IMPROVING PHISHING DETECTION VIA PSYCHOLOGICAL TRAIT SCORING

Sadat Shahriar, Arjun Mukherjee and Omprakash Gnawali
University of Houston, 4800 Calhoun Rd, Houston, TX 77004, USA

ABSTRACT

Phishing emails exhibit some unique psychological traits which are not present in legitimate emails. From empirical analysis and previous research, we find three psychological traits most dominant in Phishing emails – **A Sense of Urgency**, **Inducing Fear by Threatening**, and **Enticement with Desire**. We manually label 10% of all phishing emails in our training dataset for these three traits. We leverage that knowledge by training BERT, Sentence-BERT (SBERT), and Character-level-CNN models and capturing the nuances via the last layers that form the Phishing **Psychological Trait (PPT)** scores. For the phishing email detection task, we use the pretrained BERT and SBERT model, and concatenate the PPT scores to feed into a fully-connected neural network model. Our results show that the addition of PPT scores improves the model performance significantly, thus indicating the effectiveness of PPT scores in capturing the psychological nuances. Furthermore, to mitigate the effect of the imbalanced training dataset, we use the GPT-2 model to generate phishing emails (Radford et al., 2019. Our best model outperforms the current State-of-the-Art (SOTA) model's F1-score by 4.54%. Additionally, our analysis of individual PPTs suggests that Fear provides the strongest cue in detecting phishing emails.

KEYWORDS

Phishing, Email, BERT, Psychology

1. INTRODUCTION

Phishing is a technique used in electronic messaging to deceive the reader, where the phisher camouflages the message with a legitimate facade to access sensitive information or monetary gain (Vishwanath et al., 2011; Bose and Leung 2009). As the phishers prey on the vulnerability of the users, they often persuade people to take on some actions which may lead to undesirable consequences. Phishing attacks increased significantly in recent years and although researchers exploited several Natural Language Processing and Machine Learning techniques to detect phishing emails, the phishers evolved over time, making it harder to detect phishing emails (Almomani et al., 2013; Khonji, Iraqi, and Jones, 2013; FBI, 2021). Hence, phishing email detection systems must be smart enough to cope with the evolving nature of phishing techniques. In this work, we propose that all phishing emails exhibit some unique psychological traits, and detection of these traits can play a significant role in improving phishing vs. legitimate email classification.

Researchers analyzed the phishing attack from psychological perspectives, such as how persuasion is conducted (Akbar, 2014; Cialdini, 2001), the human factors in phishing attack (Stajano and Wilson, 2011; Jakobsson, 2007), and psychological mechanism in the effectiveness of phishing attacks (Luo et al., 2013). Research suggests that phishing messages often exhibit psychological cues, which can be crucial for their successful detection (Jones et al., 2019; Jakobsson, 2007). However, the current research is not adequate to quantify psychological traits expressed through the body of text. Consequently, how these traits play into detecting phishing emails is still an unexplored area of research. Nevertheless, for a smart detection of phishing emails, it is of immense importance to incorporate psychological attributes of the email's text, along with the linguistic model. We define *Phishing Psychological Traits (PPT)* as the psychological attributes evident in phishing emails. We claim that three major psychological attributes are evident in the phishing emails–based on whether the email sounds rushed (*a Sense of Urgency*), if the email induces fear *(Inducing Fear by Threatening*), and if there is an enticement through that email (*Enticement with Desire*). These traits can appear standalone or with a combination of any two and even three.

In this research, we capture the psychological traits by modeling a BERT, Sentence-BERT (SBERT) and Char-CNN network (Devlin et al., 2019; Reimers and Gurevych, 2019; Zhang, Zhao, and LeCun, 2015). We use these models to compute the softmax probability score (PPT score) for every phishing and legitimate emails. Next, we use pretrained BERT and SBERT model to find the feature-embedding (768-D) from text and concatenate the PPT scores with these embeddings. The concatenated features are fed to a fully-connected neural network to predict the email being phishing or legitimate. Our best performing model achieves the F1-score of 88.04%, which outperforms the current SOTA by 4.54%. We also observe a significant improvement of F1- score by up to 2.62% for the PPT-based model over the PPT-less model. The key to the consistent performance improvement is the PPT scores which provide reliable and unique cues by capturing the subtlety of psychological aspect expressed in the emails.

The novelty of this research is that our work is the first one to quantify the underlying psychological cues and leverage them for phishing email detection. We further investigate how the PPT scores help boosting the classification performance by providing t-SNE-based visualization (van der Maaten and Hinton, 2008). Furthermore, we analyze the contribution of individual PPT score and effectiveness of PPT scores in low-training-data situations. Our research provides important insights into the unique psychological attributes of phishing emails, which can create a new research direction in the phishing email detection paradigm.

2. RELATED WORKS

Phishing emails have been adversely affecting the internet world since 1996 (Salloum et al., 2021). Different NLP techniques were used to extract semantic, syntactic and contextual features which played important role for phishing detection along with classical machine learning techniques (Cui et al., 2020; Verma and Hossain, 2013; Park and Taylor, 2015; Blanzieri and Bryl, 2009; Gansterer and Pölz, 2009; Feng et al., 2016). However, the evolving nature of phishing emails entails more sophisticated techniques as often they do not contain the malicious code or word choices of known attacks (Lee, Saxe, and Harang, 2020a). The transformer-based models and their variants proved to have superiority over the traditional deep learning models mostly due to their transfer learning capabilities and they were successfully used in many deceptive text detection tasks (Vaswani et al., 2017; Shahriar et al., 2021). Lee et al. proposed a BERT-based phishing email detection model where they pruned half of the transformer blocks to better capture the semantics (Lee, Saxe, and Harang, 2020b). However, none of these works consider the unique psychological aspect of phishing emails, that can provide useful features to detect them. Naidoo suggested the *urgency* to be a dominant psychological feature in phishing email, but an ML-based automatic detection strategy is not present in their work (Naidoo, 2015). We address these research gaps by leveraging the context embedding of the BERT and SBERT model and using the psychological traits to detect phishing emails.

3. METHODOLOGY

Figure 1 explains the whole process in brief. We start by selecting 10% of the phishing emails and manually label them as 1 or 0 based on the presence of each PPTs. Next, we train a BERT, SBERT and Char-CNN model for each PPTs and use the trained model to compute the PPT scores of all emails. We concatenate these PPT scores with the fine-tuned BERT model and pretrained SBERT model and feed them to a Deep Neural Net model to detect the phishing email.

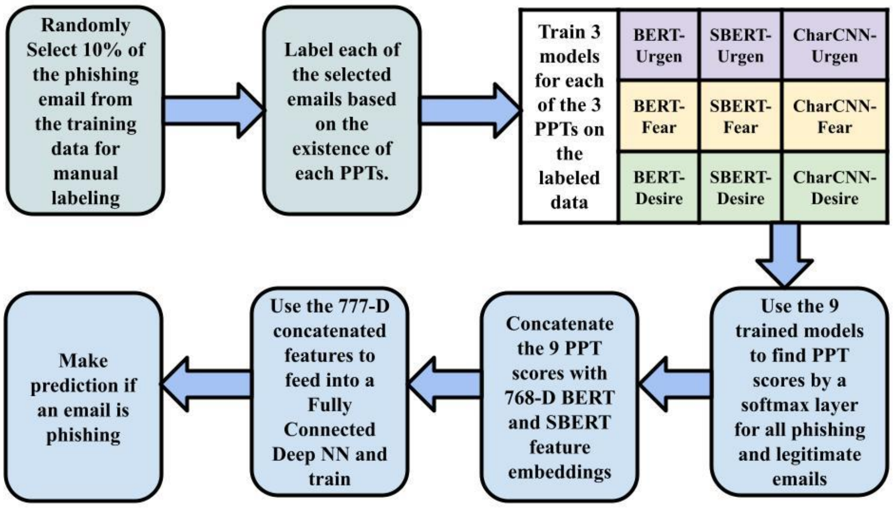

Randomly Select 10% of the phishing email from the training data for manual labeling
Label each of the selected emails based on the existence of each PPTs.
Train 3 models for each of the 3 PPTs on the labeled data
BERT-Urgen
SBERT-Urgen
CharCNN-Urgen
BERT-Fear
SBERT-Fear
CharCNN-Fear
BERT-Desire
SBERT-Desire
CharCNN-Desire
Use the 9 trained models to find PPT scores by a softmax layer for all phishing and legitimate emails
Concatenate the 9 PPT scores with 768-D BERT and SBERT feature embeddings
Use the 777-D concatenated features to feed into a Fully Connected Deep NN and train
Make prediction if an email is phishing

Figure 1. The complete flow diagram for phishing email detection using Phishing Psychological Traits (PPT)

3.1 Phishing Psychological Traits (PPT)

We claim that one or more persuasion strategies are implemented with the psychological traits expressed in the text. PPTs work as a broader umbrella and all phishing emails exhibit one or more of these PPTs.

3.1.1 A Sense of Urgency

One of the crucial characteristics of many phishing messages is the expression of urgency. Having a time constraint can induce stress in the readers' mind even if they are capable to do so within the stated time (Ordonez and Benson III, 1997). The phishing emails often urge the reader to take some action with promptness, thus reducing the time for the reader to reason or report (Aggarwal, Kumar, and Sudarsan, 2014). Additionally, it can induce impulsive behavior in the recipient that can lead to an error of judgment (Cui et al., 2020). Therefore, it is imperative to detect urgency from an email. We observe that urgency expressed in the phishing messages are direct and expressive, as the attackers want to maximize the possibility of a reader's response.

3.1.2 Inducing Fear by Threatening

Research shows that Fear by threatening is one of the most frequently exploited emotional trigger by the attackers (Sharma and Bashir, 2020, Halevi, Lewis, and Memon, 2013; Ferreira and Lenzini, 2015). The threat can be of many forms, for example, being locked out or blocked from one's account, losing access to information, getting hacked, stealing information, stealing currency, and individually targeted attack (Wang et al., 2012). Notably, Bitaab et al. stated that during the COVID-19 pandemic, the readers' fear is exploited by the attackers, which led to a high increase of phishing attacks (Bitaab et al., 2021). Hence it is evident that a direct or indirect threat can be a significant cue of phishing emails.

3.1.3 Enticement with Desire

Phishing email often lures by enticing the readers' personality trait of openness that can make them to be greedy and curious which can result in getting phished (Ding et al., 2015; Halevi, Lewis, and Memon, 2013). Phishing emails often contain a lucrative financial reward in exchange for clicking on some links, providing personal details or credit card information, and so on. Stajano and Wilson maintained that "Need and Greed" is one of the seven basic principles of scams, where people can be a victim of a lottery scam or a sexy swindler (Stajano and Wilson, 2011). Hence, the enticement with greed or curiosity can be an important signal of phishing emails.

3.2 Experimental Setup

To obtain the PPT scores of all emails, we train three models for each of the PPTs – BERT, SBERT and Char-CNN. We split all the manually-labeled-PPT emails in 80% for training and 20% for validation and repeat the experiments for three different splits. For the training of phishing email detection task, we use the BERT and SBERT model and apply the same train-validation split. We find the best set of hyperparameters by observing the performance on the validation set.

4. DATASET

The dataset we used was provided in the Anti-Phishing Pilot at ACM IWSPA 2018 (Verma, Zeng, and Faridi, 2019). Our test data size is 4300, from the first shared task, where emails are provided without the headers (*IWSPA_NH*). The IWSPA_NH training set has 5092 legitimate and 629 phishing emails. We also added 4082 legit, 503 phish emails from the IWSPA header-added dataset (*IWSPA_H*). Additionally, to examine how our trained model performs on other datasets, we curated a new small dataset called *UNIV_Phish*, containing 326 emails. We collected 163 phishing email from three different university websites: 72 emails from Stanford University, 68 from Lehigh University and 23 from University of Washington. We also added 163 emails from Enron "ham" emails. Notably, we made sure, none of these emails appeared in the IWSPA training set.

5. RESULT AND DISCUSSION

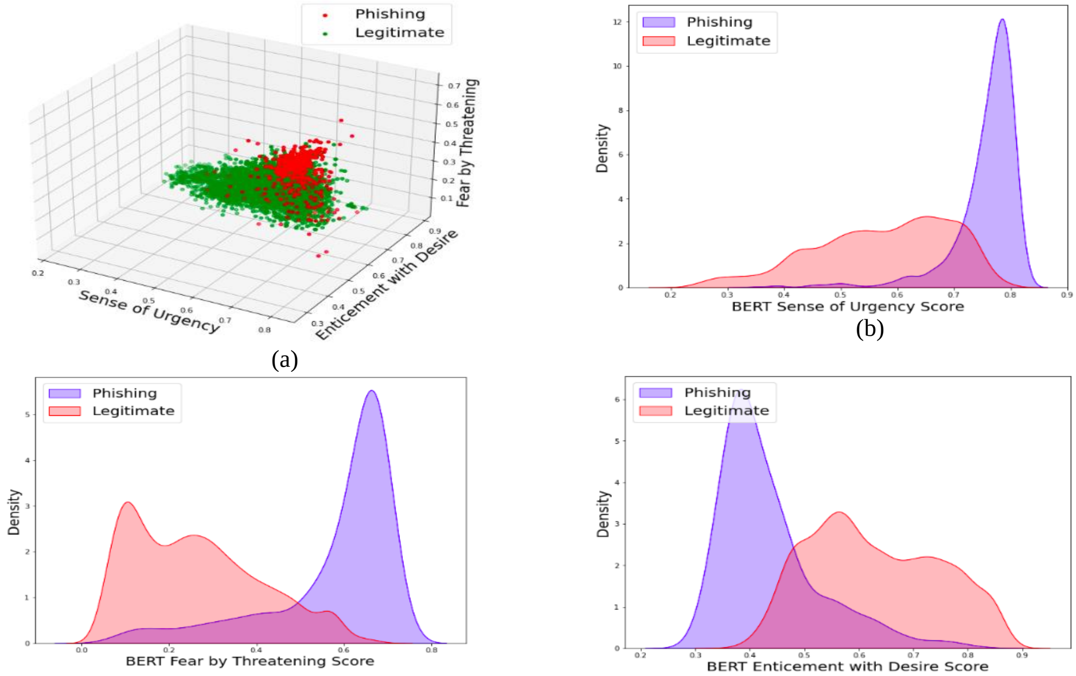
We hypothesize that when we add the PPTs along with the language model, the performance of phishing email detection improves. Therefore, we first need to find the PPT scores for all emails and then use these scores to detect phishing emails.

5.1 Detection of Phishing Psychological Traits

The randomly selected phishing emails are labeled by one of the authors as 1 or 0 for each PPTs. We find 82.54% emails are labeled as Urgent, where 71.42% emails are labeled as both Urgent and Fear. Only 4.76% emails exhibit all the three traits. We use BERT, SBERT, and Char-CNN to train on the manually labeled emails. Then, we make prediction on rest of the email using these models' last softmax layer and use the softmax output as PPT score.

Figure 2 depicts the distribution of BERT-based PPT scores in phishing vs legitimate emails. We observe that the PPT scores create distinguishable clouds for phishing and legitimate emails in Figure 2(a). Figure 2(b) shows the kernel density plot of individual PPTs, which represents the continuous probability density curve for these traits. For *Sense of Urgency* and *Fear by Threatening*, we observe a high density of Phishing emails on the right side of the curve which indicates the high probability of *urgency* and *fear* in the phishing emails. However, contrary to our intuition, phishing emails have a lower probability in the *Desire* score than the legitimate emails. One reason could be the lack of phishing emails in our dataset with this trait, which may have caused the failure to capture the *Desire* trait in the emails.

We explore how different words are related to the three PPTs. We observe "immediately", and "soon" are two most recurring adverbs in the *Urgent*-labeled emails. The *Desire*-labeled emails are mostly the offers of a free upgrade of some subscriptions. However, since the main goal of phishing emails is to persuade the user to take some action by clicking on a phishing link, we observe the words "link" and "click" frequently in all the phishing emails.



(a)

(b)

(c)

(d)

Figure 2. (a) 3-dimensional plot of PPT scores for BERT. (b) Kernel Density Estimation (KDE) plot of BERT-based PPT Score for Sense of Urgency (c) KDE plot Fear by Threatening (d) KDE plot for Enticement with Desire

5.2 Phishing Email Detection

From Table 1, we observe that, for every case, the performance improves when PPT scores are added to the training process. While trained on IWSPA_NH training data for the BERT model, we found that adding the PPTs improves the accuracy by 0.70% and F1-score significantly by 2.62% (p-value=0.04). We find the same trend for SBERT model as well. Adding the psychological traits improves the accuracy by 1.31% and F1-score by 1.34%.

Next, we added the additional training data for IWSPA with header information (IWSPA_H). The additional training data helps improve the performance. However, adding the PPTs again improves the performance. For BERT model, the improvement is by 0.63% in accuracy and 2.19% (p-value=0.03) in F1-score. Similarly, for the SBERT model, F1-score has a significant improvement of 2.50% (P-value=0.02). It may be noted that the current SOTA F1-score for this test set is 83.5%. Adding the PPTs with IWSPA_NH and IWSPA_H training data, we outperform the current SOTA (85.16% for BERT and 83.88% for SBERT).

A major challenge in our task is the lack of training data in the phishing email category due to the corporations being reluctant and individuals being ashamed to share such sensitive data (Aassal et al., 2018). In order to balance the training dataset, we tried different approaches like SMOTE (Chawla et al., 2002), cost-sensitive learning methods (Thai-Nghe, Gantner, and Schmidt-Thieme, 2010), and the addition of GPT- 2-generated phishing emails (Radford et al., 2019). However, we did not find any performance improvement for the first two. For GPT-2, we observe the performance boost for BERT models by 1.02% in accuracy and 3.59% in F1-score. When we added the psychological trait features with GPT- 2-generated emails, it also improved the accuracy by 0.45%, and F1-score by 1.38% .

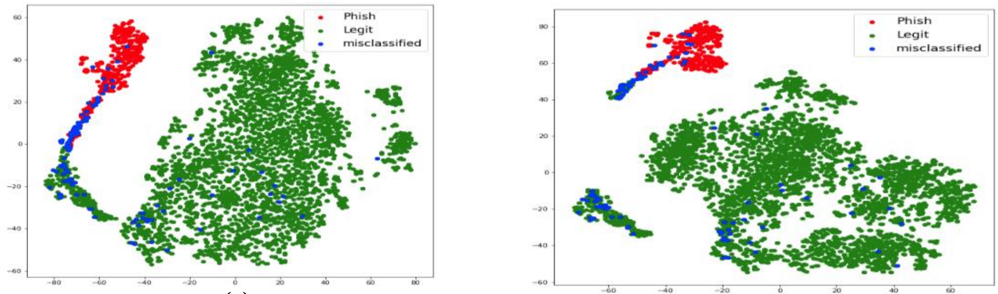
Table 1. Performance of the IWSPA test set and UNIV_Phish dataset while trained on different training data. We observe the best performance is found when we use IWSPA header-less, header-added data, GPT-2-generated phishing emails, and PPT features added with the BERT model

		BERT		BERT + PPT		SBERT		SBERT + PPT	
Tested On	***Training Data***	**Acc (%)**	**F1 (%)**	**Acc (%)**	**F1 (%)**	**Acc (%)**	**F1 (%)**	**Acc (%)**	**F1 (%)**
IWSPA Test set	*IWSPA_NH*	95.11	79.34	95.81	81.96	95.39	79.29	95.70	80.63
	IWSPA_NH + IWSPA_H	96.00	83.07	96.63	85.61	96.23	81.38	96.54	83.88
	IWSPA_NH+IWSPA_H+GPT2	97.02	86.66	97.47	**88.04**	94.32	77.19	95.04	79.41
UNIV_ Phish	*IWSPA_NH*	85.27	84.71	86.81	85.61	81.29	79.38	82.82	80.70
	IWSPA_NH + IWSPA_H	86.50	85.23	87.11	86.17	82.21	80.01	83.74	82.15
	IWSPA_NH+IWSPA_H+GPT2	87.42	86.98	88.03	**87.77**	80.06	76.63	82.82	80.55

However, contrary to the BERT model, the addition of GPT-2 generated data created a significant decline in the SBERT performance. The reason could be the poor coherence of some of the generated data. While for the BERT model, we use the output from the [*cls*] token, SBERT uses a pooling strategy, leading to poor sentence embedding of non-coherent texts. Additionally, as suggested in Reimers and Gurveych (Reimers and Gurevych, 2019), since SBERT cannot be used to update all the internal layers of BERT architecture, it may not be well suited for transfer learning.

We further test our model performance on UNIV_Phish dataset. From Table 1, we observe that added PPT improves the performance up to 4.02% in F1-score. We also observe the similar performance boost with added IWSPA_H set (up to 1.45%) and added GPT-2-generated email set (up to 1.75%). Hence, the performance of UNIV_Phish dataset further strengthens our model validity.

Next, we analyze the embedding representation of the emails using t-SNE plot (Maaten and Hinton, 2008). Figure 3 shows that the cloud of misclassified samples is mainly in the overlapped region of phishing and legitimate emails, which indicates the lack of better embedding representation of these emails. However, we observe that the distance between the center of phishing email and the legitimate email cloud increased by adding the PPTs by 9.56%, and 15.29% using Euclidean and Manhattan distance, respectively. Thus, the addition of psychological traits seems to improve the embedding representation.



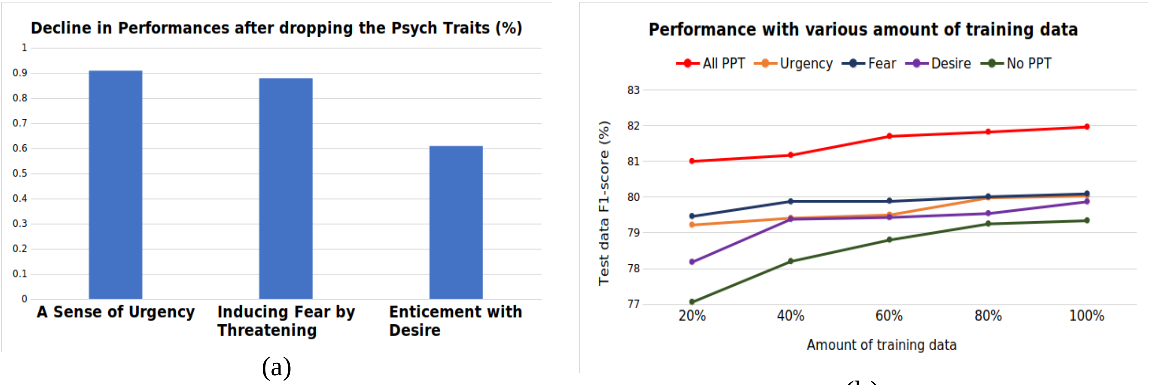
(a)

(b)

Figure 3. t-SNE representation of Phishing and Legitimate email with (a) BERT-based feature only (b) BERT-based + Phishing Psychological Trait features. The figure shows the misclassification zones in blue

We further examine the effectiveness of the Phishing Psychological Traits model by ablation experiments (Meyes et al. 2019). From Figure 4 (a), we observe that the performance decreases the most, when we remove the *Urgency* trait (0.91% in F1-score, 1.01% in accuracy), followed by the *Fear* trait (0.88% in F1-score, 0.99% in accuracy). *Desire* had the least effect on performance (0.60% in F1-score and 0.71% in accuracy), which is consistent with our previous analysis.

Finally, we vary the training data proportion to examine the effect of PPT when we have a small amount of training data. Figure 4(b) shows that while with 100% training data, PPT scores improve the F1-score by 2.62%, with only 20% training data, the F1-score improvement is by 3.94% indicating the effectiveness of PPTs even with insufficient training data. Figure 4(b) also demonstrates the effect of adding PPTs individually. We observe that as a standalone PPT, Fear by Threatening has a better impact on performance than the others. Nevertheless, the three PPTs combined provide the best cue for detecting phishing emails.



(a)

(b)

Figure 4. (a) Decline in performance after dropping the PPTs one at a time (b) Performance in test data at varying proportion of training data

6. CONCLUSION

Quantifying the psychological traits of an email can provide key signals which help improve the phishing email detection performance. In this paper, we define, analyze, and quantify the PPTs that can successfully capture the nuances of an email's intent and show promising results. Hence, our work may provide potential research direction to win the battle against evolving nature of phishing. However, we still have limitations and room for further improvement. First, we will obtain ground truth for the PPTs by labeling them with multiple human raters enabling us to measure the kappa statistics for testing inter-rater reliability, which in turn can provide a more accurate estimation of PPTs. Second, further research might be required to understand the flow of psychological traits in conversational turns to detect more organized phishing than single email-based phishing. Finally, an investigation of how the individual PPTs contribute to forming a phishing email can provide valuable insight that can be utilized for more efficient phishing email detection.

ACKNOWLEDGMENT

Research was supported in part by grants NSF 1838147, ARO W911NF-20-1-0254. The views and conclusions contained in this document are those of the authors and not of the sponsors. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

REFERENCES

Aassal, A. E., Moraes, L. F., Baki, S., Moraes, L., & Verma, R. (2018). Anti-Phishing Pilot at ACM IWSPA 2018 Evaluating Performance with New Metrics for Unbalanced Datasets. *1st Anti-Phishing Shared Task at 4th ACM IWSPA (IWSPA-AP)*. Retrieved from http://ceur-ws.org/Vol-2124/invited_paper_1.pdf

Aggarwal, S., Kumar, V., & Sudarsan, S. D. (2014). Identification and Detection of Phishing Emails Using Natural Language Processing Techniques. *Proceedings of the 7th International Conference on Security of Information and Networks* (pp. 217–222). New York, NY, USA: Association for Computing Machinery. doi:10.1145/2659651.2659691

Agrawal, A., Venkatraman, J., Leonard, S., & Paepcke, A. (2015). YouEDU: Addressing Confusion in MOOC Discussion Forums by Recommending Instructional Video Clips. *EDM*.

Akbar, N. (2014, October). Analysing Persuasion Principles in Phishing Emails. *Analysing Persuasion Principles in Phishing Emails*. Retrieved from http://essay.utwente.nl/66177/

Akoury, N., Wang, S., Whiting, J., Hood, S., Peng, N., & Iyyer, M. (2020). STORIUM: A Dataset and Evaluation Platform for Machine-in-the-Loop Story Generation. *arXiv preprint arXiv:2010.01717*.

Aljofey, A., Jiang, Q., Qu, Q., Huang, M., & Niyigena, J.-P. (2020). An Effective Phishing Detection Model Based on Character Level Convolutional Neural Network from URL. *Electronics, 9*. doi:10.3390/electronics9091514

Almomani, A., Gupta, B. B., Atawneh, S., Meulenberg, A., & Almomani, E. (2013). A Survey of Phishing Email Filtering Techniques. *IEEE Communications Surveys Tutorials, 15,* 2070-2090. doi:10.1109/SURV.2013.030713.00020

Bitaab, M., Cho, H., Oest, A., Zhang, P., Sun, Z., Pourmohamad, R., . . . others. (2021). Scam Pandemic: How Attackers Exploit Public Fear through Phishing. *arXiv preprint arXiv:2103.12843*.

Blanzieri, E., & Bryl, A. (2009). A survey of learning-based techniques of email spam filtering. *Artificial Intelligence Review, 29,* 63-92.

Bose, I., & Leung, A. C. (2009). Technical opinion What drives the adoption of antiphishing measures by Hong Kong banks? *Communications of the ACM, 52,* 141–143.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002, June). SMOTE: Synthetic Minority over-Sampling Technique. *J. Artif. Int. Res., 16,* 321–357.

Cialdini, R. B. (2001). The science of persuasion. *Scientific American, 284,* 76–81.

Cui, X., Ge, Y., Qu, W., & Zhang, K. (2020). Effects of Recipient Information and Urgency Cues on Phishing Detection. In C. Stephanidis, & M. Antona (Ed.), *HCI International 2020 - Posters* (pp. 520–525). Cham: Springer International Publishing.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Minneapolis: Association for Computational Linguistics. doi:10.18653/v1/N19-1423

Ding, K., Pantic, N., Lu, Y., Manna, S., & Husain, M. I. (2015). Towards building a word similarity dictionary for personality bias classification of phishing email contents. *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015),* (pp. 252-259). doi:10.1109/ICOSC.2015.7050815

Fang, Y., Zhang, C., Huang, C., Liu, L., & Yang, Y. (2019). Phishing Email Detection Using Improved RCNN Model With Multilevel Vectors and Attention Mechanism. *IEEE Access, 7,* 56329-56340. doi:10.1109/ACCESS.2019.2913705

FBI. (2021). *Internet Crime Report 2020.* Retrieved from https://www.ic3.gov/Media/PDF/AnnualReport/2020_IC3Report.pdf

Feng, W., Sun, J., Zhang, L., Cao, C., & Yang, Q. (2016). A support vector machine based naive Bayes algorithm for spam filtering. *2016 IEEE 35th International Performance Computing and Communications Conference (IPCCC),* (pp. 1–8).

Ferreira, A., & Lenzini, G. (2015). An analysis of social engineering principles in effective phishing. *2015 Workshop on Socio-Technical Aspects in Security and Trust,* (pp. 9-16). doi:10.1109/STAST.2015.10

Gansterer, W., & Pölz, D. (2009, April). E-Mail Classification for Phishing Defense., (pp. 449-460). doi:10.1007/978-3-642-00958-7_40

Goldberg, L. R. (1982). From Ace to Zombie: Some explorations in the language of personality. *Advances in personality assessment, 1,* 203–234.

Halevi, T., Lewis, J., & Memon, N. (2013). A Pilot Study of Cyber Security and Privacy Related Behavior and Personality Traits. *Proceedings of the 22nd International Conference on World Wide Web* (pp. 737–744). New York, NY, USA: Association for Computing Machinery. doi:10.1145/2487788.2488034

Jakobsson, M. (2007). The human factor in phishing. *Privacy & Security of Consumer Information, 7,* 1–19.

Jones, H. S., Towse, J. N., Race, N., & Harrison, T. (2019). Email fraud: The search for psychological predictors of susceptibility. *PloS one, 14,* e0209684.

Kejriwal, M., & Zhou, P. (2019). Low-Supervision Urgency Detection and Transfer in Short Crisis Messages. *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 353–356). New York, NY, USA: Association for Computing Machinery. doi:10.1145/3341161.3342936

Khonji, M., Iraqi, Y., & Jones, A. (2013). Phishing Detection: A Literature Survey. *IEEE Communications Surveys Tutorials, 15,* 2091-2121. doi:10.1109/SURV.2013.032213.00009

Lee, Y., Saxe, J., & Harang, R. (2020). CATBERT: Context-Aware Tiny BERT for Detecting Social Engineering Emails. *CATBERT: Context-Aware Tiny BERT for Detecting Social Engineering Emails*.

Li, Q., Cheng, M., Wang, J., & Sun, B. (2020). LSTM based Phishing Detection for Big Email Data. *IEEE Transactions on Big Data,* 1-1. doi:10.1109/TBDATA.2020.2978915

Luo, X. (., Zhang, W., Burd, S., & Seazzu, A. (2013). Investigating phishing victimization with the Heuristic–Systematic Model: A theoretical framework and an exploration. *Computers & Security, 38,* 28-38. doi:https://doi.org/10.1016/j.cose.2012.12.003

Maaten, L. V., & Hinton, G. E. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research, 9,* 2579-2605.

Meyes, R., Lu, M., Puiseau, C. W., & Meisen, T. (2019). Ablation Studies in Artificial Neural Networks. *ArXiv, abs/1901.08644.*

Microsoft (2021) *The quiet evolution of phishing*. Available at: https://www.microsoft.com/security/blog/2019/12/11/ the-quiet-evolution-of-phishing (accessed August 20, 2021).

Naidoo, R. (2015, February). Analysing urgency and trust cues exploited in phishing scam designs *in 10th International Conference on Cyber Warfare and Security (p. 216).*

Ordonez, L., & Benson III, L. (1997). Decisions under time pressure: How time constraint affects risky decision making. *Organizational Behavior and Human Decision Processes, 71*, 121–140.

Park, G., & Taylor, J. M. (2015). Using Syntactic Features for Phishing Detection. *CoRR, abs/1506.00037*. Retrieved from http://arxiv.org/abs/1506.00037

Parrish Jr, J. L., Bailey, J. L., & Courtney, J. F. (2009). A personality based model for determining susceptibility to phishing attacks. *Little Rock: University of Arkansas*, 285–296.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., & others. (2019). Language models are unsupervised multitask learners. *OpenAI blog, 1*, 9.

Reimers, N., & Gurevych, I. (2019, November). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics. Retrieved from https://arxiv.org/abs/1908.10084

Salloum, S., Gaber, T., Vadera, S., & Shaalan, K. (2021). Phishing Email Detection Using Natural Language Processing Techniques: A Literature Survey. *Procedia Computer Science, 189,* 19-28. doi:https://doi.org/10.1016/j.procs.2021.05.077

SBERT (2021) *SBERT Pretrained Models.* Available at: https: //www.sbert.net/docs/pretrained models.html (accessed July 29, 2021).

Shahriar, S., Mukherjee, A., & Gnawali, O. (2021, September). A Domain-Independent Holistic Approach to Deception Detection. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021).* (pp. 1308-1317)

Sharma, T., & Bashir, M. (2020). An Analysis of Phishing Emails and How the Human Vulnerabilities are Exploited. In I. Corradini, E. Nardelli, & T. Ahram (Ed.), *Advances in Human Factors in Cybersecurity* (pp. 49–55). Cham: Springer International Publishing.

Sheng, S., Wardman, B., Warner, G., Cranor, L., Hong, J. I., & Zhang, C. (2009). An Empirical Analysis of Phishing Blacklists. *CEAS 2009.*

Stajano, F., & Wilson, P. (2011, March). Understanding Scam Victims: Seven Principles for Systems Security. *Commun. ACM, 54*, 70–75. doi:10.1145/1897852.1897872

Thai-Nghe, N., Gantner, Z., & Schmidt-Thieme, L. (2010). Cost-sensitive learning methods for imbalanced data. *The 2010 International Joint Conference on Neural Networks (IJCNN),* (pp. 1-8). doi:10.1109/IJCNN.2010.5596486

van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research, 9,* 2579–2605. Retrieved from http://www.jmlr.org/papers/v9/vandermaaten08a.html

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is All you Need. *ArXiv, abs/1706.03762*.

Verma, R. M., & Hossain, N. (2013). Semantic Feature Selection for Text with Application to Phishing Email Detection. *ICISC.*

Verma, R. M., Zeng, V., & Faridi, H. (2019). Data Quality for Security Challenges: Case Studies of Phishing, Malware and Intrusion Detection Datasets. *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security* (pp. 2605–2607). New York, NY, USA: Association for Computing Machinery. doi:10.1145/3319535.3363267

Vishwanath, A., Herath, T., Chen, R., Wang, J., & Rao, H. R. (2011). Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model. *Decision Support Systems, 51*, 576-586. doi:https://doi.org/10.1016/j.dss.2011.03.002

Wang, J., Herath, T., Chen, R., Vishwanath, A., & Rao, H. R. (2012). Research article phishing susceptibility: An investigation into the processing of a targeted spear phishing email. *IEEE transactions on professional communication, 55*, 345–362.

Williams, E. J., Beardmore, A., & Joinson, A. N. (2017). Individual differences in susceptibility to online influence: A theoretical review. *Computers in Human Behavior, 72*, 412-421. doi:https://doi.org/10.1016/j.chb.2017.03.002

Workman, M. (2008, February). Wisecrackers: A Theory-Grounded Investigation of Phishing and Pretext Social Engineering Threats to Information Security. *J. Am. Soc. Inf. Sci. Technol., 59,* 662–674.

Zhang, N., & Yuan, Y. (2012). Phishing Detection Using Neural Network.

Zhang, X., Zhao, J. J., & LeCun, Y. (2015). Character-level Convolutional Networks for Text Classification. *CoRR, abs/1509.01626.* Retrieved from http://arxiv.org/abs/1509.01626