Systematic Literature Review on the Semantic Analysis of Cybersecurity Indicators

Nic Recasens
Computer Science
Georgia Southern University
Statesboro, Georgia, USA
jr38088@georgiasouthern.edu

Nigel Smith
Computer Science
Georgia Southern University
Statesboro, Georgia, USA
ns15468@georgiasouthern.edu

I. Introduction

In the modern world, attackers are a step ahead of all Cybersecurity professionals. Professionals need to be able to adapt to every threat out there, while attackers only need to be able to adapt to a single fatal weakness. As such, the realm of machine learning and predictive analytics is of known importance in Cybersecurity research. For decades, machine learning analysis has helped keep defenders on top of their game by preventing novel attacks from gaining footholds. In the 90s, when the number of attackers was smaller, professionals were able to get away with only blocking known threats. But with Ransomware and Malware becoming billion dollar multi-national black markets, we need to be able to apply risk trends to novel data. At the forefront of this prevention is the identification and blocking of spam and phishing. An individual datum related to spam and phishing is called an indicator of attack (IoA), and they let us know if someone is targeted by an attacker. And if they are used successfully, they become indicators of compromise (IoC). IoCs have the negative reputation of being indicative of failed Cybersecurity protection, but in relevance to this project, they are the second half of the puzzle that is semantic analysis of indicators. Via the correlation of IoAs and/or IoCs, numerous researchers have pushed the boundary forward over the decades since the aforementioned 1990s. The attackers may be numerous in their successes, but they leave behind their own value aswell: data. With that data, we can see how pervasive cyber threats are, train models to simulate said attackers, and allow the very computers they attack help mitigate the risks themselves. Through this literature review, we will go over 8 papers that show how researchers achieved these goals, and show through their research how important cyber-risk mitigation truly is in our modern interconnected world.

II. Methodology

At the most basic level, the 3 keywords we used to guide our literature selection process were "spam", "phishing", and "semantic". Spam and Phishing are the strongest indicators Cybersecurity professionals have as indicators of attack. In fact, Business Email Compromise or (BEC) can be identified as the largest threat to the modern digital business [2]. As such, once we have papers scoped to the huge risk spam and phishing compose, the "semantic" keyword allows us to predominantly find Computer Science papers detailing how semantic analysis can help with this cause. Some linguistic journals do appear with these keywords, but with the current rise of the Large-Language Model, semantic research is seeing a heavy Computer Science edge.

For searching these terms, Google Scholar was the primary resource, though GALILEO also was utilized for accessing some non-free files. When on our Georgia Southern University network, some research journals allow our IP to access full papers without logon, which was also useful for reading full texts. Publication Year and citation information was considered and prioritized past the above keyword filtering, and was based on the Google Scholar data for those indicators.

And to wrap up our selection process, our primary exclusion criteria was mentioned earlier, and that is a Computer Science focus. Some articles were industry paid, so were more like Sociology in nature. And for other articles, they were focusing on the behavioral aspects of phishing and spam, which while useful, are not in scope for this review. And speaking of scope, the primary *inclusion* criteria was related to this course's guidelines, and that is publication year. Many useful articles at the beginning of semantic analysis for spam prevention were made in the late 2000's, but they were specifically asked to be ignored for this paper by the paper requirements. As such, only 'recent' research was provided, meaning publication dates past 2020 as much as it could be helped. These criteria, alongside with the above requirements and toolset mindsets, allowed us to choose the below 8 papers to systematically review.

III. Literature Review

A. 2019: iMCircle: Automatic Mining of Indicators of Compromise from the Web

To begin, we start with our oldest paper from 2019, as this paper shows the limitations of the research into spam indicators before the usage of modern semantic ML techniques. Authors Zhang, Ya, Liu, Li, Shi, and Gu for this paper prioritized the fact that before we can remediate and prevent spam and phishing, we have to know what spam and phishing actually *is* via data collection. And while that is a valorous cause for research, there are significant limitations with traditional Cyber Threat Intelligence (CTI) gathering regarding those indicators.

Modern CTI which often relies on a "fixed-point" monitoring of curated data feeds from sources like security organizations and white papers [4]. It also should be noted that "fixed-point" here refers to when detection systems make a baseline "secure" configuration, and an institution then marks all modifications to that baseline as insecure. This conventional approach is described as time-consuming and heavily complicated by the authors because it required researchers to clean data from multiple disparate sources, and to wait for information to be released by data providers as public. Consequently, current methods may suffer from incomplete coverage and data delays, meaning this traditional methods fail to keep pace with the rapidly evolving landscape of cyber threats, as every second counts when triaging.

To solve this, the authors developed a system called iMCircle. The iMCircle system employs an active recursive architecture consisting of four major components: a search engine crawler, an information preprocessor, an IoC checker and generator, and a new "seed" extractor (with seeds being the inputs for the next run of the iMCircle recursion) [4]. Unlike former passive systems, iMCircle takes spam and phishing indicators (like domains and IoCs) and combines them with a "target threat field" (i.e. "APT attack") to create search engine queries for exploratory data collection. It uses Google to retrieve relevant public information, which is then processed using Natural Language Processing (NLP) techniques to extract semantic features from site titles, URLs, and page snippets [4]. Machine learning models like Support Vector Machines (SVM) and Decision Trees then classify whether an indicator is a legitimate threat. In the final step of the chain, Validated IoCs are then used to extract new seeds from their associated webpages, restarting the cycle.

The study's primary contribution with this cyclical approach is the development of the first system (according to the authors) that mines IoCs continuously and automatically by checking suspicious indicators against open-source threat information. Since the time of this paper as well, we have seen many security scraping efforts be prioritized as well [4]. During a two-month real-world evaluation, the system demonstrated significant efficacy: the SVM model achieved a precision of 96.70% and an F1 score of 90.46% in identifying threat indicators. Furthermore, iMCircle successfully discovered 2,941 unique data sources, including more niche and varied public sources that traditional monitoring might overlook. That ability to get fresh data, rather than recycled, was of considerable satisfaction for the authors [4].

But, while iMCircle showed impressive performance, the paper did still acknowledge several limitations. Firstly, the system is entirely dependent on open-source information available on the Web; it cannot identify or verify suspicious indicators that have not been publicly discussed or indexed by search engines. This also means it was not tested on private repository information either for organization specific results [4]. There is also a disproportionate reliance on the quality and accuracy of search engine results, with the authors noting that crawling restrictions and empty results can reduce and fluctuate the usable sample size [4]. And lastly, the current implementation primarily focuses on a few metrics like domain names and Advanced Persistent Threat attacks (APT). So while the authors suggest that iMCircle could be adapted for other indicators like IP addresses or botnets, novel combinations like those would require further validation to prove the system's comprehensiveness across the entire threat landscape.

B. 20XX: TODO

C. 20XX: TODO

D. 2020: A semantic-based classification approach for an enhanced spam detection

The 2020's were a boom in semantic research, as we saw more powerful AI models hit impressive feats. As such, the research of Saidani, Adi and Allili in the beginning of this year addresses how the persistent and evolving threat of email spam can be mitigated by incorporating these new paradigms. Spam continues to account for over 56% of total email traffic, but traditional spam detection systems largely rely on the provenly ineffectual "bag-of-words" model, which evaluates emails as an unstructured collection of independent words without considering grammar, context, or word order. Semantic analysis can help prevent one of the biggest risks of BoW that the authors identified, which is polysemy [2]. Polysemy is a linguistic problem where a single word like "bow" can have drastically different meanings depending on the context. Spammers purposefully exploited the weaknesses of basic keyword filters through obfuscation, leading to the semantic classification employed in the paper.

To combat these tactics, the authors proposed a novel two-level semantic analysis framework. In the first level, incoming emails are categorized into distinct subject domains (such as Health, Finance, Education, Computer, and Adult). This establishes a tailored dataset of spam indicators for each specific sector [2]. In the second level, the researchers build a "feature space" for each domain by combining regular expression formatted hyperparameters with automatically extracted rules generated via the CN2-SD top-down search algorithm [2]. These rules are then merged and applied against test messages, with various machine learning algorithms being tested. These included Adaboost, Random Forest, and Naïve Bayes, with the goal to construct classifiers that accurately differentiate between spam and legitimate email indicators semantically.

Moving past lexical analysis to semantics led to the central finding of the study, which was that the domain-specific semantic approach significantly outperforms general-purpose classification models, including enhanced topic-based vector space models (eTVSM) (which were the high standard before this paper). Some domains were able to be trained successfully enough to reach 98% accuracy [2]. A major concept that this research contributed is proving that while spammers can always manipulate spelling and utilize synonyms to trick BoW classification, they cannot easily disguise the underlying intent or meaning of their emails (as of when this paper was published) [2].

Now, despite the high accuracy in multiple domains, the proposed semantic framework has areas requiring further development. The authors acknowledge that achieving "word sense disambiguation" could clearly improve their work, with word sense referring to the ability of models to disambiguate context around a phrase. Word sense is an ongoing challenge, and to the authors represents the next frontier for future work in spam detection [2]. While domain categorization narrows down the context, teaching a computer system to definitively resolve nuanced word meanings based purely on complex sentence structures is still difficult. Additionally, the reliance on manually specified rules as part of the hybrid approach means that experts must continuously update their knowledge engineering to maintain efficiency and keep pace with the rapidly evolving tactics of modern spammers [2].

E. 2020: Detecting Phishing SMS Based on Multiple Correlation Algorithms

Now that we've started getting the foundations of this research topic laid, this new paper from Sonowal and Gunikhan's shows how breadth of study for semantic spam blocking entails so much more than just emails. When we discussed spam indicators earlier, email shows up as the most common vector, but these authors have a notable paper showing how cellular communications also hurt end users. This is introduced as "smishing", or SMS phishing. Importantly, it is characterized as a disproportionately dangerous form of cyberattack due to the high trust users place in text messaging. Unlike traditional email delivery methods, which see open rates of roughly 20-30%, SMS messages boast a staggering 98% open rate, making them an ideal delivery system for malicious and unwanted messages [3]. Traditional defense mechanisms like static blacklists and keyword filters are weakened by the modern attacker trends of URL shorteners, swapping similar-looking characters like O and 0, and social engineering through falsely promoting urgency and legitimacy. And to top off the challenges SMS face, the study also identifies a central technical hurdle: high-accuracy detection is difficult to scope when we are talking about a mobile device. A desktop computer has a powerful full-sized CPU, and is chained to a wall. But a phone has much more stringent power, heat, and battery constraints when the topic of machine learning is concerned [3].

To combat the aforementioned newer tactics, the researchers went away from the older "bag of words" keywords, and instead used a more dynamic "features" paradigm to profile text. They identified 52 distinct features, including textual metadata like message length and capitalization ratios, and more complex linguistic metrics like the Gunning Fog and Coleman-Liau Indices [3]. A notable metric the authors uncovered was the dichotomy between standard business email template and the erratic failed syntax of spammers, with it allowing high consistency in detection.

To process this data, they utilized the AdaBoost (Adaptive Boosting) classifier, an iterative ML model that prioritizes learning from its own classification errors to distinguish between the differences in true and false positives. And as for the ranking algorithm, Kendall rank correlation provided superior accuracy over other correlation algorithms in prioritizing categorized results. These together helped lower the 52 features from earlier down to just 20 . That is a 61.53% decrease in dimensions, all while still achieving an impressive 98% accuracy [3]. That reduction in dimensionality was key in making the model viable for mobile devices' lower power draw.

But, despite all the successes in the SMS field the authors enjoyed, there is still work to be done in non-email based instances. First, as generative AI tools become more indistinguishable when mimicking the professional readability scores of companies, the "readability dissonance" the researchers rely on may diminish [3]. Additionally, while the paper focuses on the positives of tracking features, in terms of user privacy, the surveillance concerns of an AI analyzing a user could limit further adoption of the author's security measures.

F. 2022: An Effective Cybersecurity Awareness Training Model: First Defense of an Organizational Security Strategy

This moves us into the year 2022, and here, with the effect of spam and phishing fully realized in the previous papers, there should be no surprise that a continued effort exists for preventing attacks from successfully landing. For the majority of the papers reviewed, we have seen how computers can protect us from threats. However, when it comes to protecting ourselves, protecting the human factor, the work of Dash and Ansari shows how computers can help us improve proactively, rather than rely on content filtering alone. For false negatives, user training is a key part of security, and in a way, can be seen as teaching humans the ways to spot the indicators ML saw earlier. Despite advancements in defensive systems like firewalls and email filtering, "naïve information security practices" as the authors put it accounts for for between 72% and 95% of all cyber threats [1].

To address these vulnerabilities, the study employs a novel qualitative research design method utilizing the Technology Acceptance Model (TAM) as its primary theoretical framework. TAM was created in 1989 to aid human-computer collaboration, but the authors add modifications that incorporate "behavioral intention" to relate it to modern training needs [1]. This helps us to understand how intervention through training can successfully influence employee behavior. Additionally, the paper evaluates the National Initiative for Cybersecurity Education (NICE) and the viCyber model. The viCyber model is a cloud-based AI system that uses visual mapping and an inference engine to develop a customized cybersecurity curriculum based on specific competency needs, while NICE is the framework the approach relies on [1].

The study found in order to be effective, security strategy must adopt a bottom-up approach. This means making every employee accountable for the security of their own representative areas, rather than blaming software vendors or IT [1]. A significant contribution of this research past that is the endorsement of AI-enabled training tools like viCyber.

AI tools like this can provide real-time feedback, and can even adapt to each user's understanding level, allowing the aforementioned accountability to come more naturally [1]. The authors found that adaptable systems like these are more effective than traditional "one-size-fits-all" Massive Online Open Courses (MOOCs) because they can be tailored in efficient and individualized ways [1].

But, the paper still did not completely solve the user education crisis. While they do advocate for robust personalized security training to protect users from phishing, the authors also identified a literature gap regarding AI-based security training efficacy, which the study itself barely begins to fill. Another notable limitation here is the warning that high expectations from executives can create an unhealthy security climate if technical defense is neglected. Just because better training supports security teams doesn't mean that other security factors can just be let go [1]. Additionally, the paper mentions that the viCyber tool is still a "work in progress", and acknowledges that many organizations struggle to implement complex frameworks like NICE due to a lack of domain experts on payroll. This all indicates that while the concepts are indeed sound, they are just expectedly not a silver bullet without further development.

G. 20XX: TODO

H. 20XX: TODO

IV. Synthesis and Conclusion

References

[1] Bibhu Dash and Meraj Farheen Ansari. An effective cybersecurity awareness training model: First defense of an organizational security strategy. *International Research Journal of Engineering and Technology*, 9:2395–0056, 04 2022.

[2] Nadjate Saidani, Kamel Adi, and Mohand Saïd Allili. A semantic-based classification approach for an enhanced spam detection. *Computers & security.*, 94, 2020-07.

[3] Sonowal and Gunikhan. Detecting phishing sms based on multiple correlation algorithms. *SN Computer Science*, 1, 11 2020.

[4] Panpan Zhang, Jing Ya, Tingwen Liu, Quangang Li, Jinqiao Shi, and Zhaojun Gu. imcircle: Automatic mining of indicators of compromise from the web. In *2019 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–6, 2019.