Theory-Guided Feature Selection in Cybercrime Data Science

Rennie Naidoo and Shiven Naidoo
University of the Witwatersrand, Johannesburg, South Africa

Rennie.Naidoo@wits.ac.za
shiven.naidoo@students.wits.ac.za

Abstract: Cybercrime data science is being significantly hampered by the presence of 'noisy' features within vast and complex datasets. We draw from the theoretical insights of the behavioural sciences to propose a feature selection model to enrich and improve the value and interpretability of cybercrime intelligence datasets. We piloted our theory-guided feature selection approach on a subset of intelligence datafeeds provided by a global fraud and cybercrime tracking firm. The results of the proposed social influence feature selection model show significant improvement in the interpretability of the machine learning-based exploratory analysis and advanced visualization techniques in an experimental setting. The feature selection model yielded rich insights about cybercriminal psychological tactics from social engineering scam data and has potential applicability in the areas of cyberthreat response and cybersecurity awareness training. Our study shows the value of an interdisciplinary theory-guided approach to cybercrime data analytics that integrates scientific knowledge from the behavioural sciences and data science expertise. Our paper concludes by suggesting avenues for future research on theory-guided feature selection seeking to incorporate behavioural science knowledge in cybercrime data science. We intend to refine, automate, evaluate, and scale our model in future research to assess its effectiveness in producing insights about cybercriminal activities and informing decision-making in a naturalistic and real-time setting. In future research efforts, we aim to automate the encoding of features and apply a wider range of machine learning tools and evaluation metrics to extract more meaningful insights into cybercriminal psychological tactics. We also intend to refine our model on larger datasets to enhance its efficiency and responsiveness to real-time cybercrime data. We call on data scientists and cybercrime domain experts to work together to apply theory-guided feature selection to improve processes of knowledge discovery that enhance our cybersecurity capabilities.

Keywords: Behavioural science, Feature selection, Cybercrime data analytics, Cybersecurity, Theory-Guided data science

1. Introduction

We are drowning in information and starving for knowledge.

–Rutherford D. Roger

Recent advances in data science and big data analytics (BDA) holds a lot of promise in finding innovative ways to tackle the scourge of cybercrime (Kamenov, 2018; Rawat et al., 2021; Sarker et al., 2020). While cybersecurity units and their data science teams receive vast amounts of data from various external and internal sources to monitor and respond to cyberincidents and cyberthreats (Suraj et al., 2018; Wang and Jones, 2021), the noise content in this information can adversely affect the quality of analysis and lead to misleading conclusions if not properly addressed (Gupta and Gupta, 2019; Zhu et al., 2003). We agree with scholars that the noise challenges in these datasets are further exacerbated by the spurious relationships detected in the data-driven models that have little or no scientific basis (Zhu and Wu, 2004).

Despite rapid advancements in computing power, databases, statistical methods and algorithms that can process and analyse large volumes of structured and unstructured data (Chen et al., 2012; Hastie et al., 2001; McAfee et al., 2012), we agree with scholars that much of the popular hype about the 'end of theory' and 'the death of the scientific method' in the popular press as a consequence of data science has been hasty (Karpatne et al., 2017). The concept of theory-guided data science (TGDS) suggests a complementary approach that seeks to integrate scientific knowledge and data science thereby uniting the world of data scientists and domain experts (Bechor and Jung, 2019; Downton et al., 2020; Viaene, 2013). This approach values the existing concepts, models, measures, and hypotheses that form part of the existing scientific knowledge base that explains a phenomenon and examines their fit with the available data instead of over relying on often spurious correlations (Wagner and Rondinelli, 2016). At the same time, theory-based models are susceptible to making assumptions that oversimplify complex phenomena (Downton et al., 2020; Faghmous and Kumar, 2014). Notwithstanding, according to Karpatne et al. (2017) theory-guided data science (TGDS) "uses the unique capability of data science models to automatically learn patterns and models from large data, without ignoring the treasure of accumulated scientific knowledge." This necessitates more effective inter-disciplinary collaborations between data scientists and domain experts (Mao et al., 2019). Similarly, we also believe that collaborations that combine theory-based models and data science models can significantly expand the potential of generating useful insights from complex cybercrime intelligence datasets.

There is a tremendous opportunity to jointly apply theory-based models and data science models to analyse cybercrime intelligence datasets, to more effectively detect, prevent, and respond to cyber threats and cybercrime (Foroughi and Luksch, 2018). Given the deluge of cybercrime data and the high noise potential in these large, diverse and complex cybercrime intelligence datasets, traditional data science and big data analytics on their own are arguably not sufficient to provide reliable and valuable insights into decision making (Rassam et al., 2017). Scam data involving social engineering techniques, where scammers manipulate individuals into divulging sensitive information, contains a significant amount of noisy content that does not contribute to successful scams but forms a substantial part of the overall dataset (Gupta and Gupta, 2019; Naidoo, 2020). As a subcomponent of TGDS, theory-guided feature selection (TGFS) is the process of incorporating theoretical advances from the problem domain to select relevant features for TGDS models (Bhowmik and Bhowmik, 2022; Miao and Niu, 2016; Slawski et al., 2010). By leveraging existing theories, concepts, principles, and expert insights, behavioural scientists and data scientists can work together to apply theory-guided feature selection to reduce noise in complex cybersecurity datasets (Mao et al., 2019). For example, behavioural scientists can apply theories and principles to recognize the relevant features in scam content. However, discerning relevant features from vast and diverse cybercrime datasets to develop more effective countermeasures remains a major challenge (Heartfield et al., 2016). This pilot study embarks on the exploration of theory-guided feature selection techniques on a small subset of a cybercrime intelligence datafeed provided by a global fraud and cybercrime company tracking cybercrime incidents. More specifically, we assess the utility of social influence theory in filtering out noise and focusing on the relevant content of social engineering scams, streamlining the dataset for analysis and interpretation (Cialdini and Goldstein, 2004; Wright et al., 2014). Thus, this study addresses the following question: *How can the application of theory-guided feature selection techniques using social influence concepts and principles enhance the interpretability of cybercrime intelligence feeds containing social engineering content?*

This paper is organized as follows: Section 2 provides a brief literature review of the proposed social influence theory-guided feature selection approach, highlighting the key social influence concepts and principles used in our feature selection process. Section 3 details our pilot case study methodology for feature engineering the cybersecurity intelligence datafeeds, detailing our collaboration architecture and process for theory-guided feature selection. Section 4 presents the results of our exploratory data analysis (EDA) and visualization of the social influence features and discusses their implications. While the findings from these results are limited to a pilot study, they will serve as a foundation for more extensive research in the future. Section 5 concludes the pilot study, highlighting the contributions of theory-guided feature selection to the field of cybercrime analytics and avenues for future research. This study contends that theory-guided feature selection techniques, grounded in domain-specific knowledge of the behavioural sciences, are instrumental in advancing cybercrime data science.

2. Social Influence Theory-Guided Feature Selection

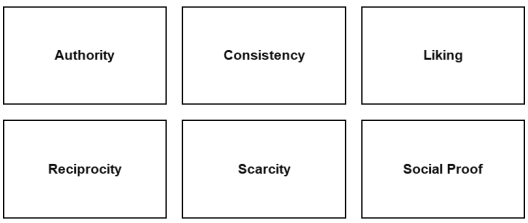

Figure 1: The Social Influence Feature Set

Cialdini's six persuasion principles, rooted in the psychology of compliance, serve as the primary knowledge base for the proposed theory-guided feature selection model, depicted as the social influence feature set in Figure 1. Drawing upon the concepts and principles of authority, consistency, liking, scarcity, reciprocity, and social proof (Cialdini and Goldstein, 2004; Wright et al., 2014), we craft our theory-guided data science model.

In Table 1 we show how our theory-guided feature selection process recasts the abovementioned social influence principles and definition as features and feature marker patterns respectively. The first feature is 'Authority,' and the feature pattern is where individuals tend to unquestioningly accept statements from perceived authorities (Milgram, 1963). 'Consistency' signifies the strong human inclination to align actions with commitments (Ferreira et al., 2015). 'Scarcity' heightens the value of items limited in availability, generating

increased demand. 'Reciprocation' captures the human tendency to feel obliged to return favours (Ferreira et al., 2015). 'Liking' highlights friendship pressures and people's readiness to assist friends, especially in online interactions, showcasing cooperative behaviour patterns (Ferreira et al., 2015). The last feature in our model, 'Social Proof' illustrates the impact of similar others on individual behaviour, leading to the adoption of similar actions or choices (Stajano and Wilson, 2011). Our goal is to enhance the interpretability of complex social engineering scams using the proposed social influence feature set model, to provide richer insights into cybercriminal tactics.

Table 1: Social Influence Feature Set Model with Feature Flags

Features	Feature Marker Patterns	Flag
Authority (A)	I follow the advice and obey the instructions of authority figures.	Present: (yes=1, no=0)
Consistency (C)	Once I have made a commitment, I do it.	Present: (yes=1, no=0)
Liking (L)	I like the agent therefore I am more inclined to believe them.	Present: (yes=1, no=0)
Reciprocity (R)	When the agent does me a favour, I feel inclined to return this favour.	Present: (yes=1, no=0)
Scarcity (S)	If the opportunity is for a limited time, I would act quickly to receive it.	Present: (yes=1, no=0)
Social Proof (SP)	If other people complied, I am more likely to comply.	Present: (yes=1, no=0)

3. Pilot Study: Social Influence Feature Engineering for Cybercrime Datasets

We used a small subset of data received from FraudWatch International's Cyber Intelligence Datafeed. These datafeeds were collected over a one-month period from March to April, 2020. Table 1 shows how the feature selection process began with the development of a coding template. The template served as a structured guide for identifying relevant features. This systematic approach not only ensured consistency in the analysis but also facilitated a deductive reasoning process, allowing for the targeted selection of features. Figure 2 shows how the researchers analysed photos and images iteratively. 185 unique documents were analysed. In the process of applying our theory-guided feature selection model, we implemented feature flags or binary indicators to operationalize the manually selected features. This critical step involved assigning a binary value of 1 (indicating 'yes') to each of the manually selected features, effectively flagging their presence within each scam. Conversely, a value of 0 (indicating 'no') was assigned to features not selected, creating a distinct flag variable for each underlying social influence principle (Li et al., 2018; Raschka et al., 2020). The authors independently coded the content and achieved consensus in case of any discrepancies. A cybersecurity expert, not familiar with project, acted as an independent auditor and reviewed the key categories and operational definitions, and provided reasonable verification of the accuracy of the coding procedure. The data underwent classification based on the specific online technologies affected, including but not limited to email, various social media platforms, video telephony and online chat services, cloud file hosting platforms, and streaming media services. Furthermore, categorization was performed according to the types of organizations subjected to impersonation attempts, encompassing entities such as banks, technology brands, product labels, and government agencies.

Code	Features	Binary Encoding
A	Authority	1
C	Consistency	0
L	Liking	1
R	Reciprocity	0
S	Scarcity	1
SP	Social Proof	1

Figure 2: Manually Selected Social Influence Features

Our collaboration architecture in Figure 3 provides an overview of the theory-guided feature selection components for analysing cybercrime datasets. The architecture typically begins with intelligence from both internal and external cybercrime feeds. This data is then subjected to data ingestion and preprocessing. The insights from both domain experts and data scientists are used for feature identification and engineering. Subsequently, the selected features are collated into a dataset, which then undergoes modelling and analysis using machine learning models and evaluation tools. The final step of the workflow involves generating documentation and reports. Throughout the process, there is a continuous cycle of monitoring and refinement, ensuring that the models remain updated and relevant. As illustrated, a collaborative platform and shared workspace serve as the backbone for theory-guided data science. Figure 4 depicts a phased approach to data analysis comprising three distinct phases. In Phase 1, the focus lies on preparing the feature-selected dataset. Phase 2 is the exploratory data analysis stage. In Phase 3, known as training and evaluation, which lies beyond the scope of this project, will involve the application of automated selection techniques to train the models.

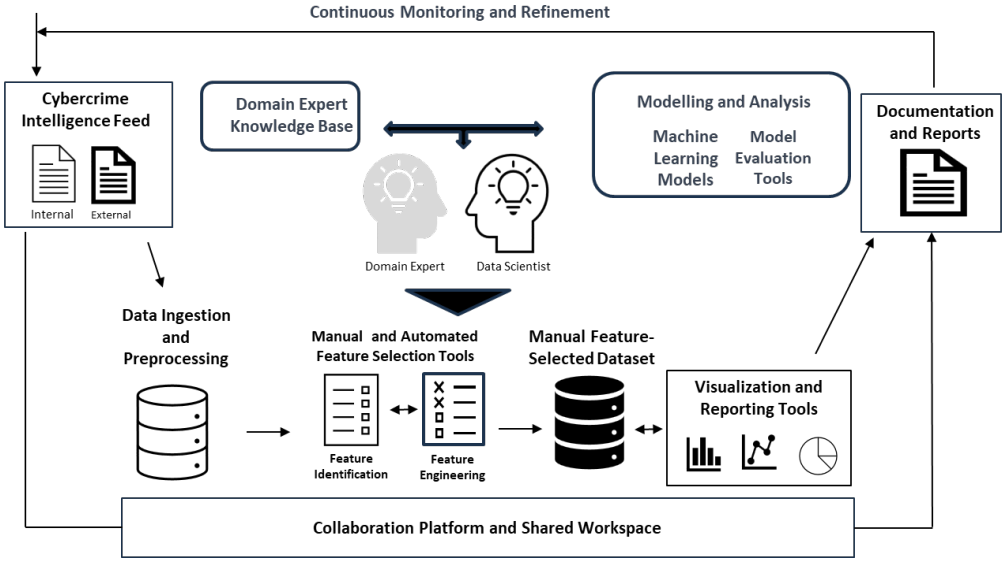

Figure 3: Collaboration Architecture for Theory-Guided Feature Selection

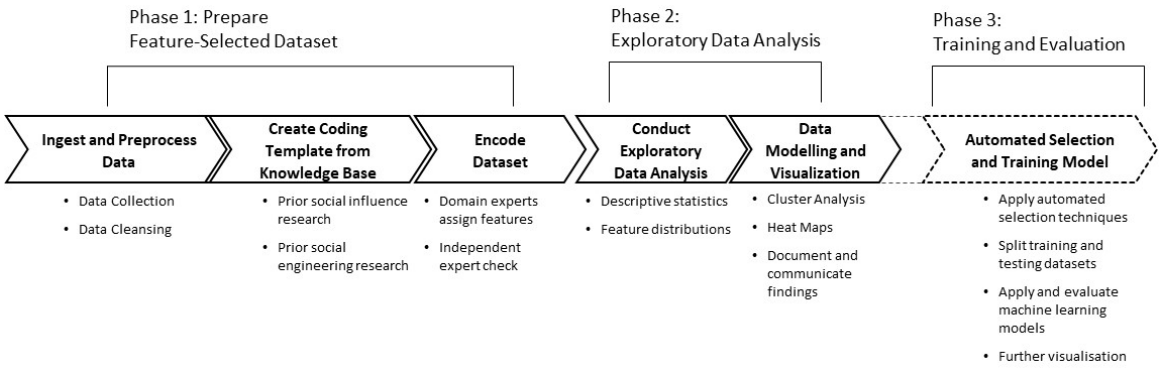

Figure 4: Manual Feature Selection Process

Python 3.0 was used for data processing, exploratory data analysis, clustering and visualizations. First, feature and scam category distributions were visualized using tools from the matplotlib library. The sklearn library was then used to perform and evaluate agglomerative clustering of the six social influence features. Next, the seaborn library was used to create a heatmap visualizing the usage of social influence features in different scam types. Lastly, the sklearn library was used to visualize the clustering of scam types based on their usage of social influence features using a t-SNE (t-distributed Stochastic Neighbor Embedding) map (Hastie et al., 2001; Raschka et al., 2020). We acknowledge that the lack of completeness of the dataset is a limitation of this study. In some cases, only phishing website information was provided and not the accompanying phishing email. Furthermore, it was difficult to assess to what extent the phishing email and phishing website were using the consistency and reciprocity tactics. Moreover, the archive documents were limited to active scams.

Notwithstanding, our approach ensured clear demarcation and identification of the social influence features, laying the groundwork for the analytical processes and results, discussed next.

4. Pilot Results and Discussion

4.1 Exploratory Data Analysis

Theory-guided EDA is important for understanding the distribution of characteristics for selected categories and features within a dataset. Figure 5 explores the distribution of 185 scams annotated according to two categories: Scam Types and Organization Types. Phishing was the dominant scam type, making up 60.9% of scams from the cybersecurity intelligence datafeed. The most common types of organizations impersonated by scammers included Social Networking Sites (SNSs) (39.1%), banks (23.9%), and technology organizations (21.2%). In contrast, there were relatively fewer scams impersonating government (8.2%) and intergovernmental (2.2%) organisations. The distributions of scam types and organisation types reveals important insights into scammer behaviour. Firstly, scammers tend to impersonate highly reputable institutions, as shown by the majority of scams impersonating banks, tech firms, government and intergovernmental organisations. Secondly, scammers may be intentionally impersonating larger international organisations in order to direct their scams to a wider, more global victimization base. This is supported by the large number of scams impersonating SNSs, technology organizations and banks.

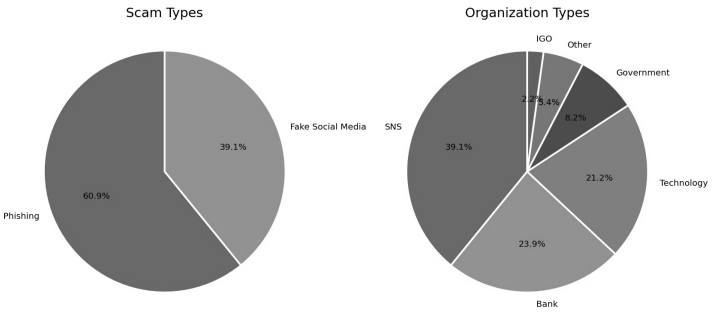

Figure 5: EDA; Distributions for Scam Categories

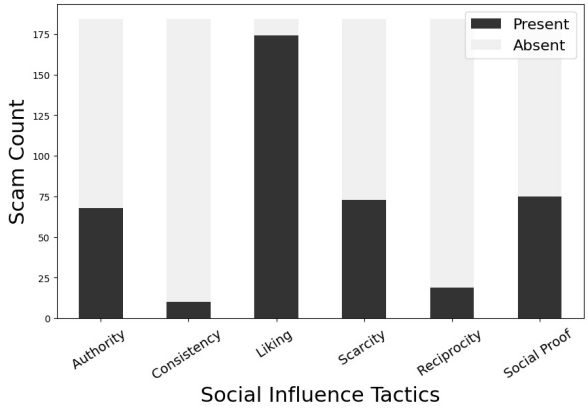

Figure 6: Prevalence of Social Influence Features in Scams

Figure 6 shows the prevalence of Cialdini's six social influence features in scams. Liking was the most employed tactic by scammers. Its widespread use suggests that scammers may primarily deceive users by impersonating people and institutions who are familiar or trustworthy, which is reflected in the types of organizations predominantly used by scammers in Figure 5. In contrast to liking, features like consistency and reciprocity were very rarely observed in scams. However, the cybercrime intelligence datafeed contained mostly active scams. Thus, there was little data showing scammer-user interactions, where these tactics are more likely to be employed. The remaining features, authority, scarcity and social proof, were all used in similar proportions but were still employed significantly less than liking.

4.2 Visualization of Social Influence Features

Visualizations were used to explore the relationships between selected categories and features (Embarak, 2018). Figure 7 shows the presence (1) or absence (0) of social influence features in scams. While liking seems

to be used ubiquitously, other features like authority, scarcity and social proof appear to be associated with specific scam types. For example, both authority and scarcity are commonly used in phishing scams. In contrast, social proof is commonly used in fake social media scams. Furthermore, consistency and reciprocity appear very rarely, but when they did appear it was almost always in phishing scams. When comparing the joint use of social influence features, it can be seen that authority and scarcity are often used together. Notably, social proof seems to be mutually exclusive with most social influence features, with the exception of liking.

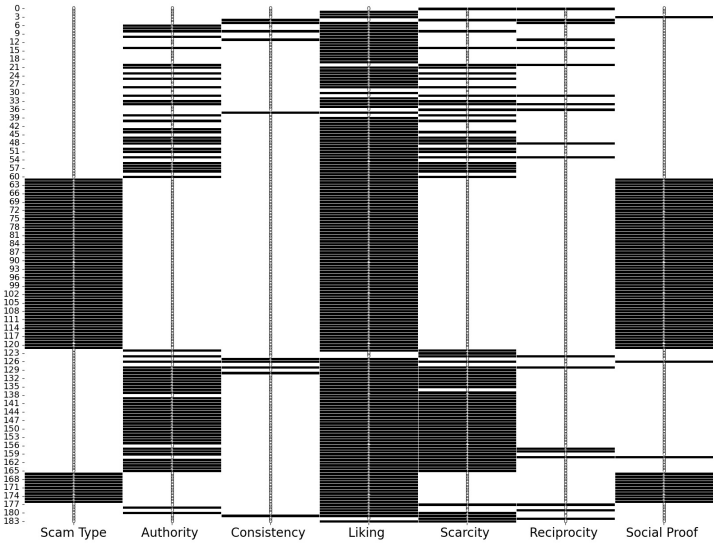

Figure 7: Heatmap showing Social Influence Features by Scam Types

Figure 8 below investigates the relationships between social influence features further using agglomerative clustering. The plot to the left shows the silhouette score for different n_cluster values when tuning the agglomerative clustering algorithm. Six clusters were used to optimise the clustering algorithm as it had the greatest silhouette score, suggesting the best cluster quality. To accommodate the data's binary encoding, Jaccard distance was used to measure the dissimilarity between clusters. Thus, clusters with Jaccard scores closer to 0 represent features which often employed in the same scams, while Jaccard scores closer to 1 represent features which are rarely employed alongside other features.

$$Jaccard's\ Distance = \frac{(A \cup B) - (A \cap B)}{A \cup B}$$

The hierarchical clustering reveals valuable insights into the combinations of social influence features scammers employ. Notably, scarcity and authority are closely clustered, suggesting their concurrent use by cybercriminals. In contrast, reciprocity and consistency clustered with high Jaccard distances. It is likely that the low number of scams using these tactics greatly penalized their clustering. More comprehensive datasets may see more prevalent use of these features and generate more useful insights into their relationships with other social influence features.

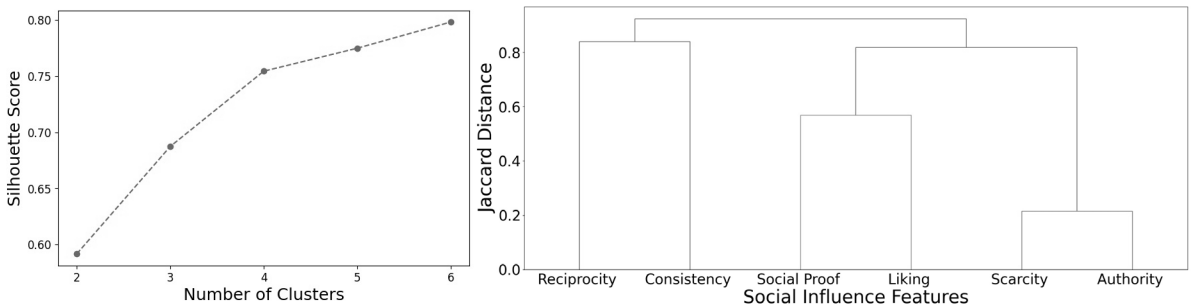

Figure 8: Agglomerative Clustering of Social Influence Features

Figure 9 shows a t-distributed Stochastic Neighbour Embedding (t-SNE) map, which is a valuable tool for mining insights from highly dimensional datasets. Here, t-SNE has been used to reduce the dimensions for the presence and absence of all six social influence tactics into two dimensions. Due to the binary nature of the

encoded data, there are many overlapping datapoints, represented by darker icons. This visualization reveals two key insights into the relationship between social influence tactics and scam types. Firstly, scam types form two distinct clusters, suggesting that the scam type informs which social influence tactics cybercriminals will employ. Secondly, phishing scams use a more complex mixture of social influence tactics compared to fake social media. This is shown by the tight clustering shown by fake social media scams compared to a widely spread distribution for phishing scams.

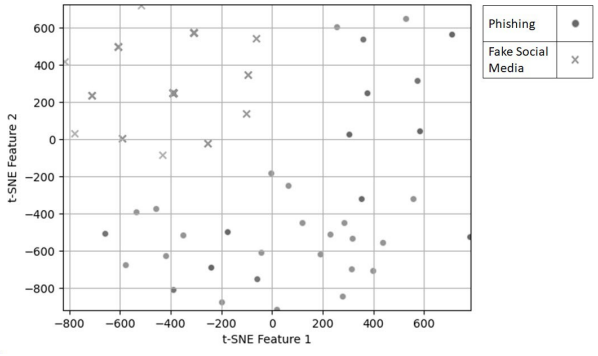

Figure 9: Dimensionality Reduction of Social Influence Features

4.3 Theoretical, Practical and Methodological Implications

Similar to prior studies on cybercrime, using the social influence concepts principals as guiding features for data analysis revealed various insights into psychological tactics employed by cybercriminals from a noisy cybersecurity dataset (Ferreira et al., 2015; Wright et al., 2014). Our study also suggests that integrating features developed by behavioural scientists and techniques from data science can produce high quality and usable insights from noisy scam datasets (Karpatne et al., 2017). In our case, selecting features from the social influence model enriched and improved the value and interpretability of our cybercrime intelligence dataset. For instance, the close clustering of scarcity and authority we found suggests that cybercriminals have a sophisticated understanding of human psychology and are seeking to both exploit the victim's perceived rarity of opportunities (scarcity) and manipulate their perceived trust in certain institutions and personalities (authority). Moreover, cybercriminals employing SNS scams focus their deception efforts on appearing familiar and similar to potential victims so that they appear to be friends (liking). These insights can be used to bolster the security posture of organizations. For instance, our theory-guided feature selection can play a crucial role in offender behaviour analytics (OBA) by enabling analysts to predict the type of cybercriminal social engineering activities from emerging threat patterns that can better inform cyberthreat response (Bhowmik and Bhowmik, 2022). For instance, by prioritising the most prevalent features at a point in time, cybersecurity experts can develop more robust techniques to identify unusual activities and initiate timely incident responses. Moreover, they can use the insights from feature selection to shape cybersecurity awareness training programs. By identifying common attack vectors and common tactics employed by cybercriminals, training can become more targeted and directly applicable to the most recent cybersecurity challenges faced by individuals and organizations. For example, our cluster analysis findings illustrated above suggest that awareness initiatives should highlight the interplay between scarcity and authority, by emphasising common scenarios where these tactics are employed. Educating users about these dual tactics can empower them to recognize and resist these types of manipulative social engineering schemes. We also introduced a robust collaboration architecture that outlines a systematic approach to theory-guided feature selection in the analysis of cybercrime datasets. This framework integrates insights from domain experts and data scientists, ensuring a holistic exploration of internal and external cybercrime feeds (Mao et al., 2019). Furthermore, we also propose a phased approach to feature selected data analysis, which includes preparation of the feature-selected dataset, exploratory data analysis where key patterns and insights are uncovered, and although not within the scope of this study, the application of automated selection techniques for model training and evaluation (Hastie et al., 2001). As found in other disciplines, we believe that theory-guided feature selection will contribute to the accuracy and effectiveness of machine learning models and that models trained on theory-guided features will generally outperform generic data science models for complex cyber intelligence datasets (Faghmous and Kumar, 2014; Karpatne et al., 2017). Finally, our findings underscore the need for multifaceted countermeasures in cybersecurity that focus not only on technical defences but also incorporate psychological and behavioural approaches.

5. Conclusions and Future Work

In this pilot study, we explored the feasibility of theory-guided feature selection in cybercrime data science. One of the challenges of cybercrime intelligence datafeeds and other vast and complex big data sets used in cybersecurity intelligence and analytics is the high severity of 'noisy' features. Our investigation suggests that integrating behavioural science theories to identify and emphasize features that are theoretically relevant to the problem at hand can reduce the noise levels in cybercrime data science models. Towards this end, we applied the psychological and behavioural insights of social influence theory to enrich and improve the value and interpretability of a dataset of scams involving complex social engineering techniques (Cialdini and Goldstein, 2004; Ferreira et al., 2015). Despite the small scale of our analysis, the improved interpretability of our dataset suggests that the proposed methodology could be further explored and refined in future research with larger datasets. We observed the noticeable relevance of selected features such as liking and social proof and their relationship with scam types, such as phishing and fake social media scams. Cybersecurity practitioners can apply the theory-guided feature selection approach to gain insights about cybercrime patterns and apply these to enhance the effectiveness of cyberthreat response and cybersecurity awareness training (Korpela, 2015). In offender behaviour analytics (OBA), feature selection methods can also help identify the social engineering modus operandi that are relevant to cybercrime victimization for different scam types. Our approach also underscores the importance of fostering collaboration between behavioural scientists and data scientists in cybersecurity data science teams to enhance data-driven decision-making processes. This study builds upon prior research by emphasizing the need for theory-guided cybersecurity data science. While the existing cybersecurity literature acknowledges the importance of feature selection, our contribution lies in the systematic integration of social influence theory to scam data involving social engineering techniques to enhance the overall interpretability and predictive power of cybercrime data science models. One of the main limitations of our pilot study is the notably small size and scope of our dataset which limits the generalizability of our results. Another limitation is that our study is also dependent on our expert understanding of social influence theory which is also evolving. Moreover, the binary encoding of features restricted the types of data science models that we could apply. Future research could use scales with increased granularity to encode the social influence tactics in scams so that the data could be analysed and visualized using more robust techniques such as linear regression and PCA, to provide more detailed and nuanced insights. These limitations highlight the need for continued refinement and adaptation of our social influence feature selection methodology. In future research, we intend to develop an automated pipeline for identifying and encoding social influence features from cybercrime intelligence datafeeds and scam images. We will refine and train our model on larger datasets, using automated feature selection algorithms and test the ranking and importance of our features as well as model performance (Raschka et al., 2020). Our model will also be expanded to use a wider range of machine learning tools and evaluation metrics to extract more meaningful insights from cybercrime intelligence datafeeds. We call for the further development of theory-guided data science methods that harnesses the contextual understanding and explanatory power of the behavioural sciences with the predictive power of data science to improve decision-making in cybersecurity. We hope that this pilot study will serve as a stepping stone for the further exploration of theory-guided data science and theory-guided feature selection to fruitfully harvest the ever-expanding, increasingly complex and noisy cybercrime and cybersecurity datasets.

Acknowledgements

This work has been supported by the National Research Foundation (NRF).

References

Bechor, T., Jung, B., 2019. Current State and Modeling of Research Topics in Cybersecurity and Data Science. Systemics, Cybernetics and Informatics 17, 129–156.

Bhowmik, P., Bhowmik, P.C., 2022. A Machine Learning Approach for Phishing Websites Prediction with Novel Feature Selection Framework, in: Hossain, S., Hossain, Md.S., Kaiser, M.S., Majumder, S.P., Ray, K. (Eds.), Proceedings of International Conference on Fourth Industrial Revolution and Beyond 2021, Lecture Notes in Networks and Systems. Springer Nature, Singapore, pp. 357–370.

Chen, Chiang, Storey, 2012. Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly 36, 1165–1188.

Cialdini, R.B., Goldstein, N.J., 2004. Social Influence: Compliance and Conformity. Annual Review of Psychology 55, 591–621.

Downton, J.E., Collet, O., Hampson, D.P., Colwell, T., 2020. Theory-guided data science-based reservoir prediction of a North Sea oil field. The Leading Edge 39, 742–750.

Embarak, Dr.O., 2018. Data Analysis and Visualization Using Python: Analyze Data to Create Visualizations for BI Systems. Apress, Berkeley, CA.

Faghmous, J.H., Kumar, V., 2014. A Big Data Guide to Understanding Climate Change: The Case for Theory-Guided Data Science. Big Data 2, 155–163.

Ferreira, A., Coventry, L., Lenzini, G., 2015. Principles of Persuasion in Social Engineering and Their Use in Phishing, in: Tryfonas, T., Askoxylakis, I. (Eds.), Human Aspects of Information Security, Privacy, and Trust, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 36–47.

Foroughi, F., Luksch, P., 2018. Data Science Methodology for Cybersecurity Projects, in: Computer Science & Information Technology. Presented at the 5th International Conference on Artificial Intelligence and Applications, Academy & Industry Research Collaboration Center (AIRCC), pp. 01–14.

Gupta, S., Gupta, A., 2019. Dealing with Noise Problem in Machine Learning Data-sets: A Systematic Review. Procedia Computer Science 161, 466–474.

Hastie, T., Friedman, J., Tibshirani, R., 2001. The Elements of Statistical Learning, Springer Series in Statistics. Springer, New York, NY.

Heartfield, R., Loukas, G., Gan, D., 2016. You Are Probably Not the Weakest Link: Towards Practical Prediction of Susceptibility to Semantic Social Engineering Attacks. IEEE Access 4, 6910–6928.

Kamenov, D., 2018. Intelligent Methods for Big Data Analytics and Cyber Security. ISIJ 39, 255–262. https://doi.org/10.11610/isij.3921

Karpatne, A., Atluri, G., Faghmous, J.H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., Kumar, V., 2017. Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data. IEEE Trans. Knowl. Data Eng. 29, 2318–2331.

Korpela, K., 2015. Improving Cyber Security Awareness and Training Programs with Data Analytics. Information Security Journal: A Global Perspective 24, 72–77.

Li, C., Xie, W., Zhou, K., 2018. Efficient Binary-Encoding Access Control Policy Combination for Large-Scale Collaborative Scenarios, in: 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE). Presented at the 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), pp. 560–566.

Mao, Y., Wang, D., Muller, M., Varshney, K.R., Baldini, I., Dugan, C., Mojsilović, A., 2019. How Data ScientistsWork Together With Domain Experts in Scientific Collaborations: To Find The Right Answer Or To Ask The Right Question? Proc. ACM Hum.-Comput. Interact. 3, 1–23.

McAfee, A., Brynjolfsson, E., Davenport, T.H., Patil, D.J., Barton, D., 2012. Big Data: The Management Revolution. Harvard business review 90, 60–68.

Miao, J., Niu, L., 2016. A Survey on Feature Selection. Procedia Computer Science 91, 919–926.

Milgram, S., 1963. Behavioral Study of obedience. The Journal of Abnormal and Social Psychology 67, 371–378.

Naidoo, R., 2020. A multi-level influence model of COVID-19 themed cybercrime. European Journal of Information Systems 29, 306–321.

Raschka, S., Patterson, J., Nolet, C., 2020. Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence. Information 11, 193.

Rassam, M.A., Maarof, M.A., Zainal, A., 2017. Big Data Analytics Adoption for Cyber- security: A Review of Current Solutions, Requirements, Challenges and Trends. Journal of Information Assurance and Security. 11, 124–145.

Rawat, D.B., Doku, R., Garuba, M., 2021. Cybersecurity in Big Data Era: From Securing Big Data to Data-Driven Security. IEEE Trans. Serv. Comput. 14, 2055–2072.

Sarker, I.H., Kayes, A.S.M., Badsha, S., Alqahtani, H., Watters, P., Ng, A., 2020. Cybersecurity data science: an overview from machine learning perspective. J Big Data 7, 41.

Slawski, M., Zu Castell, W., Tutz, G., 2010. Feature selection guided by structural information. Ann. Appl. Stat. 4.

Stajano, F., Wilson, P., 2011. Understanding scam victims: seven principles for systems security. Commun. ACM 54, 70–75.

Suraj, M.V., Kumar Singh, N., Tomar, D.S., 2018. Big data Analytics of cyber attacks: a review, in: 2018 IEEE International Conference on System, Computation, Automation and Networking (ICSCA). Presented at the 2018 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), IEEE, Pondicherry, pp. 1–7.

Viaene, S., 2013. Data Scientists Aren't Domain Experts. IT Prof. 15, 12–17.

Wagner, N., Rondinelli, J.M., 2016. Theory-Guided Machine Learning in Materials Science. Front. Mater. 3, 1–9. https://doi.org/10.3389/fmats.2016.00028

Wang, L., Jones, R., 2021. Big Data Analytics in Cyber Security: Network Traffic and Attacks. Journal of Computer Information Systems 61, 410–417.

Wright, R.T., Jensen, M.L., Thatcher, J.B., Dinger, M., Marett, K., 2014. Research Note—Influence Techniques in Phishing Attacks: An Examination of Vulnerability and Resistance. Information Systems Research 25, 385–400.

Zhu, X., Wu, X., 2004. Class Noise vs. Attribute Noise: A Quantitative Study. Artificial Intelligence Review 22, 177–210.

Zhu, X., Wu, X., Chen, Q., 2003. Eliminating Class Noise in Large Datasets, in: Proceedings of the Twentieth International Conference on Machine Learning. Presented at the ICML-2003, Washington DC, pp. 1–18.