UNIVERSITY OF LEÓN

DEPARTMENT OF ELECTRICAL, SYSTEMS AND AUTOMATIC ENGINEERING

ANALYSIS AND CLASSIFICATION OF SPAM EMAIL USING ARTIFICIAL INTELLIGENCE TO IDENTIFY CYBERTHREATS

A dissertation supervised by

PROF. DR. ROCÍO ALAIZ RODRÍGUEZ,

PROF. DR. VÍCTOR GONZÁLEZ CASTRO

and submitted by

FRANCISCO JÁÑEZ MARTINO

in fulfillment of the requirements for the Degree of

PHILOSOPHIÆDOCTOR (PH.D.)

León, October 2023

UNIVERSIDAD DE LEÓN

DEPARTAMENTO DE INGENIERÍA ELÉCTRICA Y DE SISTEMAS Y AUTOMÁTICA

ANÁLISIS Y CLASIFICACIÓN DE CORREO ELECTRÓNICO NO DESEADO MEDIANTE INTELIGENCIA ARTIFICIAL PARA LA IDENTIFICACIÓN DE CIBERAMENAZAS

Tesis doctoral dirigida por

PROF. DR. ROCÍO ALAIZ RODRÍGUEZ,

DR. VÍCTOR GONZÁLEZ CASTRO

y desarrollada por

FRANCISCO JÁÑEZ MARTINO

a fin de optar al grado de

DOCTOR POR LA UNIVERSIDAD DE LEÓN DEL PROGRAMA
DE INGENIERÍA DE PRODUCCIÓN Y COMPUTACIÓN

León, Octubre de 2023

Abstract

In this Thesis, we propose new models, methodologies, approaches and datasets to analyze and identify rising cybertreats in spam emails. Motivated by our collaboration with the Spanish National Institute of Cybersecurity (INCIBE), we focus our efforts on developing applications and conducting studies to improve the earlier detection of these risky and harmful emails. Several of the contributions presented in this dissertation are planned to be incorporated in tools developed by INCIBE to launch more detailed and earlier warnings to organizations and citizens about potential risks associated with spam emails. Our approach heavily relies on the application of Natural Language Processing, as well as Machine and Deep Learning techniques, mainly centred around supervised learning methods.

First, we aimed at employing text classification methods to classify spam emails related to cybersecurity topic for the first time in the literature. Our supervised approaches have lead us to building custom and novel datasets for each contribution. In this case, we created a dataset called *SP*am *EM*ail Classification dataset (SPEMC), a novel dataset that includes eleven classes of spam emails based on cybersecurity topics. SPEMC is composed of two sub-datasets, i.e., SPEMC-E-15K and SPEMC-S-15K, which contain emails written in English and Spanish, respectively. We used SPEMC to evaluate the combination of four text representation techniques along with four Machine Learning models. The combination of Term Frequency - Inverse Document Frequency (TF-IDF) with Logistic Regression (LR) achieved the highest performance in the assessment done with the emails in English, 0.953 of Macro F1-score, while TF-IDF with Naïve Bayes (NB) achieved 0.945 in the Spanish dataset. In both languages, TF-IDF with LR was the fastest combination with 2.0 ms and 2.2 ms per email, English and Spanish respectively.

Secondly, we aimed at understanding the role of persuasion in spam emails to combat cybersecurity threats more effectively. We developed intelligent systems to detect persuasion and used techniques through Natural Language Processing at three granularity levels: full emails, sentences, and specific text spans (i.e., a group of one or more words shorter than a sentence). We replicated the *Proppy* (Barrón-Cedeño et al., 2019) classifier to spot persuasion in full emails and built our binary and multilabel models on top of RoBERTa (Liu et al., 2019) for sentence and text spans classification (based on Chernyavskiy et al. (2020)). We created a novel dataset called Persuasion Sentence in Spam Emails (PerSentSE) containing annotated sentences based on binary, i.e., persuasion or not, and multilabel classification. For the multilabel approach, we considered eight persuasion techniques: *Appeal to authority*, *Appeal to fear/prejudice*, *Doubt*, *Exaggeration or minimization*, *Flag-waving*, *Loaded Language*, *Name Calling or Labeling* and *Repetition*. We collected spam emails from the Bruce Guenter repository.

Lastly, our objective was to create an intelligent system capable of detecting potentially risky spam emails for both individuals and organizations. We created Spam Email Risk Classification (SERC-4K), a novel dataset encompassing spam emails classified in two categories based on the potential risk for users due to their content, low and high risk, as well as a continuous value from 1 to 10. The dataset is composed of two subdatasets, one with spam emails shared by INCIBE (SERC-I) and another collected from the Bruce Guenter repository, Spam Archive (SERC-BG). SERC-I contains English and Spanish emails, while in the case of SERC-BG almost all of them are written in English. Firstly, our approach attempted to extract potentially worthy features from headers, text, attachments, URLs and protocols (56 features in total). Then, the sets of features along with three popular Machine Learning classifiers were evaluated resulting in Random Forest as the highest classifier-performance (0.914 of F1-score). Regarding regression approach, the Random Forest Regressor achieved the lowest MSE (0.579).

Our work also included a feature evaluation to determine the importance of each feature and set. In the design of our methodologies, we have considered the influence of the dataset shift, as well as the spam domain is and adversarial environment. Our email processing sought to overcome some spammer strategies such as image-based spam and hidden text.

Keywords: *Spam email filtering, Text Classification, Machine Learning, Attention models, Natural Language Processing, Persuasion detection, Feature extraction, Risk classification, Cybersecurity, Cyberawareness*

Resumen

En esta Tesis, proponemos nuevos modelos, metodologías, enfoques y conjuntos de datos para analizar e identificar las crecientes ciberamenazas en los correos electrónicos no deseados, conocidos como correos spam. Motivados por nuestra colaboración con el Instituto Nacional de Ciberseguridad (INCIBE), concentramos nuestros esfuerzos en desarrollar aplicaciones y llevar a cabo estudios para mejorar la pronta detección de estos peligrosos correos electrónicos. Varias contribuciones entre las presentadas en esta Tesis están preparadas para una futura incorporación en las herramientadas desarrolladas por INCIBE a la hora de lanzar avisos más detallados y rápidos a organizaciones y ciudadados sobre el potencial riesgo de un correo spam. Nuestros enfoques se basan sobre todo en la aplicación de técnicas del Procesamiento del Lenguaje Natural, así como de Aprendizaje Automático y Profundo centrado principalmente en modelos de aprendizaje supervisado.

Primero, nuestro objetivo fue emplear métodos de clasificación de texto para clasificar los correos spam de acuerdo a su temática de ciberseguridad por primera vez en la literatura. Nuestros enfoques supervisados nos han dirigido a la creación de nuevos y personalizados conjuntos de datos para cada contribución. En este caso, hemos creado **SP**am **EM**ail Classification dataset (SPEMC), un novedoso conjunto de datos que incluye once clases de correo spam correspondientes a temas de ciberseguridad. SPEMC está compuesto de dos subconjuntos, SPEMC-E-15K y SPEMC-S-15K, que contienen emails escritos en inglés y en español, respectivamente. Usando SPEMC, evaluamos la combinación de cuatro descriptores de texto junto con cuatro modelos de Aprendizaje Automático. La combinación de TF-IDF y Regresión Logística alcanzó el mejor valor de Macro F1-score (0.953). Por otro lado, la combinación de TF-IDF con Naïve Bayes logró 0.945 en el conjunto de datos en español. En ambos idiomas, TF-IDF con Regresión Logística fue la combinación más rápida con 2.0 y 2.2 ms por email, en inglés y en español, respectivamente.

Después, buscamos comprender el rol de la persuasión en los correos spam para luchar contra las amenazas de cibersecuridad más eficientemente. Desarrollamos sistemas inteligentes para detectar la persuasión y sus técnicas usadas mediante Procesamiento del Lenguaje Natural en tres niveles de granularidad: correo completo, oraciones y fragmentos específicos de texto (una o más palabras siempre menores a una oración). Replicamos el clasificador Proppy (Barrón-Cedeño et al., 2019) para detectar la persuasión en el correo completo y construimos un modelo binario y otro multietiqueta basado en RoBERTa (Liu et al., 2019) para la clasificación a nivel de oración y fragmento (basado específicamente en Chernyavskiy et al. (2020)). Creamos un nuevo conjunto de datos llamado Persuasive Sentences in Spam Emails (PerSentSE), que contiene oraciones etiquetadas de manera binaria, es decir, si contiene persuasión o no, y multietiqueta. Para este último enfoque, consideramos ocho técnicas de persuasión: *Apelar a la Autoridad*, *Apelar al miedo/prejuicio*, *Duda*, *Exageración o minimización*, *Patriotismo*, *Lenguaje Cargado*, *Descalificación o Etiquetado* y *Repetición*.

Por último, nuestro objetivo fue desarrollar un sistema inteligente capaz de detectar los correos potencialmente peligrosos para los individuos y las organizaciones. Construimos un novedoso conjunto de datos llamado Spam Email Risk Classification (SERC-4K) que incluye correos spam divididos en dos clases basadas en un potencial riesgo para los usuarios debido a su contenido, bajo o alto riesgo, así como una valoración del riesgo de 1 a 10. El corpus está compuesto de dos conjuntos, uno con correos spam compartidos por INCIBE (SERC-I) y otro recolectado del repositorio público de Bruce Guenter, Spam Archive (SERC-BG). SERC-I contiene correos tanto en inglés como en español, mientras que en SERC-BG casi todos están escritos en inglés. Primero, nuestro enfoque busca extraer 56 características de las cabeceras, texto, adjuntos, URLs y protocolos de los correos spam. Después, los conjuntos de características junto con tres populares modelos de Aprendizaje Automático fueron evaluados, dando como resultado que Random Forest obtuvo el F1-score más alto (0.914). En cuanto al enfoque de regresión, el estimador Random Forest Regressor consiguió el MSE más bajo (0.579). Nuestro trabajo también incluye una evaluación de las características para determinar la importancia de cada una individualmente y de los grupos de características.

Nuestras metodologías consideran la influencia del cambio en el conjunto de datos y el entorno contra un adversario (la persona que crea y envía correos spam, llamado spammer) para sus diseños. Nuestro procesamiento del correo electrónico buscó superar algunas estrategias creadas por spammers, por ejemplo correos con mensaje spam en las imágenes o texto oculto.

Palabras clave: *Detección de correos spam, Clasificación de texto, Aprendizaje Automático, Modelos de atención, Procesamiento del Lenguaje Natural, Detección de la persuasión, Extracción de Características, Predicción del riesgo, Ciberseguridad, Ciberavisos*

Contents

List of Figures

List of Tables

Índice general

Agradecimientos

Todo viaje tiene su final. Atrás quedan los días y las noches de trabajo, el esfuerzo, frustracciones y, por suerte, buenas alegrías que un doctorado conlleva. Porque un doctorado es una mezcla de todo, un salto hacia arriba y una bajada. Una vuelta a empezar y un intentarlo de nuevo. Una decepción y una alegría. A menudo se tiende a pensar que un doctorado es un viaje solitario. Sin embargo, un doctorado es menos duro en los días malos y más bonito en los días buenos cuando te rodeas de tu gente. Como en todo viaje, como en la vida. Quisiera agradecer a través de estas líneas a todas aquellas personas e instituciones que de una forma u otra me han ayudado durante este viaje. En especial quiero agradecer:

Al Instituto Nacional de Ciberseguridad (INCIBE) por permitirme realizar esta tesis bajo su apoyo y colaboración. Asi como ofrecerme los datos, la ayuda y las aplicaciones para la mayoría de los modelos de Inteligencia Artificial presentados en esta tesis. Gracias especialmente a Alejandro, Santi y María por vuestra ayuda durante muchas partes de esta tesis.

A la Junta de Castilla y León por haberme permitido disfrutar de una ayuda para la Formación de Personal Predoctoral durante los dos últimos años, sin la cual esta tesis hubiera sido más complicada de llevar a cabo.

A mis directores, Rocío Alaiz y Víctor González, por darme la oportunidad de trabajar y realizar mi doctorado bajo su supervisión. No tengo palabras para agradeceros todo lo que me habéis enseñado durante estos años, semana tras semana, esfuerzo tras esfuerzo. Cuando era pequeño y soñaba con obtener un doctorado, no imaginaba la suerte que tendría con mis directores, y lo importantes que se convertirían. Modelos a seguir para mi, en lo profesional y sobre todo, en lo personal. Gracias por todo.

A Enrique Alegre por confiar en mi hace ya muchos años atrás. Tus consejos y confianza también me han acompañado durante este viaje para conseguir ser mejor persona e investigador. Tu liderazgo nos ha guiado siempre en este camino. A Eduardo Fidalgo, quién empezó a construir la base del investigador que soy a día de hoy. Quién siguió a mi lado ayudándome en cada momento que lo he necesité, dándome alas y responsabilidad para seguir creciendo. Gracias de corazón a los dos.

A Alberto Barrón por acogerme tan amablemente no solo una, si no dos veces, en la

Università di Bologna. Permitiéndome aprender de él lo mucho que sabe sobre procesamiento del lenguaje natural, y además, enseñarme los mejores platos de pasta y pizza de la Emilia-Romagna, siempre con interesantes conversaciones.

A David Guerra por su profesionalidad en el laboratorio, siempre atento y preparado para echar un mano, con todos los equipos e información necesaria para que nuestras tareas se cumplan. A Wesam, Deisy y Rubel por transmitirme vuestra experiencia durante estos años. A Manu por compartir café, bromas y laboratorio durante mis primeros años. A todo el equipo de GVIS, a los que están y estuvieron, Andrés, Aitor, Javi, Pablo, Alex, Alicia, Luis Ángel, Surajit, Laura, Felipe, Juanjo, Dani y Oscar, ha sido un placer. Al Equipo de Etiquetado, en especial a aquellos con los que tuve la oportunidad de trabajar, Marta, Iria, Eva, Víctor, Diego, Felix, Pablo, Pedro y Laura, gracias por vuestro trabajo, indispensable para crear nuestros datasets.

A Javier Álvarez por ser mi bastón al caminar. A Arturo Armario por convertirse en el hermano que nunca tuve. A Paula Flórez y Laura Arias por escucharme y motivarme siempre. A Hugo Martínez por tus discursos motivacionales en cada punto bajo. A la gente que me ha llenado de alegría durante este camino, Víctor, Patri, Elena, Iván, Tomás, Mary, a mis amigos de siempre de Carrizo, al CDF de Carrizo, a la gente que me acompañó en Bologna, a mi familia.

Por último, a las personas más importantes que tengo en mi vida, a mis cuatro pilares que me han hecho la persona que soy. A mi madre Julia Martino, quién a diario me demuestra lo mucho que me quiere, cuidándome, escuchándome, guíandome y nunca dejándome solo en la oscuridad. Eres mi luz en este mundo. A mi padre Francisco Jáñez por mostrarme el amor implícito en las cosas pequeñas y lo que significa querer. Guiándome en mi caos, siempre en la dirección correcta. A mis abuelos Julia Pérez y Francisco Martino por enseñarme a nunca rendirme y cuidar de mi siempre. Sé que no solo se cumple uno de mis sueños, si no también el vuestro, recordarlo siempre. Os quiero.

No quiero olvidarme de aquellos que siempre me dieron su compañía, que hacían entretenidas mis reuniones online, que me dieron un amor puro. Jisco, Mini, Kiko, me hicisteis muy feliz.

Por los que están, por los que se fueron. Gracias a todos. Este viaje termina, pero el fin es un principio necesario.

Francisco Jáñez Martino
León
30th October 2023

Chapter 1

Introduction

1.1. Motivation

Electronic mail (email) has been one of the most popular means of communication for organizations and citizens for over three decades. Besides social media, email remains a popular communication system because it allows for the direct and private communication between peers, often through a free service, and potentially in an anonymous fashion. Nevertheless, *malicious* users take advantage of these characteristics to massively distribute advertisement or bothersome messages, typically known as spam. Spam email is an unsolicited, unwanted or unhelpful message, which appeared since the early days of email.

Billions of spam emails are sent and received everyday[1]. According to the reports of Cisco Talos[2] and Kaspersky Lab[3], spam emails represent between 55% and 85% of the daily total volume of worldwide emails. Spam can cause productivity loss, distrust in email service, annoyance or service bottlenecks that limit memory space and speed of computers. These issues result in an economic cost for organisations that is steadily increasing. As a consequence of all the above, a decade ago spam was estimated to cost companies around twenty billion dollars annually (Mohammad, 2020). This cost is likely to surpass 250 billion dollars in a couple of years (Mohammad, 2020).

All email clients have a spam folder which automatically collects undesired content, often going unnoticed. Thanks to these spam folders, unsolicited emails are less annoying and they do not swamp users' mailboxes. However, it is important to note that spam emails can still be adapted for specific targeted attacks, bypassing spam filters and jumping to our main inboxes (Wang et al., 2022). They may just contain advertisements and company promotions and, although this may be annoying, it is harmless (Bhowmick and Hazarika, 2018; Ferrara, 2019). Originally intended as way of advertising products and services, spam email has evolved significantly in recent times. Nowadays spam emails often

[1] `https://techjury.net/blog/how-many-emails-are-sent-per-day/`, retrieved September 2023

[2] `https://talosintelligence.com/reputation_center/email_rep` retrieved September 2023

[3] `https://www.statista.com/statistics/420391/spam-email-traffic-share/` retrieved September 2023

contain scams, phishing[4] , malware or spoofing[5] , leaving users exposed and vulnerable to the effects of cyberattacks (Jáñez-Martino et al., 2022).

Most research works in this field have been focused on the spam filtering, i.e., classification of emails as spam or not spam — also known as ham — (Dada et al., 2019). During the last decade, spammers have modified and enhanced their strategies to mislead both filter systems and users (Mohammad, 2020). From a forensic perspective, investigating spammer strategies and social engineering techniques in emails may help uncover similar disguises in other fields that suffer both an adversarial figure and digital crimes (Yu, 2015). First, Wang et al. (2013) and, then, Bhowmick and Hazarika (2018) already analysed spam trends and highlighted the dynamic nature of spam content. Wang et al. (2013) warned that the spam email is not dying, but becoming more elaborate, fancy and sophisticated.

When users open a spam email, they may encounter a message deliberately created using social engineering techniques. These techniques aim to persuade users to take certain actions —such as clicking on a link, opening an attachment or responding to the message—, often putting themselves at risk of inadvertently disclosing confidential information (Ferreira et al., 2015; Dada et al., 2019). Organizations try to improve their working environment by providing cybersecurity training for employees and using state-of-the-art spam filtering technology (Jáñez Martino et al., 2021). Moreover, Law Enforcement Agencies (LEAs) actively monitor spam in an effort to identify, manage, and launch timely warnings to inform organizations and citizens about these threats.

INCIBE, the Spanish National Cybersecurity Institute[6] collaborates with Spanish LEAs by offering services and implementing automatic systems to aid in combating and mitigating the risks posed by online criminal activities. These automatic tools play an important role in saving time and human efforts by exploring thousands of daily spam emails. However, spam emails are currently inspected in semiautomatic fashion based on a traditional spam filtering followed by a visual inspection of experts. By quickly spotting campaigns in more detail, INCIBE can effectively prioritize and analyze the most high-risk threats for recipients, discarding low hazard emails.

We have collaborated with INCIBE in developing Artificial Intelligence-based solutions for enhancing their cybersecurity tools and services. In this work, our focus has been to develop solutions to address the issue of spam email, using techniques based on Natural Language Processing (NLP) and Machine Learning.

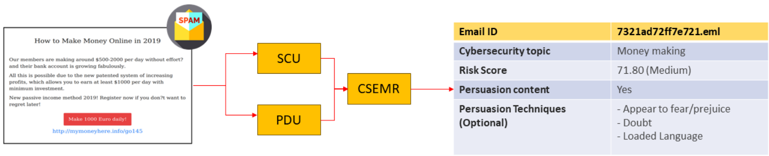
Our research can be divided into three principal components or units (see Fig. 1.1):

1. Spam Classification Unit (SCU): Given the textual content of a spam email, the SCU classifies it based on its cybersecurity topic.

[4] Phishing occurs when a cybercriminal masquerades as a legitimate organization with the intention of stealing individuals sensitive information.

[5] Spoofing occurs always when an online scammer disguises their identity as popular company or commonly familiar person.

[6] `https://www.incibe.es/`, retrieved September 2023



How to Make Money Online in 2019

Our members are making around $500-2000 per day without effort? and their bank account is growing fabulously.

All this is possible due to the new patented system of increasing profits, which allows you to earn at least $1000 per day with minimum investment.

New passive income method 2019! Register now if you don?t want to regret later!

http://mymoneyhere.info/go145

Email ID	**7321ad72ff7e721.eml**
Cybersecurity topic	Money making
Risk Score	71.80 (Medium)
Persuasion content	Yes
Persuasion Techniques (Optional)	- Appear to fear/prejuice - Doubt - Loaded Language

Figure 1.1: Spam email monitoring pipeline.

2. Persuasion Detection Unit (PDU): Given the textual content of a spam email, the PDU spots whether the email uses persuasion to convince the recipients to perform an action.

3. Classification of the Spam Emails Risk Unit (CSERU): Given a spam email, the CSERU extracts a set of features that helps in detecting the risk level of that email.

The outcomes of this work will be used by INCIBE to analyze their incoming spam emails, enabling them to detect cybersecurity threats and assign a risk score to each email.

This work can be adapted to similar applications such as fraudulent websites or Short Message Service (SMS), where users may also encounter scams like phishing (Sánchez-Paniagua et al., 2021) or smishing (Mishra and Soni, 2020). Our research is focused on NLP tasks, and we do not address malware analysis (Karbab and Debbabi, 2019) or email tracking (Haupt et al., 2018), as they are beyond the scope of this Thesis dissertation.

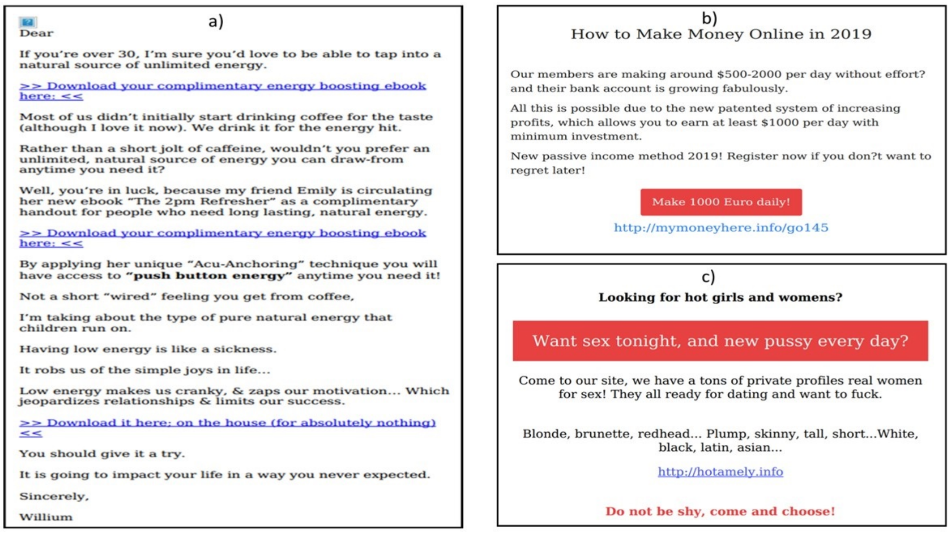
In the following Sections, we present the motivation of each unit.

1.1.1. Spam Email Classification Unit

Our objective with the Spam Email Classification Unit (SCU) is to classify spam emails according to their cybersecurity topic (Jáñez-Martino et al., 2020, 2023). Traditional approaches to spam email typically focus on the spam email filtering, that is, they address the binary categorisation of emails into spam or ham emails (Dada et al., 2019; Jáñez-Martino et al., 2022). Some recent works, however, focused on investigating the main topics or threats of spam emails. (Murugavel and Santhi, 2020; Saidani et al., 2020).

Spanish LEAs are interested in this latter approach to classify their spam emails into classes in order to prioritize certain types and issue more detailed and timely warnings. Analysing thousands of emails manually, and on a daily basis, is an impractical strategy in terms of time and human resources.

We were motivated to pursue this novel approach to parse spam emails and gaining insights into their topics, as well as considering the needs and services of LEAs. Since this is the first time in the literature that cybersecurity topics have been addressed by an Artificial Intelligence system, we firstly clustered the emails and carried out a manual inspection over the resulting groups to define the likely categories, e.g., the emails showed in Fig. 1.2.



a)

Dear

If you're over 30, I'm sure you'd love to be able to tap into a natural source of unlimited energy.

>> Download your complimentary energy boosting ebook here: <<

Most of us didn't initially start drinking coffee for the taste (although I love it now). We drink it for the energy hit.

Rather than a short jolt of caffeine, wouldn't you prefer an unlimited, natural source of energy you can draw-from anytime you need it?

Well, you're in luck, because my friend Emily is circulating her new ebook "The 2pm Refresher" as a complimentary handout for people who need long lasting, natural energy.

>> Download your complimentary energy boosting ebook here: <<

By applying her unique "Acu-Anchoring" technique you will have access to **"push button energy"** anytime you need it!

Not a short "wired" feeling you get from coffee,

I'm taking about the type of pure natural energy that children run on.

Having low energy is like a sickness.

It robs us of the simple joys in life...

Low energy makes us cranky, & zaps our motivation... Which jeopardizes relationships & limits our success.

>> Download it here: on the house (for absolutely nothing) <<

You should give it a try.

It is going to impact your life in a way you never expected.

Sincerely,

Willium

b)

How to Make Money Online in 2019

Our members are making around $500-2000 per day without effort? and their bank account is growing fabulously.

All this is possible due to the new patented system of increasing profits, which allows you to earn at least $1000 per day with minimum investment.

New passive income method 2019! Register now if you don?t want to regret later!

Make 1000 Euro daily!

http://mymoneyhere.info/go145

c)

Looking for hot girls and womens?

Want sex tonight, and new pussy every day?

Come to our site, we have a tons of private profiles real women for sex! They all ready for dating and want to fuck.

Blonde, brunette, redhead... Plump, skinny, tall, short...White, black, latin, asian...

http://hotamely.info

Do not be shy, come and choose!

Figure 1.2: Spam email examples of the different topics a) health, b) making money and c) sexual dates.

1.1.2. Persuasion Detection Unit

The Persuasion Detection Unit (PDU) is responsible for identifying persuasive elements within spam emails. PDU has the capability to detect both the presence of persuasion and the specific techniques employed.

Developing intelligent systems capable of spotting persuasion is an essential tool to enhance the security and awareness of users toward spam email for three reasons. Firstly, users can be warned about spam emails with a high persuasion load, which involves an intent of pushing for (potentially undesired) actions. Secondly, this allows researchers to get further insights on the reason why people gets pushed to interact with spam emails. Thirdly, the information about the level and kind of persuasion techniques in an email is a factor to boost the performance of spam filtering technology.

Some authors have adopted the term of phishing email for a certain group of emails. Phishing attacks steal sensitive information, extort money or cause illicit actions through usually social engineering techniques (El Aassal et al., 2020). Social engineering techniques take advantage of persuasive principles to manipulate people to carry out a specific action. In addition, phishing has become one of the most widespread cyber-incidents (Sánchez-Paniagua et al., 2021).

On the one hand, researchers attempt to develop robust models departing from anti-spam filters —but expressly designed to spot phishing email (Volkamer et al., 2017; El Aassal et al., 2020; Magdy et al., 2022; Sturman et al., 2023). On the other hand, persuasion has been approached in other fields studying their use in misinformation (Chen et al., 2021;

Ali et al., 2022) and propagandist news articles (Barrón-Cedeño et al., 2019; Da San Martino et al., 2019; Sony Dewantara and Budi, 2020), among others. The spammer is conscious about persuading the users to perform an action, in a similar fashion as creators of biased and hyperpartisan contents for disinformation and propaganda do. Some authors have related and studied the principles of persuasion along with phishing email Ferreira et al. (2015); Ferreira and Teles (2019); Lawson et al. (2020).

However, current models have not considered persuasion as an essential factor to identify spam email (Sankhwar et al., 2018; El Aassal et al., 2020; Sonowal, 2020). Moreover, weapons of influence, such as persuasive techniques, and life domains (which is a specific topic or aspect of an individual's life used by attackers), highly affect to the susceptibility of the user against phishing email (Lin et al., 2019), which may be potential features for the systems.

1.1.3. Classification of the Spam Emails Risk Unit

Once we have classified spam emails into cybersecurity topics, we use the Classification of Spam Email Risk Unit (CSERU) to extract potentially useful information from both their headers and body. CSERU is a combination of the SCU output, a novel approach to classify spam email addresses and an additional set of features.

The adversarial environment play an essential role in analyzing spam email, since spammers often introduce strategies for both bypassing filters and misleading users (Bhowmick and Hazarika, 2018). Those emails, which seek to obtain fraudulent benefits from the recipients like leaked data, system accesses or financial rewards, are designed with special attention to look like legitimate emails. Spammers currently spread cyberscams based on phishing, spoofing or ransomware distribution through spam emails. This creates the need of distinguishing into different types of spam email according to their level of risk for users. Traditional spam, known due to being annoying and unsolicited, as well as containing advertisements, is barely a loss of time and resources, but it does not provide a critical door for cybercriminals. However, fraudulent spam is increasingly used for attacks and potentially harmful (Gallo et al., 2021).

INCIBE pays a great deal of attention to the identification of this type of spam emails to issue warnings indicating the level of risk (Fig. 1.3). They expose the email screenshot and a description of its content highlighting the critical points. This warning is heavily important for informing citizens and companies. Nevertheless, cybersecurity experts semi-automatically verify the daily spam emails received on their honeypots, but it becomes challenging to address every campaign and reduce the number of accurate warnings. Hence, INCIBE is interested in developing intelligent systems to spot potentially riskier spam emails with minimal time and human effort. We follow cybersecurity experts approaches to create a set of potentially valuable features. Although previous works such as Gallo et al. (2021) and Bera et al. (2023) were highly focused on only company solutions, they set a baseline for our purpose.

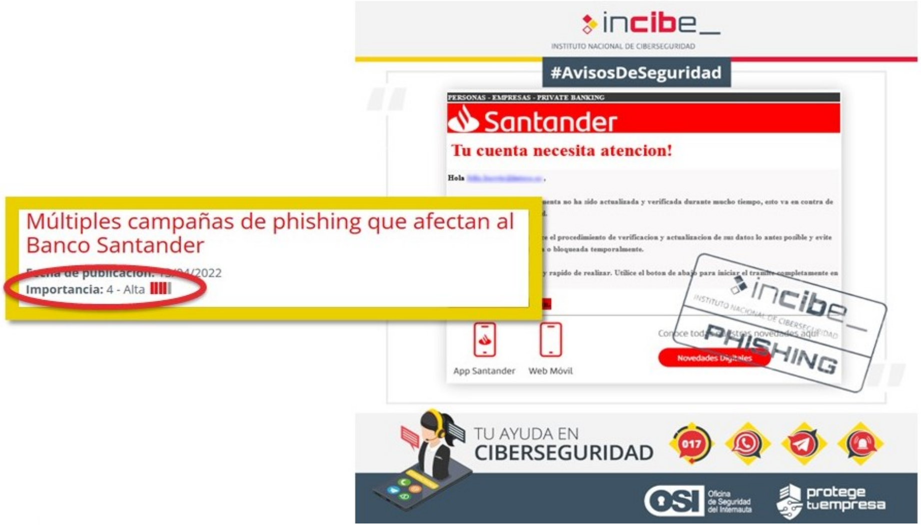

incibe_
INSTITUTO NACIONAL DE CIBERSEGURIDAD
#AvisosDeSeguridad
PERSONAS - EMPRESAS - PRIVATE BANKING
Santander
Tu cuenta necesita atencion!
Hola [redacted] ,
...enta no ha sido actualizada y verificada durante mucho tiempo, esto va en contra de
...e el procedimiento de verificacion y actualizacion de sus datos lo antes posible y evite
... o bloqueada temporalmente.
...y rapido de realizar. Utilice el boton de abajo para iniciar el tramite completamente en
Múltiples campañas de phishing que afectan al Banco Santander
Fecha de publicación: 13/04/2022
Importancia: 4 - Alta
App Santander
Web Móvil
Conoce todas nuestras novedades aquí
Novedades Digitales
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD PHISHING
TU AYUDA EN CIBERSEGURIDAD
017
OSI Oficina de Seguridad del Internauta
protege tuempresa

Figure 1.3: Example of warning published by Spanish Cybersecurity Institute (INCIBE).

1.2. Objectives

The main objective of this Thesis is to propose new methods and solutions using Artificial Intelligence to analyse and spot cyberthreats in spam emails, such as phishing, spoofing, malware, extortion, among other scams. With this goal in mind, we explore, research and propose different techniques based on Machine Learning (ML), Deep Learning (DL) and Natural Language Processing (NLP). We have designed these tools and services to be used in real-world scenarios.

To achieve the main objective, we have established the following specific objectives:

1. To review current trends in the spam email domain and its main challenges: dataset shift and spammer strategies.
2. To build a multiclass text classifier capable of finding out the cybersecurity topics within spam emails.
3. To provide an intelligent system capable of detecting and categorizing the persuasive techniques employed in spam emails.
4. To develop an automatic solution for detecting those spam emails that have a higher risk of causing a cybersecurity incident for individuals and companies.

1.3. Main contributions

The main contributions of this Thesis may be summarized as follows:

1. We presented a recent review of spam classification, dataset shift and adversarial Machine Learning in the field of spam email, as well as an empirical demonstration in Chapter 2 (Jáñez-Martino et al., 2022).

2. We built two topic-based spam email datasets using hierarchical clustering, SPam EMail Classification in English and Spanish (SPEMC-15K-E and SPEMC-15K-S). These datasets were the first multiclass datasets based on eleven cybersecurity topics, and they are presented in Section **??** (Jáñez-Martino et al., 2023).

3. We proposed a new baseline methodology based on text classification to detect the cybersecurity topic of spam emails within a taxonomy proposed by us and established an email processing system that extracts textual information from various sources, including the subject line, visual text (i.e., by bypassing hidden text strategies), and extracting messages from images in Section **??**(Jáñez-Martino et al., 2023).

4. We presented a spam email dataset in which we manually annotated a set of sentences extracted from these emails using a binary classification (persuasion or non-persuasion) and eight persuasion techniques, considering a multilabel perspective called PerSentSE (Section 4).

5. We presented a novel intelligent system to spot persuasion and those used techniques — eight in total — in spam emails at three different text granularity levels, i.e., document, sentence and fragment, which demonstrated spammers heavily rely on persuasion techniques (Section 4).

6. We proposed a novel feature set based on NLP to describe the spam email sender address with the aim of measuring their quality automatically using Machine Learning classifiers in Section **??** (Jáñez Martino et al., 2021).

7. We built and manually annotated two datasets from different sources, i.e., Bruce Guenter Reposity and private data from INCIBE, called together as Spam Email Risk Classification (SERC) and individually, SERC-BG and SERC-I, respectively. We followed two perspectives, one based on a risk score (1-10) for regression and another on two classes (low and high risk) for classification (Section **??**).

8. We proposed a new model to classify spam emails according to their potential risk for users. We followed a feature extraction study based on NLP to capture information from headers, body, attachments, URLs and protocols of spam emails with the aim of classifying them depending on their potential risk for users. It is based on 56 features clustered in five groups. We evaluated three classifiers and three estimators trained on our three datasets (SERC-BG, SERC-I and SERC) in Section **??**.

1.4. Publications and research results

This Section presents the research results obtained during the completion of this doctoral Thesis.

1.4.1. Publications related to this manuscript

- Jáñez-Martino, F., Fidalgo, E., González-Martínez, S., and Velasco-Mata, J. (2020). Classification of spam emails through hierarchical clustering and supervised learning. ArXiv preprint arXiv:2005.08773.

- Jáñez-Martino, F., Alaiz-Rodríguez, R., González-Castro, V., and Fidalgo, E. (2021). Trustworthiness of spam email addresses using Machine Learning. In Proceedings of the 21st ACM Symposium on Document Engineering, DocEng '21, page 4, New York, NY, USA. Association for Computing Machinery. (Rank B in CORE2021).

- Jáñez-Martino, F., Alaiz-Rodríguez, R., González-Castro, V., Fidalgo, E., and Alegre, E. (2022). A review of spam email detection: analysis of spammer strategies and the dataset shift problem. Artificial Intelligence Review, 56:1145–1173. (JCR: Q1 in 2023).

- Jáñez-Martino, F., Alaiz-Rodríguez, R., González-Castro, V., Fidalgo, E., and Alegre, E. (2023). Classifying spam emails using agglomerative hierarchical clustering and a topic-based approach. Applied Soft Computing, 139:110226. (JCR: Q1 in 2023).

- Jáñez-Martino, F., Barrón Cedeño A., Alaiz-Rodríguez, R., González-Castro, V., and Muti, A. On Persuasion in Spam Email: A Multi-Granularity Analysis Using Natural Language Processing. Submitted to Expert System with Applications in 2023 (JCR: Q1 in 2023).

- Jáñez-Martino, F., Alaiz-Rodríguez, R., González-Castro, V. and Alegre, E. (2023). Spam Email Classification Based on Cybersecurity Potential Risk Using Natural Language Processing. Submitted to Knowledge-Based Systems in 2023 (JCR: Q1 in 2023).

1.4.2. Other publications related to the Thesis subject

- Riesco, A., Fidalgo, E., Al-Nabki, M. W., Jáñez-Martino, F., and Alegre, E. (2019). Classifying pastebin content through the generation of pastecc labeled dataset. In 14th International Conference on Hybrid Artificial Intelligent Systems (HAIS), pages 1–12. (Rank C in 2018).

- Sánchez-Paniagua, M., Fidalgo, E., Alegre, E., Jáñez-Martino, F. (2021). Fraudulent E-Commerce Websites Detection Through Machine Learning. In: Sanjurjo

González, H., Pastor López, I., García Bringas, P., Quintián, H., Corchado, E. (eds) Hybrid Artificial Intelligent Systems. HAIS 2021. Lecture Notes in Computer Science(), vol 12886. Springer, Cham. (Rank: National Spain in 2020).

- Redondo-Gutierrez, L. A., Jáñez-Martino, F., Fidalgo, E., Alegre, E., González-Castro, V., and Alaiz-Rodríguez, R. (2022). Detecting malware using text documents extracted from spam email through Machine Learning. In Proceedings of the 22nd ACM Symposium on Document Engineering, DocEng '22, New York, NY, USA. Association for Computing Machinery. (Rank B in 2021).

1.4.3. Research projects

1. "Acuerdo de Colaboración para la puesta en marcha de un equipo de investigación aplicada en visión artificial y reconocimiento de patrones". Addendum 22 to the Framework Agreement between INCIBE (Spanish National Cybersecurity Institute).

2. "Acuerdo de Colaboración para la continuidad de los trabajos de un equipo de investigación aplicada en visión artificial y aprendizaje automático". Addendum 01 to the Framework Agreement between INCIBE (Spanish National Cybersecurity Institute).

3. "Ayuda para financiar la contratación predoctoral de personal investigador (Predoc) de la Junta de Castilla y Leon EDU/875/2021". Junta de Castilla y Leon EDU/875/2021 Predoc Grant.

4. European Union's Horizon 2020 Research and Innovation Framework Programme, H2020 SU-FCT- 2019 under the GRACE project with Grant Agreement 88334.

1.4.4. Attended conferences

- Presentation of "Trustworthiness of spam email addresses using Machine Learning" at the 22st ACM symposion on Document Engineering (DocEng 2022), Limerick, Ireland.

- Presentation of "Fraudulent E-Commerce Websites Detection Through Machine Learning" at the 22st ACM symposion on Document Engineering (DocEng 2022), Limerick, Ireland.

- Presentation of "Detecting malware using text documents extracted from spam email through Machine Learning" at the 16th International Conference on Hybrid Artificial Intelligence Systems (HAIS'21), Bibao, Spain.

1.4.5. Intellectual property registrations

- Al-Nabki, M. W., Fidalgo, E., Alegre, E., Alaiz-Rodríguez, R. and Jáñez-Martino, F., Application for the identification and classification of files related to Child Sexual Abuse. Application# Le-151-20, grant date: December 30, 2020[7].

- Al-Nabki, M. W., Fidalgo, E., Alegre, E., and Jáñez-Martino, F., Application for the name entity recognition on Darknet web Tor. Application# Le-4-21, grant date: January 5, 2021.

- Jáñez-Martino, F., Fidalgo, E., González-Castro, V., Alaiz-Rodríguez R., and Chaves D., Application for the classification of spam email addresses based on their quality. Application# 765-1013411, grant date: November 17, 2022.

- Jáñez-Martino, F., Fidalgo, E., González-Castro, V., Alaiz-Rodríguez R., and Biswas R., Application for the classification of spam email based on their topic. Application# 765-1020114, grant date: November 17, 2022.

- Jáñez-Martino, F., Fidalgo, E., Alegre E., Guerra-Vega D., Application for the classification of webpages: frauds, e-commerce, topic and entity type. Application# 765-1013266, grant date: November 17, 2022.

1.4.6. Other activities

Teaching activities

1. Electronics in the Bachelor's Degree in Computer Science at University of León.

2. Machine Learning in the Bachelor's Degree in Data Science and Artificial Intelligence at University of León.

3. Fundamentals of Deep Learning in the Master's Degree in Research in Cybersecurity at University of León.

Internship supervision

1. Lesec, Clémentine at école nationale supérieure des mines de Saint-étienne. Using Convolutional Neural Network (CNN) to cells segmentation of round cells in spermograms. Supervisors: Chaves Sánchez, D. and Jáñez-Martino, F. Full time from June 1 to August 8, 2020. University of León.

2. Redondo Gutiérrez, L.A., Supervisors: Jáñez-Martino, F. and Fidalgo Fernández, E., Internship II of 75h in Master's Degree in Research in Cybersecurity at University of León. April 2022.

[7] Spanish Patent and Trademark Office, published in Spanish

3. Redondo Gutiérrez, L.A., Supervisors: Jáñez-Martino, F. and Fidalgo Fernández, E., Internship III of 75h in Master's Degree in Research in Cybersecurity at University of León. May 2022.

4. Gómez García, J., Supervisors: Jáñez-Martino, F. and Fidalgo Fernández, E., Internship I of 75h in Master's Degree in Research in Cybersecurity at University of León. January 2023.

Co-supervisor of bachelor's final thesis projects

1. Navarro Donadios, M., Development of algorithms based on Artificial Intelligence for the detection of false messages in the field of air safety, directors Jáñez-Martino, F. and González Castro, V., Aeronautical Engineering department, University of León. February 2021.

2. Arias Martínez, L., Detection and classification of gaseous and dendritic pores in sand ceramic molds by Artificial Intelligence, directors: Jáñez-Martino, F. and Fidalgo Fernández, E. Mechanical Engineering department, University of León. September 2021.

Co-supervisor of master's final thesis projects

1. Cerviño Loira, B., Automatic roughness measurement in metal parts manufactured with SLM technology, directors: Fidalgo Fernández, E. and Jáñez-Martino, F. Mechanical Engineering department, University of León and University of Vigo. July 2021.

2. Diéguez González, L., Automatic roughness measurement in metal parts manufactured with additive manufacturing technology using Deep Learning, directors: Jáñez-Martino, F. and Fidalgo Fernández, E., Mechanical Engineering department, University of León and University of Vigo. September 2022.

3. Redondo Gutiérrez, L. A., Malware detection in spam emails using Natural Language Processing, directors: Jáñez-Martino, F. and Fidalgo Fernández, E., Computing Engineering department, University of León. July 2022

4. Castaño Ledesma, L. F., Creation of a phishing kit dataset for phishing websites identification, directors: Fidalgo Fernández, E. and Jáñez-Martino, F. Computing Engineering department, University of León. September 2022.

5. Martínez Darriba, A., Detection and classification of gas and dendritic pores in sand and ceramic molds using Deep Learning, directors: Fidalgo Fernández, E. and Jáñez Martino, F. Mechanical Engineering department, University of León and University of Vigo. July 2023.

6. Aldea Alonso, D., Evaluation of transfer learning for roughness measurement in additive manufacturing, directors: Fidalgo Fernández, E. and Jáñez-Martino, F. Mechanical Engineering department, University of León and University of Vigo. September 2023

International Mobility

1. Research stay at Department of Interpretation and Translation (DIT), Università di Bologna, Forlì, Italy, April - July 2022. Supervised by Prof. Luis Alberto Barrón Cedeño.

2. Research stay at Department of Interpretation and Translation (DIT), Università di Bologna, Forlì, Italy, February - April 2023. Supervised by Prof. Luis Alberto Barrón Cedeño.

Chapter 2

State of the Art

This chapter presents a review of spam detection and filtering, as well as its main challenges, dataset shift and adversarial figure, known as spammer. Due to copyright issues, we have removed this chapter from the thesis. It is possible to access to the related article through the following DOI: `https://doi.org/10.1007/s10462-022-10195-4`

Chapter 3

Spam Email Classification

This chapter introduces a novel approach to classify spam emails based on their cybersecurity topic. We used a hierarchical clustering and visual inspection to annotate spam emails and build two datasets in both English and Spanish. We also evaluate 16 machine learning and NLP pipelines to determine the most suitable one. Due to copyright issues, we have removed this chapter from the thesis. It is possible to access to the related article through the following DOI: https://doi.org/10.1016/j.asoc.2023.110226

Chapter 4

Detecting Persuasion in Spam Email

This chapter presents a novel empirical analysis on the identification of persuasion techniques in spam email. Our work represents the first time NLP is employed for the identification of persuasion in spam email. We explore this detection at three different granularity levels: full email, sentence, and text span. The related article is still in the publication process.

Chapter 5

Risk Classification of the Spam Emails

This chapter focuses on the classification of the risk that spam emails entail for users. We try to identify spam emails that pose a higher level of risk because they include malware, phishing, spoofing or other scams, and are carefully designed to mislead both anti-spam filters and receptors. We developed a feature extraction stage composed of five groups of feature vectors collected using NLP techniques. The related article is still in the publication process.

Chapter 6

Conclusions and Future Work

6.1. Work summary

Spam email is one of the main vectors for criminals to spread cyberthreats such as (spear) phishing, extortion, spoofing and black economy, among other scams. The growing number of cyberattacks originating in spam emails poses a risk the security and integrity of individuals and organizations. Therefore, Law Enforcements Agencies (LEAs) and cybersecurity organizations like Spanish National Cybersecurity Institute (INCIBE) attempt to provide early warnings and to detect those spam emails with potential risk using different available means, seeking to include Artificial Intelligence systems.

Currently, the analysis of spam emails is mostly conducted using manual approaches and rule-based systems. However, due to the huge volume of spam emails (big data problem), these methods are insufficient for analyzing all incoming messages. In this Thesis, we propose new methods to analyse and identify spam emails by applying NLP, Deep and Machine Learning, text classification and information retrieval techniques and algorithms. Specifically, this Thesis presents a comprehensive framework comprising three units: 1) a Spam Email Classification Unit (SCU) to categorize different spam emails into cybersecurity topics; 2) Persuasion Detection Unit (PDU) to spot persuasion and its types in spam emails at three different text granularity levels — entire message, sentences and text spans; and 3) Classification of the Spam Emails Risk Unit (CSPRU) to measure the potential risk of a spam email to both individuals and organizations. These units would help LEAs and INCIBE in analysing and spotting malicious spam emails automatically and faster.

The remainder of this Chapter provides a concise overview of the contributions made in this study, as well as outlines for future research to further expand upon our findings.

6.2. Summary of conclusions

In this Thesis, we present a framework based on three units to analyse and identify cyberthreats in spam emails. These applications and our findings are not limited to spam emails, but they can be extended to other areas with similar issues, such as social media, Short Message Services (SMS) or websites. This Section provides a summary of our conclusions to illustrate how this work can contribute to analyse and identify malicious spam emails. Our contributions are listed as follows:

- We semi-automatically labeled a novel dataset in the spam emails domain by using hierarchical clustering and visual inspection. In Chapter 3, we presented the first dataset, called Spam Email Classification (SPEMC), holding 14479 English and 14992 Spanish spam emails categorized in eleven cybersecurity topics.

- We presented a text classification model based on traditional Machine Learning algorithms to detect eleven cybersecurity topics in spam emails. In Chapter 3, we designed and evaluated a text classification pipeline. We explored four text representation approaches TF-IDF, Bag of Words, word2vec and BERT along with four commonly used Machine Learning classifier, Support Vector Machine (SVM), Naïve Bayes (NB), Logistic Regression (LR) and Random Forest (RF). Moreover, we created a processing pipeline to extract internally all textual content from spam email. In Chapter 3, we also exposed our system to extract and join the textual content considering the appearance of spammer strategies, such as image-based message or hidden text.

- We annotated sentences of spam emails based on binary and multilabel perspective. In Chapter 4, we explained our annotation procedure to label 1934 sentences into two classes, i.e. persuasion or non-persuasion, and eight persuasion techniques, i.e. *Appeal to authority*, *Appeal to fear/prejudice*, *Doubt*, *Exaggeration or minimization*, *Flag-waving*, *Loaded Language*, *Name Calling or Labeling* and *Repetition*. We called it as Persuasion Sentences in Spam Emails (PerSentSE).

- We developed intelligent systems to detect persuasion and its techniques at different levels of granularity, whole email, sentences and text spans. In Chapter 4, we explained our replication of proppy, a Machine Learning and Natural Language Processing system trained to detecting binary persuasion in outlet news. We also fine-tuned Transformer models, with RoBERTa on top, to spot persuasion sentences and which persuasion techniques among the eight — previously enumerated — techniques are used and combined. Finally, we spot those text spans containing persuasion and again, which techniques.

- We introduced a novel set of features based on Natural Language Processing to discriminate those spam emails with more potential risk for individuals and organizations. In Chapter 5, we provided a detailed description of the five feature sets, namely headers, text, attachments, URLs, and protocols. Each feature set was examined individually, outlining its content and the specific features it encompasses.

- We manually annotated two spam email datasets collected in different sources based on their potential risk. In Chapter 5, we presented two datasets, one collected by Bruce Guenter repository (1744 emails) and publicly available, and another containing private data from INCIBE (1905 emails) to capture both individuals and company spam emails. We respectively called them SPEC-BG and SPEC-I and labeled them for (i) a regression problem using a scale of risk (1-10) and (ii) a classification problem distinguishing two classes, low and high risk.

- We proposed a new model fed up with our set composed of 56 features following two perspectives: classification and regression. In Chapter 5, we described the evaluation of these models on SPEC-BG, SPEC-I and a combination of both (SPEC). We chose Support Vector Machine (SVM), Logistic Regression (LR) and Random Forest (RF) as classifiers and Support Vector Regression (SVR), Linear Regression and Random Forest Regression (RFR) as estimators.

- We conducted an analysis of feature importance for the classification approach by systematically removing or retaining one set of features at a time in Chapter 5. Additionally, we examined the impact of removing each individual feature by removing them one by one during the analysis (only for Random Forest classifier).

6.3. Open challenges and future work

Next, we would like to present a few several research lines that have emerged during this work, which may be worth exploring in the future.

- *Investigating spammer strategies based on obfuscation techniques.* One of the strategies commonly used by spammers is the use of textual obfuscation in their messages. However, this aspect has only been studied to a limited extend and is a very challenging problem (Mageshkumar et al., 2022). It is not only prevalent in spam emails but also extends to social media platforms and suspicious online forums (Álvaro Huertas-García et al., 2023).

- *Examining various attention models (Transformers)* Vaswani et al. (2017). In the domain of Natural Language Processing, attention models like BERT (Devlin et al., 2018), XLNet (Yang et al., 2020), or GPT-4 (Brown et al., 2020) have emerged as state-of-the-art solutions, demonstrating exceptional performance across many applications. Leveraging these models in our pipelines for the cybersecurity topic and risk classification can significantly enhance their effectiveness.

- *Improving the feature set of attachments in the risk classification.* Due to the low importance of some features, it is worth investing time in conducting a more thorough investigation of this particular set. This deeper analysis could potentially reveal a relationship between malware detection and textual information, as discussed in the study carried out by (Redondo-Gutierrez et al., 2022).

- *Increasing the number of persuasion techniques to spot in spam email.* Given the prevalence of persuasion techniques in spam emails, it is worth to broaden the scope of techniques identified within sentences and text spans (Ferreira et al., 2015).

- *Evaluate persuasion-based features for our risk prediction models.* Once the persuasion detection models have been improved, it is possible to incorporate features extracted from the outputs of these models, such as the percentage of persuasion in an email or which techniques have been used, into our potential risk classification model for spam emails.

- *Transfering knowledge to Short Message Service (SMS) and instant messaging.* Cybercriminals are increasingly leveraging SMS or instant messaging services to send unsolicited message as spammers do (Naqvi et al., 2023). Smishing, a form of attack through text messages, exploits social engineering techniques to deceive individuals into revealing sensitive information (Akande et al., 2023). Hence, adapting the strategies proposed in this work to fight this growing challenge can be an interesting starting step.

Capítulo 7

Conclusiones y perspectiva

7.1. Resumen del trabajo

El correo spam es uno de los principales vectores para los criminales a la hora de propagar ciberamenzas tales como phishing, extorsiones, spoofing o economía sumergida, entre otras estafas. El creciente número de ciberataques originados en los correos spam supone un riesgo para la seguridad y la integridad de personas y organizaciones. Por ello, las Fuerzas y Cuerpos de Seguridad (LEAs, de sus siglas en inglés Law Enforcement Agencies) y organizaciones de ciberseguridad como el Instituto Nacional de Ciberseguridad (INCIBE) intentan alertar tempranamente y detectar aquellos correos spam potencialmente peligrosos utilizando diferentes medios disponibles, buscando incluir sistemas de Inteligencia Artificial.

Actualmente, el análisis de correos spam se realiza principalmente mediante enfoques manuales y sistemas basados en reglas. Sin embargo, debido al enorme volumen de correos eletrónicos spam (suponiendo un problema de big data), estos métodos son insuficientes para analizar todos los mensajes entrantes. En esta Tesis, proponemos nuevos métodos para analizar e identificar correos spam mediante la aplicación de técnicas y algoritmos de Procesamiento del Lenguaje Natural, Aprendizaje Automático y Profundo, clasificación de texto y recuperación de información. Específicamente, esta Tesis presenta un marco integral que comprende tres unidades: 1) Spam Email Classification Unit (SCU) para clasificar correos electrónicos spam dependiendo de los diferentes temas de ciberseguridad contengan; 2) Persuasion Detection Unit (PDU) para detectar la persuasión y sus tipos en correos eletrónicos spam en diferentes granularidades de texto: correos completos, oraciones y fragmentos de texto; y 3) Classification of the Spam Emails Risk Unit (CSPRU) para medir el riesgo potencial de un correo electrónico spam tanto para usuarios personales como para organizaciones. Estas unidades ayudarían a las LEAs y al INCIBE a analizar y detectar correos eletrónicos spam maliciosos de forma automática y más rápida.

El resto de este capítulo proporciona una descripción general concisa de las contribuciones realizadas en este estudio, así como un esquema para futuras investigaciones para ampliar aún más nuestros hallazgos.

7.2. Conclusiones generales

En esta Tesis, presentamos un marco basado en tres unidades para analizar e identificar ciberamenazas en correos spam. Estas aplicaciones y nuestros hallazgos no se limitan a los correos spam, sino que pueden extenderse a otras áreas con problemas similares, como las redes sociales, los servicios de mensajes de texto (SMS, de sus siglas en inglés Short Message Service) o los sitios web. Esta sección proporciona un resumen de nuestras conclusiones para ilustrar cómo este trabajo puede contribuir a analizar e identificar correos spam maliciosos. Nuestras contribuciones se enumeran a continuación:

- Etiquetamos de forma semiautomática un conjunto de datos novedoso de correos spam mediante la agrupación jerárquica y la inspección visual. En el Capítulo 3, presentamos el primer conjunto de datos, llamado **SP**am **EM**ail Classification dataset (SPEMC), que contiene 14479 correos spam en inglés y 14992 en español categorizados en once temas de ciberseguridad.

- Presentamos un modelo de clasificación de texto basado en algoritmos tradicionales de Aprendizaje Automático para detectar once temas de ciberseguridad en correos spam. En el Capítulo 3, diseñamos y evaluamos un proceso de clasificación de texto. Exploramos cuatro enfoques de representación de texto: TF-IDF, Bag of Words (BOW), word2vec y BERT, junto con cuatro clasificadores de aprendizaje automático de uso común: Support Vector Machine (SVM), Naïve Bayes (NB), Logistic Regression (LR) y Random Forest (RF). Además, creamos un procesamiento para extraer internamente todo el contenido textual del correo spam. En el Capítulo 3, también expusimos nuestro sistema para extraer y unir contenido textual considerando la aparición de estrategias de spammers, como mensajes basados en imágenes o texto oculto.

- Anotamos oraciones de correos spam siguiendo dos perspectivas: binaria y multietiquetas. En el capítulo 4, explicamos nuestro procedimiento de anotación para etiquetar 1934 oraciones en dos clases: persuasión o no persuasión y ocho técnicas de persuasión (*Apelar a la Autoridad*, *Apelar al Miedo/prejuicio*, *Duda*, *Exageración o minimización*, *Patriotismo*, *Lenguaje Cargado*, *Descalificación o etiquetas* y *Repetición*). Lo llamamos Persuasive Sentences in Spam Emails (PerSentSE).

- Desarrollamos sistemas inteligentes para detectar la persuasión y sus técnicas en diferentes niveles de granularidad, correos electrónicos completos, oraciones y fragmentos de texto. En el Capítulo 4, explicamos nuestra replicación de Proppy, un sistema de Aprendizaje Automático y Procesamiento del Lenguaje Natural entrenado para detectar la persuasión binaria en los medios de comunicación. También ajustamos los modelos de Transformer, con RoBERTa como base, para detectar oraciones de persuasión y qué técnicas de persuasión entre ocho técnicas (enumeradas anteriormente) se utilizan y combinan. Finalmente, identificamos aquellos fragmentos de texto que contienen persuasión y nuevamente, qué técnicas.

- Introdujimos un nuevo conjunto de características basadas en el Procesamiento del Lenguaje Natural para discriminar aquellos correos spam con mayor riesgo potencial para usuarios personales y organizaciones. En el Capítulo 5, proporcionamos una descripción detallada de los cinco conjuntos de características (encabezados, texto, archivos adjuntos, URLs y protocolos). Cada conjunto de características se examinó individualmente, describiendo su contenido y las características específicas que abarca.

- Anotamos manualmente dos conjuntos de datos de correo spam recopilados en diferentes fuentes según su riesgo potencial. En el Capítulo 5, presentamos dos conjuntos de datos, uno recopilado por el repositorio de Bruce Guenter (1744 correos electrónicos), disponible públicamente, y otro que contiene datos privados de INCIBE (1905 correos electrónicos) para analizar correos spam tanto de usuarios personales como de organizaciones. Conjuntamente le otorgamos el nombre de Spam Email Risk Classification (SERC-4K). Llamamos SPEC-BG y SPEC-I respectivamente y los etiquetamos según un problema de regresión utilizando una escala de riesgo (1-10) y un problema de clasificación que distingue dos clases, riesgo bajo y alto.

- Propusimos un nuevo enfoque al alimentar modelos de Machine Learning con nuestro conjunto de 56 características siguiendo dos perspectivas: clasificación y regresión. En el Capítulo 5, describimos la evaluación de estos modelos en SPEC-BG, SPEC-I y una combinación de ambos (SPEC). Elegimos SVM, LR y RF como clasificadores y Support Vector Regression (SVR), regresión lineal y Random Forest Regression (RFR) como estimadores.

- Realizamos un análisis de la importancia de las características para el enfoque de clasificación, eliminando o reteniendo sistemáticamente un conjunto de características a la vez en el Capítulo 5. Además, examinamos el impacto de eliminar cada característica una por una durante el análisis (solo para el clasificador RF).

7.3. Trabajos futuros

A continuación, nos gustaría presentar algunas líneas de investigación que han surgido durante este trabajo y que quizás valga la pena explorar en el futuro.

- *Investigar estrategias de spammers basadas en técnicas de ofuscación.* Una de las estrategias comúnmente utilizadas por los spammers es el uso de ofuscación textual en sus mensajes. Sin embargo, este aspecto sólo se ha estudiado de forma limitada y es un problema muy presente en los correos spam actuales (Mageshkumar et al., 2022). No solo prevalece en los correos spam, sino que también se extiende a las plataformas de redes sociales y foros en línea sospechosos (Álvaro Huertas-García et al., 2023).

- *Examinar varios modelos de atención (Transformers)* Vaswani et al. (2017). En el ámbito del Procesamiento del Lenguaje Natural, modelos de atención como BERT (Devlin et al., 2018), XLNet (Yang et al., 2020) o GPT-4 (Brown et al., 2020) han surgido como soluciones de última generación, demostrando un rendimiento excepcional en muchas aplicaciones. Aprovechar estos modelos en nuestros modelos para la clasificación de riesgo puede mejorar significativamente su efectividad.

- *Mejorar el conjunto de características de los archivos adjuntos en la clasificación de riesgo.* Debido a la poca importancia de algunas características, vale la pena invertir tiempo en realizar una investigación más exhaustiva de este conjunto en particular. Este análisis más profundo podría potencialmente revelar una relación entre la detección de malware y la información textual, como se analiza en el estudio de (Redondo-Gutierrez et al., 2022).

- *Aumentar el número de técnicas de persuasión a detectar en correos spam.* Dada la prevalencia de las técnicas de persuasión en los correos spam, vale la pena ampliar el alcance de las técnicas identificadas en oraciones y fragmentos de texto (Ferreira et al., 2015).

- *Evaluación de características basadas en la persuasión para nuestros modelos de predicción de riesgo.* Una vez mejorados los modelos de detección de persuasión, cabe la posibilidad de incorporar características extraídas de las salidas de estos modelos, como pueden ser el porcentaje de persuasión en un correo o qué técnicas han sido usadas, a nuestro modelo de clasificación del riesgo potencial de un correo spam.

- *Transferencia de conocimientos al SMS y mensajería instantánea.* Los ciberdelincuentes aprovechan cada vez más los SMS o los servicios de mensajería instantánea para enviar mensajes no solicitados, como lo hacen los spammers en el correo electrónico (Naqvi et al., 2023). El smishing, una forma de ataque a través de mensajes de texto, explota técnicas de ingeniería social para engañar a las personas para que revelen información confidencial (Akande et al., 2023). Por lo tanto, adaptar las estrategias propuestas en este trabajo para luchar contra este desafío creciente puede ser un punto de partida interesante.

Bibliography

Akande, O. N., Gbenle, O., Abikoye, O. C., Jimoh, R. G., Akande, H. B., Balogun, A. O., and Fatokun, A. (2023). Smsprotect: An automatic smishing detection mobile application. *ICT Express*, 9(2):168–176.

Ali, K., Li, C., ul abdin, K. Z., and Muqtadir, S. A. (2022). The effects of emotions, individual attitudes towards vaccination, and social endorsements on perceived fake news credibility and sharing motivations. *Computers in Human Behavior*, 134:107307.

Barrón-Cedeño, A., Jaradat, I., Da San Martino, G., and Nakov, P. (2019). Proppy: Organizing the news based on their propagandistic content. *Information Processing & Management*, 56(5):1849–1864.

Bera, D., Ogbanufe, O., and Kim, D. J. (2023). Towards a thematic dimensional framework of online fraud: An exploration of fraudulent email attack tactics and intentions. *Decision Support Systems*, page 113977.

Bhowmick, A. and Hazarika, S. M. (2018). E-mail spam filtering: A review of techniques and trends. *Advances in Electronics, Communication and Computing*, 443:583–590.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language Models are Few-Shot Learners.

Chen, S., Xiao, L., and Mao, J. (2021). Persuasion strategies of misinformation-containing posts in the social media. *Information Processing & Management*, 58(5):102665.

Chernyavskiy, A., Ilvovsky, D., and Nakov, P. (2020). Aschern at SemEval-2020 task 11: It takes three to tango: RoBERTa, CRF, and transfer learning. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1462–1468, Barcelona (online). International Committee for Computational Linguistics.

Da San Martino, G., Yu, S., Barrón-Cedeño, A., Petrov, R., and Nakov, P. (2019). Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Dada, E. G., Bassi, J. S., Chiroma, H., Abdulhamid, S. M., Adetunmbi, A. O., and Ajibuwa, O. E. (2019). Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon*, 5(6):e01802.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805:1–16.

El Aassal, A., Baki, S., Das, A., and Verma, R. (2020). An In-Depth Benchmarking and Evaluation of Phishing Detection Research for Security Needs. *IEEE Access*, 8:1–1.

Ferrara, E. (2019). The history of digital spam. *Communications of the ACM*, 62(8):82–91.

Ferreira, A., Coventry, L., and Lenzini, G. (2015). Principles of persuasion in social engineering and their use in phishing. In Tryfonas, T. and Askoxylakis, I., editors, *Human Aspects of Information Security, Privacy, and Trust*, pages 36–47, Cham. Springer International Publishing.

Ferreira, A. and Teles, S. (2019). Persuasion: How phishing emails can influence users and bypass security measures. *International Journal of Human-Computer Studies*, 125:19–31.

Gallo, L., Maiello, A., Botta, A., and Ventre, G. (2021). 2 years in the anti-phishing group of a large company. *Computers & Security*, 105:102259.

Haupt, J., Bender, B., Fabian, B., and Lessmann, S. (2018). Robust identification of email tracking: A machine learning approach. *European Journal of Operational Research*, 271(1):341–356.

Jáñez Martino, F., Alaiz-Rodríguez, R., González-Castro, V., and Fidalgo, E. (2021). Trustworthiness of spam email addresses using machine learning. In *Proceedings of the 21st ACM Symposium on Document Engineering*, DocEng '21, page 4, New York, NY, USA. Association for Computing Machinery.

Jáñez-Martino, F., Alaiz-Rodríguez, R., González-Castro, V., Fidalgo, E., and Alegre, E. (2022). A review of spam email detection: analysis of spammer strategies and the dataset shift problem. *Artificial Intelligence Review*, 56:1145–1173.

Jáñez-Martino, F., Alaiz-Rodríguez, R., González-Castro, V., Fidalgo, E., and Alegre, E. (2023). Classifying spam emails using agglomerative hierarchical clustering and a topic-based approach. *Applied Soft Computing*, 139:110226.

Jáñez-Martino, F., Fidalgo, E., González-Martínez, S., and Velasco-Mata, J. (2020). Classification of spam emails through hierarchical clustering and supervised learning.

Karbab, E. B. and Debbabi, M. (2019). Maldy: Portable, data-driven malware detection using natural language processing and machine learning techniques on behavioral analysis reports. *Digital Investigation*, 28:S77–S87.

Lawson, P., Pearson, C. J., Crowson, A., and Mayhorn, C. B. (2020). Email phishing and signal detection: How persuasion principles and personality influence response patterns and accuracy. *Applied Ergonomics*, 86:103084.

Lin, T., Capecci, D. E., Ellis, D. M., Rocha, H. A., Dommaraju, S., Oliveira, D. S., and Ebner, N. C. (2019). Susceptibility to spear-phishing emails: Effects of internet user demographics and email content. *ACM Trans. Comput.-Hum. Interact.*, 26(5).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Magdy, S., Abouelseoud, Y., and Mikhail, M. (2022). Efficient spam and phishing emails filtering based on deep learning. *Computer Networks*, 206:108826.

Mageshkumar, N., Vijayaraj, A., Arunpriya, N., and Sangeetha, A. (2022). Efficient spam filtering through intelligent text modification detection using machine learning. *Materials Today: Proceedings*, 64:848–858. International Conference on Advanced Materials for Innovation and Sustainability.

Mishra, S. and Soni, D. (2020). Smishing detector: A security model to detect smishing through sms content analysis and url behavior analysis. *Future Generation Computer Systems*, 108:803–815.

Mohammad, R. M. A. (2020). A lifelong spam emails classification model. *Applied Computing and Informatics*, page 11.

Murugavel, U. and Santhi, R. (2020). Detection of spam and threads identification in e-mail spam corpus using content based text analytics method. *Materials Today: Proceedings*, pages 3319–3323.

Naqvi, B., Perova, K., Farooq, A., Makhdoom, I., Oyedeji, S., and Porras, J. (2023). Mitigation strategies against the phishing attacks: A systematic literature review. *Computers & Security*, page 103387.

Redondo-Gutierrez, L. A., Jáñez Martino, F., Fidalgo, E., Alegre, E., González-Castro, V., and Alaiz-Rodríguez, R. (2022). Detecting malware using text documents extracted from spam email through machine learning. In *Proceedings of the 22nd ACM Symposium on Document Engineering*, DocEng '22, New York, NY, USA. Association for Computing Machinery.

Saidani, N., Adi, K., and Allili, M. S. (2020). A semantic-based classification approach for an enhanced spam detection. *Computers & Security*, 94:101716.

Sánchez-Paniagua, M., Fidalgo, E., González-Castro, V., and Alegre, E. (2021). Impact of Current Phishing Strategies in Machine Learning Models for Phishing Detection. In Herrero, Á., Cambra, C., Urda, D., Sedano, J., Quintián, H., and Corchado, E., editors, *13th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2020)*, pages 87–96, Cham. Springer International Publishing.

Sankhwar, S., Pandey, D., and Khan, P. R. (2018). Email phishing: An enhanced classification model to detect malicious urls. *ICST Transactions on Scalable Information Systems*, 6:158529.

Sonowal, G. (2020). Phishing email detection based on binary search feature selection. *SN Computer Science*, 1:14.

Sony Dewantara, D. and Budi, I. (2020). Combination of lstm and cnn for article-level propaganda detection in news articles. In *2020 Fifth International Conference on Informatics and Computing (ICIC)*, pages 1–4.

Sturman, D., Valenzuela, C., Plate, O., Tanvir, T., Auton, J. C., Bayl-Smith, P., and Wiggins, M. W. (2023). The role of cue utilization in the detection of phishing emails. *Applied Ergonomics*, 106:103887.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

Volkamer, M., Renaud, K., Reinheimer, B., and Kunz, A. (2017). User experiences of torpedo: Tooltip-powered phishing email detection. *Computers & Security*, 71:100–113.

Wang, D., Irani, D., and Pu, C. (2013). A Study on Evolution of Email Spam Over Fifteen Years. In *9th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing*, pages 1–10.

Wang, J., Shi, J., Wen, X., Xu, L., Zhao, K., Tao, F., Zhao, W., and Qian, X. (2022). The effect of signal icon and persuasion strategy on warning design in online fraud. *Computers & Security*, page 102839.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2020). Xlnet: Generalized autoregressive pretraining for language understanding.

Yu, S. (2015). Covert communication by means of email spam: A challenge for digital investigation. *Digital Investigation*, 13:72 – 79.

Álvaro Huertas-García, Martín, A., Huertas-Tato, J., and Camacho, D. (2023). Countering malicious content moderation evasion in online social networks: Simulation and detection of word camouflage. *Applied Soft Computing*, 145:110552.