IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 18 July 2022, accepted 13 September 2022, date of publication 16 September 2022, date of current version 27 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3207272

RESEARCH ARTICLE

Avoiding the Hook: Influential Factors of Phishing Awareness Training on Click-Rates and a Data-Driven Approach to Predict Email Difficulty Perception

THOMAS SUTTER, AHMET SELMAN BOZKIR, BENJAMIN GEHRING, AND PETER BERLICH

Institute of Applied Information Technology, Zurich University of Applied Sciences, 8401 Winterthur, Switzerland

Corresponding author: Thomas Sutter (suth@zhaw.ch)

This work was supported in part by Lucy Security AG, Zug, Switzerland; and in part by Innosuisse—Swiss Innovation Agency, Bern, Switzerland.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Zurich University of Applied Sciences.

ABSTRACT Phishing attacks are still seen as a significant threat to cyber security, and large parts of the industry rely on anti-phishing simulations to minimize the risk imposed by such attacks. This study conducted a large-scale anti-phishing training with more than 31000 participants and 144 different simulated phishing attacks to develop a data-driven model to classify how users would perceive a phishing simulation. Furthermore, we analyze the results of our large-scale anti-phishing training and give novel insights into users' click behavior. Analyzing our anti-phishing training data, we find out that 66% of users do not fall victim to credential-based phishing attacks even after being exposed to twelve weeks of phishing simulations. To further enhance the phishing awareness-training effectiveness, we developed a novel manifold learning-powered machine learning model that can predict how many people would fall for a phishing simulation using the several structural and state-of-the-art NLP features extracted from the emails. In this way, we present a systematic approach for the training implementers to estimate the average ''convincing power'' of the emails prior to rolling out. Moreover, we revealed the top-most vital factors in the classification. In addition, our model presents significant benefits over traditional rule-based approaches in classifying the difficulty of phishing simulations. Our results clearly show that anti-phishing training should focus on the training of individual users rather than on large user groups. Additionally, we present a promising generic machine learning model for predicting phishing susceptibility.

INDEX TERMS Difficulty estimation, human-centered, machine learning, phishing awareness, susceptibility, phishing attack simulations.

I. INTRODUCTION

Humans are often said to be the weakest link in IT security. It is, therefore, not surprising that email phishing is still seen as one of the most significant threats in cyber security and is widely discussed in the literature [1]. In recent years, several successful phishing attacks on large companies were publicly reported [2], [3], [4]. The Federal Bureau of Investigations (FBI) estimates that more than $2.1 billion in actual losses were from business email compromises during 2014 to 2019 [5]. Moreover, during the COVID-19 pandemic phishing related cyber-crime has been increasing [6]. Malware families like Emotet have demonstrated that emails are still an effective method to deliver malicious binaries to end-users

The associate editor coordinating the review of this manuscript and approving it for publication was Binit Lukose.

and that credential-stealing attacks and ransomware often work hand in hand. Companies have long realized that phishing is a threat to be taken seriously and that traditional countermeasures like email filters and two-factor authentication cannot entirely prevent such attacks.

One of the latest trends in phishing countermeasures is to work on the weakest link, the human, by applying anti-phishing training. Companies specializing in anti-phishing training offer their customers services in simulated phishing attacks and educational training material. The effectiveness, methodology, and ethics of anti-phishing training are controversially discussed in the research community. Several researchers have focused on estimating the impact of anti-phishing training [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], the evaluation of email content and structure [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], the influence of knowledge retention [7], [10], [19], [22], [30], [31], [32], [33], [34], [35], the structure of training material and methods [7], [13], [30], [31], [36], [37], [38], [39], [40], [41], or the impact of anti-phishing training on the target group [9], [42], [43], [44], [45], [46].

One of the main actions performed when estimating the base awareness level of users is to send simulated phishing attack emails. These simulations test companies' security policies and practices to increase awareness and decrease their susceptibility to attacks. For example, in the industry, it is common practice for companies to repeatedly conduct simulated phishing attacks and then observe how many of the users would perform dangerous activities such as clicking on a link, downloading a file, or submitting account credentials. In addition, companies are using the accumulated results to track the progress of the anti-phishing training over time and to compare the performance of individual user groups to each other.

The main challenge in applying training is estimating the user's baseline awareness level so that the anti-phishing training can progress with maximal efficiency and track the users' progress over time. Currently, companies rely on performing repeated attack simulations to estimate a risk metric for their users and then decide after an arbitrary number of simulations which users imply a higher security risk. However, this approach is problematic as it ignores the fact that users perceive the difficulty of an attack simulation differently and that by conducting divergent attack simulations, we may generate biased results.

In general, not every anti-phishing simulation is suitable for every target group or individual user within a group, and not every set of anti-phishing simulations is suitable for estimating the awareness level of a group or user. For example, if we send out several anti-phishing simulations to Group A with a topic about a service that none of the users in Group A have ever heard of, likely, many of these users will not click on any of the phishing links or submit their credentials because they may not have any interest in trying out a new service and see the email as spam. In contrast, if we send out spear-phishing emails to Group B, which are individually crafted emails for every user, and in line with their daily routines, we expect a high number of clicks- and submits because the attack simulation is relevant their interest and habits. In such cases, it is misleading for companies to compare the user's performance based on these simulations. If we compare Group A's click- and submit rates to Group B's, it may look like Group B is less aware than Group A because more users performed dangerous actions. However, it must not be the case that Group B is less aware because our anti-phishing simulations were biased in their difficulty.

One way to overcome this problem is by sending out several attack simulations to the users and taking an average value. However, it is unclear how many simulations are necessary to estimate a base awareness level and companies follow different approaches. In addition, some companies would use the users' performance for their decision process on how to progress with the anti-phishing training.

Another way of avoiding this problem is by sending out the same attack simulations to all users, which may work in some cases when the number of users is relatively small, and the email content is generally in line with the user's expectations. Unfortunately, in practice, this approach results in the problem that for larger user groups, the content and structure of the email often cannot fully be aligned with the expectation of all the users. Consequently, the results of attack simulations do not fully represent the actual subspeciality level of the users, for instance, in cases where we want to test the user's ability to detect spear-phishing emails. Therefore, we conclude that the method ''For a fair selection, everyone has to take the same exam'' seems not suitable for anti-phishing training of individual users as the attack simulations have varying content and difficulty perception.

Currently, most companies neglect that not every anti-phishing simulation is suitable for every target group and conduct anti-phishing training in various ways. The examples mentioned earlier demonstrate that a generalized way of measuring the difficulty of anti-phishing simulations is a fundamental need for practical anti-phishing training. Furthermore, it shows that to estimate users' awareness levels, we need (i) the possibility to send out individually crafted anti-phishing simulations to different groups and (ii) robust measurements that allow us to compare groups with divergent anti-phishing simulations. Using biased results without robust measurements leads to a wrong perception of the user's awareness level and inadequate anti-phishing training.

Furthermore, some cases like the West Midlands Trains [47] and the Tribune Publishing case [48] showcase that wrongly applied phishing awareness training can backfire. In both cases, the companies sent a phishing email stating that their employees would receive a one-off payment due to their good work or ongoing commitment during the COVID-19 pandemic when many employees were financially struggling. After the employees found out that the emails were part of a phishing awareness training and nobody would receive any bonus, many employees felt offended and demanded that the companies pay a real bonus as a

reparation. One of the companies publicly apologized for sending misleading emails, and these examples demonstrate that anti-phishing training is often an ethical gray zone.

An important factor is how users are treated when falling for anti-phishing simulations. Several researchers have investigated the aspects of negative (e.g., punishment) and positive (e.g., reward) feedback on users falling for anti-phishing simulations [36], [49], [50], [51], [52], [53]. Bora *et al.* reported in [51] that punishment reduced the number of users clicking on phishing emails. However, punishment can have negative organizational side effects, such as users not reporting real phishing incidents because they feel they did something wrong and could get chastised. On the other hand, some industry experts [54] recommend positive reinforcement learning because, from their perspective, it encourages users to report phishing emails more often.

In literature, it is commonly agreed that the structure and content of anti-phishing simulations play a crucial role in how users perceive and react to phishing threats. Other factors like the curiosity of the users [55], the content alignment to the user's expectation [19], [23], or the alignment of the sending time [8] seem to be influential factors. Ideas to estimate a phishing mail's difficulty have been discussed using clue-based scales [56], [57], [58], [59]. However, to our knowledge, none of the proposed difficulty scales were ever used in large-scale studies.

In this paper, we propose a novel and data-driven method to estimate the difficulty perception of an anti-phishing simulation by conducting a large-scale study with 31'940 participants. We conducted 144 anti-phishing simulations and used the collected data to estimate users' susceptibility to specific phishing emails. Furthermore, we reflect on our anti-phishing training and show which factors seem to influence our participants' click behavior.

A. RESEARCH QUESTIONS

1) RQ1: HOW EFFECTIVE ARE WEB-LINK-BASED PHISHING ATTACK SIMULATIONS FOR AWARENESS TRAINING?

Our first research question aims to study whether the clicking behavior of our participants will change in regards to repeatedly applied phishing awareness training. In addition, we analyze how many of our users would not fall for any credential-based phishing simulation.

2) RQ2: HOW MANY TIMES DO PARTICIPANTS FALL FOR SIMULATED PHISHING ATTACKS?

One of our goals in this study is to estimate how often participants would fall for phishing simulations. Previous studies have often measured the total clicks for specific phishing simulations. However, our goal for the second research question is to determine the number of times specific participants clicked because we wanted to analyze the so-called ''repeated clickers'' phenomena and find out how many of the users would never fall for any of the phishing simulations.

3) RQ3: DO PARTICIPANTS WITHOUT ANY TRAINING MATERIAL FALL MORE OFTEN FOR PHISHING ATTACK SIMULATIONS THAN THOSE WITH E-LEARNING-BASED TRAINING MATERIALS?

Our third research question aims to determine if participants who get confronted with e-learning materials perform better in terms of click-rate than those without. We randomly assigned our participants into equally sized groups and provided different learning content, and our study aims to analyze if providing different training material makes a significant difference over time.

4) RQ4: CAN WE PREDICT THE EMAIL DIFFICULTY PERCEPTION USING ONLY THE EMAIL CONTENT WITH A DATA-DRIVEN APPROACH?

We aim to identify the influential factors for falling for simulated phishing attacks based on our data. Our last research goal is to develop a generic Machine Learning model that is capable of classifying a phishing email into three difficulty levels (''easy,'' ''medium,'' and ''difficult'') for the prediction of the difficulty perception.

B. CONTRIBUTIONS

By conducting a large-scale anti-phishing training, we give novel insights into the key factors of why users fall for phishing attacks and how users should effectively be trained over time. We describe our experimental setup in Section III and explain how we were able to conduct 144 anti-phishing simulations for one year with 31940 participants.

We describe our key findings for ***RQ1*** and ***RQ2*** in Section V. Our contributions show that most users do not fall for simulated phishing attacks. In addition, we show that credential-based phishing simulations are not beneficial for many users, as most of our participants in this study did not fall for any of our attack simulations.

We discuss ***RQ3*** in Section VI, were we conduct a Chi-Square analysis to test the different training methods applied. Additionally, we confirm that providing training material positively affects lowering click rates. Surprisingly, however, we show that most users never complete any training courses but still perform significantly better than users without training material.

Regarding ***RQ4*** we contribute by developing a novel data-driven Machine Learning model for predicting the perception of difficulty of anti-phishing simulations in Section VII. In Section VIII, we evaluate our Machine Learning (ML) model and illustrate the most influential factors for phishing susceptibility. We show that the alignment of email content to the user's workplace is one of the key factors in phishing difficulty perception. Using 5-fold-cross validation, our generic model has, in the best case, an accuracy of 68% and proofs that the prediction of phishing susceptibility is possible but limited due to the lack of data points.

In Section IX we summarize all of our findings regarding $RQ1 - RQ4$, and we discuss the limitations of our study in Section X.

II. RELATED WORK

In our previous study [60], we conducted a comparative literature review of anti-phishing training methods. We concluded that large-scale studies are indispensable to show the long-term effects of anti-phishing training. Despite decades of research, anti-phishing training continues to be debated among researchers, and various studies show contradictory results. For example, if user demographics such as age and technicality have an impact or not [9], [30], [42], [45], [46]. However, a closer look at the literature on anti-phishing training reveals several gaps and shortcomings. Many studies lack in the number of participants, attack simulations or the use of control groups to verify their results [12], [13], [24], [33], [34], [39], [40], [41], [44], [46], [61], [62], [63], [64], [65], [66], [67], [68], [69], [70], [71]. The literature on the efficiency of anti-phishing training is more consistent even when still many contradictions exist. For example, a recent study by Lain *et al.* [72] concluded that conducting embedded training may not provide the wished training effects or can even have negative side effects. Other studies have focused on when to reapply the training [35].

In this study, we show that providing phishing awareness emails together with educational material statistically significant impacts the number of users clicking on phishing links independent from the used training material (video, text, quiz, or illustrations). All of the provided training materials (including embedded training) seem to reduce the number of users clicking. In other words, our results show that for most users, it does not matter if embedded training or other training methods were applied. Our unexpected findings signal the need for additional studies to understand more about the effect of anti-phishing training material on human behavior.

One of the tough challenges for all researchers in this domain is that collecting the necessary data for such a study involves conducting anti-phishing training with several hundred or even thousands of participants, which holds some ethical-, legal- and organizational challenges. Nevertheless comparable studies have been conducted which show promising results in revealing the nature of anti-phishing training effects [19], [42], [45], [72], [73], [74], [75], [76]. However, to our knowledge, no other studies have been attempting to estimate the difficulty of phishing awareness training at this scale; Other papers had fewer participants and fewer training iterations [56], [57], [58], [59], or did not focus on the estimation of difficulty [72], [74].

III. TRAINING METHODOLOGY

We had to answer three practical questions before setting up the experiment. First, how to split our participants into groups? Second, how do we create attack simulations with various scenarios and different attack techniques? Third, how often and when do we send out attack simulations and training material? This section will answer these questions and give the reader an overview of our anti-phishing experiment and methodology. Finally, we will discuss the training methods recommended in the literature and continue with the ethical and organizational limitations of anti-phishing training.

A. APPLIED METHODS

Embedded training is an anti-phishing training methodology where participants are confronted with education material whenever they fail a phishing attack simulation. The basic idea of embedded training is to directly engage the participant with educational material whenever a dangerous activity such as clicking on a link or submitting credentials is performed. Often educational material in the form of videos, illustrations, games, or text is shown to participants when they fail one of the anti-phishing simulations.

In literature, the fact that embedded training is more effective than other training methods is controversial, as mentioned before. Nevertheless, embedded training seems to be the de facto standard in anti-phishing training as all major security awareness companies [77], [78], [79], [80], [81], [82], [83] offer embedded training as a service. Other training methods such as in-class training seem to be less popular. Therefore, we use an embedded training methodology for anti-phishing training. For the anti-phishing training, we use the product of a security awareness company that offers training material in the form of quizzes, videos, and texts. Quizzes would be a mix of detecting phishing emails, answering general security questions, and educational texts about how to spot a phishing email. Similar to the video and text material, the company would explain phishing and how to detect clues that reveal phishing attacks.

In our case, in-class training with all of our participants was no option due to the sheer number of participants. Therefore, we selected the six most-used training materials of the security awareness company at that time and set them up as embedded training. The anti-phishing training material would consist of four different interactive phishing quizzes and two one-pager web pages with mixed content in video and text material. A maximum of 10 minutes was required to complete any training lesson, and four of the training lessons could be completed in under five minutes.

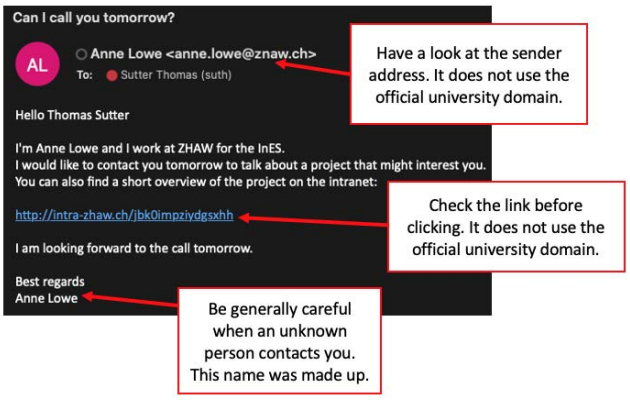

FIGURE 1. Example rubric of a phishing mail.

In addition to the provided training courses, we created our training material. The idea of our training material was to test if participants would learn better if we showed them their mistakes and our expectations directly after falling for a phishing simulation. We, therefore, set up a so-called rubric for each attack simulation. A rubric is a teaching technique that is widely used in higher education [84]. It is a learning technique where we give the students an overall expectation for the assignment and show the students how we expect them to solve the assignment. In our experiment, a rubric is a one-pager website with screenshots and descriptive text of the attack simulation. An example rubric is displayed in Figure 1. The screenshot shows how we would expect the users to detect the specific phishing email, and the text would give them some more general tips and tricks. In addition, we would always add a screenshot of the landing page, similarly to the rubric on the website. As with the other embedded training material, every time the users would fall for one of our attack simulations, we would redirect them directly to the rubric website showing two screenshots and description texts.

To answer our research questions RQ_2 we categorize our users into the four training groups: Control (C), Embedded Training (ET), Rubric (R), and Embedded Training plus Rubric (ETR). The Control Group would not receive any training material, and we would send them only the attack simulations without any further information. The Embedded Training Group would always receive one of the six training methods provided by the security awareness company as embedded training. The Rubric Group would receive only the one-pager website with the rubric. Finally, the ETR group would receive the embedded training and the rubric. Moreover, whenever a user would fall for one of the phishing emails, we would immediately send them an email with a link to the training material so that users could read the training material later.

B. ETHICS STATEMENT

This study was located at a university of applied sciences. Throughout the complete study, we followed the ethical guidelines of our university. The study was approved by the Chief Information Security Officer and the highest panel of the university, which both followed the internal university process for approval and ethics of the project. Before sending the anti-phishing emails, several information security experts reviewed the content of the study.

During the study, we had access to Personally Identifiable Information (PII), such as the student's or staff members' university email addresses or names. The PII we had access to were strictly necessary to set up the application for managing the phishing awareness training. The phishing awareness application was hosted in our universities internal IT environment and under strict security controls. None of the PII was handed out to third parties or used for purposes other than this study. Only a minimal number of staff members had access to the data. In addition, access to the PII data was removed after the study ended.

For the complete study, we had only access to the PII, which was necessary to conduct the study. During the study, we collected data from our participants, such as click- and submit actions. We followed our institution's guidelines to process and manage the collected data, including anonymization, pseudonymization, and data encryption whenever possible and reasonable.

Generally speaking, phishing awareness training continually exposes the participants to the risk of wasting their time [85]. However, as Lain *et al.* [72] stated in their study, it does not expose the participants to a greater risk than what they would encounter during their daily lives because the participants are regularly exposed to real phishing emails or spam. We acknowledge that conducting this study exposed our participants to minimal risks but similar to other researchers [72], we believe that the positive experience the participants gain merited these risks. In addition, the decision to conduct this study with all students and staff members was approved by the highest panel of our university (including the CISO) and we followed the ethical guidelines of our university to the best of our knowledge.

Furthermore, during the study, we constantly gathered feedback from our participants to verify whether our methods were appropriate. We received positive and negative feedback from participants; if wanted, participants could opt-out of the study. Moreover, at the end of the study, we sent an email to debrief all of the participants as proposed by Resnik *et al.* [86].

C. GROUP SELECTION

Our study participants were mainly university students studying for bachelor's or master's degrees 57%, students from continuing education 24%, and university staff members 19%. Overall, 47.36% of our participants were male, and 52.64% were females. Participants were selected from all university institutions and departments: Administration, Finance, HR, IT, Management, Law, Psychology, Architecture, Linguistics, Social Work, Life Sciences, and Engineering. Moreover, participants from continuing education and part time students which work in a wide variety of industry sectors participated.

We randomly selected participants and applied a stratified sampling approach to group the participants into groups of equal size. With stratified sampling, we ensure that the same percentage of university staff members and students are represented in every group. We then split every main group (G_i) into four subgroups (SG_j). We assigned for every subgroup SG_j one of the training methods C, ET, R, and ETR. Subgroups of the main group would all receive the same attack simulations to compare their results. We had 48 groups, with each group G_i having 2000 participants and each subgroup SG_j having 500 participants for the first six weeks of the experiment. The first part of the experiment involved 24000 participants, and we created for each main group attack simulation with differing content.

D. ATTACK FEATURE SELECTION

For our experiment, we focused our study on credential phishing attacks written in German and English. Such attacks are usually conducted by sending the victim an email with a link to a domain controlled by the attacker. The attacker domain usually hosts a web page that attempts to trick the users into entering their credentials or other sensitive information such as credit card details. In our cases, we set up 144 attack emails and landing pages. The landing pages were either created by cloning an original page or contained made-up content. As a basis for phishing emails, real-world emails were used, which were modified according to our requirements. We used a wide variety of emails by modifying existing university emails, using real-world phishing examples as a template, or writing fictional emails. All emails were assigned randomly to the groups. The attack simulations used several credentials phishing techniques like typo-squatting, double-barrel attacks, hidden links, and image-based attacks.

We decided to use German and English emails because these are the two main languages spoken at our university. Our participants receive mails in both languages on a daily basis. Thus, it is reasonable to use phishing emails in these languages for our experiment.

E. FREQUENCY OF TRAINING

Butavicius *et al.* showed in [14] that a pre-announcement of the anti-phishing training is not necessarily beneficial for the performance of the users. Therefore, we sent the attack simulations without informing the users in advance. We conducted twelve weeks of anti-phishing training with solely university members participating. The university has a wide variety of focus areas. The experiment included participants from engineering, management, law, psychology, linguistics, life science, facility management, civil engineering, architecture, health sciences, and university staff members.

To prevent training fatigue [87] we set a maximum of one attack simulation per week, and we spread the simulations throughout two semesters, with each semester having six weeks of training. The day and time on which the simulations were sent to the participants were randomly selected from Monday to Friday, and the phishing links would be taken offline after two weeks. Prior studies [7], [45] have shown that most responses to phishing attacks are received within 24 hours. Therefore two weeks is more than enough for participants to conduct the simulations. In addition, subgroups of the same main group would always receive their attack simulations simultaneously to prevent the effect of users warning each other.

IV. EXPERIMENT CONSTRAINTS

When conducting anti-phishing training, there are several ethical and organizational aspects to consider. The goal of every anti-phishing training should be a beneficial awareness effect for the organization and the participants. However, when creating anti-phishing simulations, we were often confronted with scenarios that could mislead participants into believing or interpreting the content of the anti-phishing simulations as real. Consequently, participants would believe the content of the anti-phishing email and behave in an unpredicted or unwanted way. For example, users would think their machine was infected with a virus and stop using the computer for several days because we sent them a phishing mail claiming their computer was infected.

Sending anti-phishing simulation emails is often an ethical grey zone that can have adverse side effects on the participants' daily routine. We refer to such adverse side effects as collateral damage, and every anti-phishing training should consider what could be the highest damage created by anti-phishing training. Such collateral damage can lead to different scenarios where users would take the information given in the anti-phishing simulation as truth and would fully believe the misinformation we sent in our anti-phishing simulations.

The following sections summarize the constraints we set for our anti-phishing training to minimize collateral damage.

A. ETHICAL AND ORGANISATIONAL CONSTRAINTS

Every anti-phishing simulation was reviewed by several information security experts to estimate possible worst-case scenarios and to minimize collateral damage. As a result, it was decided that some email topics were off-limits, such as the COVID pandemic, not to disrupt the actual communication of the university. In other cases, the email writing was often adjusted not to include actual events, persons, or institutions to reduce possible collateral damage.

Additionally, to not disrupt the research of other departments of our university or damage individuals' reputations, we were not allowed to imitate university members, such as students or professors, or research groups. Instead, we used fictional names and groups.

B. LEGAL CONSTRAINTS

Sending anti-phishing emails with the look and feel of real existing brands, such as for example, Google, Microsoft, or Meta, without the written allowance of the brand or trademark is considered a crime in our jurisdiction. Other countries may have different regulations, but sending phishing awareness emails at that scale would risk a lawsuit in our jurisdiction.

In general, such regulations make anti-phishing training challenging since brands would often refuse to give the allowance to use their brands or trademarks for anti-phishing training. Therefore, we were limited to using a subset of existing brands on the market that gave their allowance, or to fictional brands, or our university's brands.

V. EXPLORATORY DATA-ANALYSIS OF THE DATASET

This section gives an overview of the collected data and analyzes how effective our anti-phishing training was. We define a successful attack for our experiment as whenever a user clicks on one of the links provided in a phishing simulation because the risk of drive-by downloads or browser

exploitation exists. However, we set up our anti-phishing training to track if the users would also submit any of their credentials or other sensitive data. To protect the user privacy, we did not store any submitted credentials, nor did we verify if users were entering valid credentials.

TABLE 1. Overview of experiment numbers.

Fact	Number
No. of participants	31940
No. of emails sent	288000
No. of phishing simulations	144
No. of clicks in total	31707 (11.01%)
No. of submits in total	15224 (5.29%)
No. of completed training	7140

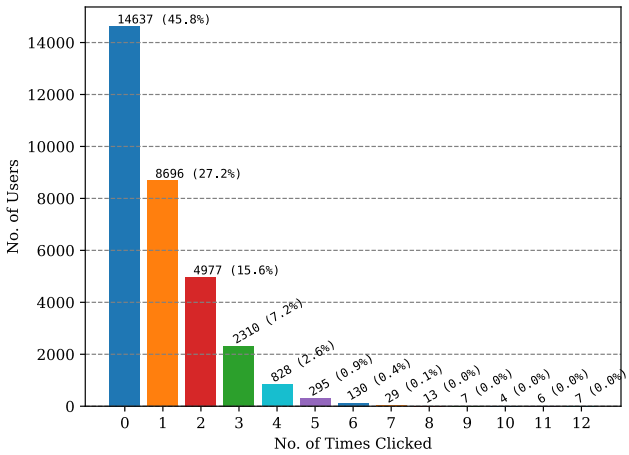

FIGURE 2. Number of times individual users clicked.

Table 1 shows our experiments key values. In total, we sent out 288000 emails over twelve weeks. A user could receive a maximum of twelve attack simulations, and in total, we had 31940 users participating in this study. Overall we registered 31707 (11.01%) clicks on phishing links for our phishing simulations. The number of times users submitted their credentials was nearly half as much, with 15224 (5,29%). We calculated how often the users would fall for phishing attacks to examine the click behavior further. Figure 2 displays the number of times the users clicked. It shows that 14637 (45%) users never clicked on any of our phishing emails, whereas 8696 (27%) clicked only one time and 4977 (16%) clicked two times, respectively 2319 (7%) clicked three times. Only a small fraction of users (1319, 4%) fell more than four times for our phishing simulations. Thus, the data support the premise that there is no need to conduct anti-phishing training for all the users because 45% of the users never clicked on any phishing simulation.

To further examine the user's click behavior, we calculated the number of times a user would submit their credentials in Figure 3. As already mentioned, we registered fewer users submitting their credentials than clicking on the phishing links, and this trend can be seen in Figure 3 as well. Two-thirds of the users, 21099, never submitted any of their credentials, and 7563 (24%) users submitted their credentials one time. Our data suggest that users are more careful when entering their credentials than when clicking on a link in an email.

Moreover, the click data supports the premise that only a few users click on all anti-phishing simulations. If we categorize users that click six or more times as "repeat clickers," we have in our population 204 (0.64%) users in this category. In case we define repeated clickers as four clicks or more, than we have 1327 (4,2%) users in this category. However, when we compare these repeated clicker numbers to the number of repeated submits, it shows that users tend to click more than they to submit their credentials. Consequently, as shown in Figure 3, we cannot see a trend in repeated submitters as only 199 (0.62%) of the users submitted their credentials four or more times, and none of the users submitted their credentials for all of our phishing simulations.

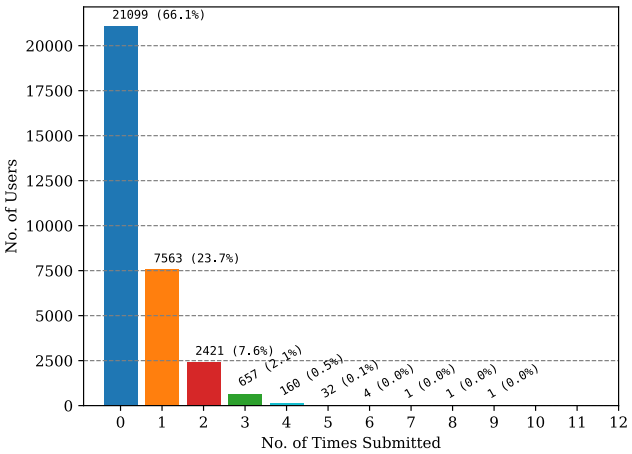

FIGURE 3. Number of submits.

Furthermore, we examine the used user agent of the users that submitted their credentials. We accumulated the user agents of the same browser vendor and showed in Figure 4 the most frequently used browsers. As Chrome is according to [88] and [89] the dominating desktop browser on the market, it is not surprisingly, that our participants used the Chrome browser most (6073, 40%) in our experiment. More interestingly seems to be the number of used mobile browsers with Mobile Safari (3148, 21%) and Chrome Mobile (1998, 13%), as it shows that one-third of the submits (34%) were from users using their mobile phones. This is insofar a vital fact as it shows that the usage of responsive designs for phishing landing pages is necessary because a large part of the users will read their emails on a mobile phone.

VI. MEASURING THE EFFECTIVENESS OF TRAINING

For the anti-phishing training evaluation, we only use the first six weeks of training since, after six weeks, we would have

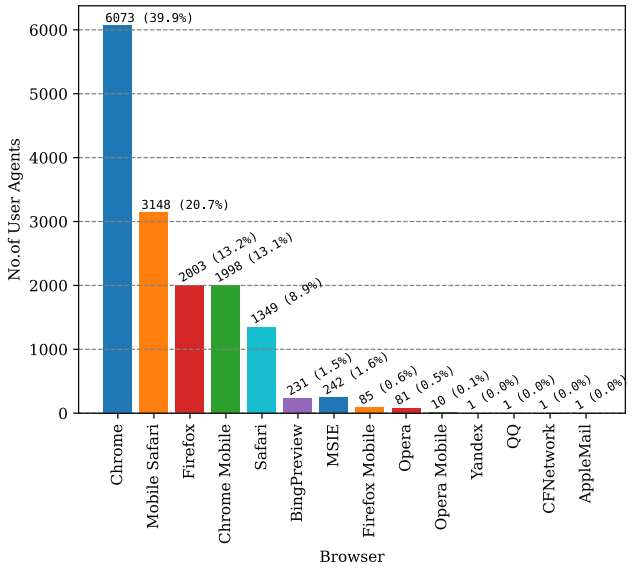

FIGURE 4. Most used browsers for submitting credentials.

a change of study semesters, and a relatively large part of the students (1/3) would finish their studies and leave the university. We used an established chi-square technique to analyze the training effectiveness.

The chi-square analysis is the most common statistical technique for analyzing $R \times C$ dimensional frequency tables (also known as multi-way contingency tables) in which row and column variables are categorical. Note that chi-square (χ^2) test statistic is a non-parametric test having the following assumptions [90], [91]:

1) The data in the contingency table cells should be the frequencies or counts of cases rather than percentages or some other transformation of the data.
2) The study groups must be independent.
3) There are two categorical variables at nominal or ordinal category.
4) The categories of the variables must be mutually exclusive. That is, a particular subject fits into one and only one level of each of the variables.
5) Each subject may contribute data to one and only one cell in the χ^2.
6) The expected values of the cells should be five or more in at least 80% of the cells.

Suppose that a contingency table of counts having R rows and C columns is as in the table below. Let n_{ij} be the observed count for the i th row ($i = 1, \ldots, R$) and jth column ($j = 1, \ldots, C$).

The χ^2 statistic is used to test the following null hypothesis:

- H_0: Row and Column variables are independent.
- H_0: There is no difference among Row 1, Row 2, …, and Row R in column variable.

The formula for calculating χ^2 test is [90], [91]:

$$\chi^2 = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \tag{1}$$

TABLE 2. Chi-Square example.

		Column Variable				
		Col. 1	Col. 2	...	Column C	Total
Row Variable	Row 1	$n_1 1$	$n_1 2$	...	$n_1 C$	$n_1.$
	Row 2	$n_2 1$	$n_2 2$	...	$n_2 C$	$n_2.$
	⋮	⋮	⋮	⋮	⋮	⋮
	Row R	$n_R 1$	$n_R 2$	...	$n_R C$	$n_R.$
	Total	$n_.1$	$n_.2$	...	$n_.C$	N

where E_{ij} is the expected count for the i th row ($i = 1, \ldots, R$) and j th column ($j = 1, \ldots, C$) calculated as $E_{ij} = \frac{n_{i.} n_{.j}}{nij}$. The χ^2 test statistic follows an asymptotic chi-square distribution with $(R-1)(C-1)$ degrees of freedom when the row and column variables are independent. If $\chi^2 \geq \chi^2_{(\alpha;(R-1)(C-1))}$, the null hypothesis H_0 is rejected. The Chi-Square test statistic in Eq. 1 can be calculated with the following formula given below:

$$\chi^2 = \sum_{i=1}^{R} \chi_i^2 = \chi_1^2 + \chi_2^2 + \ldots + \chi_R^2 \tag{2}$$

where $\chi_1^2, \chi_2^2, \ldots, \chi_R^2$ are calculated via (1) with n_{ij} and $E_{ij} (i = 1, \ldots, R)$ and $(j = 1, \ldots, C)$ for Row 1, Row 2,…, Row R, respectively. If the null hypothesis is rejected, the row which has the highest χ_i^2 ($i = 1, \ldots, R$) value is removed and the χ^2 analysis is recomputed on the contingency table with the other rows [92].

TABLE 3. Click-rate frequency table.

		Main User Groups											
Method	Training Program	G1	G2	G3	G4	G5	G6	G8	G9	G10	G11	G12	Total
A	Simulation	5	5	69	112	39	5	17	200	60	51	56	619
	Embed.Training	6	11	100	53	26	8	39	176	42	65	28	554
	Rubric	3	17	96	84	42	6	15	129	66	89	46	593
	ET+Rubric	8	10	101	96	49	5	6	152	46	64	42	579
	Total	22	43	366	345	156	24	77	657	214	269	172	2345
B	Simulation	63	28	126	5	18	140	17	20	96	89	17	619
	Embed.Training	88	23	99	6	19	149	17	20	74	87	12	594
	Rubric	91	16	91	5	26	131	15	20	70	96	12	573
	ET+Rubric	105	21	112	5	32	147	6	2	90	90	10	620
	Total	347	88	428	21	95	567	55	62	330	362	51	2406
C	Simulation	20	22	8	35	10	12	20	33	7	41	29	236
	Embed.Training	24	19	8	35	6	24	12	27	6	29	16	206
	Rubric	20	29	5	43	5	18	8	6	17	37	31	219
	ET+Rubric	19	19	12	28	4	19	12	10	11	29	29	192
	Total	83	89	33	141	25	73	52	76	41	136	105	854
D	Simulation	32	113	71	88	30	44	130	75	86	189	45	903
	Embed.Training	55	143	42	104	23	44	133	47	64	161	43	859
	Rubric	54	163	48	117	28	56	96	38	92	148	60	900
	ET+Rubric	38	149	36	79	22	52	107	51	76	123	53	786
	Total	179	568	197	388	103	196	466	211	318	621	201	3448
E	Simulation	78	57	55	39	31	96	5	6	66	60	150	643
	Embed.Training	133	69	52	25	37	101	11	8	47	49	137	669
	Rubric	115	82	52	16	30	82	5	6	59	40	156	643
	ET+Rubric	122	65	50	21	33	101	4	2	47	48	156	649
	Total	448	273	209	101	131	380	25	22	219	197	599	2604
F	Simulation	56	30	45	51	14	113	53	54	101	96	12	625
	Embed.Training	79	34	35	49	8	76	30	44	76	95	15	541
	Rubric	58	35	24	34	6	69	32	40	89	108	24	519
	ET+Rubric	71	32	35	39	8	77	21	39	86	70	8	486
	Total	264	131	139	173	36	335	136	177	352	369	59	2171

To meet the assumption of mutually exclusivity, under each Method (A, B, C, D, E, and F), we performed the χ^2 analysis for the contingency table with training programs representing row variables whereas user groups constitute the column variable. We present the click-rate frequency table collected from different weeks and main user groups in Table 3. We, here, represent the weeks as *methods* (i.e., A=1, B=2, C=3,

TABLE 4. Results of χ^2 analysis for contingency tables.

Method	Null Hypothesis	χ^2 statistic	p-value
A	H_0 : There is no difference among all Training Programs in Main User Groups	$\chi^2 = \chi^2_S + \chi^2_{ET} + \chi^2_{Rubric} + \chi^2_{ET+Rubric}$ $= 29.712 + 48.369 + 23.608 + 17.403 = 119.092$	<.001* ($\mathbf{H}_1$)
	H_0 : There is no difference among Simulation, Rubric, and ET+Rubric training programs in Main User Groups	$\chi^2 = \chi^2_S + \chi^2_{Rubric} + \chi^2_{ET+Rubric}$ $= 25.352 + 21.939 + 10.806 = 58.097$	<.001* ($\mathbf{H}_1$)
	H_0 : There is no difference between Rubric and ET+Rubric training programs in Main User Groups	$\chi^2 = \chi^2_{Rubric} + \chi^2_{ET+Rubric}$ $= 9.435 + 9.676 = 19.111$	0.039* ($\mathbf{H}_1$)
B	H_0 : There is no difference among all Training Programs in Main User Groups	$\chi^2 = \chi^2_S + \chi^2_{ET} + \chi^2_{Rubric} + \chi^2_{ET+Rubric}$ $= 17.624 + 5.205 + 7.999 + 23.336 = 54.164$	0.004* ($\mathbf{H}_1$)
	H_0 : There is no difference among Simulation, Emb.Training and Rubric training programs in Main User Groups	$\chi^2 = \chi^2_S + \chi^2_{ET} + \chi^2_{Rubric}$ $= 12.459 + 2.752 + 8.234 = 23.444$	0.268 (H_0)
C	H_0 : There is no difference among all Training Programs in User Main Groups	$\chi^2 = \chi^2_S + \chi^2_{ET} + \chi^2_{Rubric} + \chi^2_{ET+Rubric}$ $= 16.936 + 12.999 + 21.05 + 8.843 = 59.827$	0.001* ($\mathbf{H}_1$)
	H_0 : There is no difference among Simulation, Emb.Training and ET+Rubric training programs in Main User Groups	$\chi^2 = \chi^2_S + \chi^2_{ET} + \chi^2_{ET+Rubric}$ $= 9.137 + 7.867 + 12.477 = 29.481$	0.079 (H_0)
D	H_0 : There is no difference among all Training Programs in Main User Groups	$\chi^2 = \chi^2_S + \chi^2_{ET} + \chi^2_{Rubric} + \chi^2_{ET+Rubric}$ $= 36.848 + 12.011 + 19.722 + 11.105 = 79.686$	<.001* ($\mathbf{H}_1$)
	H_0 : There is no difference among Emb.Training, Rubric, and ET+Rubric training programs in Main User Groups	$\chi^2 = \chi^2_{ET} + \chi^2_{Rubric} + \chi^2_{ET+Rubric}$ $= 11.936 + 10.045 + 8.101 = 30.082$	0.069 (H_0)
E	H_0 : There is no difference among all Training Programs in Main User Groups	$\chi^2 = \chi^2_S + \chi^2_{ET} + \chi^2_{Rubric} + \chi^2_{ET+Rubric}$ $= 25.146 + 11.005 + 10.878 + 6.682 = 53.711$	0.005* ($\mathbf{H}_1$)
	H_0 : There is no difference among Emb.Training, Rubric, and ET+Rubric training programs in Main User Groups	$\chi^2 = \chi^2_{ET} + \chi^2_{Rubric} + \chi^2_{ET+Rubric}$ $= 7.847 + 7.731 + 5.011 = 20.589$	0.422 (H_0)
F	H_0 : There is no difference among all Training Programs in Main User Groups	$\chi^2 = \chi^2_S + \chi^2_{ET} + \chi^2_{Rubric} + \chi^2_{ET+Rubric}$ $= 19.080 + 6.419 + 18.860 + 10.735 = 55.094$	0.004* ($\mathbf{H}_1$)
	H_0 : There is no difference among Emb.Training, Rubric, and ET+Rubric training programs in Main User Groups	$\chi^2 = \chi^2_{ET} + \chi^2_{Rubric} + \chi^2_{ET+Rubric}$ $= 3.454 + 13.822 + 10.105 = 27.381$	0.125 (H_0)

D=4, E=5, and F=6) internally. The outcomes of χ^2 analysis are introduced in Table 4.

We see from Table 4 a statistically significant difference between four training programs in main user groups for the first week ($\chi^2 = 119.092, p < .001$). We find the largest chi-square value as the chi-square value calculated for the Emb. Training ($\chi^2_{ET} = 48.369$). In other words, Emb. Training emerges as the differentiator group among all. To find any potential training program among the remaining ones, we exclude Emb. Training and continue to investigate by computing chi-square values. The computation covering Simulation, Rubric, and ET+Rubric groups reveals a statistically significant difference ($\chi^2 = 58.097, p < .001$) along with having the Simulation group as being the distinguishing one ($\chi^2_S = 25.352$). In the next stage, we exclude the Simulation group and inspect any further differences between Rubric and ET+Rubric groups. As a result, we detect a statistically significant difference between the two of them ($\chi^2 = 19.111, p = 0.039 < 0.05$). Overall, click rates among all training programs are different when the first week is considered. For the second week, the chi-square analysis reveals a statistically significant difference among all training programs ($\chi^2 = 54.164, p = 0.004 < 0.05$) and ET+Rubric is the differentiator group. Next, we exclude ET+Rubric and continue to investigate by computing chi-square values. According to the results, however, there exists no significant difference among the other three groups (Simulation, Emb. Training, and Rubric) ($\chi^2 = 23.444, p = 0.268 > 0.05$).

Inspection of the third-week layout showed a statistically significant difference among all four training programs ($\chi^2 = 59.828, p = 0.001 < 0.05$) whereas the distinct group was found as the Rubric. However, excluding the Rubric sub-group creates no difference among the other groups ($\chi^2 = 29.481, p = 0.079 > 0.05$).

Similarly, analysis over the fourth week indicates a statistically significant difference among all four training programs

($\chi^2 = 79.686, p < .001$) while the Simulation group emerges as the differing group. ($\chi^2_S = 36.848$). Performing the chi-square analysis on the frequency table created with the Emb. Training, Rubric, and ET+Rubric training programs reveal no statistically significant difference between these three training programs within main user groups ($\chi^2 = 30.082, p = 0.069 > 0.05$). The analysis over the fifth week clearly shows a statistical difference among the training programs ($\chi^2 = 53.711, p = 0.005 < 0.05$). Moreover, the differing group is found as the Simulation ($\chi^2_S = 25.146$). Performing the chi-square analysis on the frequency table created with Emb. Training, Rubric, and ET+Rubric laid out no statistical difference among these three training programs ($\chi^2 = 20.589, p = 0.422 > 0.05$). Combining these two findings, we state that the "Simulation" group is statistically different from the other three training groups.

The analysis of the sixth week on four training programs shows a statistically significant difference among these training programs ($\chi^2 = 55.094, p = 0.004 < 0.05$) while the distinguishing one is the Simulation group. On the other hand, performing the chi-square analysis on the frequency table created with Emb. Training, Rubric, and ET+Rubric noted no difference among these groups ($\chi^2 = 27.381, p = 0.125 > 0.05$). Thus, it can be inferred that the "Simulation" group is different from the rest. Consequently, we have discovered statistically significant differences among the training programs by having different distinguishing training programs per week. The first week's data indicate differences among all groups. We should note that participants had no training before the first week. It is, therefore, possible to accept this week as the starting week because no user had any form of pre-training before (the effect of any possible awareness training for any individual prior to our study is neglected – since it is not measurable). Subsequently, during the second and third weeks, Emb. Training and Rubric groups are distinguished from the rest, respectively, which can be inferred as some kinds of training, such as embedding training and Rubric delivery, start to create a difference. Starting with the fourth week, the Simulation group has become and continues to be the distinguishing one without any exception. Training becomes meaningful for the user groups after the third week. However, we could not find any advantage in one out of three different training schemes.

Another interesting fact is that most users (4963, 15.45%) only completed one training lesson, as displayed in Figure 5. However, as the chi-square analysis shows, all groups that received training material performed better over time. Consequently, we conclude that the training material is not the driving factor for fewer click rates because most participants never completed more than one training course. Furthermore, we assume participants receiving training material were more cautious after getting caught once because they would assume that more phishing simulations would follow. On the other hand, participants without training material would not get any information that they fell for a phishing email and thus would not know that we were sending phishing simulations.

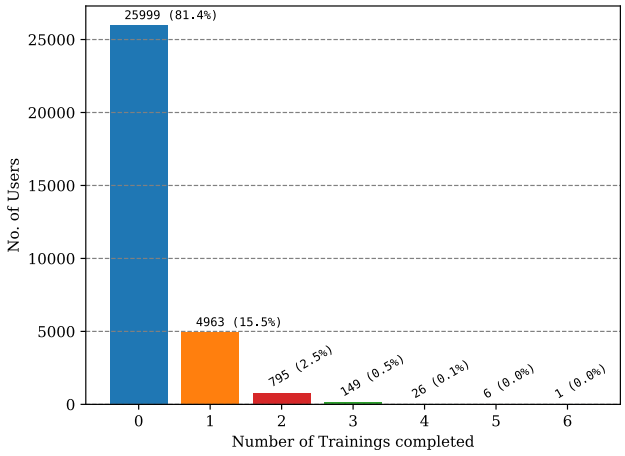

FIGURE 5. Number of times users completed a training.

VII. MACHINE LEARNING DRIVEN PHISHING EMAIL DIFFICULTY PREDICTION

The last two decades of information security field have witnessed the establishment of numerous approaches aiming at identification of phishing characteristics in several modalities such as emails [93], [94], web pages [95], URLs [94], [96], SMS messages [97], and visual contents [95], [98]. The incorporation of machine learning methodologies and the continuous progress in AI systems have resulted in the proliferation of highly accurate classifiers and detectors. These attempts employed and tested various features extracted from different sources of information mentioned above. While relatively past studies focused on manual feature crafting (e.g., presence of "HTTPS" or domain registration date), current studies have evolved so that representations are obtained through more deep architectures. In particular, inventions in the Natural Language Processing field, such as attention-based Transformers (e.g., BERT, GPT3), has revolutionized the way of semantic understanding of text data.

However, it should be noted that those studies generally focus on either phish/legitimate discrimination or brand-based identification, requiring the class labels to be obtained through automated threat services or human annotations. In this work, we follow a different path, diverging from other studies, and attempt to predict the convincing power of emails that we call *difficulty* via machine learning approaches for the first time. The primary motivation behind this idea is to explore and validate whether a system can estimate the average *difficulty* perceived by users. It is evident that the development of such a system has a great potential in (a) scheduling of awareness training by practitioners effectively and (b) efficiency gain in the perspective of CISOs [56]. This purpose relies on estimating human perception, which is pointed by [56] and diverges from others in the following aspects:

a: HUMAN FACTOR

Prediction of human perception for a particular stimulus is related to explicit and implicit human-based factors making it more challenging to model. Thus, this kind of study can be seen as probing the human brain to explore the factors playing a role in critical decision-making processes. Likewise, the perception of phishing is no exception of this. As the behavior of falling into phishing sources for numerous reasons and has direct connections to the Kahneman's System I [99] (i.e., fast and unconscious decision making), understanding the driving factors in falling into phishing needs continual research adapting to new tactics exploited by attackers.

b: DATA LABELING

As is known, the perception is individual. Nonetheless, phishing email difficulty prediction requires finding a consistent way to compute *average human perception* in order to be used by CISOs when examining communities. Our problem, therefore, needs a different labeling scheme rather than the use of conventional binary labels. So, we employ click rates and some thresholds to establish a systematic way for data labeling.

c: FEATURES USED

Traditional machine learning-based phishing recognition schemes need to employ automatically extractable features independent of the underlying methods they leverage. In other words, they need to rely on measurable features. Nonetheless, as stated above, estimating the human phishing perception requires some subjective, hard, and even impossible to measure features such as *familiarity* of an individual with a specific brand or service. The nature of our problem, thus, addresses the use of manual features besides automatically extractable features, as shown in Table 6 and Table 5

In the following sub-sections, we introduce our model proposal by presenting (i) the methodologies, features, and tools we employed, (ii) details about training and evaluation, and (iii) experimental results.

A. LEARNING-BASED METHODOLOGIES USED

This section will overview the primary methods we employed during our model design. Due to space constraints, we prefer to provide the selected methods' fundamentals, benefits, and justifications. Interested readers can refer to the given papers for further information.

1) SENTENCE TRANSFORMERS

The last decade of the NLP world has witnessed a series of advances and the increasing use of deep architectures. In simple words, a piece of textual data (i.e., sequence) has to be converted into a vector form to be processed by computers, which was first achieved by the model of *Bag of Words* (BOW), where each term is first detected and added to an extensive ordered dictionary. The absence of meaning in BOW has led to the born of *Word Embeddings* (e.g., Glove, WordVec), in which the semantic similarities and relationships were taken into account in a local manner. *Recurrent Neural Networks* (RNNs) were later developed in order to involve the positional information of given words. However, the well-known vanishing gradients problem in RNNs caused the invention of LSTMs [100] which can work unidirectional or bidirectional to capture past and future data. Next, a variety of *Attention Mechanisms* (e.g., additive, multi-head) were developed to detect the most relevant elements of a given data resulting in more accuracy gain. Although these approaches pose impressive results, their drawback of sequential processing makes it challenging when large amounts of data come into play. Besides, Word Embeddings pose a very narrow contextual window for a word because of utilizing only its nearby words [101]. Eventually, the Transformer [102] models such as BERT [103], and RoBERTa [104] have emerged and performed state-of-art results in various NLP tasks by introducing positional encodings, multi-head attention networks, and ability of parallel training.

BERT and alike models present beneficial information-rich dense representations. However, for some sentence-pair regression tasks such as semantic textual similarity, the design of those models is not well suited for constructing sentence embeddings [105]. Introduced by Reimers and Gurevych [105], Sentence-BERT (SBERT) model produces sentence-level embeddings rather than word or token level either. In addition, instead of using a cross-encoder module, SBERT leverages siamese and triplet networks to generate semantically meaningful sentence embeddings [105] that can be used for several tasks such as clustering, question & answering information retrieval, and textual similarity comparison. Fundamentally, SBERT is based on BERT model, and utilizes a siamese triplet network schema which contains a pooling layer and a soft-max classifier on top of it, as depicted in Fig.6. Given two sentences, namely A and B, the produced vectors u for A and v for B are concatenated together with the element-wise difference vector $|u-v|$ yielding the resultant vector to be fed into the softmax classifier for training given in Eq. (3), where W_t corresponds to weights

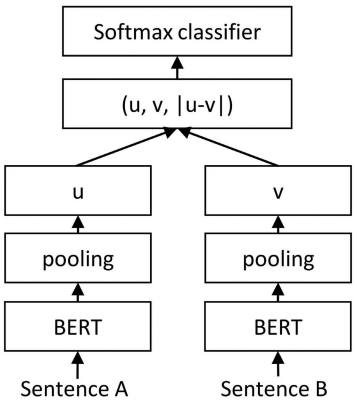

FIGURE 6. Overview of the SBERT neural model [105].

and holds for $W_t \in \mathbb{R}^{3n \times k}$ [105].

$$o = softmax(W_t(u, v, |u - v|)) \tag{3}$$

The trained weights can be used to estimate the similarity. Note that the degree of similarity of two vectors u and v can be measured via cosine similarity. To determine the similarity, the triplet loss function given in Eq.(4) is used

$$max(||s_a - s_p|| - ||s_a - s_n|| + \epsilon, 0) \tag{4}$$

where $s_a - s_p$ denotes the distance between an anchor sentence and its similar counterpart (i.e. positive). Similarly, $s_a - s_n$ represents the distance between the anchor and a dissimilar sentence (i.e. negative). Note that $||.||$ defines the metric (e.g. Euclidean) whereas ϵ is used to ensure that s_p is at least ϵ closer to s_a than s_n [105].

With the advent of Transformers, apart from their monolingual counterparts, the multi-lingual models (MLMs), which support more than 100 languages, emerged. The main goal of MLMs is to produce feature vectors as close as possible for the sentences that are the same in meaning but different in language. Thus, it is aimed to remove the language barrier for tasks associated with semantic similarity. Given a piece of text not longer than 512 tokens, the trained model can be used to generate fixed-sized sentence vectors which are semantics preserving. Consequently, we employed three different sentence transformers, namely *BERT*, *RoBERTa* and *XLM Multilingual* that are publicly served by the HuggingFace community.

2) ZERO-SHOT LEARNING

Apart from being a rising trend in recent years, zero-shot learning (ZSL) is a promising transfer learning scheme when a low data or low resource regime(s) comes into play. In simple terms, ZSL aims to build a classifier through a finite set of class labels and later perform classifications in the wild with a different set of labels that the classifier never observed before. According to [106], ZSL is learning how to identify fresh concepts by just describing them. In a similar vein, authors of [107] have shown that a comprehensively trained language model, the so-called GPT-2, performs well in other downstream tasks without any fine-tuning along with a new training dataset. Since it is a highly active field, the literature has witnessed several ZSL approaches, especially in the NLP domain.

In this work, we leveraged the approach suggested by [108], which employs a pre-trained Multi Natural Language Inference sequence-pair classifier as a zero-shot text classifier. As is known, Natural Language Inference (NLI) deals with predicting whether the given hypothesis h and premise p (a) contain any logical connection such as entailment, (b) exhibit contradiction, (c) be neutral in to each other [109]. The fundamental idea of this method is to take the sequence to be labeled as the ''premise,'' convert each candidate label into a ''hypothesis,'' and evaluate its level of ''coherence.'' In other words, the NLI model predicts whether the premise ''entails'' the hypothesis. The probability scores assigned to the prediction enable us to determine the ''best'' label among the given ones. In this way, it becomes possible to identify the coherence of a set of unknown labels with the premise.

Since emails are a universal way of communication, we intentionally adopted a language-agnostic ZSL classifier in this study. The pre-trained model of the approach we employed can be found online.[1]

3) SUPERVISED MANIFOLD LEARNING VIA UMAP

In many of the supervised ML tasks, acquiring more features for an entity is often preferred since it is expected that the more features the algorithms are fed, the more accurate models we obtain. This expectation, however, does not hold for every case due to the well-known problem of the *curse of dimensionality* which leads to confusion for the algorithms and yields poorer results. Further, the risk of overfitting emerges when the number of observations is significantly lower than the number of features to be employed. To combat those problems, apart from feature selection techniques, dimension reduction methods (e.g., PCA [110], Isomap [111], t-SNE [112]) were proposed. Recently, McInnes *et al.* [113] suggested a dimension reduction and manifold learning scheme so-called ''UMAP,'' by addressing several shortcomings of previous works such as high computational cost, large memory requirements, low speed, and the inability to preserve local and global structures.

The fundamental goal of a dimension reduction method is to project the data points lying in a high dimensional space into a lower-dimensional space by preserving the similarities and dissimilarities in the original space. From the technical point of view, a dimension reduction algorithm considers the data points are uniformly spread in a manifold that can be approximated and projected into a lower-dimensional space. Throughout this procedure, the parameters that define the topological structure (manifold) of high-dimensional space can be learned unsupervised. To do so, the UMAP first constructs a high dimensional graph structure through the concept of *fuzzy simplicial complex* which can be considered as a weighted graph storing the weights of edges indicating the degree of probability between the vertices [114]. Next, controlled by the diameter hyperparameter, it is determined to connect an observation to another, relying on whether they overlap situations within a volumetric space [115]. Then, due to the necessity of a careful consideration between obtaining very tiny or large clusters, UMAP adjusts a diameter according to the distance of nearest neighbors of each observation in the higher dimensional space to fuzzify the graph by also lowering the connections' possibility together with an increased connecting diameter [114]. Overall, UMAP optimizes this *fuzzy* graph layout iteratively through the Stochastic Gradient Descent algorithm. Finally, the projection mapping is learned, saved, and can be used later as a transformer.

[1] https://huggingface.co/facebook/bart-large-mnli

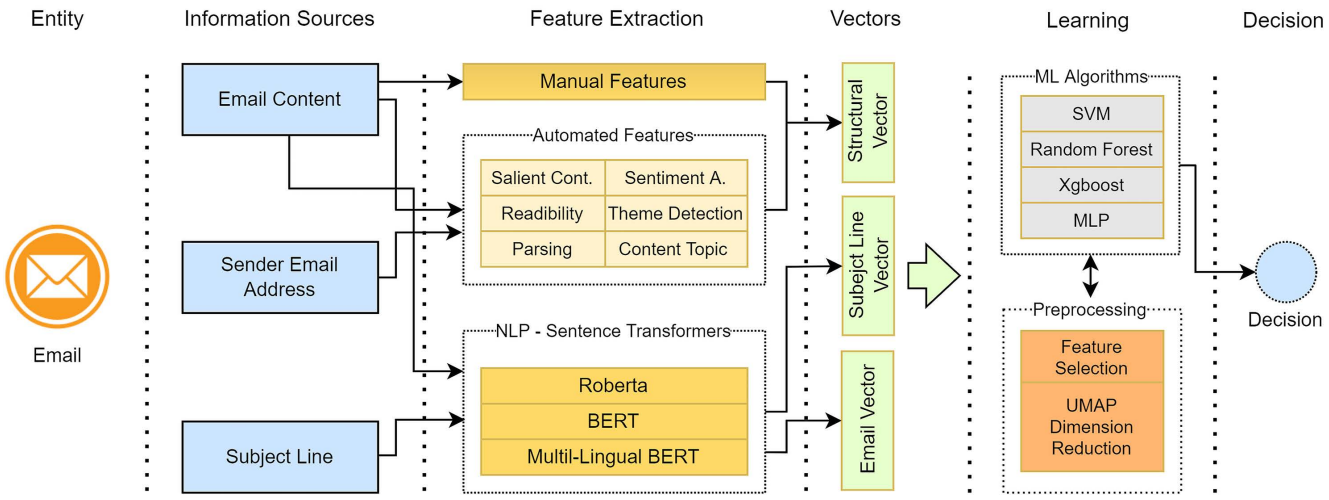

FIGURE 7. Data and work flow of the proposed machine learning based phishing email difficulty prediction.

In this study, we employed the supervised UMAP, a class label powered version of UMAP, in both feature sets to (i) increase the classification performance and robustness, (ii) acquire more discriminative lower dimensional embeddings by transforming the test set through the *learned mapper function* gained from the training set, (iii) visualize the data samples in 2-D. UMAP requires careful tuning of several vital parameters such as n_neighbors, min_dist, n_components, and metric. Nonetheless, due to space constraints, we suggest the reader [113] for further reading on more advanced explanations. As a side note, to utilize the UMAP in our system, we have used the ''umap-learn'' Python package, whose built version is ''0.5''.

B. FEATURES EMPLOYED

To date, the classification of phishing emails has been studied in many works [93], [94], [101], [123] by employing various structural, syntactical, lexical features as well as NLP based features. In particular, recent studies such as [93], [101] benefit from the natural language-based understanding. In a different vein, to assess the phishing susceptibility of users, [124] leveraged questionnaire-based information such as demographic, personality, security and knowledge experience, etc. In this context, as can be seen, many works in the anti-phishing literature mainly aim at discovering the patterns in (i) phishing contents through a wide variety of cues and (ii) user-centered information affecting the behavioral activities. Consequently, inspired by these works, our ML-driven difficulty prediction model makes use of two main feature categories, namely *structural* and *NLP based semantic* features derived from (i) email content, (ii) sender email address, or (iii) subject line as depicted in Fig. 7. However, it should always be kept in mind that our approach diverges from many studies by its problem domain. In this regard, according to our best knowledge, the closest study to ours is the work conducted by [58].

Unlike plain text-based emails, emails built via HTML-based templates often involve one or more irrelevant parts for the main message (e.g., footer, header, etc.). According to us, the impact of these parts, though cannot be completely ignored, is limited. We, therefore, decided to focus on acquiring the most ''core'' message, which will be input for the subsequent feature extractors. However, as is known, revealing the *main content* out of a more noisy text is a challenging task. Thus, we have investigated several open-source solutions and finally preferred to employ Trafilatura [125] as our main content extractor to be used in email bodies. The primary rationale behind this selection was its high F1 score continuously reported in updated benchmarks. Technically speaking, from a given HTML file of the email, Trafilatura extracts the core and *noise-cleaned* plain text content. In the second stage, we leveraged Google Translate API to translate non-English contents (i.e., email body and subject line) into English since the readability scores we compute are only available when the content is in English. Together with the use of Beautiful Soup[2] python interface, we constructed all our pre-processing mechanisms for further operations.

1) STRUCTURAL FEATURES

Throughout this study, we call the structural features for the features having binary, scalar, or multinomial variables either. Depending on the acquisition scheme, we categorize them into *automated* and *manually crafted* features which are comprehensively introduced in Table 5 and Table 6 respectively. The manual features essentially require human intervention, whereas the automated features can be computed via either machine learning or conventional scripting. Similar to the works [124], [126], [127], we assume that the behavior of falling into phishing email is highly related to the content exposed and human-centric factors. We further hypothesize

[2]https://www.crummy.com/software/BeautifulSoup/

TABLE 5. Automated features and their summary.

Feature Name	Criteria	Datatype	Distribution	References
Hyphen in harmful link URL	Existence of hyphen character in the target URL	Categorical	43.3% True - 56.6% False	[116], [117], [118], [119]
"click here" link	Existence of "click here" or a very similar text piece in any button or clickable link	Boolean	22.3% True - 77.6% False	[116]
"account" in email body text	Existence of "account" text in the message	Boolean	34.2% True - 65.7% False	[116]
"information" in email body text	Existence of "information"text in the message	Boolean	18.1% True - 81.8% False	[116]
"security" in email body text	Existence of "security" text in the message	Boolean	4.1% True - 95.8% False	[116]
Average number of dots in URLs of link(s)	Average number of dots found in anchor of link(s)	Continuous	Mean: 1.762 - S.D.: 0.628	[116], [118], [120]
Average URL length of email links	Average target URL length of found link(s)	Continuous	Mean: 52.25 - S.D.: 39.04	[116], [120]
Readability score (CLI)	CLI based score for readibility	Continuous	Mean: 19.02 - S.D.: 7.75	[116], [121]
Readability score (ARI)	ARI based score for readibility	Continuous	Mean: 24.38 - S.D.: 10.82	[116], [122]
Positive sentiment percentage of core email message	Fraction of positive sentences to the number of all detected sentences	Continuous	Mean: 0.689 - S.D.: 0.252	Ours
Negative sentiment percentage of core email message	Fraction of neutral sentences to the number of all detected sentences	Continuous	Mean: 0.311 - S.D.: 0.252	Ours
Emotion of email subject line	Email subject emotion identification through zero shot learning classification and Plutchik's wheel of emotion	Categorical	loathing: 1.3%, ectasy: 4.1%, admiration: 14.6%, amazement: 5.5%, rage 2.7%, vigilance: 32.1%, terror: 1.3%, other: 37.7%	Ours
Theme of email subject line	Zero shot learning classification based email subject theme identification through a pre-defined label set	Categorical	security: 17.4%, business: 17.4%, announcement: 29.3%, request:12.5%, urgency: 10.4%, vacation: 1.3%, romance: 1.3%, other: 10.4%	Ours

that the luring potential of the message has implicit semantical and emotional connections to the content, as pointed out by [58], [128] too.

Thus, apart from the cited features given in Table 5 and Table 6, we propose new and novel features to gain more discriminating representations together with better generalization capability. Due to space limitations, we only provide the details of the features we contributed and the ones needing clarification.

2) SENTIMENTAL DISTRIBUTION OF THE EMAIL BODY

According to [127] and [23], phishing emails often contains emotional and motivational appeals. Moreover, as pointed out by [141], the *Socio-Emotional Selectivity Theory* [142] has some potential constructs in phishing susceptibility. In simple terms, this theory proposes that older people compared to younger tend to be more influenced by emotions in their goals due to the perceived limitation of future time left. Similarly, according to [133], phishers attempt to influence victims through particular emotional triggers like fear and anticipation. Therefore, we decided to approach the problem by including sentimental analysis to explore the emotional distribution of the email body and subject line.

With this aim in mind, we leveraged a relatively recent Transformer based sentiment classifier using the "RoBERTa" language model [104] as the back-end. The model, which was fine-tuned and evaluated on fifteen data sets from diverse text sources, allows binary fashion

TABLE 6. Manually crafted features and their properties.

Feature Name	Criteria	Datatype	Distribution	References
Hidden URL in links	Is any link rendered different than real URL?	Boolean	49.6% True - 50.3% False	[116], [129], [130]
Presence of company's brand logo	Does the email contain the logo of imitated brand?	Boolean	38.4% True - 61.5% False	[98], [126], [127], [128]
Typosquating in link(s) of email body	Does the link of email contain typosquating?	Boolean	47.5% True - 52.4% False	[131], [132]
Typosquating in sender email address	Does the email sender name contain typosquating?	Boolean	51.7% True - 48.2% False	[131], [132]
Is the harmful link button (True) or text(False)	Is the harmful link rendered as button or text?	Boolean	10.4% True - 89.5% False	[123]
Is the email body HTML content (True) or plain text (False)	Does the email have HTML layout or is it plain text?	Boolean	67.8% True - 32.1% False	[94]
Authoritative tone in email	Does the email contain an authoritative tone in content?	Boolean	10.4% True - 89.5% False	[56], [75]
Urgency tone in the email	Does the email contain an urgent situation creating time pressure?	Boolean	41.2% True - 58.7% False	[56], [75], [126], [127], [133]
Misspelling(s) in the email	Does the email body contain any misspelling, mismatched plurality?	Boolean	17.4% True - 82.5% False	[126], [129], [130], [134]
Inappropriate capital letters in subject line	Does the subject line contain irregular capital letters?	Boolean	2.1% True - 97.9% False	[75]
Use of 3 or more different font face	Does the email body use at least 3 different font faces?	Boolean	11.8% True - 88.1% False	[75]
Disparity between domain name of sender and harmful link	Is the domain of the sender email consistent with the domain of harmful link's URL?	Boolean	54.5% True - 45.4% False	[123]
Personalized greetings	Does the email include personalized greeting?	Boolean	23% True - 76.9% False	[56], [126], [134], [135]
Informal language	Is the language of the email informal?	Boolean	21.6% True - 78.3% False	[94]
Homographic URLs	Does the email contain any homographic URL?	Boolean	17.4% True - 82.5% False	[94]
Familiarity	Is the email an expectable one for the user such as a company-wide email?	Boolean	46.1% True - 53.8% False	[56], [127], [136]
Prize or reward offers	Does the email offer a direct profit of something in favor of the receiver? (e.g. discount concert ticket)?	Boolean	27.4% True - 72.7% False	[56]
Consistency and coherence in visual design	Does the visual design of the email fulfill those conditions?: (a) consistent font harmony, (b) the layout the elements are coherent, (c) images are shown correctly	Categorical	82.5% Yes, 4.1 No, 13.2% Unclear	[**?**], [134], [137], [138]
Copyright and/or legal notices	Does the email involve any copyright or legal notice information at the bottom?	Boolean	6.9% True - 93% False	[126], [138]
Contact info	Does the email present any contact information of the sender person/company?	Boolean	32.8% True - 67.1 False	[56], [138]
Presence of "too good to be true offers"	Does the content present something that is too good to be true (e.g. having won a lottery, free holiday)	Boolean	13.2% True - 86.7% False	[56], [139], [140]
Circle of relevance	The general relevance rating for the audience	Categorical	Unknown company/person: 30.7%, lesser known comany: 5.5%, well-known company or friend: 11.1%, businessplace or school: 52.4%	Ours

sentiment analysis for English-language texts by outputting either positive or negative sentiment. The model is being hosted by the Hugging Face model hub.[3] We first split the content produced by Trafilatura into sentences via NLTK Language Toolkit.[4] Second, each sentence was fed into the model, and depending on the outcome, we built a histogram of sentiments for both extracted email content and the subject line. An example output presenting the ratios of positive and negative sentiments is shown in Fig. 8. Meanwhile, we employed the model for inference rather than making any prior fine-tuning to prevent any possible bias.

3) EMOTION OF THE SUBJECT LINE

In a similar vein to the sentimental distribution of the email body, we also inspected the core emotion of the subject line since it constitutes the first perceived element of the received email. Likewise, we aimed at finding out whether any emotional aspect incepted by the subject line plays an influencing role on users and followingly causes them to fall into phishing. However, unlike doing positive/negative type discrimination, we instead followed the framework of Plutchik's *wheel of emotions* [143], [144] which serves as a psycho-evolutionary classification approach for general emotional responses.

To achieve this, we selected the eight-core emotions in the Plutchik's wheel of emotions as follows *ectasy, admiration, terror, amazement, grief, loathing, rage, vigilance*. At this stage, we utilized the zero-shot learning paradigm. As mentioned before, ZSL enables performing classification from a given pre-defined set of labels without requiring any model fine-tuning or re-training. In the implementation, we let the algorithm run in multi-class mode and set a threshold of 0.5 to validate the class probability of the top-1 ranked ZSL outcome. If and only if the likelihood of the top-1 prediction exceeded the defined threshold, we accepted it as a confident prediction. Otherwise, we rejected, assigning the label of ''unknown/other.''

4) TOPIC OF THE SUBJECT LINE

We also investigated whether the topic/concept of the email subject line impacts phishing or legitimacy perception. A preliminary study [133] lists the most used themes of phishing email as ''account verification, update, confirmation, validation,'' ''document sharing,'' and ''payment, transaction or bank related issues.'' Nonetheless, the challenge of determining a highly distinguishing set of topics has led us to follow an experimental methodology. We first defined ten different topic-set, each having various topics ranging from 6 to 10. Although phishing emails originated from a more diverse set of themes as exampled in [145], we reduced it since the number of data points we have is a compelling factor. Next, we downloaded a publicly-available email dataset containing more than 350 emails and performed ZSL by examining each topic set and the subject lines. Later, we calculated the likelihood of each prediction for each topic set and observed the overall confidence of the ZSL model we used when performing each topic set. According to our experiments, the topic-set involving the following concepts, namely *announcement, business, request, romance, security, urgency, vacation* yielded the highest average confidence score. The same threshold-based label assignment verification procedure (t=0.5) was also applied.

As a result, we have employed the ZSL approach for our dataset by providing the topic-set mentioned above and labeling them as a pre-processing stage. Nonetheless, it should be kept in mind that the given predefined label sets in our ZSL-powered study may not be universally representative and needs further investigation with much larger data collected from diverse cultures.

5) READABILITY SCORES

In general, the readability score is used to measure how hard to perceive a piece of text by people. According to [116], readability is an essential part of accessibility [146], and it has direct implications for phishing emails. Regarding the readability, Song [147] points out that the complexity of the text-based stimuli has some implications on the behavior of sharing on social media sharing, such as people often tend to share easy to read contents since it is triggered by Kahneman's *System I*. We hypothesize that the readability as a metric could be engaged with textual complexity and, thus, a link between *System I* [99] and the readability score might be established. Therefore, we leveraged two different readability scores named ARI (Automated Readability Index) and CLI (Coleman Liau Index) to determine whether our hypothesis is valid. Designed for English texts, the automatic readability index [122] is used to calculate the readability score, which is formulated in Eq.(5)

$$ARI = 4.71(\frac{C}{W}) + 0.5(\frac{W}{S}) - 21.43 \tag{5}$$

where C denotes the number of numbers and letters while W shows the number of spaces and S indicates the number of sentences [116]. CLI, on the other hand, suggested by Coleman and Liau [121] is given in Eq. (6)

$$CLI = 0.0588L - 0.296S - 15.8 \tag{6}$$

where L shows the average number of letters per hundred words while S indicates the average number of sentences per hundred words, it should be noted that, in his experiments, Sonowal [116] found that ARI outcomes a more distinguishing histogram compared to the one produced with CLI. Further, in two of these metrics, legitimate emails often exhibit higher scores than phishing ones. In our study, we leveraged the *textstat*[5] Python package to measure both of these readability scores.

[3]https://huggingface.co/siebert/sentiment-roberta-large-english
[4]https://www.nltk.org/
[5]https://pypi.org/project/textstat/

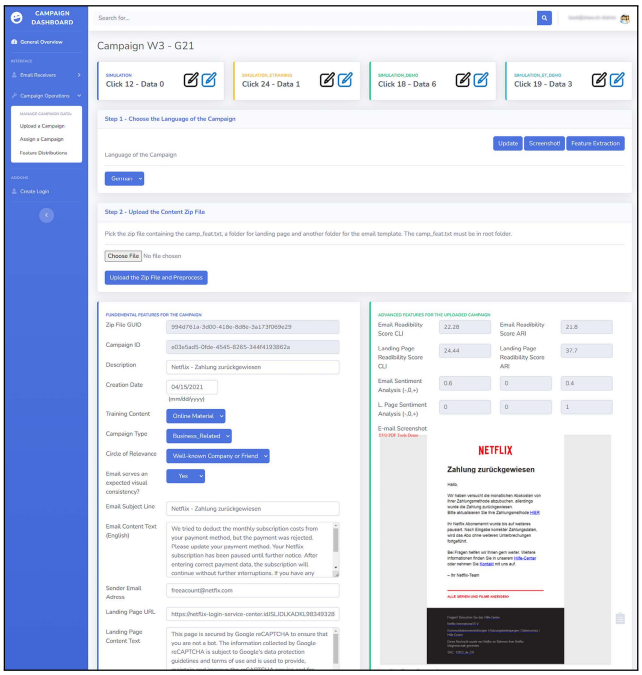

FIGURE 8. Dashboard GUI we developed for campaign management.

6) CIRCLE OF RELEVANCE

The attraction power of an email is highly correlated with the topic and relevance of content to the user. From the user's point of view, if the email is received from an unknown company, the willingness to open or read it will be less compared to an expected email. The word "expected" here plays a key role since it refers to familiarity, relevance, and even credibility to the sender. Therefore, phishing awareness trainers and even spear phishers often send highly relevant emails throughout the work or school environment. Steves *et al.* [59] name this phenomenon as a "premise alignment." According to them, the premise alignment is the adjusting the appropriateness of an email through the acquired context of the target population. To this end, they suggested three levels for this alignment process such as "high," "medium," and "low," to be defined in either a blended or formulaic approach.

However, we argue that the *alignment* categorization/scoring process in [59] is a bit complex and abstract. We, instead, propose a more concrete and easy to perceive feature, the so-called "Circle of Relevance," by taking into account the priority and context of users. We, thus, define four levels such as "Business or School Place," "Well known Company or Friend," "Lesser-known Company," and "Unknown Company or Person," depending on the contextual relevance. We believe that awareness trainers would easily label the campaign emails through this level-based categorization. Nonetheless, it should be noted that, although the *Circle of Relevance* is inherently individual, we here refer to a population-wise assignment determined by the training implementer. Furthermore, we believe that this feature could be used for a person- or group-based analysis.

7) CONTENT ORIENTED SEMANTIC FEATURES

It is an inevitable fact that the content of the subject line and email body have a significant role in individuals' behavior and response to phishing attacks. Several works [133], [145] highlight how the semantics of the emails play a crucial role in luring victims. Thus, we attempted to obtain an information-rich vectorial representation of the subject line and *core* message of the email.

For this purpose, we experimented with sentence transformers (ST) having different underlying pre-trained models such as BERT, RoBERTa, and XLM Multilingual. In the implementation stage, we first extract the *concise* message from the body. Then, we maximally take the first 512 tokens of the content. Finally, we input both subject line and email message into the mentioned models' specific tokenizers for parsing and to be further processed. For example, while BERT-based, ST generates 768-*d* vectors, the remaining two output 1024 dimensional vectors (i.e., logits). Eventually, we obtained two vectors of the given subject line-body pair to be further fed into machine learning models.

According to our best knowledge, our work is the first of its kind that employs the state of the art sentence transformers to obtain paragraph-level representations from email contents. We also assess whether single or multi-lingual-based models better fit our problem by processing both original and translated versions of the email body. It is a known phenomenon that multi-lingual transformer models are prone to perform worse when the input document belongs to a language in which the model had seen insufficient training data. Keeping in mind that our campaign emails are in English or German, our initial expectation would be to obtain similar accuracy scores across all these transformer models since both languages are sufficiently sampled from diverse sources. Nonetheless, emails belonging to less spoken languages might perform worse when multi-lingual-based models are used. Thus, as the first step, we mainly preferred to translate our contents into English through more advanced and complex GPT-3 and alike models, and then utilize the single language models like BERT and RoBERTa.

C. DEVELOPED TOOLS

At the beginning of the study, we first developed a web-based dashboard application for (1) scheduling and over-viewing all 144 campaigns over 12 weeks, (2) storing all manual features, and (3) extracting automated features. In this way, we maintained the data integrity and observed the feature-value distributions. As shown from Fig. 8, the campaign management module can store all campaign features and click rates under different awareness training programs.

Our ASP.NET v3.5-based dashboard was implemented using C#.NET and Python languages. We have utilized Microsoft SQL Server v14 as the central database server for data storage. Meanwhile, the automated feature extraction parts took advantage of various Python packages and Hugging Face inference pipelines.

VIII. EXPERIMENTATION

In previous sub-sections, along with Table 5 and Table 6, we have explained our features in detail. To create a prediction system for phishing email difficulty, we first moved our SQL-based data into a Python environment and created an experimentation and deployment pipeline. We then labeled our target variable by following the guideline in the work of [56]. Accordingly, we have assigned the label ''easy'' for the cases where the ratio of average clickers in four audience subgroups (i.e., Simulation, Embed. Training, Rubric, and ET+Rubric) is less than 10% of the corresponding audience. Similarly, the label ''medium'' was preferred for 10%-20%, whereas the label ''difficult'' was chosen for the cases where the explained ratio is higher than 20%. In this context, we obtained 83 ''easy'' samples and 39 ''medium'' cases, whereas the number of ''difficult'' samples reached 21.

In our experiments, to achieve the best ML model, we investigated the impact of (1) the input data modality or a combination of them, (2) the type of sentence transformer if employed, (3) the machine learning classifier, (4) the presence of a feature selection technique, and (5) parameters of UMAP based dimension reduction if used. As depicted in Fig. 7, we also approached the problem by considering these perspectives. Since we have a limited amount of data points, 144, we have preferred to employ 5-fold cross-validation to ensure generalization capability. Besides, due to the imbalance in our dataset, we used stratified sampling in cross-validation via shuffling. Throughout the modeling stage, we have employed Support Vector Machine (SVM), Random Forest (RF), XGBoost, and Multi-layer Perceptron (MLP) classifiers shipped with *sklearn*[6] Python package.

Furthermore, we applied a standard scaler to normalize the features when we ran MLP. For performing dimension reduction, we leveraged the supervised UMAP. As stated before, UMAP provides a variety of hyper-parameters that affect the *representation performance* (i.e., focusing on local or global similarities) during the projection of high-dimensional features into lower-dimensional vectors. Thus, we systematically performed a grid search to test various hyper-parameters of both UMAP and other ML classifiers as listed in Table 7. Particularly, the asterisk marked parameters such as *'C' and 'number of features'* belonging to a particular method were involved within the grid-search with the corresponding values given as a list. As a result, we conducted extensive experiments to find out the best settings.

We first divided the experiments into two main tracks in a way that either ''Whole Structural/Sentimental Features'' (WSSF) or ''Automated Structural/Sentimental Features'' (ASSF) are fundamentally utilized. Note that, the ASSF set shown in Table 5 is a subset of the WSSF set (i.e. composition of features both in Table 5 and Table 6) and comprises only the automated features apart from external semantic features obtained through sentence transformers. Next, we attempted to discover the outcome of each modality along with their combinations. For this reason, from the modality point of view, we built four kind of models extracting information from (1) pure structural features W/A (SSF), (2) *email content* only NLP *features* (ECF), (3) *subject line* only NLP *features* (SLF) and (4) W/A(SSF) + (ECF and/or SLF) through late fusion of vectors.

[6]https://scikit-learn.org/stable/

TABLE 7. Hyper-parameters of the employed methods. The asterisk marked parameters' values were used in grid-search.

Method	Parameter	Tested Values
SVM	Kernel	RBF
SVM	C*	[1,25,50]
XGBoost	Objective	multi-softprob
XGBoost	Eval metric	merror
MLP	Solver	lbfgs
MLP	Alpha	1e-5
MLP	Hidden layer sizes	(30,10)
Feature Selection	Number of features*	[50,100,200]
UMAP	Number of dimensions	5
UMAP	Neighborhood*	[20,40,60,80]
UMAP	Minimum distance*	[0.3, 0.5, 0.7]
UMAP	Metrics*	['euclidean','hamming', 'cosine']

After having a comprehensive experimentation period which includes 5-fold cross-validation and parameter searching, we have achieved the results given in Table 8 and Table 9. The two tables differ on only SSF and SSF+(ECF and SLF) groups since the other two categories share the same information source. Besides, the right-most column shows the best baseline scores achieved via no feature-selection or dimension-reduction technique. Consequently, our key findings are listed below, relying on the obtained scores. We elaborated on these findings in the discussion section.

- Compared to the use of automated only features, manually crafted features contributed to superior results in terms of all metrics.
- As opposed to our initial expectations, the best performing model yielding an accuracy of 68.54%, F1-score of 66.51% and AUC of 75.02% was obtained through whole SSFs concatenated with the BERT-represented subject line, built with the SVM, dimensionally reduced. Although WSSF+EC performs a relatively close score with an accuracy of 67.85%, this surprising result shows that the subject line provides slightly more specific information than email content.
- As can be seen from the Table 8 and Table 9, the UMAP supervised dimension reduction technique provided performance gain (i.e. +6.49% in average accuracy for WSSF-based experiments whereas +2.51% in ASSF-based experiments), pointing out the benefit of manifold learning in the problem domain. Furthermore, rather than automated features, incorporating all structural features yielded more gain, especially when coupled with supervised UMAP. This provides another perspective suggesting the superiority of WSSF-based features such that manifold learning performs better in class separation in the projected lower-dimensional space. Besides, the metric of *hamming* outperformed the other metrics.

TABLE 8. 5-fold cross-validation results obtained from different modalities through all features. The best model is shown with bold font-face.

Modality	Transformer	Classifier	Feat. Sel.	Dim. Red.	Accuracy - Gain	Precision	Recall	F1-Score	AUC	Base Acc.
Whole Structural/Sentimental Features (WSSF)	N/A	XGBoost	✗	✓	64.360 (+10.9%)	63.177	64.360	61.183	71.899	58.03
Email Content (Original Content)	XLM-Multilingual	MLP	✗	✓	61.552 (+1.18%)	54.740	61.552	49.895	52.664	60.83
Email Content (English Translated)	Roberta	RF	✗	✓	61.552 (+3.53%)	49.704	61.552	51.297	53.040	59.45
Email Content (English Translated)	Bert	MLP	✗	✓	61.192 (+2.93%)	48.264	62.190	50.895	58.599	59.45
Subject Line (English Translated)	Roberta	SVM	✗	✓	62.266 (+4.73%)	50.226	62.266	50.900	56.869	59.45
Subject Line (English Translated)	Bert	RF	✗	✗	60.887 (N/A)	50.241	60.887	51.595	60.946	60.88
Subject Line (Eng. T.) + Email Content (Eng. T.)	Roberta	XGBoost	✗	✓	60.148 (+3.40%)	45.891	60.148	48.343	54.251	58.17
Subject Line (Eng. T.) + Email Content (Eng. T.)	Bert	SVM	✗	✓	61.576 (+1.09%)	46.964	61.576	50.436	67.243	60.91
WSSF + Email Content (Eng. T.)	Roberta	SVM	✗	✓	65.074 (+10.7%)	66.963	65.074	62.514	73.903	58.76
WSSF + Email Content (Eng. T.)	Bert	SVM	✗	✓	67.857 (+9.11%)	65.454	67.857	65.180	70.325	62.19
WSSF + Subject Line (Eng. T.)	Roberta	XGBoost	✗	✓	65.739 (+8.01%)	62.728	65.739	59.132	70.764	60.86
WSSF + Subject Line (Eng. T.)	**Bert**	**SVM**	✗	✓	**68.547 (+12.6%)**	**68.210**	**68.547**	**66.513**	**75.025**	**60.83**
WSSF + Subject Line (Eng. T.) + Email (Eng. T.)	Roberta	XGBoost	✗	✓	63.670 (+5.86%)	54.297	63.670	57.592	72.570	60.14
WSSF + Subject Line (Eng. T.) + Email (Eng. T.)	Bert	MLP	✗	✓	66.453 (+10.4%)	63.829	66.453	63.414	71.649	60.14

TABLE 9. 5-fold cross-validation results obtained from different modalities through automated-only features. The best model is shown with bold font-face.

Modality	Transformer	Classifier	Feat. Sel.	Dim. Red.	Accuracy - Gain	Precision	Recall	F1-Score	AUC	Base Acc.
Automated Structural/Sentimental Features (ASSF)	N/A	RF	✗	✓	58.744 (+2.53%)	37.402	58.744	43.992	58.052	57.29
Email Content (Original Content)	XLM-Multilingual	MLP	✗	✓	61.552 (+1.18%)	54.740	61.552	49.895	52.664	60.83
Email Content (English Translated)	Roberta	RF	✗	✓	61.552 (+3.53%)	49.704	61.552	51.297	53.040	59.45
Email Content (English Translated)	Bert	MLP	✗	✓	61.192 (+2.93%)	48.264	62.190	50.895	58.599	59.45
Subject Line (English Translated)	**Roberta**	**SVM**	✗	✓	**62.266 (+4.73%)**	**50.226**	**62.266**	**50.900**	**56.869**	**59.45**
Subject Line (English Translated)	Bert	RF	✗	✗	60.887 (N/A)	50.241	60.887	51.595	60.946	60.88
Subject Line (Eng. T.) + Email Content (Eng. T.)	Roberta	XGBoost	✗	✓	60.148 (+3.40%)	45.891	60.148	48.343	54.251	58.17
Subject Line (Eng. T.) + Email Content (Eng. T.)	Bert	SVM	✗	✓	61.576 (+1.09%)	46.964	61.576	50.436	67.243	60.91
ASSF + Email Content (Eng. T.)	Roberta	RF	200	✓	61.552 (+1.18%)	55.411	61.552	53.372	63.773	60.83
ASSF + Email Content (Eng. T.)	Bert	RF	✗	✗	60.862 (N/A)	47.445	60.862	50.463	57.135	60.86
ASSF + Subject Line (Eng. T.)	Roberta	RF	200	✓	61.527 (+4.74%)	49.476	61.527	49.638	58.162	58.74
ASSF + Subject Line (Eng. T.)	Bert	XGBoost	✗	✓	60.862 (+1.15%)	49.418	60.862	50.174	56.763	60.17
ASSF + Subject Line (Eng. T.) + Email (Eng. T.)	Roberta	XGBoost	200	✓	60.172 (+1.21%)	40.469	60.172	46.674	56.250	59.45
ASSF + Subject Line (Eng. T.) + Email (Eng. T.)	Bert	RF	✗	✗	61.552 (N/A)	53.162	61.552	51.022	62.620	61.55

This finding is no surprise since most of our SSFs are designed as binary features, in line with the formulation of hamming distance.

- According to our observations, the RoBERTa representations consisting of 1024-*d* vectors slightly outperform 768-*d* BERT based representations for the ECF and SLF models. As opposed to this, merging NLP features with SSF switches this finding in favor of BERT representations.
- Except for three cases, the ''mutual info'' based feature selection technique could not improve the ML models in terms of generalization capability. We believe this is more likely related to the loss of essential NLP features during the training stage. Thus, in our particular problem, this finding emphasizes the feasibility of dimension reduction in favor of sustainable generalization capability.
- No significant advantage among the ML classifiers was observed. Likewise, we could not notice any noteworthy accuracy gain between multi-lingual and single-lingual transformers for our particular problem.

A plausible reason for having superior results with Subject Line Features instead of Email Content Features might be that Transformer models produce more consistent results when the number of input tokens decreases. Technically speaking, shorter sequences often provide more consistent representations. Furthermore, as subject lines usually consist of fewer words, the acquired semantic representations could be more *cohesive* and informative. Nonetheless, a careful inspection reveals that EC and SL provide equivalent gain. However, their exclusive combinations with (W)SSF yield a slightly different score. It should be noted that our finding contradicts the finding of [128] which states the email body has a more significant impact on users' decisions compared to the subject line.

We also conducted a study to reveal the importance scores of the significant SSF features. In this way, we aimed to evaluate the structural features' effectiveness and verify their validity for the phishing email difficulty estimation. Thus, to define the best strategy, we focused on the WSSFs since we are not interested in the NLP features. If we review Table 8 and Table 9, it can be seen that the scores for the topmost rows were achieved through the UMAP technique with five features only. So, we preferred to utilize a model trained with WSSF only. As a result, we retrained the models having only WSSF via all algorithms. Next, we applied the feature selection method to remove irrelevant features and picked the one having the best outcome among the others. The best

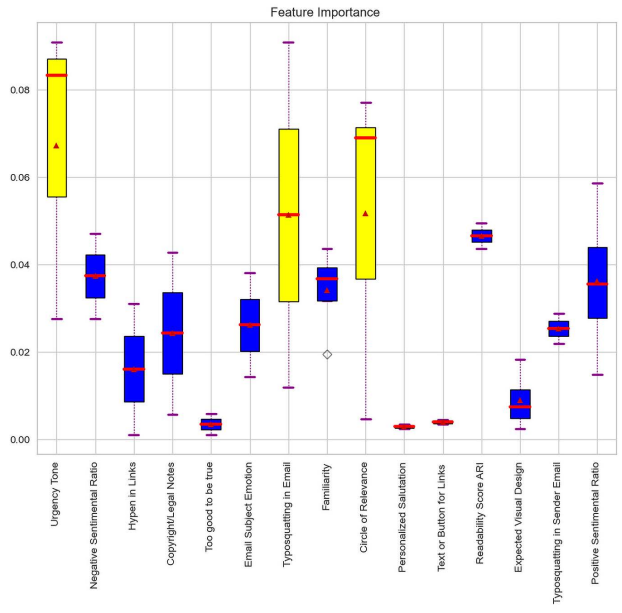

FIGURE 9. Box-plot of significant features with their importance scores for the WSSFs. The yellow colored bars represent the most important three factors. Accounting for five-folds, the triangles indicate mean values whereas red lines show median values.

model (Random Forest) achieves the accuracy of 61.08% on average through the 5-fold CV evaluation. To compute more precise importance values, we used *permutation based* feature importance module, which is a sklearn built-in package. Next, the important features were collected along with their scores and were fed into a box-whisker plot, as can be seen from Fig. 9.

The revealed feature importance scores are only used for gaining some insight into the feature effectiveness. Accordingly, the most impactful three factors were found as (1) the presence of urgency tone in the email, (2) the proposed concept - circle of relevance, and (3) the presence of typo-squatting in the email body text. Furthermore, user familiarity, copyright, legal notes, sentimental scores, and readability scores (ARI) follow them. On the other hand, ''too good to be true offers'' and the presence of personalized salutation were found ineffective in the prediction. Similarly, our evaluations stated that the remaining SSFs were also not found impactful. These results, in line with *System I* of Kahneman, clearly show that the urgency tone continues to play the most critical role in phishing vulnerability among the users. As can be seen from Fig. 10, the distribution of urgency tone among three classes indicates how this feature shows significant ratio contrasts among those difficulty levels. The second most important finding is that our new feature ''circle of relevance'' indicates the significance of familiarity and *premise alignment* [59]. In concordance with literature, we observe that users are more vulnerable when they are confronted with business or school emails due to the high relevance. The distribution of this feature among three classes is given in Fig. 11. As it can be seen, the ''difficult'' class is dominantly involves ''business place or school'' - the highest convincing relevance. Similarly, the ''medium'' class holds similarities to the ''difficult'' category in terms of distribution, implying that the higher the relevance is more likely a user will fall into a phishing email.

One another finding is to detect the Automated Readability Index as a relatively important factor. We believe that this finding has relation to the ''principle of least effort'' [148] which suggests the use of ''shortcuts'' and least effortful mode in human decision making processes. According to [149], daily e-mail reading activities is progressively shape the interaction in a way of reducing the cognitive effort. Thus, it is not very surprising to observe the impact of perceived textual complexity.

As another key finding derived from our models, *Typosquatting* continues to haunt victims. In a similar vein, users still likely fall into phishing without noticing the hyphen character in URLs and even in sender email addresses. Finally, the feature ''expected visual design'' was surprisingly lower than our initial expectation. Nonetheless, it should be kept in mind that this might be sourced because a significant portion of our campaign emails was purely text-based.

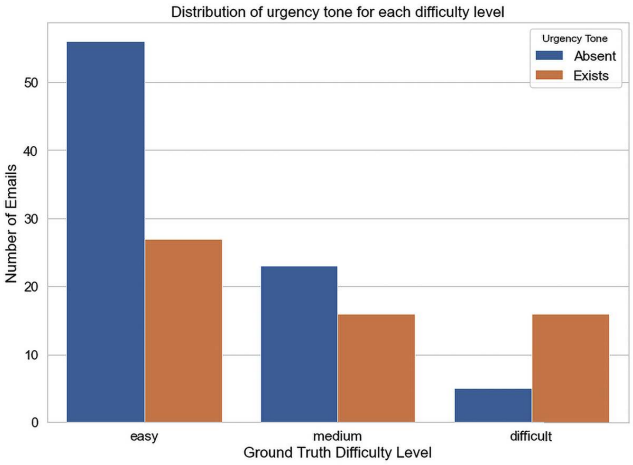

FIGURE 10. Distribution of urgency tone for each difficulty level.

IX. RESULTS

A. RESULTS REGARDING RQ1

In our experiment, we conducted anti-phishing training with a high frequency of attack simulations. Participants would receive either six or twelve emails for three months or six months. However, just over two-thirds of users (68.1%) never submitted their credentials. In addition, 45.8% of the participants never clicked on any of the attacks simulations, and we assume that especially these participants were becoming security fatigued over time as the anti-phishing training had no beneficially effect on them. Thus, credential-based phishing attack simulations are not an effective training method

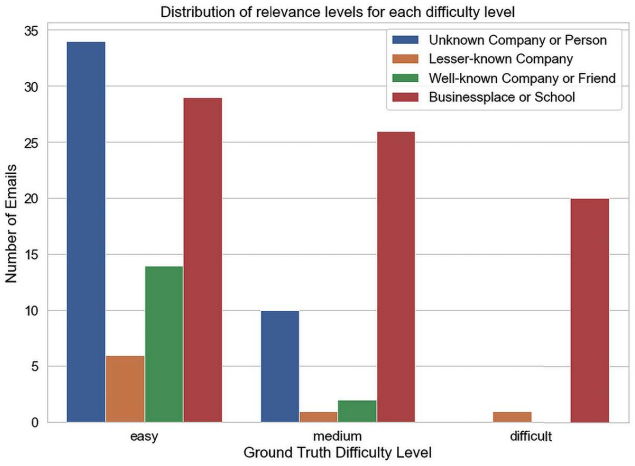

FIGURE 11. Distribution of 'circle of relevance' for each difficulty level.

for most users. This finding is in line with the feedback received by many of our participants, as many of them would complain at our service desk that they do not need such phishing awareness training.

From the received user feedback, we can say that some users were unhappy with the frequency of anti-phishing simulations, and there was a tendency of "security fatigue effect" [87]. Moreover, it shows that phishing awareness simulations should focus on enhancing their training contents towards content more in line with the user's circle of relevance. Alternatively, in other words, to enhance current phishing awareness training methods, the content of the attack simulations needs to be more tuned towards individual users instead of groups.

B. RESULTS REGARDING RQ2

We sent 144 randomly selected phishing simulations, and every participant would receive twelve different emails over twelve weeks.

Furthermore, the email sending time was randomly distributed over the weeks. The result now provides evidence to [150] that the number of times users fail to detect phishing attack simulations follows a power-law distribution. In other words, our results confirm that only a tiny fraction of users repeatedly fall (more than four times) for simulated phishing attacks. Additionally, as shown in Figure 2 there is a clear trend visible that with an increasing number of phishing simulations, the number of repeated clickers decreases.

C. RESULTS REGARDING RQ3

In Section VI we performed a Chi-Square analysis to determine further the influence of training material on the click-behavior of the participants. The analysis shows that all groups with training material performed better over time. Surprisingly, however, as shown in Figure 5 most users that had the opportunity for a training lesson did not fully complete any of our training courses. This is insofar a novel finding as it shows a contradiction. Participants who received learning material performed better over time, but most did not complete any of the training courses. Thus, we can only assume that the groups receiving training material performed better because they were expecting to receive more phishing simulations. We speculate that participants with training material would expect to receive additional attack simulations because they would receive an info mail with the training material when falling for a phishing simulation. On the other hand, participants in the control group would not receive any information about the study when falling for a phishing attack and may not realize they got phished.

Nonetheless, we believe that it is well justified to send training material as we had participants completing several training courses, and we assume that these users could benefit from the training material. Further studies are necessary to test which training courses are more accepted and beneficial for users over time.

D. RESULTS REGARDING RQ4

To our best knowledge, our machine learning model is the first to demonstrate that it is possible to predict phishing susceptibility. Our current model uses state-of-the-art ST (RoBERTa, and BERT) in combination with UMAP dimension reduction on the email content and subject line to classify an email into the three labels "easy," "medium," and "difficult." We also tested the plausibility of a multi-language model. Overall, we combined structural and semantic features. In addition, we illustrate the most significant features and importance scores in Figure 9 and show that urgency, circle of relevance, and typo-squatting are prominent factors for phishing susceptibility. Using 5-fold-cross validation, our model has an accuracy of 68% on a limited dataset of 143 emails but shows that it is possible to predict phishing susceptibility even within this scarce amount of data.

X. DISCUSSION

Unlike an algorithmic point of view, we believe that there are points to be discussed especially for the behavioral perspective and data distribution. First of all, the obtained model reflects the average behavioral patterns of the target community. In other words, the predictive models we built are specific to the community whose data were used during training. Thus, the personality [151], psychology [152], culture, technical background [72], business place, and job type [72] can be listed as significant influential factors driving the audience's response. Apart from these, according to [153], experiential and dispositional factors play a key role in decision-making processes so phishing victimization does. Second, the textual contents involved in our campaigns cover only a very tiny portion of the whole possibilities that the attackers exploit. Thus, according to us, building a much more robust model that can be universally employed requires the following: (1) a much more diverse dataset collected from different countries, cultures and companies and (2) efficient and effective use of transfer learning.

Nevertheless, as the generalization ability assessed through our 5-fold cross-validation scores show, aggregation of some manually crafted/automated features such as ''circle of relevance,'' sentimental distributions, and presence of urgency tone are helpful to capture the variance to a certain extent. At this point, one might ask ''to what extent these features could be extended and measured?''.

In a very recent study, Zhuo *et al.* [152] described the two dimensions of the quality of evidence affecting the ecological validity of the experimental setup for awareness training simulation based anti-phishing studies: (1) *experiment type* and (2) *sample size* in terms of user groups. Instead of conducting an ''email management'' approach which is cost-effective but less accurate, we followed the way of rolling out real-world phishing simulations that are much more accurate despite being restricted by legal and ethical issues. To our best knowledge, this study involves the largest campaign data (i.e. 144 exclusive emails) along with the corresponding click-rate information. Considering the above-mentioned two criteria, the collected data emerges as the largest one. However, from the ML point of view, it is obvious that there is a need for more data. Thus, for cost-efficient future research, we believe that it is beneficial to explore proper strategies for aggregating *simulation* and *plausible survey* data such as the utilization of survey data for weak supervision. Besides, for large companies and institutions, it seems reasonable to take the advantage of active learning and fine-tuning through the supervision of a domain expert. In this way, a pre-trained model created within a similar ecology could be adapted to the needs of the target company/institution. Moreover, the validity of the models could perhaps be checked against cross-datasets.

Ever-growing anti-phishing literature has sought answers for the underlying social-psychological and behavioral factors driving the behaviour of falling into phishing. Studies like [23], [72], [128], [138], [151], [152] aimed at finding or verifying influential factors triggering users to fall into phishing contents. For instance, a recent study [72] points out that the type of computer use is more decisive than the amount of usage. Likewise, [128] states that the use of proper logo and design elements that give an exact look-and-feel immerse victims into clicking. Again, a recent study by Frauenstein and Flowerday [151] attempts to explore the impact of the big five personality traits on phishing susceptibility and reveals the negative correlation between heuristic processing and conscientious users confirming that this personality type is less susceptible to phishing attacks on social media platforms. It is well known fact that the works in the literature often contradicted each other in numerous aspects. The main reasons of these contradictions and limitations that can be listed as (1) number and type of users, (2) number of applied campaigns, (3) duration of the experiment, and (4) experiment environment. However, we argue that the main limitation is being unable to measure individual factors changing over time. Thus, cognitive properties of individuals should also be taken into the account since the perception of email phishing is highly subjective and dynamic. As pointed out by [152], most of the anti-phishing literature has put much effort into the technical aspects leaving the factors for phishing susceptibility not explored yet. In line with our opinions, [152] explained the largest research gap - *user's situational factors* - in phishing susceptibility by exampling some key factors such as ''in the moment emotion'', ''stress'', ''mental fatigue'', ''distraction'' each of which requires careful and realistic future studies. Likewise, it is obvious that measuring these factors are currently not straightforward yet presence of any of these actors will likely contribute to our model. We also believe that studies in human brain interfacing and other biological sensors could open new fields in this problem domain in near future.

As stated before, the demographics of the users were not acquired due to the regulations. The legal, ethical, and organizational constraints mentioned in Section IV are further limitations that cannot easily be overcome. As other researchers may have experienced before, phishing simulations are subject to legal constraints that limit the usage of branded or trademarked email content. Instead of using emails with known trademarks such as Microsoft, Google, or Meta, we mainly used emails with content related to the working space of the participants.

Retrospectively, we think these limitations were beneficial for our study as they forced us to be more creative and use phishing simulations that were more in line with the expectation of our participants. In addition, it did not impair our study experiment as we speculate that most of our students would anyways not use their university mail for most of the trademarked services that we would impersonate in a phishing attack simulation. For example, we see it as unlikely that many of our students would use their students email address for their personal social media accounts because of the fact that the student email address is revoked after graduation.

XI. CONCLUSION

This study has made significant progress towards an automated estimation of phishing susceptibility through the data collected from a large amount of users. The developed ML model shows that a generic and automated estimation is feasible. Nonetheless, to further enhance our model, more data points from different countries, companies, or universities are necessary for increasing the variety of samples and diversity of user profiles and the model's accuracy. Moreover, apart from being promising, the obtained performance scores clearly indicate the necessity of human-centered features towards an ultimate and universal model.

The feature importance analysis revealed the superiority of some features, supporting some existing works in the literature. In this study, we also proposed and evaluated some easy to use features like ''circle of relevance'', ''emotion of subject line''. Furthermore, our approach clearly demonstrated the usability and feasibility of Sentence Transformers and ZSL paradigm in the task of meaning extraction from both for subject line and email body.

Additionally, our study results indicate that most users either easily detect simulated phishing emails or that many users are sufficiently aware of phishing in general. Therefore, it is likely that future phishing awareness training will use ML-based models to further enhance training quality and effectiveness.

XII. ACKNOWLEDGMENT

(Thomas Sutter and Ahmet Selman Bozkir contributed equally to this work.)

REFERENCES

[1] M. Humayun, M. Niazi, N. Jhanjhi, M. Alshayeb, and S. Mahmood, "Cyber security threats and vulnerabilities: A systematic mapping study," *Arabian J. Sci. Eng.*, vol. 45, no. 4, pp. 3171–3189, Apr. 2020.

[2] T. H. Jr. (Dec. 2019). *Lithuanian Man Sentenced to 5 Years in Prison for Theft of Over $120 Million in Fraudulent Business Email Compromise Scheme*. [Online]. Available:https://www.justice.gov/usao-sdny/pr/lithuanian-man-sentenced-5-years-prison-theft-over-120-million-fraudulent-business

[3] B. Krebs. (Jan. 2016). *Firm Sues Cyber Insurer Over $480K Loss*. [Online]. Available: https://krebsonsecurity.com/2016/01/firm-sues-cyber-insurer-over-480k-loss/

[4] E. Kovacs. (Jan. 2016). *Cybercriminals Steal $54 Million From Aircraft Parts Maker*. [Online]. Available: https://www.securityweek.com/cybercriminals-steal-54-million-aircraft-parts-maker

[5] Federal Bureau of Investigations Public Service Announcements. (Apr. 2020). *Cyber Criminals Conduct Business Email Compromise Through Exploitation Of Cloud-Based Email Services, Costing Us Businesses More Than $2 Billion*. [Online]. Available: https://www.ic3.gov/Media/Y2020/PSA200406

[6] H. S. Lallie, L. A. Shepherd, J. R. C. Nurse, A. Erola, G. Epiphaniou, C. Maple, and X. Bellekens, "Cyber security in the age of COVID-19: A timeline and analysis of cyber-crime and cyber-attacks during the pandemic," *Comput. Secur.*, vol. 105, Jun. 2021, Art. no. 102248.

[7] P. Kumaraguru, J. Cranshaw, A. Acquisti, L. Cranor, J. Hong, M. A. Blair, and T. Pham, "School of phish: A real-world evaluation of anti-phishing training," in *Proc. 5th Symp. Usable Privacy Secur.*, New York, NY, USA, 2009, pp. 3.1–3.12.

[8] C. Iuga, J. R. C. Nurse, and A. Erola, "Baiting the hook: Factors impacting susceptibility to phishing attacks," *Hum.-Centric Comput. Inf. Sci.*, vol. 6, no. 1, pp. 1–20, Jun. 2016, doi: 10.1186/s13673-016-0065-2.

[9] S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor, and J. Downs, "Who falls for phish?: A demographic analysis of phishing susceptibility and effectiveness of interventions," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2010, pp. 373–382.

[10] T. N. Jagatic, N. A. Johnson, M. Jakobsson, and F. Menczer, "Social phishing," *Commun. ACM*, vol. 50, no. 10, pp. 94–100, 2007.

[11] T. Halevi, N. Memon, and O. Nov, "Spear-phishing in the wild: A real-world study of personality, phishing self-efficacy and vulnerability to spear-phishing attacks," *SSRN Electron. J.*, pp. 1–10, Jan. 2015. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2544742#

[12] K. W. Hong, C. M. Kelley, R. Tembe, E. Murphy-Hill, and C. B. Mayhorn, "Keeping up with the joneses: Assessing phishing susceptibility in an email task," in *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, 2013, vol. 57, no. 1, pp. 1012–1016.

[13] K. Greene, M. Steves, M. Theofanos, and J. Kostick, "User context: An explanatory variable in phishing susceptibility," in *Proc. Workshop Usable Secur.*, San Diego, CA, USA, 2018, pp. 1–14.

[14] M. Pattinson, C. Jerram, K. Parsons, A. McCormac, and M. Butavicius, "Why do some people manage phishing e-mails better than others?" *Inf. Manag. Comput. Secur.*, vol. 20, no. 1, pp. 18–28, 2012.

[15] O. P. John and S. Srivastava, "The big five trait taxonomy: History, measurement, and theoretical perspectives," in *Handbook of Personality: Theory and Research*, 2nd ed. New York, NY, USA: Guilford Press, 1999, pp. 102–138.

[16] R. Dodge, K. Coronges, and E. Rovira, "Empirical benefits of training to phishing susceptibility," in *Information Security and Privacy Research*, vol. 376, D. Gritzalis, S. Furnell, and M. Theoharidou, Eds. Berlin, Germany: Springer, 2012, pp. 457–464.

[17] K. Parsons, A. McCormac, M. Pattinson, M. Butavicius, and C. Jerram, "Phishing for the truth: A scenario-based experiment of users behavioural response to emails," in *Security and Privacy Protection in Information Processing Systems*, L. J. Janczewski, H. B. Wolfe, and S. Shenoi, Eds. Berlin, Germany: Springer, 2013, pp. 366–378.

[18] G. D. Moody, D. F. Galletta, and B. K. Dunn, "Which phish get caught? An exploratory study of individuals susceptibility to phishing," *Eur. J. Inf. Syst.*, vol. 26, no. 6, pp. 564–584, Nov. 2017.

[19] H. Siadati, S. Palka, A. Siegel, and D. McCoy, "Measuring the effectiveness of embedded phishing exercises," in *Proc. 10th USENIX Workshop Cyber Secur. Experimentation*, Aug. 2017, pp. 1–8.

[20] E. J. Williams and D. Polage, "How persuasive is phishing email? The role of authentic design, influence and current events in email judgements," *Behaviour Inf. Technol.*, vol. 38, no. 2, pp. 184–197, Feb. 2019.

[21] P. Patel, D. M. Sarno, J. E. Lewis, M. Shoss, M. B. Neider, and C. J. Bohil, "Perceptual representation of spam and phishing emails," *Appl. Cognit. Psychol.*, vol. 33, no. 6, pp. 1296–1304, Nov. 2019, doi: 10.1002/acp.3594.

[22] J. Jansen and R. Leukfeldt, "How people help fraudsters steal their money: An analysis of 600 online banking fraud cases," in *Proc. Workshop Socio-Tech. Aspects Secur. Trust*, Jul. 2015, pp. 24–31.

[23] K. Parsons, M. Butavicius, M. Pattinson, D. Calic, A. Mccormac, and C. Jerram, "Do users focus on the correct cues to differentiate between phishing and genuine emails?" in *Proc. Australas. Conf. Inf. Syst.*, May 2015, pp. 1–10.

[24] K. Pfeffel, P. Ulsamer, and N. H. Müller, "Where the user does look when reading phishing mails—An eye-tracking study," in *Learning and Collaboration Technologies. Designing Learning Experiences*, P. Zaphiris and A. Ioannou, Eds. Cham, Switzerland: Springer, 2019, pp. 277–287.

[25] H. S. Jones, J. N. Towse, N. Race, and T. Harrison, "Email fraud: The search for psychological predictors of susceptibility," *PLoS ONE*, vol. 14, no. 1, Jan. 2019, Art. no. e0209684. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/30650114

[26] B. Harrison, A. Vishwanath, Y. J. Ng, and R. Rao, "Examining the impact of presence on individual phishing victimization," in *Proc. 48th Hawaii Int. Conf. Syst. Sci.*, Jan. 2015, pp. 3483–3489.

[27] J. Andric, D. Oreski, and T. Kisasondi, "Analysis of phishing attacks against students," in *Proc. 39th Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, May 2016, pp. 1423–1429.

[28] D. D. Caputo, S. L. Pfleeger, J. D. Freeman, and M. E. Johnson, "Going spear phishing: Exploring embedded training and awareness," *IEEE Secur. Privacy*, vol. 12, no. 1, pp. 28–38, Jan. 2014.

[29] G. Canova, M. Volkamer, C. Bergmann, and B. Reinheimer, "NoPhish app evaluation: Lab and retention study," in *Proc. NDSS Workshop Usable Secur.*, Jan. 2015, pp. 1–10.

[30] P. Kumaraguru, Y. Rhee, S. Sheng, S. Hasan, A. Acquisti, L. Cranor, and J. Hong, "Getting users to pay attention to anti-phishing education: Evaluation of retention and transfer," in *Proc. Anti-Phishing Working Groups 2nd Annu. Ecrime Researchers Summit*, vol. 269, Jan. 2007, pp. 70–81.

[31] J. Schroeder, *Persistent Training*. Berkeley, CA, USA: Apress, 2017, pp. 25–32, doi: 10.1007/978-1-4842-2835-7_4.

[32] P. Kumaraguru, S. Sheng, A. Acquisti, L. Cranor, and J. Hong, "Lessons from a real world evaluation of anti-phishing training," in *Proc. eCrime Researchers Summit*, Nov. 2008, pp. 1–12.

[33] P. Kumaraguru, Y. Rhee, A. Acquisti, L. F. Cranor, J. Hong, and E. Nunge, "Protecting people from phishing: The design and evaluation of an embedded training email system," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, Apr. 2007, pp. 905–914.

[34] C. Jackson, D. Simon, D. Tan, and A. Barth, "An evaluation of extended validation and picture-in-picture phishing attacks," in *Proc. Int. Conf. Financial Cryptogr. Data Secur.*, vol. 4886, Feb. 2007, pp. 281–293.

[35] B. Reinheimer, L. Aldag, P. Mayer, M. Mossano, R. Duezguen, B. Lofthouse, T. Von Landesberger, and M. Volkamer, "An investigation of phishing awareness and education over time: When and how to best remind users," in *Proc. 16th Symp. Usable Privacy Secur.*, Aug. 2020, pp. 259–284. [Online]. Available: https://www.usenix.org/conference/soups2020/presentation/reinheimer

[36] A. Carella, M. Kotsoev, and T. Truta, "Impact of security awareness training on phishing click-through rates," in *Proc. IEEE Int. Conf. Big Data*, Dec. 2017, pp. 4458–4466.

[37] M. M. Al-Daeef, N. Basir, and M. Hukins, "Security awareness training: A review," in *Proc. World Congr. Eng.*, vol. 1, 2017, pp. 1–6.

[38] D. Akhawe and A. P. Felt, "Alice in warningland: A large-scale field study of browser security warning effectiveness," in *Proc. 22nd USENIX Secur. Symp.*, vol. 13, 2013, pp. 257–272.

[39] S. Egelman, L. F. Cranor, and J. Hong, "You've been warned: An empirical study of the effectiveness of web browser phishing warnings," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, Apr. 2008, pp. 1065–1074.

[40] W. Yang, A. Xiong, J. Chen, R. W. Proctor, and N. Li, "Use of phishing training to improve security warning compliance: Evidence from a field experiment," in *Proc. Hot Topics Sci. Secur., Symp. Bootcamp*, New York, NY, USA, 2017, pp. 52–61.

[41] Z. A. Wen, Z. Lin, R. Chen, and E. Andersen, "What.Hack: Engaging anti-phishing training through a role-playing phishing simulation game," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA: Association for Computing Machinery, May 2019, pp. 1–12.

[42] R. Taib, K. Yu, S. Berkovsky, M. Wiggins, and P. Bayl-Smith, "Social engineering and organisational dependencies in phishing attacks," in *Human-Computer Interaction—INTERACT*, D. Lamas, F. Loizides, L. Nacke, H. Petrie, M. Winckler, and P. Zaphiris, Eds. Cham, Switzerland: Springer, 2019, pp. 564–584.

[43] W. Li, J. Lee, J. Purl, F. Greitzer, B. Yousefi, and K. Laskey, "Experimental investigation of demographic factors related to phishing susceptibility," in *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, 2020, pp. 1–10. [Online]. Available: http://hdl.handle.net/10125/64015

[44] J. Rastenis, S. Ramanauskaite, J. Janulevicius, and A. Cenys, "Credulity to phishing attacks: A real-world study of personnel with higher education," in *Proc. Open Conf. Electr., Electron. Inf. Sci.*, 2019, pp. 1–5.

[45] J. G. Mohebzada, A. E. Zarka, A. H. Bhojani, and A. Darwish, "Phishing in a university community: Two large scale phishing experiments," in *Proc. Int. Conf. Innov. Inf. Technol. (IIT)*, Mar. 2012, pp. 249–254.

[46] O. A. Zielinska, R. Tembe, K. W. Hong, X. Ge, E. Murphy-Hill, and C. B. Mayhorn, "One phish, two phish, How to avoid the internet phish: Analysis of training strategies to detect phishing emails," in *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, 2014, vol. 58, no. 1, pp. 1466–1470.

[47] G. Topham. (May 2021). *Train Firm'S 'Worker Bonus' Email is Actually Cybersecurity Test*. [Online]. Available: https://www.theguardian.com/uk-news/2021/may/10/train-firms-w orker-bonus-email-is-actually-cyber-security-test

[48] A. Picchi. (Sep. 2020). *Tribune Workers Got an Email Dangling a Bonus—But It Was a Hoax From Their Employer*. [Online]. Available: https://www.cbsnews.com/news/tribune-bonus-email-hoax-cybersecurity-test/

[49] J. M. Blythe, A. Gray, and E. Collins, "Human cyber risk management by security awareness professionals: Carrots or sticks to drive behaviour change?" in *HCI for Cybersecurity, Privacy and Trust*, A. Moallem, Ed. Cham, Switzerland: Springer, 2020, pp. 76–91.

[50] M. Volkamer, A. Sasse, and F. Boehm, "Analysing simulated phishing campaigns for staff," in *Proc. Eur. Symp. Res. Comput. Secur.*, Dec. 2020, pp. 312–328.

[51] B. Kim, D.-Y. Lee, and B. Kim, "Deterrent effects of punishment and training on insider security threats: A field experiment on phishing attacks," *Behav. Inf. Technol.*, vol. 39, pp. 1–20, Aug. 2019.

[52] M. L. Jensen, R. Wright, A. Durcikova, and S. Karumbaiah. (Jul. 2020). *Building the Human Firewall: Combating Phishing Through Collective Action of Individuals Using Leaderboards*. [Online]. Available: https://ssrn.com/abstract=3622322

[53] W. Yeoh, H. Huang, W.-S. Lee, F. Al Jafari, and R. Mansson, "Simulated phishing attack and embedded training campaign," *J. Comput. Inf. Syst.*, vol. 62, no. 4, pp. 802–821, Jul. 2022, doi: 10.1080/08874417.2021.1919941.

[54] B. Krebs, "Should failing phish tests be a fireable offense?" KrebsOnSecurity, Arlington, VA, USA, May 2019, p. 1. Accessed: Sep. 1, 2022. [Online]. Available: https://krebsonsecurity.com/2019/05/should-failing-phish-tests-be-a-fireable-offense/

[55] Z. Benenson, F. Gassmann, and R. Landwirth, "Exploiting curiosity and context: How to make people click on a dangerous link despite their security awareness," *BlackHat USA*, 2016. Accessed: Sep. 19, 2022. [Online]. Available: https://www.youtube.com/watch?v=ThOQ63CyQR4

[56] S. Dawkins, J. Jacobs, and K. Greene, "The NIST phish scale: Method for rating human phishing detection difficulty," in *Proc. Messaging, Malware Mobile Anti-Abuse Work. Group (MAAWG) 51st Gen. Meeting*, Gaithersburg, MD, USA, 2021. [Online]. Available: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=931736

[57] F. Barrientos, J. Jacobs, and S. Dawkins, "Scaling the phish: Advancing the NIST phish scale," in *HCI International Posters*, C. Stephanidis, M. Antona, and S. Ntoa, Eds. Cham, Switzerland: Springer, 2021, pp. 383–390.

[58] M. Steves, K. Greene, and M. Theofanos, "Categorizing human phishing difficulty: A phish scale," *J. Cybersecur.*, vol. 6, no. 1, Jan. 2020, Art. no. tyaa009, doi: 10.1093/cybsec/tyaa009.

[59] M. Steves, K. Greene, and M. Theofanos, "A phish scale: Rating human phishing message detection difficulty," in *Proc. Workshop Usable Secur. (USEC)*, San Diego, CA, USA, 2019, pp. 1–14. [Online]. Available: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=927333

[60] D. Jampen, G. Gür, T. Sutter, and B. Tellenbach, "Don't click: Towards an effective anti-phishing training. A comparative literature review," *Hum.-Centric Comput. Inf. Sci.*, vol. 10, no. 1, pp. 1–41, Dec. 2020.

[61] C. B. Mayhorn and P. G. Nyeste, "Training users to counteract phishing," *Work*, vol. 41, pp. 3549–3552, Jan. 2012.

[62] W. R. Flores, H. Holm, G. Svensson, and G. Ericsson, "Using phishing experiments and scenario-based surveys to understand security behaviours in practice," *Inf. Manag. Comput. Secur.*, vol. 22, no. 4, pp. 393–406, Oct. 2014.

[63] A. Neupane, M. L. Rahman, N. Saxena, and L. Hirshfield, "A multi-modal neuro-physiological study of phishing detection and malware warnings," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, Denver, CO, USA, 2015, pp. 479–491.

[64] A. K. Welk, K. W. Hong, O. A. Zielinska, R. Tembe, E. Murphy-Hill, and C. B. Mayhorn, "Will the 'Phisher-Men' reel you in?: Assessing individual differences in a phishing detection task," *Int. J. Cyber Behav., Psychol. Learn. (IJCBPL)*, vol. 5, no. 4, pp. 1–17, 2015.

[65] I. Kirlappos and M. A. Sasse, "Security education against phishing: A modest proposal for a major rethink," *IEEE Secur. Privacy Mag.*, vol. 10, no. 2, pp. 24–32, Mar. 2012.

[66] N. A. G. Arachchilage, "User-centred security education: A game design to thwart phishing attacks," 2015, *arXiv:1511.03459*.

[67] D. J. Lemay, R. B. Basnet, and T. Doleck, "Examining the relationship between threat and coping appraisal in phishing detection among college students," *J. Internet Serv. Inf. Secur.*, vol. 10, no. 1, pp. 38–49, Feb. 2020.

[68] M. S. B. O. Mustafa, M. N. Kabir, F. Ernawan, and W. Jing, "An enhanced model for increasing awareness of vocational students against phishing attacks," in *Proc. IEEE Int. Conf. Autom. Control Intell. Syst. (ICACIS)*, Jun. 2019, pp. 10–14.

[69] Y. Li, K. Xiong, and X. Li, "Understanding user behaviors when phishing attacks occur," in *Proc. IEEE Int. Conf. Intell. Secur. Inform. (ISI)*, 2019, p. 222.

[70] G. Baral and N. A. G. Arachchilage, "Building confidence not to be phished through a gamified approach: Conceptualising user's self-efficacy in phishing threat avoidance behaviour," in *Proc. Cybersecur. Cyberforensics Conf. (CCC)*, 2019, pp. 102–110.

[71] K. Yu, R. Taib, M. A. Butavicius, K. Parsons, and F. Chen, "Mouse behavior as an index of phishing awareness," in *Human-Computer Interaction—INTERACT*, D. Lamas, F. Loizides, L. Nacke, H. Petrie, M. Winckler, and P. Zaphiris, Eds. Cham, Switzerland: Springer, 2019, pp. 539–548.

[72] D. Lain, K. Kostiainen, and S. Capkun, "Phishing in organizations: Findings from a large-scale and long-term study," 2021, *arXiv:2112.07498*.

[73] E. Leukfeldt, "Phishing for suitable targets in The Netherlands: Routine activity theory and phishing victimization," *Cyberpsychol., Behav. Social Netw.*, vol. 17, pp. 551–555, Aug. 2014.

[74] W. J. Gordon, A. Wright, R. Aiyagari, L. Corbo, R. J. Glynn, J. Kadakia, J. Kufahl, C. Mazzone, J. Noga, M. Parkulo, B. Sanford, P. Scheib, and A. B. Landman, "Assessment of employee susceptibility to phishing attacks at us health care institutions," *JAMA Netw. Open*, vol. 2, no. 3, Mar. 2019, Art. no. e190393, doi: 10.1001/jamanetworkopen.2019.0393.

[75] A. Baillon, J. De Bruin, A. Emirmahmutoglu, E. Van De Veer, and B. Van Dijk, "Informing, simulating experience, or both: A field experiment on phishing risks," *PLoS ONE*, vol. 14, no. 12, 2019, Art. no. e0224216.

[76] Z. Benenson, F. Gassmann, and R. Landwirth, "Unpacking spear phishing Susceptibility," in *Financial Cryptography and Data Security* (Lecture Notes in Computer Science), M. Brenner, K. Rohloff, J. Bonneau, A. Miller, P. Y. Ryan, V. Teague, A. Bracciali, M. Sala, F. Pintore, and M. Jakobsson, Eds. Cham, Switzerland: Springer, 2017, pp. 610–627.

[77] Lucy Security. *Lucy Security Powered by ThriveDX*. Accessed: Sep. 1, 2022. [Online]. Available: https://lucysecurity.com/

[78] Sophos. *Sophos Phish Threat*. Accessed: Sep. 1, 2022. [Online]. Available: https://www.sophos.com/en-us/products/phish-threat

[79] Kaspersky. *Kaspersky Security Awareness*. Accessed: Sep. 1, 2022. [Online]. Available: https://www.kaspersky.com/enterprise-security/security-awareness
[80] Infosec Institute. *Security Awareness Training Built to Educate & Engage*. Accessed: Sep. 1, 2022. [Online]. Available: https://www.infosecinstitute.com/iq/security-awareness-training/
[81] ProofPoint. *A Targeted, Data-Driven Approach to Making Users Resilient*. Accessed: Sep. 1, 2022. [Online]. Available: https://www.proofpoint.com/us/products/security-awareness-training
[82] Knowbe4. *Find Out How Effective Our Security Awareness Training is*. Accessed: Sep. 1, 2022. [Online]. Available: https://www.knowbe4.com/
[83] Cofense. *Phishing Awareness Training, Anti-Phishing Tools and Threat Simulations*. Accessed: Sep. 1, 2022. [Online]. Available: https://cofense.com/product-services/phishme/
[84] Y. M. Reddy and H. Andrade, "A review of rubric use in higher education," *Assessment Eval. Higher Educ.*, vol. 35, no. 4, pp. 435–448, Jul. 2010, doi: 10.1080/02602930902862859.
[85] P. Finn and M. Jakobsson, "Designing ethical phishing experiments," *IEEE Technol. Soc. Mag.*, vol. 26, no. 1, pp. 46–58, Mar. 2007.
[86] D. Resnik and P. Finn, "Ethics and phishing experiments," *Sci. Eng. Ethics*, vol. 24, pp. 1241–1252, Aug. 2018.
[87] B. Stanton, M. F. Theofanos, S. S. Prettyman, and S. Furman, "Security fatigue," *Professional*, vol. 18, no. 5, pp. 26–32, Sep. 2016.
[88] L. S. Vailshery. (Feb. 2022). *Market Share Held by Leading Desktop Internet Browsers in the United States From January 2015 to December 2021*. [Online]. Available: https://www.statista.com/statistics/272697/market-share-desktop-internet-browser-usa/
[89] Statscounter. (Jun. 2022). *Browser Market Share Worldwide*. [Online]. Available: https://gs.statcounter.com/browser-market-share
[90] A. Agresti, *An Introduction to Categorical Data Analysis*. Hoboken, NJ, USA: Wiley, 2007.
[91] A. Agresti, *Categorical Data Analysis* (Wiley Series in Probability and Statistics). Hoboken, NJ, USA: Wiley, 2013. [Online]. Available: https://books.google.ch/books?id=6PHHE1Cr44AC
[92] R. A. Fisher, "The conditions under which X^2 measures the discrepancy between observation and hypothesis," *J. Roy. Stat. Soc.*, vol. 87, pp. 442–450, May 1924.
[93] Y. Fang, C. Zhang, C. Huang, L. Liu, and Y. Yang, "Phishing email detection using improved RCNN model with multilevel vectors and attention mechanism," *IEEE Access*, vol. 7, pp. 56329–56340, 2019.
[94] S. Salloum, T. Gaber, S. Vadera, and K. Shaalan, "Phishing email detection using natural language processing techniques: A literature survey," *Proc. Comput. Sci.*, vol. 189, pp. 19–28, Jan. 2021.
[95] F. C. Dalgic, A. S. Bozkir, and M. Aydos, "Phish-IRIS: A new approach for vision based brand prediction of phishing web pages via compact visual descriptors," in *Proc. 2nd Int. Symp. Multidisciplinary Stud. Innov. Technol. (ISMSIT)*, Oct. 2018, pp. 1–8.
[96] P. Maneriker, J. W. Stokes, E. G. Lazo, D. Carutasu, F. Tajaddodianfar, and A. Gururajan, "URLTran: Improving phishing URL detection using transformers," in *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, Dec. 2021, pp. 197–204.
[97] G. Sonowal, "Detecting phishing SMS based on multiple correlation algorithms," *Social Netw. Comput. Sci.*, vol. 1, no. 6, pp. 1–9, Nov. 2020.
[98] A. S. Bozkir and M. Aydos, "LogoSENSE: A companion hog based logo detection scheme for phishing web page and E-mail brand recognition," *Comput. Secur.*, vol. 95, Aug. 2020, Art. no. 101855.
[99] D. Kahneman, *Thinking, Fast and Slow*. London, U.K.: Macmillan, 2011.
[100] R. C. Staudemeyer and E. R. Morris, "Understanding LSTM—A tutorial into long short-term memory recurrent neural networks," 2019, *arXiv:1909.09586*.
[101] J. Lee, F. Tang, P. Ye, F. Abbasi, P. Hay, and D. M. Divakaran, "D-fence: A flexible, efficient, and comprehensive phishing email detection system," in *Proc. IEEE Eur. Symp. Secur. Privacy*, Sep. 2021, pp. 578–597.
[102] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[103] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[104] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.
[105] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," 2019, *arXiv:1908.10084*.
[106] B. Romera-Paredes and P. Torr, "An embarrassingly simple approach to zero-shot learning," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2152–2161.
[107] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.
[108] W. Yin, J. Hay, and D. Roth, "Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach," 2019, *arXiv:1909.00161*.
[109] B. MacCartney, *Natural Language Inference*. San Francisco, CA, USA: Stanford Univ., 2009.
[110] J. E. Jackson, *A Users Guide to Principal Components*. Hoboken, NJ, USA: Wiley, 2005.
[111] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000.
[112] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.
[113] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," 2018, *arXiv:1802.03426*.
[114] A. Coenen and A. Pearce, "Understanding UMAP," Tech. Rep., 2019.
[115] A. S. Bozkir, E. Tahillioglu, M. Aydos, and I. Kara, "Catch them alive: A malware detection approach through memory forensics, manifold learning and computer vision," *Comput. Secur.*, vol. 103, Apr. 2021, Art. no. 102166.
[116] G. Sonowal, "Phishing email detection based on binary search feature selection," *Social Netw. Comput. Sci.*, vol. 1, no. 4, pp. 1–14, 2020.
[117] R. Basnet, S. Mukkamala, and A. H. Sung, "Detection of phishing attacks: A machine learning approach," in *Soft Computing Applications in Industry*. Berlin, Germany: Springer, 2008, pp. 373–383.
[118] R. M. Mohammad, F. Thabtah, and L. McCluskey, "An assessment of features related to phishing websites using an automated technique," in *Proc. Int. Conf. Internet Technol. Secured Trans.*, 2012, pp. 492–497.
[119] Y. Zhang, J. I. Hong, and L. F. Cranor, "Cantina: A content-based approach to detecting phishing web sites," in *Proc. 16th Int. Conf. World Wide Web*, 2007, pp. 639–648.
[120] M. He, S.-J. Horng, P. Fan, M. K. Khan, R.-S. Run, J.-L. Lai, R.-J. Chen, and A. Sutanto, "An efficient phishing webpage detector," *Expert Syst. With Appl.*, vol. 38, no. 10, pp. 12018–12027, Sep. 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417411000662, doi: 10.1016/j.eswa.2011.01.046.
[121] M. Coleman and T. L. Liau, "A computer readability formula designed for machine scoring," *J. Appl. Psychol.*, vol. 60, no. 2, p. 283, 1975.
[122] R. Senter and E. A. Smith, "Automated readability index," Cincinnati Univ. OH, USA, Tech. Rep., 1967.
[123] A. A. Akinyelu and A. O. Adewumi, "Classification of phishing email using random forest machine learning technique," *J. Appl. Math.*, vol. 2014, pp. 1–7, Apr. 2014.
[124] R. Yang, K. Zheng, B. Wu, D. Li, Z. Wang, and X. Wang, "Predicting user susceptibility to phishing based on multidimensional features," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–11, Jan. 2022.
[125] A. Barbaresi, "Trafilatura: A web scraping library and command-line tool for text discovery and extraction," in *Proc. Joint Conf. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process., Syst. Demonstrations*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 122–131. [Online]. Available: https://aclanthology.org/2021.acl-demo.15
[126] S. Furnell, "Phishing: Can we spot the signs?" *Comput. Fraud Secur.*, vol. 2007, no. 3, pp. 10–15, 2007.
[127] D. Kim and J. H. Kim, "Understanding persuasive elements in phishing e-mails: A categorical content and semantic network analysis," *Online Inf. Rev.*, vol. 37, no. 6, pp. 835–850, Nov. 2013.
[128] M. Blythe, H. Petrie, and J. A. Clark, "F for fake: Four studies on how we fall for phish," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, May 2011, pp. 3469–3478.
[129] C. I. Canfield, B. Fischhoff, and A. Davis, "Quantifying phishing susceptibility for detection and behavior decisions," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 58, no. 8, pp. 1158–1172, Dec. 2016.
[130] J. S. Downs, M. B. Holbrook, and L. F. Cranor, "Decision strategies and susceptibility to phishing," in *Proc. 2nd Symp. Usable Privacy Secur.*, 2006, pp. 79–90.
[131] G. Chen, M. F. Johnson, P. R. Marupally, N. K. Singireddy, X. Yin, and V. Paruchuri, "Combating typo-squatting for safer browsing," in *Proc. Int. Conf. Adv. Inf. Netw. Appl. Workshops*, May 2009, pp. 31–36.

[132] A. Moubayed, E. Aqeeli, and A. Shami, "Ensemble-based feature selection and classification model for DNS typo-squatting detection," in *Proc. IEEE Can. Conf. Electr. Comput. Eng. (CCECE)*, Sep. 2020, pp. 1–6.
[133] T. Sharma and M. Bashir, "An analysis of phishing emails and how the human vulnerabilities are exploited," in *Proc. Int. Conf. Appl. Hum. Factors Ergonom.* Cham, Switzerland: Springer, 2020, pp. 49–55.
[134] K. Parsons, A. McCormac, M. Pattinson, M. Butavicius, and C. Jerram, "Phishing for the truth: A scenario-based experiment of users behavioural response to emails," in *Proc. IFIP Int. Inf. Secur. Conf.* Berlin, Germany: Springer, 2013, pp. 366–378.
[135] P. Rajivan and C. Gonzalez, "Creative persuasion: A study on adversarial behaviors and strategies in phishing attacks," *Frontiers Psychol.*, vol. 9, p. 135, Feb. 2018.
[136] E. J. Williams, J. Hinds, and A. N. Joinson, "Exploring susceptibility to phishing in the workplace," *Int. J. Hum.-Comput. Stud.*, vol. 120, pp. 1–13, Dec. 2018.
[137] B. J. Fogg, "Prominence-interpretation theory: Explaining how people assess credibility online," in *Extended Abstracts on Human Factors in Computing Systems*. Lausanne, Switzerland: Frontiers in Psychology, 2003, pp. 722–723.
[138] M. Jakobsson, "The human factor in phishing," *Privacy Secur. Consum. Inf.*, vol. 7, no. 1, pp. 1–19, 2007.
[139] S. Grazioli, "Where did they go wrong? An analysis of the failure of knowledgeable internet consumers to detect deception over the internet," *Group Decis. Negotiation*, vol. 13, no. 2, pp. 149–172, 2004.
[140] A. Karakasiliotis, S. Furnell, and M. Papadaki, "Assessing end-user awareness of social engineering and phishing," in *Proc. 7th Austral. Inf. Warfare Secur. Conf.*, 2006, pp. 1–11.
[141] T. Lin, D. E. Capecci, D. M. Ellis, H. A. Rocha, S. Dommaraju, D. S. Oliveira, and N. C. Ebner, "Susceptibility to spear-phishing emails: Effects of internet user demographics and email content," *ACM Trans. Comput.-Hum. Interact.*, vol. 26, no. 5, pp. 1–28, Oct. 2019.
[142] L. L. Carstensen, D. M. Isaacowitz, and S. T. Charles, "Taking time seriously: A theory of socioemotional selectivity," *Amer. Psychologist*, vol. 54, no. 3, p. 165, 1999.
[143] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Theories of Emotion*. Amsterdam, The Netherlands: Elsevier, 1980, pp. 3–33.
[144] R. Plutchik, "A psychoevolutionary theory of emotions," *Social Sci. Inf.*, vol. 21, nos. 4–5, pp. 529–553, 1982, doi: 10.1177/053901882021004003.
[145] D. M. Sarno, J. E. Lewis, C. J. Bohil, and M. B. Neider, "Which phish is on the hook? Phishing vulnerability for older versus younger adults," *Hum. Factors*, vol. 62, no. 5, pp. 704–717, Aug. 2020. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31237787/
[146] E. M. S. Norberto, J. Taylor, R. Salvador, A. Revilla, B. Merino, and C. Vaquero, "The quality of information available on the internet about aortic aneurysm and its endovascular treatment," *Revista Española de Cardiología*, vol. 64, no. 10, pp. 869–875, 2011.
[147] S. Song, "Sharing fast and slow: The psychological connection between how we think and how we spread news on social media," *Nieman Journalism Lab*, vol. 15, Nov. 2013.
[148] G. K. Zipf, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Barcelona, Spain: Ravenio, 2016.
[149] A. Vishwanath, T. Herath, R. Chen, J. Wang, and H. R. Rao, "Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model," *Decis. Support Syst.*, vol. 51, no. 3, pp. 576–586, 2011.
[150] M. Canham, C. Posey, D. Strickland, and M. Constantino, "Phishing for long tails: Examining organizational repeat clickers and protective stewards," *SAGE Open*, vol. 11, no. 1, pp. 1–11, 2021, doi: 10.1177/2158244021990656.
[151] E. D. Frauenstein and S. Flowerday, "Susceptibility to phishing on social network sites: A personality information processing model," *Comput. Secur.*, vol. 94, Jul. 2020, Art. no. 101862.
[152] S. Zhuo, R. Biddle, Y. S. Koh, D. Lottridge, and G. Russello, "SoK: Human-centered phishing susceptibility," 2022, *arXiv:2202.07905*.
[153] G. Norris, A. Brookes, and D. Dowell, "The psychology of internet fraud victimisation: A systematic review," *J. Police Criminal Psychol.*, vol. 34, no. 3, pp. 231–245, 2019.

THOMAS SUTTER was born in Sankt Gallen, Switzerland, in 1991. He received the B.Sc. and M.Sc. degrees in computer science from the Zurich University of Applied Sciences (ZHAW), Zürich, Switzerland, in 2021. He is currently a Research Associate with the Information Security Group, ZHAW, and a part of several research projects. His research interests include operating system security, malware analysis, vulnerability research, applied cryptography, anomaly detection, and network security.

AHMET SELMAN BOZKIR was born in Muğla, Turkey, in 1983. He received the B.S. degree in computer engineering from Eskişehir Osmangazi University, in 2002, and the M.Sc. and Ph.D. degrees in computer engineering from Hacettepe University, in 2009 and 2016, respectively. He is currently working as a Research Associate with the Zurich University of Applied Sciences (ZHAW). He has more than 40 technical publications covering the fields, such as information security, human–computer interaction, information retrieval, and machine learning.

BENJAMIN GEHRING received the B.S. degree in computer science from the Zurich University of Applied Sciences (ZHAW), Switzerland, in 2020, where he is currently pursuing the M.S. degree in computer science with the MSE Program. Since 2020, he has been a Research Assistant in the area of information security with the Institute of Applied Information Technology, ZHAW. His research interests include the development of secure software systems as well as the research of cyber attacks and defense.

PETER BERLICH received the Diploma degree in physics and the Ph.D. degree from Freiburg University, Germany, in 1990 and 1997, respectively. He is currently working as a Principal Consultant for cybersecurity at Zühlke Engineering, Schlieren, Switzerland. From 2018 to 2022, he was a Senior Lecturer of information security with the Zurich University of Applied Sciences, Winterthur, Switzerland. He has held a variety of positions in the IT Security industry. His publications include "Privacy enhancing identity management," *Information Security Technical Report* (2004), contributions to the *Official (ISC)*^{2} *Guide to the CISSP CBK* (Auerbach, 2006), and "Executive Career Paths in Information Security Management," *ISSE 2013 Securing Electronic Business Processes*. His research interests include security careers and the human aspects of security as well as privacy.