Human Language Technologies

# Phishing Attacks in the Context of Natural Language Processing and Natural Language Generation

Shane Cooke

Student ID: 17400206

A research report submitted in part fulfilment of the module of

**Human Language Techonologies (COMP40020)**

UCD School of Computer Science
University College Dublin

April 30, 2022

# Table of Contents

# Chapter 1: **Introduction**

### What is Phishing?

Phishing is a form of social engineering attack which is most often used to steal user credentials such as login credentials, credit card numbers and cryptocurrency keys. A phishing attack can occur in many forms, however the most common technique is for an attacker to pose as some sort of trusted authority such as a bank, government body or customer support agent. Posing as this trusted entity, the attacker will then send some form of communication to the victim, prompting them to enter their sensitive information or to click a malicious link. Once the victim has entered the details or clicked the link, the attacker will have access to any or all data exposed, and will be able to use this data for their own personal gain.

### Rates of Phishing Attacks

Even though phishing is a relatively old technique with the first email phishing attacks being detected in 1995[1], it is still to this day one of the most frequent and reliable forms of attack carried out in the online sphere. In the APWG's (Anti-Phishing Working Group) 'Phishing Activity Trends Report' for the 1st Quarter of 2021[2], it is reported that phishing attack rates doubled in 2020, and that an all-time high of 245,771 attacks had been detected in January of 2021. The constantly improving technology in the past decades, has led to both increasingly sophisticated phishing attack methods, and increasingly sophisticated phishing detection methods, and natural language processing has been at the forefront of both efforts.

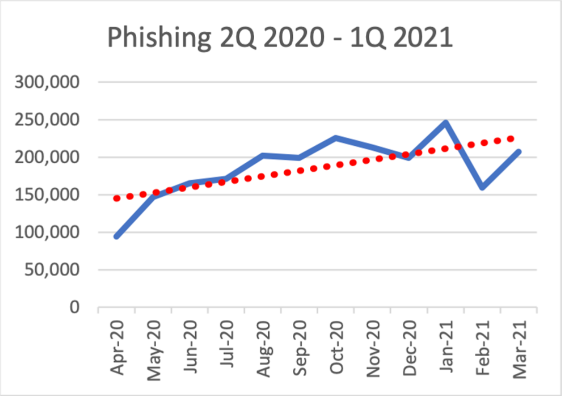


Figure 1.1: An Anti-Phishing Work Group graph illustrating the rates of phishing attacks from Q2 of 2020 to Q1 of 2021.

### Phishing Attacks in the Context of Natural Language Processing and Natural Language Generation

Natural language processing refers to the automatic manipulation of natural language through the use of computational methods, and natural language generation refers to the generation of natural language through computational methods. Due to the nature of phishing attacks as a massively language-based and communication-based attacks, both of these technologies have massive roles in both the creation and prevention of phishing attacks. Aspects of a phishing attack such as target selection, the generation of phishing communications, and the detection of phishing communications, all heavily benefit from the use of NLP and NLG, and both the attacker and the preventer benefit from the same technologies.

# Chapter 2: Target Selection for Phishing Attacks using Natural Language Processing

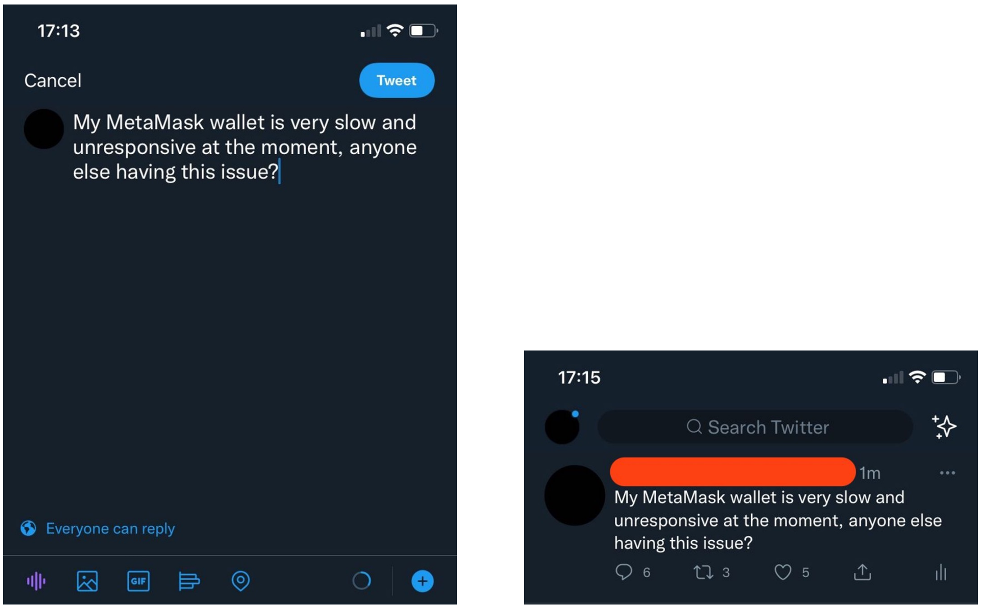
Target selection for phishing attacks can come in a variety of different forms such as 'Mass Phishing' where thousands or even millions of phishing communications are sent randomly to a vast array of different users, or 'Targeted Phishing' (also known as Spear Phishing) where an attack will be carefully created and designed for a small group or single user, based on their credentials, interests or needs[3].

For the target selection section of this report, I am going to focus on a very modern and specific form of targeted phishing which heavily relies on natural language processes for the selection of attack targets, and this attack goes by the name of "Angler Phishing"[4]. Angler phishing is a type of phishing attack which aims at social media users specifically. Bots who are disguised as customer service agents, or some other form of authority aim to mislead a user into revealing user account details, clicking on a malicious link, or entering some other for of compromising information.

In this form of attack, bots are programmed to search for specific key-words in posts, commonly financial institutions such as PayPal or cryptocurrency exchanges and wallets such as MetaMask. Once a post is found with the ley-word, the bot then uses natural language processing to analyse the language used in the post and evaluates it to try and compute what the best possible response would be. A response will then be sent, either from an account disguised as an authority, or an account disguised as a random stranger trying to do good, prompting the user to click a link, or reveal sensitive information. An example of this form of attack would be a user posting about their dissatisfaction with a service such as PayPal on an online social media platform. A pre-programmed bot will then find this post while scouring the social media channel and identify the keyword 'PayPal'. The bot then processes the language used in the post, comes to a conclusion as to what the best possible phishing response is, and will reply to this comment with a link to a malicious webpage indicating that this webpage will solve the original posters problems with PayPal.

The target selection process that I have outlined above is an extremely prevalent technique used on the Twitter social media platform. In order to demonstrate this fact, I decided to set up a dummy account on Twitter for testing purposes. Upon creation of this new user account, I made a public post which detailed my issues with the cryptocurrency wallet 'MetaMask', as shown in Figure 2.1(a) and 2.1(b). Within seconds, malicious bots that were scouring Twitter found my post, used NLP to process the content and language used in my post, and identified me as a potential target. In Figure 2.2, you can see the final step of the process, which is the phishing bots replying to my post with a phishing attack catered for my specific post, with statements such as "I had a similar Issue but they resolved it. Write to their official email for assistance at <fake phishing email account>".

As you can see, angler phishing is an extremely prevalent attack in todays social-media dominated world, and relies heavily on natural language processes for all stages of the attack. Posts are first found through key-word based searches or other methods, and then the content of the post is processed and analysed in order for the bot to produce the best possible response. These responses made by the bots may be human-made or hard coded into the bot, however it is very possible that these posts are created by natural language generation processes also, which I will discuss in the next chapter.



17:13

My MetaMask wallet is very slow and unresponsive at the moment, anyone else having this issue?

17:15

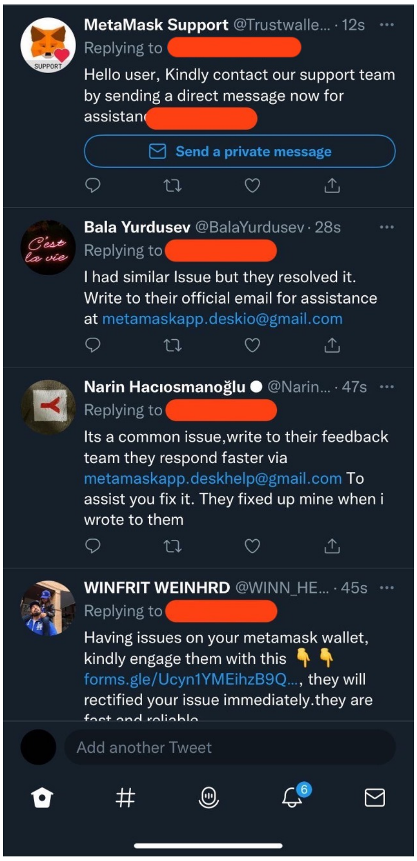
[redacted] 1m

My MetaMask wallet is very slow and unresponsive at the moment, anyone else having this issue?

(a) Original Post

(b) Instant Replies

Figure 2.1: Post I made on Twitter which alerts potential phishing attackers.



**MetaMask Support** @Trustwalle... · 12s

Replying to [redacted]

Hello user, Kindly contact our support team by sending a direct message now for assistan[redacted]

**Bala Yurdusev** @BalaYurdusev · 28s

Replying to [redacted]

I had similar Issue but they resolved it. Write to their official email for assistance at metamaskapp.deskio@gmail.com

**Narin Hacıosmanoğlu** ● @Narin... · 47s

Replying to [redacted]

Its a common issue,write to their feedback team they respond faster via metamaskapp.deskhelp@gmail.com To assist you fix it. They fixed up mine when i wrote to them

**WINFRIT WEINHRD** @WINN_HE... · 45s

Replying to [redacted]

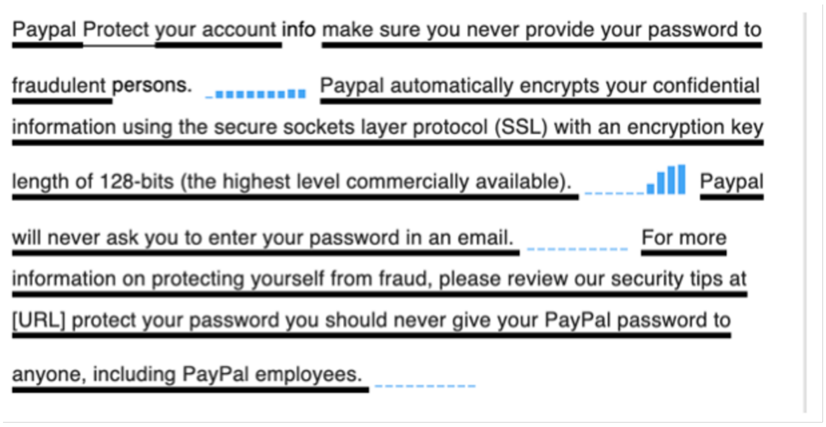
Having issues on your metamask wallet, kindly engage them with this 👇 👇 forms.gle/Ucyn1YMEihzB9Q..., they will rectified your issue immediately.they are fast and reliable

Figure 2.2: Phishing attempts sent to me after attackers used NLP to process the tweet I had posted.

# Chapter 3: Generation of Phishing Communications using Natural Language Generation

In recent years, language models are increasingly being used to generate both compelling and detection-evading phishing communications. At the Black Hat and Defcon security conferences in August of 2021, a team from Singapore's Government Technology Agency conducted an experiment where one AI language model created phishing email and one human created phishing email where sent to 200 of their fellow employees[5]. Upon analysis of which emails had been clicked most frequently, it was found that more people clicked the AI language modelled phishing emails by a significantly large margin when compared to the human-created phishing emails.

In a 2021 paper by Duskin K. et al., titled 'Evaluating and Explaining Natural Language Generation with GenX'[6], a system for the generation of phishing emails through NLG (Natural Language Generation) is proposed. In this system, 9234 emails are used to train and test a GPT-2 language model in order to produce 500 unique generated phishing emails. The output generated by this natural language generation was found to be "highly coherent" and boasted a wide diversity in approaches between emails, which is perfectly suited for mass phishing attacks. As you can see in Figure X, the emails created by this natural language generation process where extremely realistic bar a few minor errors, which may not be picked up by detection systems.



Paypal Protect your account info make sure you never provide your password to fraudulent persons. Paypal automatically encrypts your confidential information using the secure sockets layer protocol (SSL) with an encryption key length of 128-bits (the highest level commercially available). Paypal will never ask you to enter your password in an email. For more information on protecting yourself from fraud, please review our security tips at [URL] protect your password you should never give your PayPal password to anyone, including PayPal employees.

Figure 3.1: A Natural Language Generated email from the Duskin K. et al., study.

While the above paper by Duskin K. et al., proposes a technical system for the natural language generation of emails for the purpose of mass phishing, a paper by Giaretta A. and Dragoni N., titled "A Middle Ground Between Massive and Spear Phishing through Natural Language Generation"[7] proposes a conceptual system for the purpose of targeted phishing (spear phishing). The NLG model proposed in this paper is a "Template-Driven" model, in which Natural Language Generation is used to fill in "blanks" in a phishing email template. First, a target for the phishing attack is chosen and categorised, then depending on the target, the NLG model will generate the most accurate and appropriate language to fill in the blanks of the pre-made phishing email. The ultimate goal of this system is to take advantage of the reduced computation power needed due to the use of email templates rather than full NLG email generation, while still adding personalised and convincing language elements to the phishing email.

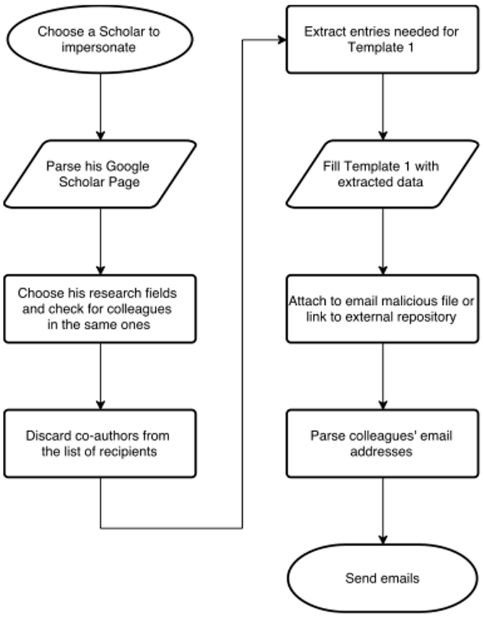
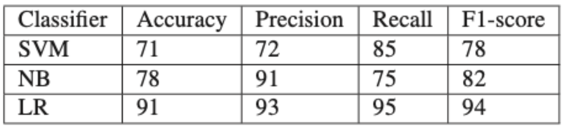

Choose a Scholar to impersonate
Parse his Google Scholar Page
Choose his research fields and check for colleagues in the same ones
Discard co-authors from the list of recipients
Extract entries needed for Template 1
Fill Template 1 with extracted data
Attach to email malicious file or link to external repository
Parse colleagues' email addresses
Send emails

Figure 3.2: The system architecure proposed in the Giaretta A. and Dragoni N., study for the creation of a Natural Language Generated email.

Confirming that the efforts of the above papers where not in vain, a study done by Baki, S., et al., titled "Scaling and Effectiveness of Email Masquerade Attacks: Exploiting Natural Language Generation"[8], found that participants performed close to random at a rate of 52% when differentiating between a communication (email or text) that was generated using natural language generation, and a communication that was created by a human. It was also found that 17% of participants could not identify a single signal that identified the communication as natural language generated, and that the knowledge and background of a user in terms of email use had no effect on their ability to differentiate NLG emails from human-made emails. This study focused on human-ability to identify natural language generated emails, however a 2019 paper by Das, A., et al., titled "Automated email Generation for Targeted Attacks using Natural Language"[9] focused on computer ability to identify natural language generated malicious emails. The authors of this paper found that even though the emails generated through language models were said to have an "incoherent nature", the classifiers used in this study only achieved a maximum accuracy of between 71% and 91% when differentiating between NLG malicious emails and human-made emails.



| Classifier | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| SVM | 71 | 72 | 85 | 78 |
| NB | 78 | 91 | 75 | 82 |
| LR | 91 | 93 | 95 | 94 |

Figure 3.3: Results achieved by classifiers in the Das, A., et al., paper when differentiating between Natural Language Generation phishing emails and human-made emails.

The above papers serve to illustrate both the methodology behind generated malicious emails, and the massive risks involved when NLG technology is being used in this manner. As natural language technologies evolve and advance, so too does the potential for NLG communications to avoid detection and cause significant harm. For this reason, sophisticated and advanced phishing detection methods must be used in communication channels, which I will delve into in the next chapter of this report.

# Chapter 4: Detection of Phishing Communications using Natural Language Processing

One of the most prevalent and successful techniques for the detection of phishing communications in the online sphere is through the use of natural language processing. In order to accurately and efficiently detect malicious online communications, a methodology is required which has the ability to detect and evaluate subtle flags and signs that the language in use may be malicious and deceptive in nature. Natural language processing techniques combined with language models are at the forefront of this effort, and aim to prevent the frequency and viability of phishing attacks through email, social media, and forums.

A paper by Abdelaziz, O., et al., titled "A Novel Phishing Email Detection Algorithm based on Multinomial Naïve Bayes Classifier and Natural Language Processing"[10], proposes a system combining the bag-of-words NLP model with a Bayesian classifier to detect phishing email communications. In this study, the IWSPA[11] dataset of 5091 legit emails and 628 phishing emails is used as training and testing data for the phishing detection models. The emails in this dataset are first pre-processed using a wide array of data cleaning techniques, and then converted to features using the bag-of-words model. A Naïve Bayes classifier is then used to produce classifications for these emails, and a phishing detection language model is created.

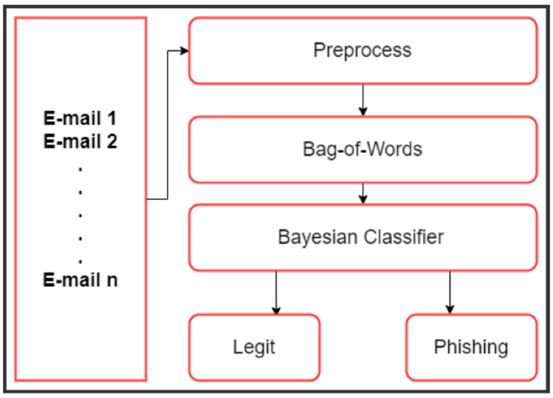
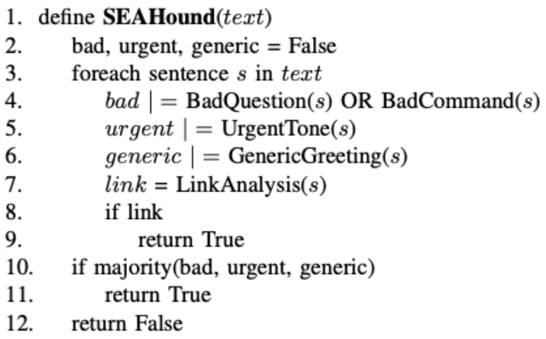


Figure 4.1: The system architecure proposed in the Abdelaziz, O., et al., study for the detection of phishing emails using Natural Language Processing.

The model proposed in this system returned extremely positive results, with a 96.03% overall accuracy and a misclassification rate of less than 4% achieved in the detection of malicious phishing emails. This was a benchmark study in the use of NLP processes for the detection of phishing communications, and serves as a perfect example for real-world uses of Natural Language Processing technologies.

While this is a viable and effective system for the detection of phishing emails, there are a wide array of solutions for this multi-faceted problem. A paper by Peng, T., et al., titled "Detecting Phishing Attacks Using Natural Language Processing and Machine Learning"[12], proposes a system in which a document or email is read one sentence at a time, and returns "True" if a social engineering attack is detected. A black-list is generated using a machine learning classifier, which specifies a list of verb-object pairs that signify the potential for a phishing attack. The black-list was then used in combination with the sentence reading algorithm, to analyse 5,000 regular emails and 5,000 phishing emails in order to determine its ability to detect possible phishing attacks.

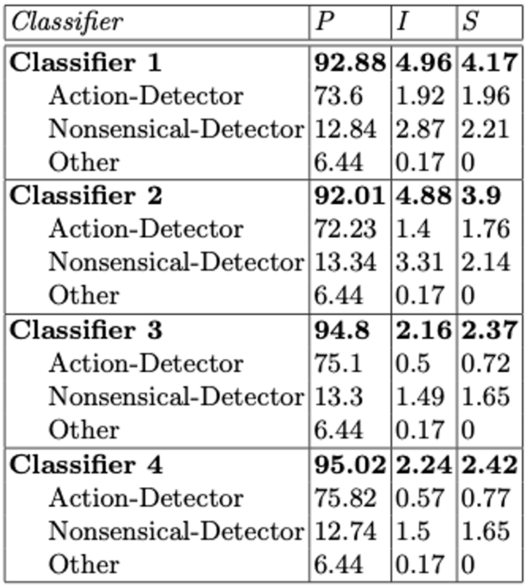


```
define SEAHound(text)
    bad, urgent, generic = False
    foreach sentence s in text
        bad |= BadQuestion(s) OR BadCommand(s)
        urgent |= UrgentTone(s)
        generic |= GenericGreeting(s)
        link = LinkAnalysis(s)
        if link
            return True
    if majority(bad, urgent, generic)
        return True
    return False
```

Figure 4.2: Example of the SEAHound algorithm used in the Peng, T., et al., paper.

This paper focused on the precision and recall evaluation metrics to judge the performance of their algorithm, and the results where again extremely positive. The SEAHound algorithm achieved a precision of 95% and a recall of 91%, which shows highly successful classification of non-phishing and phishing emails. This was a slight decrease on the precision achieved in the Abdelaziz, O., et al., paper proposed above (95.27%) , however serves to show that the problem of detecting phishing emails using Natural Language processing has multiple possible viable solutions.

Confirming this statement, A paper by Verma, R., et al., titled "Sematic Feature Selection for Text with Application to Phishing Email Detection"[13] proposes an entirely new solution to the two previous papers. In this study, natural language processes such as lexical analysis, part-of-speech tagging, stemming and stop word removal are central to the overall system. In this study, emails are first processed using these aforementioned NLP steps, and then a collection of four language models are used for the detection of phishing emails, including Pattern Matching (PM), PM and POS tagging, PM, POS and Word Senses, and finally POM, POS, Word Senses and WordNet.



| *Classifier* | *P* | *I* | *S* |
|---|---|---|---|
| **Classifier 1** | **92.88** | **4.96** | **4.17** |
| Action-Detector | 73.6 | 1.92 | 1.96 |
| Nonsensical-Detector | 12.84 | 2.87 | 2.21 |
| Other | 6.44 | 0.17 | 0 |
| **Classifier 2** | **92.01** | **4.88** | **3.9** |
| Action-Detector | 72.23 | 1.4 | 1.76 |
| Nonsensical-Detector | 13.34 | 3.31 | 2.14 |
| Other | 6.44 | 0.17 | 0 |
| **Classifier 3** | **94.8** | **2.16** | **2.37** |
| Action-Detector | 75.1 | 0.5 | 0.72 |
| Nonsensical-Detector | 13.3 | 1.49 | 1.65 |
| Other | 6.44 | 0.17 | 0 |
| **Classifier 4** | **95.02** | **2.24** | **2.42** |
| Action-Detector | 75.82 | 0.57 | 0.77 |
| Nonsensical-Detector | 12.74 | 1.5 | 1.65 |
| Other | 6.44 | 0.17 | 0 |

Figure 4.3: Results achieved by the four detection systems used in the Verma, R., et al., paper.

As you can see in Figure 4.3, the first three detection methods of choice in this paper produced slightly worse precision scores than the first two papers mentioned in this section (92.88%, 92.01% and 94.80%). The fourth classifier (Classifier 4) however, achieved a 95.02% precision, which means it performed slightly better than the detection system used in the Peng, T., et al., paper (95%), and a small margin worse than the detection system used in the Abdelaziz, O., et al., paper (95.27%). The authors concluded that while most phishing detection systems only use natural language processing on the email body, this study had great success in extending NLP processes to the header of the emails also, which again is a unique and novel solution for phishing email detection.

In the studies carried out above, there are a wide array of different methodologies employed to tackle the challenge of phishing communication detection. While the specific methodologies used may vary in these studies however, there is always a common theme running through these detection efforts, and that theme is the heavy use of, and reliance on natural language processes. The use of language in phishing communications is one of the most vital aspects of the whole attack, and through the categorisation, evaluation and analysis of this language using NLP techniques, researchers are beginning to form extremely proficient and accurate phishing detection systems.

# Chapter 5: **Summary and Conclusions**

Phishing attacks involve the use of language to social engineer and manipulate a victim into carrying out some action that is against their best interest such as entering sensitive details, clicking a link, or downloading an unsafe file. Due to the nature of this language-based attack, natural language processing and natural language generation have played, and continue to play a massive role in all aspects of phishing attacks. An attacker can use natural language processing in efforts to find attack targets, and can use natural language generation to quickly and effectively produce phishing communications such as emails, texts, or social media messages. A preventer can use natural language processes to counteract the attackers efforts, and to rapidly and accurately detect phishing communications before they have reached their victim.

This paper aims to outline one of the most prevalent and important use-cases of natural language technologies in the online security field today. Every time we send or receive emails, interact with others on social media, or share links through forums, natural language processes are efficiently and quietly being used in the background to keep our sensitive and confidential information safe from the most frequent form of attack carried out in the online sphere today, phishing attacks.

# Bibliography

1. Cofense. History of Phishing. https://cofense.com/knowledge-center/history-of-phishing/ (2021).
2. apwg. *Phishing Activity Trends Report* (Anti-Phishing Work Group, 2021). https://docs.apwg.org/reports/apwg_trends_report_q1_2021.pdf.
3. PureCloud. THE DIFFERENCE BETWEEN MASS PHISHING  SPEAR PHISHING. https://www.purecloudsolutions.co.uk/the-difference-between-mass-phishing-spear-phishing/ (2020).
4. UK, I. G. What is angler phishing? https://www.itgovernance.co.uk/blog/beware-of-angler-phishing (2019).
5. Wired. AI Wrote Better Phishing Emails Than Humans in a Recent Test. https://www.wired.com/story/ai-phishing-emails/ (2021).
6. Duskin, K. & Sharma, S. Evaluating and Explaining Natural Language Generation with GenX. https://aclanthology.org/2021.dash-1.12.pdf (2021).
7. Giaretta, A. & Dragoni, N. A Middle Ground Between Massive and Spear Phishing through Natural Language Generation. https://arxiv.org/pdf/1708.07342.pdf (2019).
8. Baki, S. & Verma, R. Scaling and Effectiveness of Email Masquerade Attacks: Exploiting Natural Language Generation. https://dl.acm.org/doi/pdf/10.1145/3052973.3053037 (2019).
9. Das, A. & Verma, R. Automated email Generation for Targeted Attacks using Natural Language. https://arxiv.org/pdf/1908.06893.pdf (2019).
10. Abdelaziz, O. A Novel Phishing Email Detection Algorithm based on Multinomial Naive Bayes Classifier and Natural Language Processing. https://www.scitepress.org/Papers/2020/104126/104126.pdf (2018).
11. Verma, R. M. Data Quality for Security Challenges: Case Studies of Phishing, Malware and Intrusion Detection Datasets. https://dl.acm.org/doi/abs/10.1145/3319535.3363267 (2019).
12. Peng, T. Detecting Phishing Attacks Using Natural Language Processing and Machine Learning. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8334479 (2021).
13. Verma, R. & Hossain, N. Semantic Feature Selection for Text with Application to Phishing Email Detection. https://cs.rochester.edu/u/nhossain/icisc-2013-phishing.pdf (2019).