Winter 12-13-2018

Natural Language Processing as a Weapon

Jordan Shropshire
University of South Alabama

Natural Language Processing as a Weapon

Jordan Shropshire[1]
School of Computing
University of South Alabama
Mobile, Al, USA

ABSTRACT

Natural Language Processing (NLP) is a science aimed at computationally interpreting written language. This field is maturing at an extraordinary pace. It is creating significant value and advancing a number of key research fronts. However, it also enables highly sophisticated phishing attacks. Given a large enough text sample, an NLP algorithm can identify and replicate defining characteristics of an individual's communication patterns. This facilitates programmatic impersonation of trusted individuals. A natural language processor could interpret incoming text messages or email and improvise responses which approximate the language of a known contact. The recipient could be tricked into sharing sensitive information. Just how vulnerable are we? This paper reviews the state of the art of natural language processing and social engineering. It also describes a test which empirically assesses our ability to discern legitimate communications from algorithmically-produced forgeries.

Keywords: natural language processing, social engineering, information security, phishing

INTRODUCTION

Natural Language Processing (NLP) is an interdisciplinary science aimed at computationally deriving meaning from written and spoken language. Natural language processors impose a hierarchy on language in order to extract meaning: words form phrases,

[1] Corresponding author.

phrases form sentences, and sentences contain ideas. They also translate concepts back into meaningful language. This field has made a number of major advances in recent years. Natural language processors are capable of observing the vocabulary and sentence structure of the language they are interpreting, and incorporating these characteristics in response text (Goldberg 2016).

Natural language processing is of significant benefit to both organizations and individuals. However, this technology could be used in highly advanced phishing attacks. Given a large enough text sample, an NLP algorithm will learn to craft language which closely resembles a specific person's communication patterns and style (Baki et al. 2017). It can then emulate that individual's specific language within a phishing attack (Sidorova et al. 2014). When communicating over written channels such as text or email, it could be difficult for a third party to determine if they are communicating with someone they know or with an algorithm masquerading as that individual (Salem et al. 2011). An unsuspecting person might assume they are communicating with a spouse, family member, friend, or coworker and end up sharing sensitive information.

Just how vulnerable are we? The purpose of this research-in-progress is to determine if NLP-enhanced phishing attacks are more effective than standard phishing attempts. It describes a test which empirically assesses our ability to discern legitimate communications from algorithmically-produced forgeries.

The remainder of this manuscript is organized as follows. The following section is the background. It introduces the concept of natural language processing in more detail. It also reviews end-user cyber security attacks which focus on phishing. After the background, the conceptual development is introduced. This section contains a series of hypotheses concerning

the relationship between NLP algorithm characteristics and human vulnerability. Once the hypotheses are introduced, the methods are described. This includes the sample, measures, and procedure. Finally, concluding comments offered and future research goals are explained.

BACKGROUND

Natural Language Processing

Natural language processors are often used in conjunction with machine learning algorithms to perform tasks such as automatic text summarization, translation, named entity recognition, sentiment analysis, relationship mapping, response suggestion, and automatic question answering (Cambria et al. 2014). A number of open source natural language processors are available. They include ApacheNLP, Natural Language Toolkit, Stanford NLP, and Mallet. Although these package include different algorithms and varied corpuses, they tend to include the same types of tools (see Table 1).

Part-of-Speech Tagger	This tool reads in text and assigns parts of speech (noun, verb, adjective, etc.) to each word.
Named Entity Recognizer	Recognizers labels sequences of words in a text which are the names of things. For instance, a person, company, bank account could be recognized. These names are extracted and reserved for model training.
Parser	The Parser evaluates the grammatical structure of sentences. The parser uses knowledge gained from previous training to produce the most likely analysis of new sentences. It can be paired with machine learning tools to create the structure for a machine-generated sentence.
Conference Resolution Finder	This application finds all expressions that refer to the same entity in a text. Individuals often use different words to describe the same object. For instance, mom, mommy, and mother all refer to the same entity. The Resolution finder is useful for patterning an individual's communication preference
Sentiment Analysis	The purpose of this tool is to identify and measure affective states and information. Understanding ingrained attitude is an important part of emulating speech.
Pattern-based Information Extraction	This tools learns pattern using labeled entities. The labeled entities are based on extractions of learned patterns. This recursive process provide higher level understandings of text.

Table 1: Common Attributes of Natural Language Processors

Phishing Attacks

Phishing is used to coerce individuals into performing certain actions or divulging sensitive information. The attackers may appear to be normal, credible, trustworthy people. By asking questions, they may piece together enough information to gain access to personal accounts for banking, email, business, or shopping. Traditionally, phishing attacks involve written communications sent via email (Khonji et al. 2013). In some cases, they are sent via text messages. The attacker massively distributes the messages, with little to modest personalization of message contents.

Phishing attacks are becoming more effective. Spear-phishing attacks are increasingly common. These attacks are targeted. The attackers use details gleaned from other sources in order to make the message seem more credible (Khonji et al. 2013; Neupane et al. 2015). For instance, they may include the customer name, items purchased, and even the name of the bank the customer uses in order to gain his or her confidence. Although this type of attack takes additional time and manpower to perpetuate, it is more successful than generic messages broadcasted to potential victims (Sheng et al. 2010).

CONCEPTUAL DEVELOPMENT

The present study holds that a new evolution in social engineering is on the horizon (Baki et al. 2017). This generation of exploit will use natural language processing and machine learning to take craft messages which take on the persona of a known, trusted contact. Malware embedded in mobile devices could analyze existing text message sequences and then use natural language processing to craft a series of message of text messages appearing to come from a

trusted contact (Verma et al. 2012). The malicious messages would emulate the language of the trusted person. The victim, believing that he or she is communicating with the trusted contact, would divulge confidential information. See Figure 1 for details.

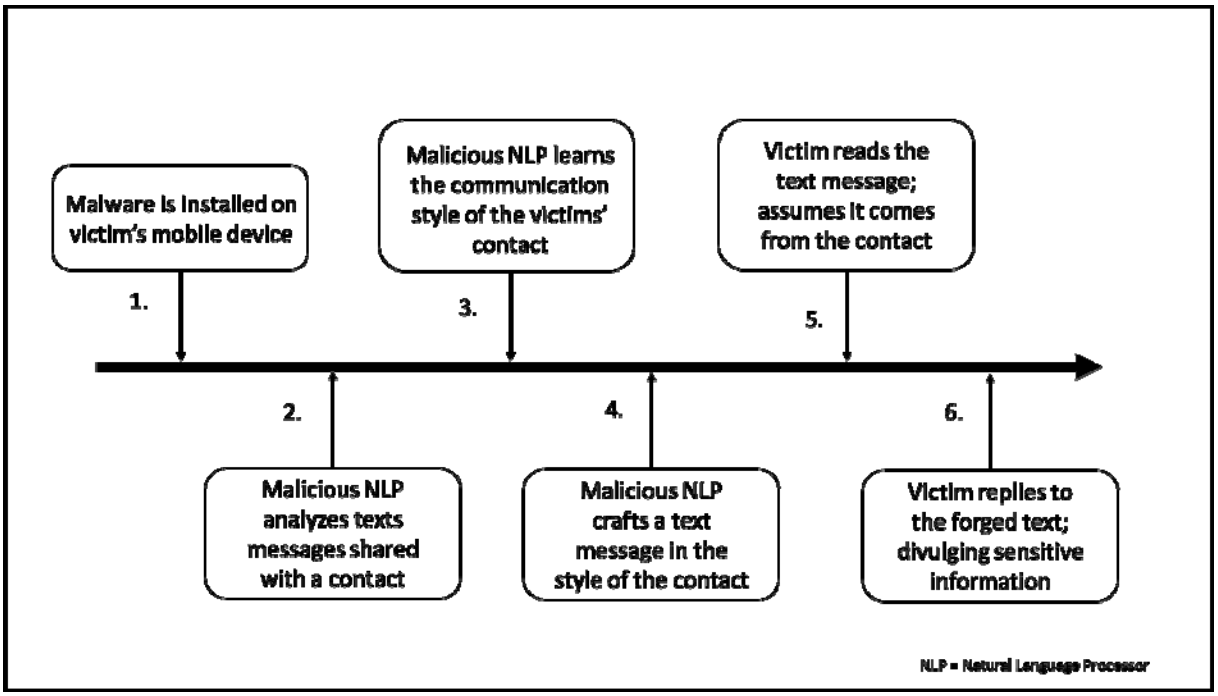

Figure 1. Social Engineering Attack Enhanced by Natural Language Processing

Compared to current phishing practices, the above exploitation enhanced by natural language processing is expected to be more effective and harder for victims to detect. It has four advantages: (1) The fraudulent communication appears to come from a specific person that the victim knows. Most phishing attackers attempt to convince their targets that the incoming message is from an authority such as a bank, enterprise, or credit card company. The target is more familiar and has more rapport with a close contact than a large, faceless organization. Hence, he or she is more likely to respond (Koppel et al. 2004). (2) Most credible businesses do not ask their customers to provide sensitive information via email. They use other communication channels. On the other hand, a close contact such as a relative or spouse may occasionally ask for sensitive information over text or email (Maxion et al. 2004; Sheng et al.

2010). (3) NLP-enhanced messages replicate the language patterns of that individual in all communications with the target. This builds confidence in the authenticity of the message (Cambria et al. 2014; Maxion et al. 2002). By contrast, most phishing attempts use neutral, unremarkable language that does not breed familiarity on within the victim (Neupane et al. 2015; Sheng et al. 2010; Verma et al. 2013). (4) Because the NLP-enhanced message appears to come from a close contact, perceived social obligation may compel the victim to respond (Maxion et al. 2004; Sidorova et al. 2014). In contrast, it is easier to ignore calls, letters, emails, and text messages from generic sources. Based on this evidence, the following hypothesis is offered:

H1: NLP-enhanced phishing attacks will be more effective than traditional fishing attacks.

METHODS

The purpose of this research is to assess the effectiveness of NLP-enhanced phishing attacks. It is hypothesized that this type of attack will be more effective than traditional phishing attacks. To test the proposed hypothesis, subjects are asked to look at groups of text messages sent to them by a known contact and identify messages which do not appear to come from the individual in question. The design of this test mirrors that of similar projects with related goals (Stringhini et al. 2015).

Procedure

Each subject is asked to review a list of 100 text messages that appear to come from a known contact on his or her phone and then identify suspicious text messages. Of these 100 messages 80 are authentic while 20 are forgeries. (Subjects were not told how many messages were fake.) Of the 20 forgeries, 10 are generically worded, standard phishing messages. The other 10 messages were crafted using the Stanford Core Natural Language Processor. To create the latter messages, it was necessary for the NLP software to review all messages sent by the

individual whose language was replicated. This allows for algorithm training. After analyzing the text messages, the NLP software modifies generic messages in order to replicate the style of the original sender. For instance, in place of using generic salutations such as "greetings," it uses phrases the sender would normally use, such as "sup bro" or "hi mom," or simply "hey." Each of the customized messages is based on one of the standard phishing messages.

Sample

For the pilot phase of this research-in-progress, 36 individuals were included in the sample. These individuals were willing to grant the researchers access to the SMS messages on their mobile devices. Subjects were interns filling non-technical positions at the US offices of a multinational organization. These individuals were primarily employed in the sales, management, logistics, and marketing departments. A total of 104 subjects were originally invited to participate in the study, resulting in a response rate of 35.7%.

Measurement

Each participant allowed the research team to copy the SMS messages from his or her mobile device to a secure computer for analysis. For each subject, the analysis focused on the communication stream which contained the most sent and received text messages. The natural language processors analyzed the 300 most recently received messages within this stream. This number was selected because all the subjects had at least 300 text messages from a single sender. For each subject, 80 randomly-selected text messages were drawn from the 300 most recent messages. These were included in the study. An additional 10 neutrally-worded phishing texts were added. A further 10 messages were included. These messages were modified using the Standard NLP Suite. They reflect the unique vocabulary, sentiment, and grammar observed within the 300 text messages analyzed during algorithm training.

CONCLUSIONS

This research-in-progress is still underway. The results of the pilot study will be presented at the conference. The results will be scored in terms of false positives (incorrectly classifying a legitimate message as a phishing attempt) and false negatives (incorrectly classifying a phishing attempt as a legitimate message). Further, the percentage of correctly classified standard phishing and NLP-enhanced phishing messages will be calculated. If the results of the pilot tests are promising and no procedural issues are evident, then the study will be conducted at full scale. The results of this research will be instructive. If people equally wary of generic and NLP-customized phishing attempts, then there will be little cause for concern. However, it is language-modified text messages will be significantly harder to detect than neutrally-worded communications.

REFERENCES

Baki, S., Verma, R., Mukherjee, A., and Gnawali, O. 2017. "Scaling and Effectiveness of Email Masquerade Attacks: Exploiting Natural Language Generation," in *ACM on Asia Conference on Computer and Communications Security*: Abu Dhabi, UAE.

Cambria, E., and White, B. 2014. "Jumping NLP Curves: A Review of Natural Language Processing Research " *IEEE Computational Intelligence Magazine* (9:2), pp 48-57.

Goldberg, Y. 2016. "A Primer on Neural Network Models for Natural Language Processing," *Journal of Artificial Intelligence Research* (57:1), pp 345-420.

Khonji, M., Iraqi, Y., and Jones, J. 2013. " Phishing detection: a literature survey," *IEEE Communications Surveys & Tutorials* (15:4), pp 2091-2121.

Koppel, M., and Schler, J. 2004. "Authorship verification as a one-class classification problem," in *Proc. 21th ICML,*: Alberta, CA.

Maxion, R., and Townsend, T. 2002. "Masquerade detection using truncated command lines," in *International Conference on Dependable Systems and Networks*: Washington, DC, USA.

Maxion, R., and Townsend, T. 2004. "Masquerade detection augmented with error analysis," *IEEE Transactions on Reliability* (53:1), pp 124-147.

Neupane, A., Rahman, N., Saxena, N., and Hirshfield, L. 2015. "A multi-modal neuro-physiological study of phishing detection and malware warnings," in *Proceedings of the 22nd ACM SIGSAC*: Denver, CO, USA.

Salem, M., and Stolfo, S. 2011. "Modeling User Search Behavior for Masquerade Detection," in *RAID: Proceedings of the 14th international conference on Recent Advances in Intrusion Detection*: Menlo Park, CA, USA, pp. 181-200.

Sheng, S., Holbrook, M., Kumaraguru, P., Cranor, L., and Downs, J. 2010. "Who Falls for Phish?: A Demographic Analysis of Phishing Susceptibility and Effectiveness of Interventions," in *CHI '10 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*: Atlanta, Georgia, USA.

Sidorova, G., Velasqueza, F., Stamatatos, E., Gelbukha, A., and Chanona-Hernández, L. 2014. "Syntactic N-Grams as Machine Learning Features for Natural Language Processing," *Expert Systems with Applications* (41:3), pp 853-860.

Stringhini, G., and Thonnard, O. 2015. "That ain't You: Blocking Spearphishing through Behavioral Modelling," in *DIMVA 2015 Proceedings of the 12th International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*: Milan, Italy.

Verma, R., and Hossain, N. 2013. "Semantic Feature Selection for Text with Application to Phishing Email Detection," in *International Conference on Information Security and Cryptology*: Dalian, China.

Verma, R., Shashidhar, N., and Hossain, N. 2012. "Detecting Phishing Emails the Natural Language Way," in *European Symposium on Research in Computer Security*: Pisa, Italy.