making a dent in the spammer economy by wasting their time and preventing them from making further attacks on those more susceptible to fraud or identity theft.

With todays technology and algorithms, it is possible to automate these messages using natural language understanding techniques. Exploring this area would also allow machines to extrapolate based on spammers past actions, and provide more robust spam filtering systems that not only classify spam emails, but distract the attacker with human-like responses.

II. RESEARCH METHODOLOGIES

The research will be a hybrid study with both qualitative and quantitative data. There will be two primary aspects of the project, one being experimental and the other being observational.

The first stage will involve building and training a mathematical model to extract patterns from a large amount of textual data and understand the conversational aspects to become language aware. This will be experimental as there will be a lot of fine-tuning of the model and adjusting of the hyperparameters that will be used to generate more accurate outputs. The goal is not to recognize whether or not an email is spam, but to generate a context-aware response given a query in any setting, not just limited to spam conversations. Using an LSTM (Long Short Term Memory) neural network, an algorithm for making sequence-based predictions, gleaning context and formulating responses can be achieved using a variety of learning and mathematical techniques. These networks are optimized for context-driven data because of their unique, sequence-based architecture. The model will be trained using the standard data set for training conversation models: the Cornell Movie Dialog data set. Once the model is trained, the accuracy can be assessed

[1]G. Mittal is a student at Henry M. Gunn High School in the Advanced Authentic Research Program; Gunn High School, 780 Arastradero Rd, Palo Alto, CA, United States gautam@mittal.net

[2]M. Tory is a Senior Research Scientist at Tableau Software; Tableau Softare, 260 California Ave, Palo Alto, CA, United States

*This work was supported by the Palo Alto Unified School District's Advanced Authentic Research Program

*Research source code is available at https://github.com/gmittal/aar-2017

*Citation: Stanford AIR Research Journal, Vol. I. 2017.

using general statistical techniques, such as mean squared error, against a dataset of testing examples not found in the original training dataset. Extensive testing of the neural network in an offline repository will take place prior to the second phase of the project.

The second phase will be observational and will include deploying the project on a live email server and having randomly sampled attackers interact with the language-aware system. Part of this process will include acquiring a spammer audience under a pseudonymous email address and engaging in conversations with spammers. This process will not only attract spammers for the system to converse with, but will provide key insight into how the conversation model should react to spam queries in addition to providing additional test data to better fine-tune the parameters of the model. The spammers will then be classified by the kind of attack that was attempted and observations regarding the interactions between the spammer and the neural network will be ranked accordingly.

Conversational data will be collected as spammer interactions occur and will be logged to a private database. As the data is received, a computer program will redact all sensitive information such as physical addresses, email addresses, names, etc. The conversations will then be sorted by their length, duration, and the number of responses exchanged by both parties. Responses will be determined significant based on model performance and observations regarding the content of the conversation.

The timeline for the research runs as follows: The final corpora, type of model, machine learning framework and tools will be chosen and finalized by January 1, 2017. The implementation, fine-tuning of parameters, and model training should be completed by March 4, 2017. The extensive local testing of the models responses and sampling accuracy in an offline repository should be completed no later than April 4, 2017. The deployment of the algorithm and observations regarding spammer interactions shall run from April 5, 2017 to May 13, 2017.

III. DATA ANALYSIS & RESULTS

Based on spam email conversations that took place under a pseudonymous email address between January 2017 and May 2017, the primary intent of most attackers is to acquire ones monetary assets in some form or another. There are three kinds of spam email messages have been identified as a result of the data collection processes: monetary offers or rewards, requests for monetary assistance or investment partners, and disguised transactions from both domestic and international banking organizations. Although these three categories of scam emails are well known among the many different kinds of cyber attacks, it is important to note that for the scope of this research, the computer program to automate the process of replying to spammers should be suited for these three primary types of attacks. Examples of each kind of spam email attack is displayed in Fig. 1.

As shown by all three examples, and the raw data collected had many more messages like this, the end goal of the attacker is to get a users bank or other financial credentials. However, in most scenarios, attackers do not receive an email response. This causes spammers to generally react quickly when a response is received, as that brings spammers a little closer to their goal of acquiring someones financial assets. The goal of writing a computer program that interacts with spammers is to keep the spammer engaged in a running conversation while never allowing them to reach their goal of attaining assets. This in essence exploits much of the spammers time and prevents them from targeting other users.

Monetary Offers & Rewards	**Personal Transaction & Investment Partners**	**Domestic & International Bank Transactions**
Subject: *"Urgent"* ***Message:*** *"I am [redacted] from [redacted], married to [redacted] who worked with [redacted] in [redacted] for 20 years. Before his death, he deposited the sum of 5.6M Dollars with a bank In [redacted] and this fund is presently with the bank awaiting my disbursement as beneficiary and next of kin to the funds. Recently, my Doctor told me that I would not last for the next Seven months due to Lung cancer Problem. Since I know my condition I decided to donate this fund to a Charity organization or good person that will use this fund to build an orphanage home which was the dream of my late husband.I took this decision because I don't have any child that will inherit this money. I kept this deposit secret till date; this is why I am taking this decision. ... confidentiality of this transfer"*	***Subject:*** *"MY TRUSTEE THIS IS THE ATTACHED STATEMENT OF ACCOUNT OF SAID AMOUNT IN THE UBA BANK"* ***Message:*** *"My trustee so firstly I am here in [redacted] were my father deposit 10.5M Dollars in my name as the beneficiary and I have gone to the [redacted] Bank in [redacted] to clear the deposit but the Branch Manager advised me to appoint a reliable investment partner to handle the transaction on my behalf due to my refugee status does not warrant me to make the claim myself for safety purposes and best utility of the fund I am presently staying in refugee camp under the protection of [redacted] here in [redacted]. So I am emailing you from the office of the coordinator in the refugee Mission Camp, I told the coordinator [redacted] about my Communication with you and he permitted me to access my email in his office computer twice a day ..."*	***Subject:*** *"OFFICIAL QUESTIONNAIRE, FILL IT AND SEND IT BACK TO THIS BANK IMMEDIATELY"* ***Message:*** *"DIRECTOR OF OPERATIONS. [redacted] OFFICIAL NOTICE FOR CLAIM. ATTN; BENEFICIARY [redacted] LETTER OF ACKNOWLEDGEMENT. WE WISH TO NOTIFY YOU THAT THE MANAGEMENT HAD AN EMERGENCY MEETING THIS MORNING CONCERNING YOUR CLAIM ON OUR DECEASED CUSTOMER WHICH YOU HAVE SENT AN APPLICATION LETTER AND THE MANAGEMENT HAS ACKNOWLEDGED YOUR APPLICATION. SEE THE ATTACHED OFFICIAL PROOF OF RELATIONS DOCUMENT FORM AND RESPOND WITHIN 72 HOURS OR FACE CANCELLATION"*

Fig. 1. Examples of spam email attacks.

The computer program that was used for generating conversation responses was built using a deep long-short term memory neural network, an algorithm that analyzes patterns in data that can be modeled using sequences, such as conversations. The program was trained for fifteen hours on Google Compute Engine NVIDIA Tesla K80 hardware on the Cornell Movie-Dialog Corpus, allowing the program to analyze and learn from more examples with greater speed and efficiency. The resulting program communicates effectively for very primitive conversations, and examples using spammer queries (S) derived from original human-spammer email conversations and the computer program output (A) are shown in Table 2.

Monetary Offers & Rewards Conversation	**Personal Investment Partners Conversation**	**Domestic & International Bank Transactions Conversation**
S: *Hello.* ***A:*** *Hello.* ***S:*** *I would like to give $5M.* ***A:*** *Thank you.* ***S:*** *Please send me a copy of your bank statement.* ***A:*** *What do you want?* ***S:*** *A copy of your bank statement.* ***A:*** *Let me tell you something, i'm afraid.* ***S:*** *No need to be afraid.* ***A:*** *Would you like to see my father?* ***S:*** *What is his name?* ***A:*** *Ken.*	***S:*** *Hello, I need to transfer $10M.* ***A:*** *How's the fuss about.* ***S:*** *I am stuck in a refugee camp.* ***A:*** *You're not good at it.* ***S:*** *I need a beneficiary for the bank.* ***A:*** *Can i make a look for that?* ***S:*** *Yes. It would be a great help.* ***A:*** *... you have to wonder, too bad.* ***S:*** *Please.* ***A:*** *I just wanted to see you.* ***S:*** *Then let's move forward with the details.* ***A:*** *I'd like to be alone.*	***S:*** *You have an unfinished bank transaction.* ***A:*** *You think i'm fake, aren't you?* ***S:*** *It needs to be process urgently.* ***A:*** *Or how will i be going to london?* ***S:*** *Your billing info has expired.* ***A:*** *Your father must be a plastic man.* ***S:*** *Do not bring my father into this.* ***A:*** *Oh, but it's worth it.* ***S:*** *Finish the transaction, sir.* ***A:*** *Gonna be a good man.*

Fig. 2. Examples of machine-generated responses.

The machine responds appropriately in most situations,

however the sentence structure and grammar is very primitive. The reasonableness of a response is somewhat difficult to gauge from a computational standpoint, because the network was trained on movie dialogue, and therefore responds to queries in relation to patterns gleaned from the corpus. Based on these results, it can be concluded that the network is able to converse in mostly legitimate English, but the responses are not always appropriate given the query because of specific patterns gleaned from the movie dialogue corpus.

In addition, spammers stopped trying to attack the pseudonymous email address after a period of time, and the final stage of implementation which involved having the computer send real email responses to spam email chains was unsuccessful. However, the findings from building and testing the automated computer model and gathering observations from interactions with spammers was beneficial for future research.

IV. DISCUSSION

The results of this research are of significance to cybersecurity enterprises, primarily because the industry is primarily interested in finding ways to prevent leaked email addresses and credentials from reaching attackers in the first place, however, there are very few solutions to counter spammers directly after they have sent an email to an unsuspecting target user. As a result, observations regarding trends in spammer feedback loops and the viability of an automated conversation system to exploit spammers has utility value in a billion dollar market that is trying to make human-computer interaction safer.

In addition, the use of deep neural networks, an area of machine learning and computer science that is quickly proliferating across many different enterprises, reinforces the effectiveness of these algorithms at carrying out complicated tasks that were only capable by humans. In addition, the data set and observations gleaned allows many researchers in machine learning to easily pick up on this research and improve the algorithm as more computing power becomes available, indicating that as time progresses and computers get faster, the sophistication and ability of the computer program will only increase.

However, there were some limitations in the research approach. The use of the Cornell Movie Dialog data set for training may have affected the results of the data, because the research was primarily concerned with responses in spam email conversations. But, there were no other known data sets of the same scale or quality that could have been used for training a natural language understanding model of this sophistication. In addition, a possibility for future research would be to explore the use of models other than long-short term memory (LSTM) neural networks. However, LSTM networks are best designed to model temporal and sequence-based data, which makes them useful for modeling natural conversation.

V. HUMAN SUBJECTS

The project will eventually be deployed onto a small scale email server, where the language-aware system will converse with and reply to emails sent by real spammers. The human subjects that will be dealt with will be spammers, most of whom will be anonymous and located outside of the United States. All conversations that are logged will keep the spammers identities anonymous and will have all physical addresses, email addresses, names, and other sensitive information redacted. All logs will be kept in a secure private repository for the duration of the project and thereafter will be open-sourced to the public for future research to build on.

ACKNOWLEDGMENTS

I would like to give a special thanks to Melanie Tory, Google TensorFlow, Google Compute Engine, and anonymous spammers for helping make this research possible.

References are important to the reader; therefore, each citation must be complete and correct. If at all possible, references should be commonly available publications.

REFERENCES

[1] MacCartney, Bill. Understanding Natural Language Understanding. Review. The Stanford Natural Language Processing Group, Stanford University, 16 July 2014, nlp.stanford.edu/ wcmac/papers/20140716-UNLU.pdf. Accessed 6 Nov. 2016.

[2] Nielsen, Michael A. Neural Networks and Deep Learning. Website ed., Determination Press, 2015.

[3] Socher, Richard. CS224d: Deep Learning for Natural Language Processing. Stanford University CS224d: Deep Learning for Natural Language Processing, edited by Andrej Karpathy, Stanford University, cs224d.stanford.edu. Accessed 6 Nov. 2016.

[4] Spam Statistics. The University of Texas at El Paso Information Security Office, U of Texas at El Paso, admin.utep.edu/Default.aspx?tabid=64462. Accessed 6 Nov. 2016.

[5] Veitch, James. This is what happens when you reply to spam email — James Veitch. 1 Feb. 2016. TED, 1 Feb. 2016. Accessed 25 Sept. 2016. Speech.