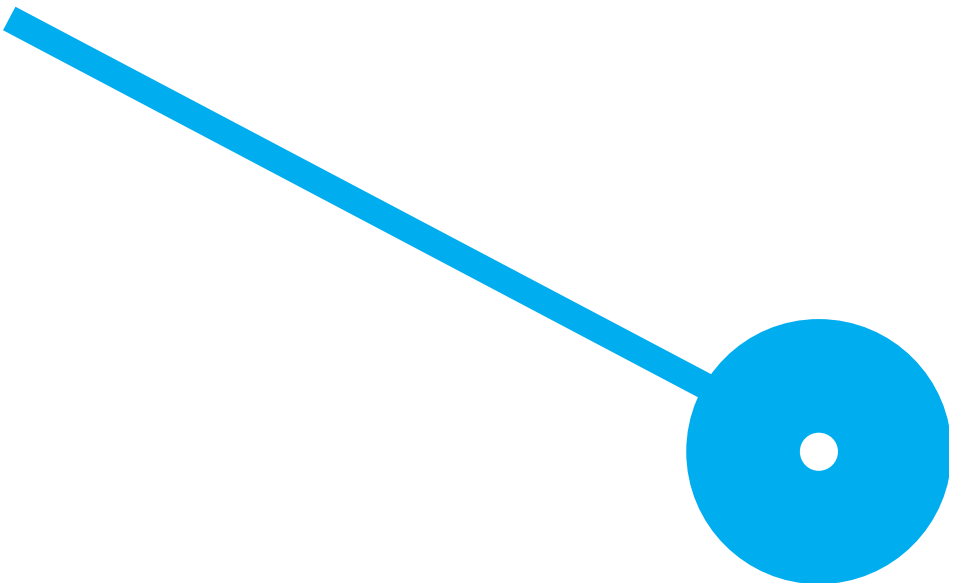
SCHOOL OF MANAGEMENT AND TECHNOLOGY
POLYTECHNIC OF PORTO

MASTER
ENGENHARIA INFORMÁTICA

Phishing Email Classification through the Lens of Psychology and Machine Learning

Sara Patrícia Santos Rodrigues

OCTOBER/**2023**



SCHOOL OF MANAGEMENT AND TECNOLOGY
POLYTECHNIC OF PORTO

MASTER
ENGENHARIA INFORMÁTICA

Phishing Email Classification through the Lens of Psychology and Machine Learning

Sara Patrícia Santos Rodrigues
8200796

Advisor(s)

Associate Professor João Paulo Ferreira de Magalhães
Invited Professor Carmen Susana Cardoso Nunes

Dissertation submitted in fulfilment of the requirements for the Master's degree in Engenharia Informática in the School of Management and Technology of the Polytechnic of Porto.

OCTOBER/**2023**

Integrity Statement

I, Sara Patrícia Santos Rodrigues, student nº 8200796, of the Master's Degree in Engenharia Informática of the School of Management and Technology of the Polytechnic of Porto, declare that I have not plagiarized or self-plagiarized, therefore the work entitled " Phishing Email Classification through the Lens of Psychology and Machine Learning" is original and of my own authorship, not having been used previously for any other purpose. I further declare that all sources used are cited, in the text and in the final bibliography, according to the referencing rules adopted in the institution.

SCHOOL OF
MANAGEMENT AND
TECHNOLOGY
POLYTECHNIC OF
PORTO

P.PORTO

M

MASTER
ENGENHARIA INFORMÁTICA

Phishing Email Classification through the Lens of Psychology and Machine Learning

Sara Patrícia Santos Rodrigues

João Paulo Ferreira de Magalhães, Associate Professor

Carmen Susana Cardoso Nunes, Invited Professor

Acknowledgments

I wish to express my profound gratitude to my advisors, Professor João Paulo and Professor Susana, who promptly accepted my invitation to be my mentors. Their collective guidance, expertise, and unwavering support were instrumental in the successful completion of this work.

I would also like to extend my heartfelt thanks to all the participants who generously dedicated their time and insights to this study, as well as to my friends and family for their continual encouragement.

Lastly, I am appreciative of everyone who has been a pillar of support throughout this academic journey. Your assistance and motivation have been priceless.

Abstract

This dissertation delves into the complex domain of email security, investigating the interplay of personality traits, decisional patterns, emotions, valence, and polarity in email content evaluation. The study employs a diverse set of analytical tools, ranging from human evaluators to AI tools, to comprehensively explore email classification and its associated psychological and emotional dimensions. Emotional analysis presents contrasting results between human and AI-generated assessments, underscoring the complexity of interpreting emotional nuances within emails. The dissertation further dissects email valence and polarity, showcasing how legitimate emails tend to maintain a neutral valence, while phishing emails employ strongly negative cues. The low degree of consensus in the emphasizes the critical need for incorporating phishing literacy training as a crucial component of a company's cybersecurity strategy. The research holds significant implications for email security and communication analysis, providing a foundation for future advancements in this critical domain.

Keywords: Phishing, Personality, Emotion, Decision Making, Machine Learning, AI tools

Resumo

Esta dissertação explora o complexo domínio da segurança de emails, investigando a interação entre traços de personalidade, padrões de tomada de decisão, emoções, valência e polaridade na avaliação de conteúdo de emails. O estudo utiliza uma variedade de ferramentas analíticas, desde avaliadores humanos até ferramentas de IA, para explorar de forma abrangente a classificação de emails e suas dimensões psicológicas e emocionais associadas. A análise emocional apresenta resultados contrastantes entre avaliações humanas e geradas por IA, destacando a complexidade na interpretação de nuances emocionais em emails. A dissertação também analisa a valência e polaridade dos emails, mostrando como os emails legítimos tendem a manter uma valência neutra, enquanto os emails de phishing empregam indicadores fortemente negativos. O baixo grau de consenso destaca a necessidade crítica de incorporar a formação em literacia de phishing como um componente essencial da estratégia de cibersegurança de uma empresa. A pesquisa possui implicações significativas para a segurança de emails e análise de comunicações, estabelecendo uma base para futuros avanços neste domínio crítico.

Palavras-chave: Phishing, Personalidade, Emoção, Tomada de decisão, Machine Learning, Ferramentas de AI

Contents

Acknowledgments ii

Abstract iii

Resumo iv

Acronyms viii

Glossary ix

List of Figures x

List of Tables xii

Chapter 1 – Introduction 1

Chapter 2 - State-of-the-art 3

2.1. Systematic Literature Review methodology 3

2.1.1. Selection of Research Articles 3

2.1.2. Search Criteria and Exclusion Process 3

2.1.3. Article Screening and Selection 3

2.1.4. Data Extraction and Analysis 3

2.2. Systematic Literature Review results 4

2.2.1 Search Results and Publication Identification 4

2.2.2 Screening and Exclusion Process 4

2.2.3 Study Design and Categorization 5

2.3. Systematic Literature Review - Discussion of the results 10

2.3.1. Best performing Algorithms 10

2.3.2. Emotions 11

2.3.3. Personality 11

2.4. Cyberattacks / Social engineering attacks 12

2.4.1. Smishing and vishing 12

2.4.2. Scareware 12

2.4.3. Tailgating/Piggybacking 12

2.4.4. Watering Hole 13

2.4.5. Business Email Compromise (BEC) 13
2.4.6. Phishing 13
2.4.7. Whaling 15
2.4.8. Baiting 15
2.4.9. Honey Trap 15
2.4.10. Pretexting 16
2.4.11. Quid Pro Quo (i.e., tech support scams) 16
2.5. Artificial Intelligence 16
2.5.1. Machine Learning Algorithms 17
2.5.2. Natural Language Processing (NLP) 23
2.5.3. Named Entity Recognition (NER) 24
2.5.4. AI tools in the scope 25
a) ChatGPT 25
b) IBM Watson 25
2.6. Psychological characteristics and predisposition to be a victim of phishing 26
2.6.1. Model of emotion of Paul Ekman 26
2.6.2. Personality Inventory NEO-FFI-20 27
2.6.3. Melbourne Decision Making Questionnaire (Portuguese Version) 28
Chapter 3 - Methodology 29
3.1. Selected Emails 29
3.1.1. Legitimate emails 29
3.1.2. Phishing Emails 38
3.2. Research inquiry 46
3.2.2. Sample 49
3.3. Tools Research 52
3.3.1. IBM Watson 52
3.3.2. Chat GPT 52
Chapter 4 – Results 53

4.1. AI Tools Research 53
4.1.1. IBM Watson 53
4.1.2. Chat GPT 70
4.2. Inquiry results 84
4.2.1. Legitimate identification concordance 84
4.2.2. Phishing Identification concordance 85
4.2.3. Valence evaluation 85
4.2.4. Activation Evaluation 86
4.2.5. Emotion Evaluation 86
4.2.6. Personality traits Evaluation 88
4.2.7. Decisional patterns evaluation 91
4.3. Result Analysis 92
Chapter 5 – Discussion 94
5.1. Polarity 94
5.2. Concordance 94
5.3. Valence 95
5.4. Activation 95
5.5. Emotions 96
5.6. Personality traits 97
5.7. Decision patterns 97
Chapter 6 – Limitations and recommendations 99
Chapter 7 – Conclusions 100
References 102
Appendix I – Inventário de Personalidade NEO-FFI-20 108
Appendix II – Melbourne Decision Making Questionnaire (Portuguese Version) 110
Appendix III – Questionário de Melbourne de Tomada de Decisão Authorization 112
Appendix IV –Inventário de Personalidade NEO-FFI-20 Authorization 113

Acronyms

BEC – Business Email Compromise

AI – Artificial Intelligence

ML – Machine Learning

IC3 – Internet Crime Complaint Center

IAPS – International Affective Picture Set

NLTK – Natural Language Toolkit

LIWC – Linguistic Inquiry and Word Count

NB – Naive Bayes

MP – Multilayer Perceptron

EOAL – Ensemble One-vs-All Learning

CNN – Convolutional Neural Network

NER – Named Entity Recognition

NLP – Natural Language processing

RNN – Recurrent Neural Network

SMS – Short Messaging Service

SVM – Support Vector Machines

Glossary

BERT: BERT (Bidirectional Encoder Representations from Transformers) is a powerful natural language processing model developed by Google that comprehends the context and meaning of words in a sentence by considering the surrounding words in both directions.

NEO PI-R: NEO PI-R (Revised NEO Personality Inventory) is a widely used psychological assessment tool designed to measure an individuals personality traits. It is based on the Five-Factor Model (FFM) of personality, which includes five major personality dimensions: Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness. The NEO PI-R provides a comprehensive assessment of these personality traits and is frequently employed in psychological research, clinical settings, and personality studies to better understand an individuals personality characteristics.

S-BERT: S-BERT, or Sentence-BERT, is an extension of the BERT (Bidirectional Encoder Representations from Transformers) model, specifically tailored for understanding the semantic similarity and context between sentences. It creates meaningful sentence embeddings by encoding sentence pairs together, making it a valuable tool for various NLP tasks, such as text classification, paraphrase identification, and information retrieval.

List of Figures

Figure 1 PRISMA 2020 flow diagram 5
Figure 2 Machine Learning Algorithms 17
Figure 3 Regression vs Classification [49] 18
Figure 4 Actor Tim Roth portraying facial expressions and their explanation [50] 27
Figure 5 Legitimate Email CTT 29
Figure 6 Legitimate Email Techonline 30
Figure 7 Legitimate Email "moey!" 30
Figure 8 Legitimate Email Fidelidade 31
Figure 9 Legitimate Email Flixbus 32
Figure 10 Legitimate email MEO 33
Figure 11 Legitimate email Santander 34
Figure 12 Legitimate Email Santander 35
Figure 13 Legitimate Email Galp 36
Figure 14 Legitimate Email Meu Presente 37
Figure 15 Phishing Email CTT 38
Figure 16 Phishing Email Nespresso 39
Figure 17 Phishing Email Worten 40
Figure 18 Phishing Email Moey! 41
Figure 19 Phishing Email MAPFRE insurances 42
Figure 20 Phishing Email Info S 43
Figure 21 Phishing Email CTT 43
Figure 22 Phishing Email Nivea 44
Figure 23 Phishing Email Grupo8 45
Figure 24 Phishing Email Info- Levarte SL 46
Figure 25 Inquiry methodology workflow 47
Figure 26 Inquiry demographic data section workflow 48
Figure 27 IBM Watson Classification results – Sentiment analysis for CTT Tolls email 53
Figure 28 IBM Watson Classification results – Sentiment analysis for Meu Presente 54
Figure 29 IBM Watson Classification results – Sentiment analysis for Techonline 55
Figure 30 IBM Watson Classification results – Sentiment analysis for Moey! 56
Figure 31 IBM Watson Classification results – Sentiment analysis for Fidelidade 57
Figure 32 IBM Watson Classification results – Sentiment analysis for MEO 58

Figure 33 IBM Watson Classification results – Sentiment analysis for FlixBus 59
Figure 34 IBM Watson Classification results – Sentiment analysis for Santander extract 60
Figure 35 IBM Watson Classification results – Sentiment analysis for Santander payment 61
Figure 36 IBM Watson Classification results – Sentiment analysis for MAPFRE 62
Figure 37 IBM Watson Classification results – Sentiment analysis for CTT Package 63
Figure 38 IBM Watson Classification results – Sentiment analysis for Nescafé 64
Figure 39 IBM Watson Classification results – Sentiment analysis for Worten 65
Figure 40 IBM Watson Classification results – Sentiment analysis for Moey! 66
Figure 41 IBM Watson Classification results – Sentiment analysis for InfoS 66
Figure 42 IBM Watson Classification results – Sentiment analysis for CTT Incomplete address 67
Figure 43 IBM Watson Classification results – Sentiment analysis for Info - Levarte SL 68
Figure 44 IBM Watson Classification results – Sentiment analysis for Grupo8 69
Figure 45 Frequency of emotions in retrieved by ChatGPT for Phishing emails vs Legitimate emails 83
Figure 46 Emotional Frequency Comparison: Phishing vs. Legitimate Emails 88
Figure 47 Summary of Chat-GPT vs Inquiry emotion classification 93

List of Tables

Table 1 – Summary of analyzed articles topics 6
Table – Inquiry Results – Sociodemographic data 52
Table 3 – IBM Watson Analysis of Email Polarity and Points 70
Table 4 – Analysis of Global Agreement Legitimate emails 84
Table 5 – Analysis of Global Agreement Phishing emails 85
Table 6 – Student T test of Email Valence and Phishing Classification 86
Table 7 – Email Activation and Phishing Classification 86
Table 8 – Correlation Between Personality Traits and Email Classification 90
Table 9 – Correlation Between Personal Decisional Patterns and Email Phishing Classification 92

Chapter 1 – Introduction

Social engineering encompasses a wide spectrum of malicious activities achieved through human interactions, with its core objective being the utilization of psychological manipulation to deceive individuals into making security errors or disclosing sensitive information. These deceptions encompass actions like revealing personal details, clicking on web links, or opening potentially malicious attachments. The realm of social engineering introduces a dozen distinct attack types, including Smishing, vishing, Scareware, Tailgating/Piggybacking, Watering Hole, Business Email Compromise (BEC), Whaling, Baiting, Honey Trap, Pretexting, Quid Pro Quo, as well as two variations of Phishing: Angler Phishing and Spear Phishing. Among these, phishing emerges as a dominant threat; according to the Internet Crime Complaint Center (IC3) of the FBI, phishing scams account for a significant portion of cybercrime activities, culminating in approximately 3.4 billion (short scale) phishing emails dispatched each day. This staggering volume underscores the prevalence of phishing as the most pervasive form of social engineering attack. Typically employing spoofed email addresses and deceptive links, phishing endeavors to dupe individuals into disclosing login credentials, credit card information, or other personal data, making it an immensely lucrative criminal enterprise. Predictions for 2023 suggest that over 33 million records will be extorted through such malicious activities. In most cases, phishing emails employ vague and generic content, often invoking a sense of urgency and emanating from unfamiliar sources. These emails usually carry unrecognizable attachments or links, coaxing recipients to provide personal or financial data on deceptive web pages or pop-up windows. Phishing campaigns have grown increasingly sophisticated, leveraging current events and trends to manipulate emotions, invoking economic uncertainties, or soliciting contributions to charities. Malicious actors are adept at exploiting the human element, preying on emotions and vulnerabilities to deceive victims. Understanding these emotional triggers and vulnerabilities is paramount in developing effective mitigation strategies. Artificial Intelligence (AI) and Machine Learning (ML) have emerged as invaluable tools for processing vast datasets and deriving novel insights. When applied to combat phishing, AI and ML enhance the creation of improved mitigation solutions. This study presents a systematic review of various approaches employing ML to detect phishing emails based on emotional content.

This document comprises seven chapters, each serving a distinct purpose. The Introduction chapter sets the stage for the dissertation, providing an overview of the subsequent chapters. The second chapter, State-of-the-Art, conducts a comprehensive literature review and introduces key elements central to this research, including cyberattacks and social engineering, artificial intelligence, Paul Ekman emotion model,

the NEO-FFI-20 personality inventory, and the Melbourne Decision Making Questionnaire in its Portuguese version.

Chapter three, Methodology, is characterized by three main components: the selection of emails, the research inquiry, and the investigation of AI tools. The fourth chapter, Results, is further divided into two subchapters, describing the outcomes of the AI tools research and the inquiry results.

The fifth chapter, Discussion, consist of eight subchapters, each delving into specific aspects: Polarity, Concordance, Valence, Activation, Emotions, Personality Traits, Decision Patterns, and the Limitations of the research. These subchapters collectively explore and analyze the research findings and implications.

Chapter six summarizes the Limitations and recommendations, offering an opportunity for further exploration in the field.

Finally, Chapter seven summarizes the Conclusions, summarizing the key learnings from the research.

Chapter 2 - State-of-the-art

This chapter comprises the methodology, results and discussion of the state of the art. It also includes an introduction to the categorization of cyberattacks, social engineering attacks, artificial intelligence, and the Paul Ekman model of emotion.

2.1. Systematic Literature Review methodology

The Systematic Literature Review PRISMA methodology encompasses the following components: Research Article Selection, Search Criteria and Exclusion Process, Article Screening and Selection, and Data Extraction and Analysis.

2.1.1. Selection of Research Articles

In this study, a comprehensive analysis of the state-of-the-art literature on phishing, email, machine learning, and emotion was conducted. Articles published until January 16, 2023, were considered for this research. Extensive searches were performed on reputable databases including Scopus, Web of Science, PubMed, and B-ON. The search was based on specific keywords and their variants, such as phishing, email, machine learning, and emotion.

2.1.2. Search Criteria and Exclusion Process

The search expression used across all databases encompassed various forms and combinations of the selected keywords. Exclusion criteria were applied to filter out irrelevant articles, which included wrong topic, wrong study, wrong target, wrong outcome, wrong publication type, foreign language, and background articles. Additionally, articles published before 2015 were not considered, and only journal articles were included while excluding other publication types like conferences.

2.1.3. Article Screening and Selection

All the articles identified through the search process were systematically reviewed. The author R., S., and co-orienteer C., S. used the Ryyan platform to individually screen and select relevant articles based on their titles and abstracts. The resulting articles selected by each author underwent a cross-check analysis to ensure the final list included the most significant works, adhering to the predefined selection criteria.

2.1.4. Data Extraction and Analysis

Data from the selected articles were extracted and organized in a tabular format for further analysis. The extracted information included author and year, the source of phishing emails, the number of phishing emails used, the number of non-spam emails, machine learning algorithms employed, machine learning

techniques utilized, relevant packages used, phishing emotions studied, personality traces investigated, and the obtained results.

2.2. Systematic Literature Review results

This subsection covers the Search Results and Publication Identification, the Screening and Exclusion Process, and the Study Design and Categorization.

2.2.1 Search Results and Publication Identification

In the year 2023, a systematic literature review PRISMA was conducted across multiple databases, including Web of Science, Scopus, PubMed, and B-on. A total of 310 publications were identified: 8 from Web of Science, 27 from Scopus, 0 from PubMed, and 304 from B-on. After excluding 39 duplicate records, the initial pool of publications for screening consisted of 271 unique records.

2.2.2 Screening and Exclusion Process

The screening process involved multiple steps to identify relevant publications. In the first step, abstracts of the 271 publications were screened, resulting in the inclusion of 196 publications. The second step involved retrieving the full-text reports for further evaluation. In the third step, based on predefined criteria, a total of 87 publications were excluded due to a wrong study design, 47 due to being unrelated to the topic, 12 due to incorrect outcomes, 11 for being of the wrong publication type, 4 for being in a foreign language, and 2 for being background articles. Figure 1 illustrates the screening process and the number of publications at each stage. Ultimately, 14 publications were included for analysis.

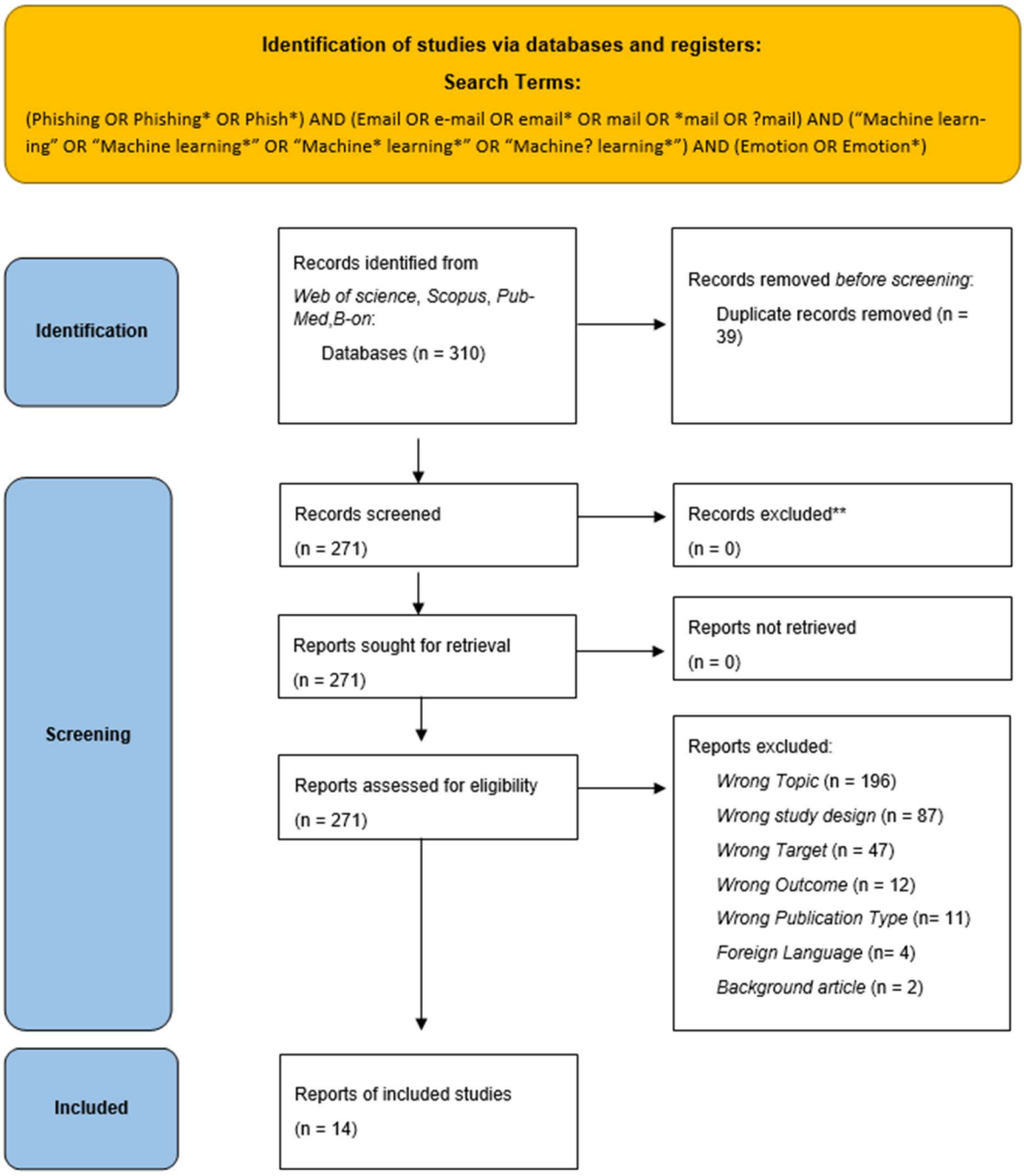


Figure 1 PRISMA 2020 flow diagram

2.2.3 Study Design and Categorization

The included publications were categorized based on their focus and study design. Four studies explored phishing emotions, two studies investigated personality aspects, and one study examined both emotions and personality. Additionally, three studies combined machine learning (ML) techniques with emotions, while three studies combined ML with personality. One study was solely dedicated to phishing

classification using ML. A comprehensive summary of the results, categorized by their respective topics, is presented in Table 1.

Author, year	Participants	Emails	ML algorithms	ML Techniques	ML Packages	Emotions	Personality
[1], 2016			✖			✖	
[2], 2018			✖	✖			✖
[3], 2019	✖						✖
[4], 2019				✖	✖	✖	
[5], 2019			✖	✖			✖
[6], 2020		✖				✖	
[7], 2021			✖	✖	✖		
[8], 2021	✖					✖	
[9], 2021	✖					✖	
[10], 2021	✖						✖
[11], 2021	✖					✖	✖
[12], 2022			✖	✖	✖	✖	
[13], 2022	✖					✖	
[14], 2022			✖				✖

Table 1 – Summary of analyzed articles topics

i. Studies exploring phishing emotions

One study conducted by researchers [9] focused on exploring the influence of emotional congruence and regulatory fit on phishing susceptibility. The study included 1000 participants from Amazon Mechanical Turk and examined how individuals' emotional states and motivational framing in phishing messages can affect their susceptibility to phishing attacks. The study also investigated the interplay between motivational framing and individuals' motivation orientation by drawing on the regulatory fit theory. The participants were asked to describe their hopes, aspirations, duties, and obligations, and their incidental emotional state was elicited using the International Affective Picture Set (IAPS) database. The findings of the study provided insights into how phishing exploits individuals' emotions, laying the groundwork for more effective strategies to mitigate such threats.

Another study conducted by [8] explored the effects of anxiety, stress, and fear caused by the COVID-19 pandemic on phishing attacks. The research aimed to understand the relationship between users' general risk-taking preferences, demographics, and the success of COVID-19 and common phishing attacks during the pandemic. The study recruited 198 participants, including regular internet users from online communities and Amazon Mechanical Turk. Participants were assessed for their phishing detection abilities through a roleplay task, where they were presented with various emails and asked to identify phishing, common phishing, or legitimate emails. The study findings indicated that fear of COVID-19 influenced the success of COVID-19 specific themed phishing scams, while anxiety, stress, and risk-taking influenced the success of both COVID-19 themed and common phishing scams. The participants' education level was also found to impact common phishing attacks during the pandemic.

In a different study conducted by [6], a content analysis of phishing emails was performed to understand the human factors and psychology behind phishing attacks. The study collected a total of 217 phishing emails from Berkeley Information Security Office and SecureIT-Kent State University between 2015 and 2019. The researchers aimed to identify the most frequent types of phishing emails and the emotional pleas used in these patterns. Using knowledge-based techniques and domain knowledge, the study analyzed the subject, contents, date, time, and compelling words in the phishing emails. The results revealed that fear, anticipation, and trust were the most frequent emotional triggers used in the dataset, and the emails often used subjects related to users' online accounts to capture their attention.

Additionally, a study conducted by [13] focused on developing a phishing awareness training service through voice-based interaction with an intelligent voice assistant (Amazon Alexa). The researchers aimed to explore how the concept of phishing materializes within an intelligent voice assistant environment and how users' trust in these devices affects their susceptibility to phishing attacks. The study included 120 participants who were exposed to three tests: no training, facts-and-advice training, and interaction-based training. The participants assessed a vignette of 28 emails and determined whether each email was phishing or not. The study findings indicated that participants who received interaction-based training on the principles of influence through Alexa performed significantly better in detecting phishing emails. The training had a positive impact on participants' understanding of phishing risks and their resistance to potential phishing attacks.

ii. Studies exploring phishing personality

A study conducted in [10] aimed to investigate the role of distinct message influences, cognitive processing, and their interaction in phishing susceptibility. The researchers surveyed 273 individuals to understand how message influences can lead to phishing susceptibility and how cognitive processing affects susceptibility. The study proposed a theoretical conceptualization of message influence and explored the stimulus-response (S-R) and stimulus-interpretation-response (S-I-R) logic in the context of

phishing. The research findings suggested that phishing messages can trigger both mindless responses and mindful interpretations simultaneously, indicating that the process between influence and consequence is not simply an alternative route.

In another study [3], the researchers designed a game-based solution to prevent spear-phishing attacks and educate individuals on identifying phishing emails. The study involved 63 participants from the School of Software, Tsinghua University, Beijing, China. The objective was to create a game that would enhance understanding of the spear-phishing process and improve users' ability to identify signs of phishing emails. The researchers performed an in-depth literature review on social engineering, phishing, game design, learning functions, human interaction, and game-based learning. Based on the findings, a board game was designed and evaluated. The results showed that the game effectively improved students' awareness of spear-phishing risks and their resistance to potential first attack contacts. The game-based learning approach proved to be an effective method for increasing cyber-related awareness among players.

iii. Studies combining Phishing Emotions and Phishing Personality

In the study presented in [15], a sample of American college students was used to investigate the factors that contribute to users' failure in recognizing red flags and falling victim to cyber events. The objective was to predict such behaviors by examining pre-experimental trust and cyber-attack knowledge. Participants were surveyed to assess their propensity to trust in technology and were then presented with six different scenarios. After reading each scenario, participants answered questions related to situational trust, emotional valence, arousal, and knowledge of attack type. This process continued until all scenarios were presented, and demographic information was collected. The findings presented in [15] explored the relationship between positive/negative emotions, emotional arousal, situational trust, and users' knowledge level. The study revealed that users with higher knowledge levels tended to have a greater propensity to trust technology, exhibited more sensitive emotional responses to risk manipulation, and displayed lower situational trust when faced with cyber-attack scenarios.

iv. Studies using only Machine Learning for Phishing Detection

In [16], a study was conducted using 3685 phishing emails and 4894 legitimate emails to investigate the effectiveness of utilizing a graph convolutional network (GCN) and natural language processing (NLP) for phishing detection in email body text. The main goal of the work presented in [16] was to improve the accuracy and speed of phishing detection. The study involved collecting and preparing data from phishing and legitimate emails for training and testing. A detection model based on deep learning GCN algorithms was constructed using the training data. The model was then tested using the testing data to validate its performance. The results presented in [16] demonstrate that the proposed classifier, which incorporated NLP, deep learning algorithms, and GCN, effectively detected phishing emails in the body text. The

classifier achieved a high accuracy rate of 98.2% and a low false-positive rate of 0.015, outperforming existing detection methods. The study concluded that the supervised learning approach, employing machine learning techniques, significantly improved the accuracy and performance of phishing detection.

v. Studies combining Phishing Detection with Machine Learning and Phishing Emotions

The research presented by authors in [1] aimed to classify a dataset of 1412 emails based on the sentiments expressed in them, utilizing data mining approaches such as k-means clustering, fuzzy c-means clustering, and backpropagation neural network. In this survey described in [1] authors collected raw emails and performed preprocessing steps, including the removal of stop words. Effective features were extracted from the emails to identify hidden sentiments. Normalized feature vectors were created to establish a database, and these vectors were used to generate feature vectors based on word counts in different categories. The emails were then classified as positive, negative, or neutral using various algorithms. Authors employed K-Means Clustering, Fuzzy C-Means Clustering, and Backpropagation Neural Network algorithms for classification. The results indicated that the backpropagation neural network algorithm performed efficiently in classifying all three categories (positive, negative, and neutral), while the clustering algorithms showed lower performance due to difficulties in identifying negative emails. The study concluded that supervised training approaches, such as backpropagation neural networks, were more effective for classifying emails in this domain. Additionally, another study presented in [4] used 102 emails from the Enron Database to explore text analytics approaches in distinguishing phishing emails from other types of emails. The study employed sentiment analysis using NLTK (Natural Language Toolkit) to determine the sentiment of phishing emails. It was found that positive sentiment significantly differed from negative sentiment in phishing messages. The study also developed models based on friends, work, money, and achievement, which are factors associated with phishing, to further identify phishing messages through statistical analysis. Furthermore, models were built to differentiate between Enron emails and phishing emails based on LIWC (Linguistic Inquiry and Word Count) variables. The findings of [4] demonstrated that text analysis, including sentiment analysis and linguistic analysis, can contribute to the field of cybersecurity and serve as a basis for future research analyzing phishing emails.

vi. Studies combining Phishing Emotions and Phishing Personality

The research presented in [17] involved analyzing approximately 450 suspected phishing emails to examine psychological aspects in phishing emails using natural language processing techniques. The study aimed to detect and understand these aspects. Initially, the researchers defined psychological nomenclature and expanded the vocabulary using WordNet. They then created data augmentation for emphatic training based on psychology vocabulary, categorized and tested the emails. The study explored

two main domains of personality: Principles of Influence (e.g., authority, commitment, liking, reciprocity) and Life Domains (e.g., financial, health, legal, security, social). These domains were considered important due to their influence on human behavior and interactions. The research employed an Ensemble One-vs-All Learning (EOAL) Classification, utilizing various machine learning algorithms such as Naïve Bayesian, SVM, Random Forest, Logistic Regression, Perceptron, and Deep Neural Network. The study found that the Multilayer Perceptron achieved the best accuracy of 96.67% among the models tested. It was concluded that supervised machine learning approaches, combined with data augmentation and negative sampling, were effective in classifying phishing emails based on psychological traits. The study presented in [5] aimed to understand the correlation between user behaviors and phishing victimization. The research collected experimental data and developed a model to predict the likelihood of a user becoming a victim based on their profile and behaviors. The study involved online and on-site studies, with experiments conducted in a lab environment and online. The authors analyzed user behavior factors, such as intervention, phishing types, and monetary incentives, to investigate user responses to phishing email attacks and potential prevention mechanisms. Four machine learning models were built: Decision Tree-J48, Naive Bayes (NB), Support Vector Machine (SVM), and Multilayer Perceptron (MP), with training and testing performed using 10-fold cross-validation. The results indicated that participants with intervention and monetary incentives performed better, while fatigue influenced performance negatively in later rounds. The Multilayer Perceptron model achieved the highest accuracy of 96.67%, but attribute selection or reduction was required due to the small dataset. Lastly, in [14] authors utilized 1295 phishing emails and 9337 legitimate emails to quantify psychological cues for phishing email detection. The study employed BERT, Sentence-BERT (SBERT), and Char-CNN networks to model the psychological traits and calculate SoftMax probability scores (PPT scores) for each email. The addition of PPT scores improved the detection model performance significantly. The study identified three dominant psychological traits in phishing emails: a sense of urgency, inducing fear through threats, and enticement with desire. The analysis revealed that fear provided the strongest cue for detecting phishing emails. The distribution of PPT scores demonstrated distinct patterns between phishing and legitimate emails, indicating the effectiveness of PPT scores in capturing psychological nuances.

2.3. Systematic Literature Review – Discussion of the results

This section delves into the discussion of the Best performing Algorithms, the Emotions and the Personality.

2.3.1. Best performing Algorithms

Among the six articles [16], [1], [12], [5], [14], [17] that utilized Machine Learning, five of them [16], [1], [12], [5], [17] employed algorithms, while the remaining article [14] combined BERT

and SBERT with an algorithm. When considering the classification of spam emails, supervised algorithms demonstrated the highest accuracy rates. The Recurrent Neural Network (RNN) Algorithm used in [12] achieved the best performance with an accuracy rate of 99%. The Deep-Learning Graph Convolutional Network (GCN) in [16] achieved the second-highest accuracy rate of 98.2%. The Backpropagation Neural Network used in [1], the Deep Neural Network presented in [17], and the Multilayer Perceptron described in [5] achieved accuracy rates of 97.91%, 97.8%, and 95.5% (survey data) and 96.67% (real data), respectively. The combination of BERT and SBERT with an algorithm presented in [14] resulted in an accuracy rate of 97.47%.

2.3.2. Emotions

Although the studies revealed the significance of emotions in phishing attacks, the exact impact of emotions on responses remains unclear due to variations in the emotions studied and the lack of consistent measurable methods across all studies, except the one presented in [6]. [6] concludes that fear, anticipation, and trust were the most commonly occurring emotional triggers, with rates of 29.17%, 23.61%, and 20.83%, respectively. [15] demonstrated that participants experienced more negative emotions in high-risk scenarios, and manipulation of risk levels influenced arousal and emotional valence. Incorporating Affect Intensities, a lexicon measuring sentiment based on emotions, as shown in study [12], improved performance compared to using the algorithm without the emotion feature. Additionally, authors in [13] found a correlation between participants' ability to identify phishing emails and persuasion principles such as authority, commitment, liking, scarcity, and their combinations. These findings suggest that incorporating such principles into phishing training programs can enhance individuals' ability to detect and avoid phishing emails.

2.3.3. Personality

Although the reviewed studies lacked sufficient information for conclusive results, they indicated that personality traits can improve the effectiveness of detecting phishing emails. However, due to variations in the personality traits examined, it was challenging to establish a clear correlation between effectiveness and personality ([10], [3], [5], [15], [14], [17]). The results presented in [17] revealed that phishing emails frequently utilized principles of influence, with reciprocation being the most prevalent at 73.62%. Liking, authority, scarcity, commitment, and consistency, social proof, and perceptual contrast were also utilized to varying degrees. Authors in [3] focused on using psychology and compliance principles to create more effective phishing emails, while in [5] authors found that participants who received intervention and a monetary incentive performed better. The study presented in [10] demonstrated that phishing susceptibility can be influenced by attractive and coercive tactics, while systematic cognitive processing is negatively correlated with susceptibility. The resulted presented in [15] indicated that arousal, emotional valence, and situational trust were not highly correlated, and emotional valence was a more significant

predictor of situational trust than arousal. Furthermore, in [14] the results highlighted that measuring the psychological traits of an email can enhance the effectiveness of phishing email detection. Future research should focus on employing Machine Learning techniques that incorporate emotional content to detect phishing emails. The existing literature on this topic demonstrates considerable heterogeneity, highlighting the need for meaningful comparisons and addressing the research gap regarding the correlation between personality traits, emotions in phishing emails, and Machine Learning.

2.4. Cyberattacks / Social engineering attacks

This section aims to present the primary forms of social engineering attacks, which include Smishing and Vishing, Scareware, Tailgating/Piggybacking, Watering Hole, Business Email Compromise (BEC), various types of Phishing (such as Angler Phishing and Spear Phishing), Whaling, Baiting, Honey Trap, Pretexting, and Quid Pro Quo.

2.4.1. Smishing and vishing

SMiShing attacks are a form of cyber-attack that occurs through short messaging services (SMS) or text messages. The term "Vishing" combines "voice" and "Phishing" and involves receiving a phone call where the attacker tricks the recipient into providing personal information for malicious purposes. For example, a customer might receive a call outside of banking hours, claiming to be from their bank contact center. The information obtained by the attacker can be used for criminal activities like identity theft or fraud. Typically, victims are deceived into revealing this information either by submitting it through a web form or by unknowingly downloading and installing malicious software.[18]

2.4.2. Scareware

Scareware refers to a type of malicious software specifically created to deceive users into believing that their computer system is infected with malware. It uses fear tactics to pressure the user into purchasing and downloading additional antivirus applications that are actually fake and harmful, such as malware, spyware, or adware. Typically, this deceptive software displays frequent warnings about supposed infections and demands payment for licensing to remove them.[19]

2.4.3. Tailgating/Piggybacking

Access control refers to a system that regulates entry to a secure area, such as a bank or an office, by verifying the credentials provided by individuals. While traditional access control systems use physical

keys, electronic access control can involve various types of credentials like passcodes or fingerprints. When authorized credentials are provided, the door opens to grant access, and when access is denied, the door remains closed. The system keeps a record of entry and exit in a database.[20]

However, many current accesses control systems face a challenge in controlling the number of people entering a restricted area. This gives rise to security vulnerabilities known as "tailgating" and "piggybacking." Tailgating occurs when an unauthorized person follows someone through the door without their knowledge, while piggybacking happens when the authorized person intentionally allows someone to enter with them through the secure door. Both tailgating and piggybacking pose significant security risks.[20]

2.4.4. Watering Hole

Watering Hole attacks are a sophisticated variant of drive-by download attacks, in which a malicious actor implants malware onto websites frequented by a particular individual or a specific group. The intention is to target and exploit the user or group when they visit one of these trusted websites. The attacker demonstrates persistence by infecting multiple websites that are known to be regularly visited by the targeted individual or group. [19]

2.4.5. Business Email Compromise (BEC)

Business Email Compromise (BEC) is a serious cyber threat that targets organizations and their partners, regardless of their size. With the rapid advancement of technology and the increasing reliance on it for transactions, BEC has become a significant risk for companies involved in financial activities. BEC involves the use of email phishing techniques to deceive individuals and gain unauthorized access to financial resources. This form of attack has seen a substantial rise, resulting in significant financial losses for companies, particularly during the period of remote work and the COVID-19 crisis. BEC attacks have increased by 94% in the third quarter of this year. Unlike other cyber threats, BEC doesn't necessarily require advanced technical skills. Instead, it relies on social engineering tactics to manipulate individuals. In this context, it is crucial to understand the nature of BEC, its methods of occurrence, strategies for prevention or mitigation, and its impact on organizations.[21]

2.4.6. Phishing

Phishing is a deceptive technique used by scammers to obtain valuable personal information by means of fraudulent emails, texts, or imitation websites. The goal is to trick individuals into sharing sensitive data like

login credentials and Social Security numbers. Once scammers acquire this information, they can gain unauthorized access to personal devices and install malware, potentially locking users out of their own programs. Phishing attacks are not specifically targeted at individuals; rather, they cast a wide net, hoping to lure any unsuspecting victim. While some attempts may be tailored to target higher-value individuals, success is not guaranteed as it relies on chance. Phishing works effectively because it preys on emotional manipulation, using tactics like false claims of account freezes or threats to family members. It is crucial to conduct proper research before clicking on suspicious messages to avoid falling victim to phishing attacks, which can expose vulnerabilities in computer systems.[22]

There are two main types of phishing, namely:

- Angler phishing – Angler phishing is a form of phishing attack where attackers leverage social media platforms to deceive individuals and carry out their malicious activities. They exploit various communication channels, such as conversations, voice and video calls, public posts, and websites, to manipulate and persuade people into revealing sensitive information or downloading malware. By taking advantage of the trust and familiarity associated with social media interactions, the attackers aim to trick individuals into compromising their personal data or unwittingly installing harmful software on their devices. [4]
- Spear phishing – Spear phishing is a targeted form of phishing that focuses on a specific individual, typically someone with a high public profile or in a position of authority, such as upper management or executives. Unlike regular phishing attacks, spear phishing involves sending an email that appears to come from a trusted source but actually leads the recipient to a fake website filled with malware. The objective of spear phishing is to steal sensitive data or install malicious software. The attackers invest significant time and effort in conducting extensive research on the targeted organization or individuals life. They gather information about people in their circle, including executives, employees, family members, and hobbies. This personalized approach makes the email appear authentic and familiar, resembling messages previously received from friends, coworkers, or employers. Victims are more likely to click on such emails and willingly provide information promptly because the content often appears urgent or important. Spear phishing attacks are not limited to cybercriminals; government-sponsored hackers and hacktivists may also engage in such activities, aiming to resell confidential data to governments or private companies. These highly personalized and tailored messages make spear phishing attacks particularly effective.[22]

2.4.7. Whaling

Whaling refers to a type of cyber-attack that is an extension of spear phishing, but with a specific target: senior executives and other high-profile individuals. The attackers carefully craft text messages or emails tailored to the targeted individual, usually someone in an upper management position, with the aim of achieving a specific goal within the company. Whaling attacks often utilize content that is relevant to executives, such as fake customer complaints, to lure them into taking action. These attacks exploit the trust and authority associated with high-ranking positions to deceive and manipulate individuals for malicious purposes.[23]

2.4.8. Baiting

Baiting phishing is a type of cyber-attack that exploits the victim curiosity by using enticing baits to lure them into compromising their confidential information. These baits can be physical items like USB drives or CDs with company logos. The attacker strategically places these baits in common areas such as parking lots or bathrooms, hoping that the victim will pick them up and insert them into their work or home computer. Once the bait is inserted, the device gets automatically infected with malware. Baiting phishing takes advantage of people desire for free things, such as downloading popular songs, movies, or apps. Scammers use baiting techniques to offer these free goods in exchange for personal information. When victims provide the requested sensitive data, hackers download viruses onto their computers. Illegal torrents, which bypass copyright laws and are often free, are commonly used as bait and are notorious for containing malware and viruses.[24]

2.4.9. Honey Trap

Honeytraps refer to systems, specifically Honeypots or Honeynets, that are intentionally designed to be compromised by intruders. Honeypots are individual host systems created to attract intruders by mimicking known vulnerabilities. They serve as modified production systems that create controlled environments for monitoring and documenting intruder actions. Honeypots do not contain real or valuable data; their primary purpose is to capture and analyze data to gain insights into the activities of the Blackhat community. On the other hand, Honeynets are a network of interconnected production and honeypot nodes. They divert the attention of intruders away from the actual target, protecting the production resources. Similar to Honeypots, Honeynets are designed to gather information from intruders and potential intruders while restricting the operations they can perform. [25]

2.4.10. Pretexting

Pretexting is a social engineering technique used by attackers to deceive and manipulate individuals by creating a convincing story or excuse. The purpose is to gain the trust of the victim and persuade them to provide valuable information or grant access to specific systems or goals. This type of attack relies on the manipulation of trust and authority, where the attacker pretends to be an authorized person with legitimate access to the information of the target. Pretexting attacks typically involve extensive research and preparation before initiating contact with the victim. The attacker gathers information to create a tailored pretext that enhances the credibility of their deception. The more information the attacker possesses, the higher the chances of successfully persuading the victim to disclose sensitive information. Pretexting attacks can be carried out through various means, including email, phone calls, or physical media.[26]

2.4.11. Quid Pro Quo (i.e., tech support scams)

Quid Pro Quo phishing is a type of cyber-attack where scammers promise a service instead of a tangible good, unlike baiting phishing. The most typical Quid Pro Quo scheme involves the scammer posing as an IT consultant or customer support representative who calls back an employee. Eventually, they target an employee who genuinely needs assistance. The scammer offers a "fix" for the problem but requires the employee to disclose sensitive information such as passwords and other credentials. In less sophisticated Quid Pro Quo plots, scammers may use tactics like workplace contests or surveys to entice individuals into revealing sensitive information. The key element of Quid Pro Quo phishing is the promise of a service in exchange for the confidential data of the victim.[24]

2.5. Artificial Intelligence

This section is divided into five subsections: Machine Learning Algorithms, which provides a summary of the classification of existing algorithms and introduces some key ones; it also introduces two crucial techniques for this study, namely Natural Language Processing (NLP) and Named Entity Recognition (NER), along with AI Systems like ChatGPT and IBM Watson.

2.5.1. Machine Learning Algorithms

This subsection is organized into five parts. The initial part, titled "Supervised Learning," provides an introduction to Linear Regression and Classification. The second part, "Unsupervised Learning," covers Clustering and Dimensional Reduction. The third part delves into algorithms that can fall into multiple categories, such as Supervised Learning, Unsupervised Learning, and Reinforcement Learning. These algorithms include Support Vector Machines, K-Nearest Neighbor, Decision Trees, and Random Forest. The fourth part, "Deep Learning and Neural Networks," delves into topics like Neurons and Hidden Layers, Convolutional Neural Networks, and Recurrent Networks. Lastly, the fifth part introduces Reinforcement Learning. Figure 2 illustrates a diagram depicting various algorithms and their classifications.

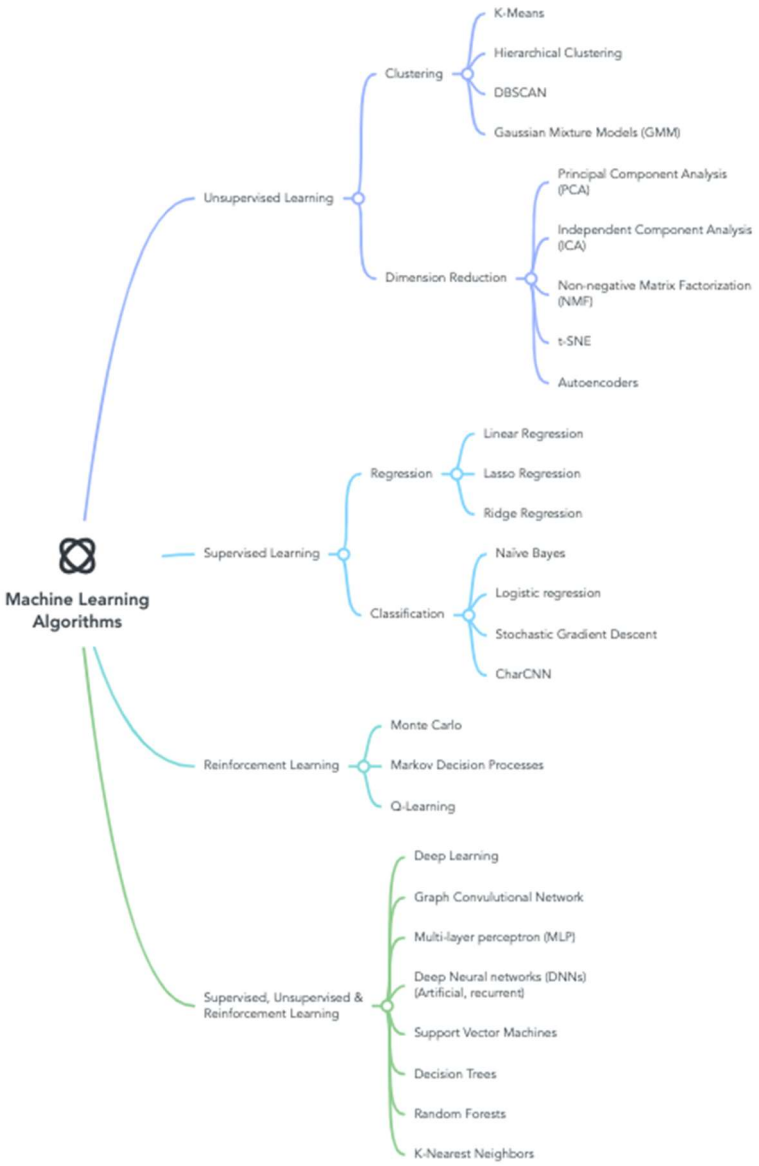


Figure 2 Machine Learning Algorithms

i. Supervised learning

This section discusses Linear Regression and Classification algorithms.

A. Linear regression

Regression is a supervised learning approach used to model continuous variables and make predictions. Linear regression, the simplest form of regression, fits a straight line or hyperplane to a dataset when a linear relationship exists between the variables. It has advantages such as simplicity, interpretability, and the ability to avoid overfitting through regularization. However, it oversimplifies real-world problems, struggles with non-linear relationships and complex patterns, and may not accurately capture the true relationship between variables. Despite these limitations, linear regression can be valuable for learning about data analysis processes and is applicable in domains such as real estate pricing, sales forecasting, and stock price movement prediction.[27]

B. Classification

Classification, as a data mining and machine learning approach, is utilized to predict the group membership of data instances. While there are several techniques available for machine learning, classification stands out as the most commonly employed method. It is highly regarded and considered an important task in machine learning, particularly for future planning and knowledge discovery.[28]

ML classification is widely used in various domains, including image recognition, speech recognition, natural language processing, fraud detection, and sentiment analysis. The accuracy of a classification model can be evaluated using metrics such as accuracy, precision, recall, and F1 score. Figure 3 shows a comparison between regression and classification.

Naïve Bayes is a supervised classification algorithm that uses conditional probability. It is simple to implement and performs well with less training data. It can handle both continuous and discrete data, and scales linearly with the number of predictors and data points. Naïve Bayes is robust against irrelevant features and provides probabilistic predictions. However, it may be outperformed by more complex models, struggles with continuous variables, and requires retraining the model with all data. It may not

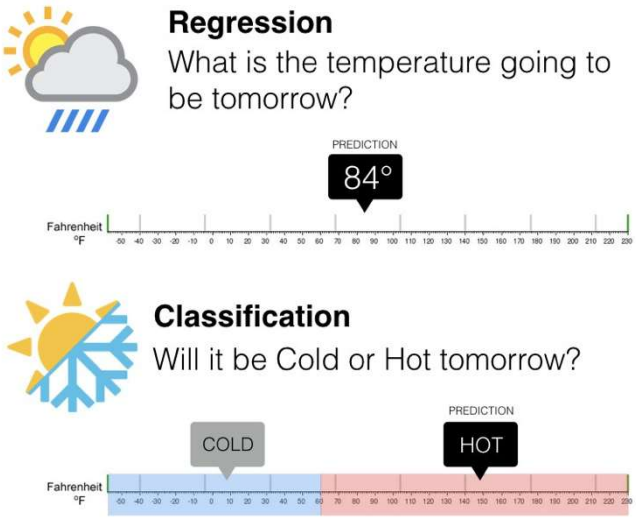


Figure 3 Regression vs Classification [49]

scale well with a high number of classes and consumes more memory compared to other algorithms. Despite these limitations, Naïve Bayes finds applications in Recommendation Systems and cancer progression forecasting after Radiotherapy.[27]

Logistic regression is another classification algorithm that predicts the binomial outcome, such as the probability of an event occurring (0 or 1), based on input variable values. It can also handle multinomial and ordinal outcomes. In contrast, linear regression deals with predicting continuous variable values, like real estate prices. Logistic regression offers advantages such as simplicity, computational efficiency, ease of regularization, and no scaling requirement for input features. It is commonly used for solving large-scale industry problems. However, logistic regression has limitations, including its inability to solve non-linear problems, potential for overfitting, and the need for identifying all independent variables. Practical applications of logistic regression include disease risk prediction, cancer diagnosis, mortality prediction for injured patients, and engineering applications for failure probability prediction. [27]

ii. Unsupervised learning

This section explores Clustering and Dimensional Reduction.

A. Clustering

Clustering algorithms, as a category of unsupervised machine learning (ML) algorithms, are widely present in contemporary data science and serve as a crucial component in various learning-based application pipelines.[29]

The K-Means Clustering Algorithm, a widely known and common type of clustering, is frequently used for solving clustering problems in unsupervised learning. It offers several advantages, such as computational efficiency for large variable sets and the ability to produce tight clusters with globular shapes and small K values. It is easy to implement and interpret the clustering results, with a computational complexity of O(K*n*d), making it efficient. However, there are disadvantages to consider. Predicting the optimal K value can be challenging, performance can suffer with globular clusters, and different initial partitions may lead to different final clusters. Performance also degrades when there are variations in the size and density of clusters in the input data. The assumption of uniform effects can result in clusters with similar sizes, even if the input data have different cluster sizes. Spherical assumption may not hold if there are correlations between features. K-Means is sensitive to outliers, initial points, and local optima, and there is no unique solution for a given K value, requiring multiple runs to find the best results. K-Means Clustering finds applications in various domains, including document classification, customer segmentation, rideshare data analysis, automatic clustering of IT alerts, call record details analysis, and insurance fraud detection.[27]

B. Dimensional reduction

In various domains such as climatology, biology, astronomy, medicine, economy, and finance, the increasing volume of data poses significant challenges for existing machine learning systems, which struggle to process such massive datasets efficiently and effectively. The curse of high dimensionality and data sparsity further compounds the problem, making it costly to find the global optimum. To address these issues, dimensionality reduction algorithms have emerged as a solution to reduce computational costs by reducing the number of dimensions in the data. By eliminating redundant and irrelevant information, these algorithms enhance result accuracy. Operating in an unsupervised manner, dimensionality reduction techniques leverage implicit data structures to uncover patterns. Several algorithms, including Multidimensional Scaling (MDS), Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Principal Component Regression (PCR), and Linear Discriminant Analysis (LDA), can be adapted and integrated with classification and regression algorithms to improve their performance.[30]

iii. Supervised learning, Unsupervised learning and reinforcement learning

This section explores Support Vector Machines, K-Nearest Neighbor, Decision Trees, and Random Forest Algorithms.

A. Support Vector Machines

Support Vector Machines (SVM) offer a versatile solution for both classification and regression tasks. This method involves defining a hyperplane as the decision boundary to separate objects belonging to different classes. In cases where linear separation is not possible, SVM utilizes kernel functions to separate objects with complex mathematical functions. The goal of SVM is to accurately classify objects based on training examples. The advantages of SVM include its ability to handle both structured and semi-structured data, its capacity to handle complex functions through appropriate kernel selection, its generalization ability that reduces the risk of overfitting, scalability with high-dimensional data, and avoidance of local optima. However, there are also disadvantages to consider. The performance of the SVM may decrease with large datasets due to increased training time, finding an appropriate kernel function can be challenging, it is sensitive to noisy datasets, and it does not provide probability estimates. Interpreting the final SVM model can also be difficult. SVM finds practical applications in various domains, including cancer diagnosis, credit card fraud detection, handwriting recognition, face detection, and text classification. When considering logistic regression, decision trees (Random Forests), and SVM, it is recommended to start with logistic regression, followed by decision trees, and then explore SVM when dealing with a high number of observations and features. [27]

B. K-nearest Neighbor

Support Vector Machines (SVM) offer a versatile solution for both classification and regression tasks. This method involves defining a hyperplane as the decision boundary to separate objects belonging to different classes. In cases where linear separation is not possible, SVM utilizes kernel functions to separate objects with complex mathematical functions. The goal of SVM is to accurately classify objects based on training examples. The advantages of SVM include its ability to handle both structured and semi-structured data, its capacity to handle complex functions through appropriate kernel selection, its generalization ability that reduces the risk of overfitting, scalability with high-dimensional data, and avoidance of local optima. However, there are also disadvantages to consider. The performance of the SVM may decrease with large datasets due to increased training time, finding an appropriate kernel function can be challenging, it is sensitive to noisy datasets, and it does not provide probability estimates. Interpreting the final SVM model can also be difficult. SVM finds practical applications in various domains, including cancer diagnosis, credit card fraud detection, handwriting recognition, face detection, and text classification. When considering logistic regression, decision trees (Random Forests), and SVM, it is recommended to start with logistic regression, followed by decision trees, and then explore SVM when dealing with a high number of observations and features.[27]

C. Decision Trees

The Decision Tree is a supervised machine learning approach used to solve classification and regression problems by iteratively splitting data based on specific parameters. The decisions are represented in the leaves, and the data is partitioned at the nodes. In a Classification Tree, the decision variable is categorical (e.g., Yes/No outcome), while in a Regression Tree, the decision variable is continuous. Decision Trees offer several advantages, such as suitability for both regression and classification problems, ease of interpretation, handling of categorical and quantitative values, ability to handle missing attribute values by assigning the most probable value, and high performance due to efficient tree traversal algorithms. However, Decision Trees may encounter challenges such as instability, difficulty controlling the size of the tree, susceptibility to sampling errors, and providing locally optimal solutions rather than globally optimal solutions. To address overfitting, an ensemble modeling approach like Random Forest can be used. Decision Trees find applications in various domains, including predicting future library book usage and tumor prognosis problems.[27]

D. Random Forest

Random Forest is an ensemble learning method used for classification and regression. It creates multiple decision trees using random subsets of data and combines their outputs to make predictions. The

algorithm consists of two stages: building the random forest by generating decision trees and using the random forest to make predictions.[30]

Random Forest is a supervised learning algorithm known for its flexibility, ease of use, and absence of hyperparameters. It is highly effective in classification tasks. The algorithm requires a minimum number of trees to classify all data, with the exact number depending on the dataset. The number of predictor attributes affects the minimum number of trees, and increasing the number of trees generally improves accuracy. However, there is a limit to the accuracy improvement, after which the accuracy stabilizes. The accuracy of Random Forest is also influenced by the number of predictor attributes, with using all available attributes resulting in lower accuracy.[31]

iv. Deep Learning and neural Networks

In this section, we will delve into the concepts of Neurons and Hidden Layers, explore Convolutional Neural Networks, and understand Recurrent Neural Networks.

A. Neurons and Hidden Layers

Neural networks consist of interconnected processing units called neurons, which work together to perform tasks like pattern recognition. These neurons receive inputs, apply weights to them, and generate outputs through activation functions. The structure of a neural network often includes hidden layers, which are layers of neurons between the input and output layers. Determining the number of hidden layers and neurons in each layer is crucial for designing an effective network. While problems requiring two hidden layers are rare, neural networks with one hidden layer are commonly used and can handle a wide range of practical tasks.[32]

B. Convolutional Neural Networks

A Convolutional Neural Network (CNN) is a powerful and widely used technology in computer vision and machine learning. It has demonstrated exceptional performance in various tasks. It is self-contained, covering the necessary mathematical background and explaining the derivations in detail. CNNs are particularly useful in image-related applications such as image classification, semantic segmentation, and object detection. This note specifically focuses on image classification, where images are categorized based on the identity of their main objects, such as dogs, airplanes, birds, etc.

C. Recurrent neural networks

Recurrent Neural Networks (RNNs) are a specific type of neural network architecture primarily used for detecting patterns in sequential data. This data can include handwriting, genomes, text, or numerical time series typically generated in industrial settings like stock markets or sensors. RNNs can also be applied to images by decomposing them into patches and treating them as a sequence. RNNs have various applications, including Language Modeling & Text Generation, Speech Recognition, Image Description Generation, and Video Tagging.[33]

The key distinction between Recurrent Neural Networks and Feedforward Neural Networks, also known as Multi-Layer Perceptrons (MLPs), lies in how information flows through the network. While Feedforward Networks transmit information without any loops, RNNs have cyclic connections that allow information to be passed back into itself.[33]

v. Reinforcement learning

Reinforcement learning (RL) is a branch of machine learning that deals with a class of learning problems and the algorithms used to solve them. In RL, an agent exists in a set of states and takes actions to transition between these states. The agent receives rewards or punishments based on its actions and aims to optimize a specific objective: the expected sum of future rewards, considering the discounting factor.[34]

To achieve this objective, RL offers a diverse range of algorithms, including model-free RL. Model-free RL involves estimating the values of actions by predicting the discrepancies between expected and actual rewards. These algorithms enable RL agents to learn and make decisions that maximize their long-term cumulative rewards in various domains.[34]

2.5.2. Natural Language Processing (NLP)

Natural Language Processing (NLP) is a field of study that involves the computerized analysis of text using a combination of theories and technologies. It aims to achieve human-like language processing for a wide range of applications. NLP encompasses computational techniques used to analyze and represent naturally occurring texts, regardless of language or mode. It focuses on understanding the different levels of linguistic analysis and utilizing them to process language in a way that resembles human comprehension.[35]

The goal of NLP is to accomplish human-like language processing, which involves tasks such as paraphrasing, translation, answering questions, and drawing inferences from text. While NLP has made

progress in tasks like paraphrasing and translation, true Natural Language Understanding (NLU) is yet to be fully achieved. Practical goals for NLP include improving information retrieval systems, machine translation, question-answering, and other specific applications where language comprehension is crucial.[35]

The origins of NLP can be traced back to multiple disciplines, including linguistics, computer science, and cognitive psychology. Linguistics focuses on formal models of language, computer science deals with data representation and processing, while cognitive psychology explores language as a reflection of human cognitive processes. These different backgrounds have influenced the development and practices of NLP.[35]

Within NLP, there are two distinct focuses: language processing and language generation. Language processing involves analyzing text to produce meaningful representations, while language generation involves producing language based on a given representation. NLP primarily focuses on language analysis, but natural language generation requires planning capabilities to generate language in specific contexts. Speech understanding, a related field, deals with the analysis of oral language and incorporates acoustics and phonology.[35]

Overall, NLP plays a significant role in achieving effective language processing, enabling tasks like information retrieval and machine translation, with the ultimate goal of achieving true natural language understanding.[35]

2.5.3. Named Entity Recognition (NER)

Named Entity Recognition (NER) is a fundamental task in natural language processing that involves identifying and categorizing specific words or phrases in text known as named entities. These entities can belong to various predefined semantic types, such as person, location, organization, gene, protein, drug, or disease names. NER is crucial for applications like information extraction, text understanding, question answering, and knowledge base construction.[36]

NER has evolved over time, starting with its introduction at the Message Understanding Conference (MUC-6). It initially focused on identifying names of organizations, people, geographic locations, and expressions like currency and time. Researchers have reached a consensus on the types of named entities to recognize, including generic categories like person and location, as well as domain-specific entities like proteins and genes.[36]

Different techniques have been employed in NER, including rule-based approaches, unsupervised learning approaches, feature-based supervised learning approaches, and deep learning-based approaches. Rule-

based approaches use hand-crafted rules, unsupervised learning approaches rely on unsupervised algorithms, and feature-based supervised learning approaches involve careful feature engineering. Deep learning-based approaches leverage neural networks to automatically extract features and make predictions. These approaches have gained significant attention and have shown promising results in NER tasks.[36]

2.5.4. AI tools in the scope

AI tools have recently gained popularity like never before. In this subsection, a small introduction is made about two tools/technologies that enhance the use of AI to assist humans in decision making.

a) ChatGPT

ChatGPT, developed by OpenAI, is an advanced NLP system that generates human-like conversations by understanding the context and generating appropriate responses. It utilizes the GPT-3 deep learning model, trained on a large dataset of conversations. ChatGPT has features such as multilingual support and the ability to generate responses in different styles, making it a powerful NLP system.[37]

The benefits of ChatGPT include increased efficiency and improved accuracy. Businesses can automate conversations, saving time and resources. ChatGPT quickly and accurately understands customer queries, providing personalized experiences. Its advanced NLP technology has helped businesses enhance customer service and efficiency, focusing on growth.[37]

However, ChatGPT also presents challenges. Security concerns arise from the potential for adversarial attacks and misuse for spreading misinformation. Additionally, ChatGPT has limitations - it can only generate responses based on provided input and lacks access to external information or internet browsing. Users should be mindful of these limitations and the possibility of biased or offensive language.[37]

b) IBM Watson

IBM Watson is an AI-based expert system that solves specific problems in specialized areas using artificial intelligence. It goes beyond simple yes/no questions and can tackle tasks like logic games, financial investing, legal research, and medical research by leveraging fuzzy logic. The DeepQA software, originally developed for the game show Jeopardy!, has been repurposed for cancer diagnosis and treatment. Watson analyzes and reasons about vast amounts of oncological information from medical journals, textbooks, and other sources to provide doctors with treatment options and recommendations. However, the success of Watson varies across jurisdictions and cancer types, with concordance rates between Watson and

doctors ranging from high percentages to lower percentages due to different treatment philosophies and training variations.[38]

While Watson has shown promise in supporting oncologists and improving cancer diagnosis, its effectiveness is not consistent. Concordance rates have been high in some cases, but lower in others, highlighting limitations in the current technology of expert systems. Issues may arise from programming errors, training variations, and the inherent unpredictability of machine data processing. The fallibility of Watson exposes it to potential legal challenges and calls for regulation. Despite its limitations, Watson remains an important tool in the field of AI and continues to evolve as technology advances.[38]

2.6. Psychological characteristics and predisposition to be a victim of phishing

This section presents the state of the art regarding personality tests applied in the field of psychology. Considering that the dissertation combines psychological aspects with phishing attacks, it is important to mention the most common tests and their potential.

2.6.1. Model of emotion of Paul Ekman

Model of emotion of Paul Ekman, often referred to as the "six basic emotions" theory, suggests that there are a number of universal emotions that are experienced by people regardless of their cultural, regional, or ethnic differences [39], [40]. These emotions are:

- Anger
- Disgust
- Fear
- Enjoyment (also referred to as happiness)
- Sadness
- Surprise

Research of Ekman shows the strongest evidence to date of a seventh emotion, which is contempt [39].

According to Ekman, emotions are a process influenced by our evolutionary and personal past, in which we sense that something important to our welfare is occurring, and a set of psychological changes and emotional behaviors begins to deal with the situation[39]. These emotional responses are an unbidden occurrence, meaning that we don't choose to feel them, they just happen to us automatically[39].

Each of these universal emotions has distinctive signals, physiologies and timelines[39]. While they vary in their onset duration and decline, emotions typically don't last longer than an hour[39]. If an emotion

persists for an extended amount of time without interruption, it is more likely that the emotion can be categorized as a mood or a disorder[39].

Later, Ekman expanded his list of basic emotions to include such things as pride, shame, embarrassment, and excitement[41].

Figure 4 exhibits actor Tim Roth portraying facial expressions and their explanation.

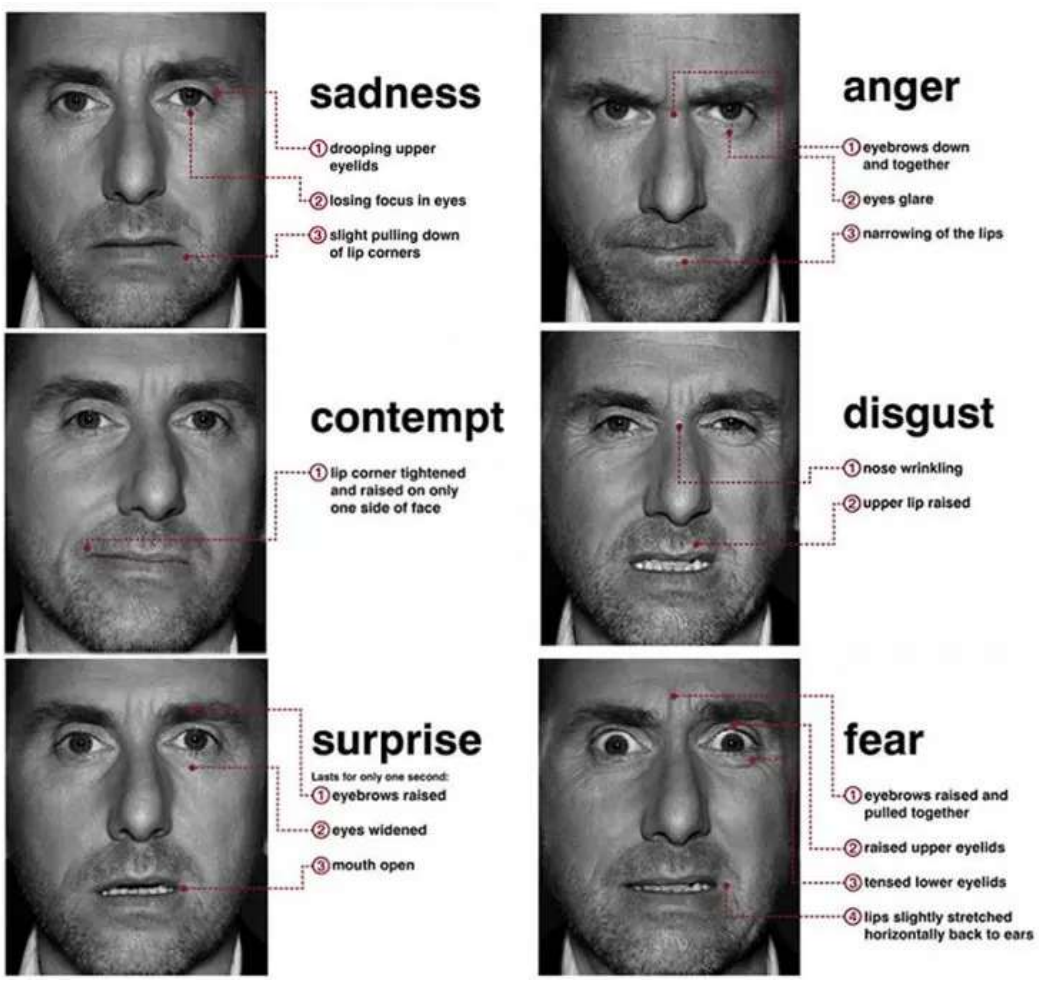


Figure 4 Actor Tim Roth portraying facial expressions and their explanation [50]

2.6.2. Personality Inventory NEO-FFI-20

The NEO-FFI-20 is a reduced version of the NEO PI-R personality inventory, consisting of 20 items[42]. This inventory assesses an individual in five personality dimensions: Neuroticism, Extroversion, Openness to Experience, Agreeableness, and Conscientiousness[42]. Each of these dimensions is assessed by 4 items: Neuroticism is assessed by items 1, 6, 11, and 16; Extroversion is assessed by items 2, 7, 12, and 17; Openness to Experience is assessed by items 3, 8, 13, and 18; Agreeableness is assessed by items 4, 9, 14, and 19; and Conscientiousness is assessed by items 5, 10, 15, and 20.[42].

In 0 is showed the questionnaire of the Personality Inventory NEO-FFI-20.

2.6.3. Melbourne Decision Making Questionnaire (Portuguese Version)

The Melbourne Decision Making Questionnaire (MDMQ) is an instrument for measuring patterns for coping with decisional conflict [43]. The Portuguese version of the MDMQ, known as "Questionário de Melbourne de Tomada de Decisão", was developed by Filipe, L., Alvarez, M.-J., Roberto, M. S., & Ferreira, J. A. in 2020 [44].

The MDMQ is used to assess decision-making abilities, which are among the most important abilities for everyday functioning. [45]

This questionnaire is composed by 22 questions, as shown on Appendix II. The Melbourne Decision Making Questionnaire (MDMQ) is used to assess four main strategies of decision-making: vigilance, buck-passing, procrastination, and hypervigilance [44]. The vigilance value es calculated by the questions 2, 4, 6, 8,12 and 16; the buck-passing is evaluated in questions number 3, 9, 11, 14, 17 and 19; the procrastination is calculated by the questions number 5, 7, 10, 18 and 21; and the hypervigilance is calculated by the questions number 1, 13, 15, 20 and 22.

This project aims to conduct a comprehensive investigation into the domain of Phishing Email Classification, leveraging insights from both Psychology and Machine Learning. Accordingly, our specific research objectives encompass:

- Investigating the participants' ability to differentiate between legitimate emails and phishing emails.
- Analyzing and comparing the emotional valence within legitimate emails and phishing emails.
- Scrutinizing the activation levels associated with legitimate emails versus phishing emails.
- Identifying and examining the emotional responses elicited by participants when engaging with legitimate emails versus phishing emails.
- Exploring the potential connections between a personality trait of the individual and their proficiency in detecting phishing emails.
- Investigating the influence of decision-making patterns on the effectiveness of phishing email detection.

By addressing these objectives, this study aims to enhance our understanding of the multifaceted factors impacting phishing email detection, thereby contributing to the development of more robust mitigation strategies.

Chapter 3 – Methodology

This chapter is divided into three main sections. The first section discusses the selected emails, including both phishing and legitimate emails, that were chosen for this research. The second section outlines the methodology of the research inquiry. The final section describes the methodology of the AI tools used in this research, including IBM Watson, Chat GPT, and Hume AI.

3.1. Selected Emails

For this research it was gather a total of 10 phishing emails and 10 non-phishing emails. These emails are from the author (Sara Rodrigues). The phishing emails were selected based on the daily institutions that users must deal with which some of them are usually used by phishers.

3.1.1. Legitimate emails

The email illustrated in Figure 5 displays the tolls that can be paid. These payments are managed by CTT, the main Portuguese Mail Company, and this figure serves as an exemplar of the emails received for unsettled payments.

Figure 5 Legitimate Email CTT

Figure 6 presents an illustrative email that showcases the process of sending an invoice from a technology reseller (Techonline). Within the email, there is a link provided, allowing for the download of the invoice.

[Techonline Portugal] Envio de documento | Fatura T-3352 | 2023-07-17

geral@techonline.pt <geral@techonline.pt>
seg, 17/07/2023 15:02
Para:<parati10@hotmail.com <<parati10@hotmail.com>

2023-07-17

Válter Costa, Unipessoal, Lda

Documentos para download

Exmo(s) Snr(s) **Daniel Oliveira .**,
Vimos por este meio enviar-lhe o documento abaixo para download.

Com os melhores cumprimentos,
Válter Costa, Unipessoal, Lda

Documentos para download

Tipo de Documento	Número	Data de Emissão	Data Vencimento	Valor	Downloads
Fatura/Recibo	T-3352	2023-07-17	2023-07-17	22.8€	Download

Válter Costa, Unipessoal, Lda
Calçada Fonte da Horta, nº 59
E-mail:

Figure 6 Legitimate Email Techonline

In Figure 7, a password reset process for "moey!" is depicted. "moey!" represents a digital banking solution in Portugal. The email includes a link to initiate the procedure for setting a new password.

Daniel esqueceste-te do teu código de acesso à app?

mymoey@moey.pt <mymoey@moey.pt>
qua, 16/08/2023 16:57
Para:parati10@hotmail.com <parati10@hotmail.com>

moey! CA Crédito Agrícola

Olá,

Parece que te esqueceste do teu código de acesso à app.

Não te preocupes. Clica no botão e segue os passos indicados
Atenção: tens de clicar no link a partir do teu telemóvel.

DEFINIR NOVO CÓDIGO

https://moey.pt

Em caso de dúvida, por favor contacta-nos através do número +351 210 521 500 (chamada para a rede fixa nacional)
Por favor não respondas a este e-mail, pois foi-te enviado automaticamente.
moey! é uma marca do Grupo Crédito Agrícola.

Figure 7 Legitimate Email "moey!"

The email presented above, featured in Figure 8, concerns a direct debit transaction linked to the indicated account. The sender of the email is Fidelidade, a Portuguese insurance company, and it encompasses information such as the policy number, IBAN, and the specified amount.


Tem novos documentos dos seus seguros Fidelidade disponiveis

Fidelidade <no-reply@fidelidade.pt>

seg, 10/07/2023 14:24

Para:parati10@hotmail.com <parati10@hotmail.com>

FIDELIDADE

SEGUROS DESDE 1885

Olá, **DANIEL MANUEL OLIVEIRA**

Aceda a área de cliente MyFidelidade, no seu computador ou App, e consulte todos os seus documentos em qualquer altura, de uma forma rápida e organizada.

NOVOS DOCUMENTOS DISPONÍVEIS:

Aviso-Recibo de Prémio
Apólice nº 0000000751156553

INFORMAÇÕES RELATIVAS AO SISTEMA DE COBRANÇA POR "DÉBITO DIRETO":

IBAN	**PT50********************92519**
Valor	**228,97 €**

Débito na conta bancária partir de: 27/08/2023

ACEDER

Vantagens

Na área de cliente MyFidelidade, no seu computador ou na App, pode ter acesso à sua documentação digital e ainda:

- Ter acesso a toda a sua documentação digital para consultar e fazer download;
- Atualizar os seus contactos e morada.
- Consultar o detalhe das suas apólices;

Estamos ao seu dispor para o ajudar e esclarecer.

Com os nossos cumprimentos,

MY FIDELIDADE

A gestão dos seus seguros quando e onde quiser.

Aceda em ou

Conheça os nossos seguros em fidelidade.pt

Fidelidade - Companhia de Seguros, S.A.
NIPC e Matrícula 500 918 880 na CRC Lisboa - Sede: Largo do Calhariz 30, 1249-001 Lisboa
Portugal - Capital Social EUR 509.263.524

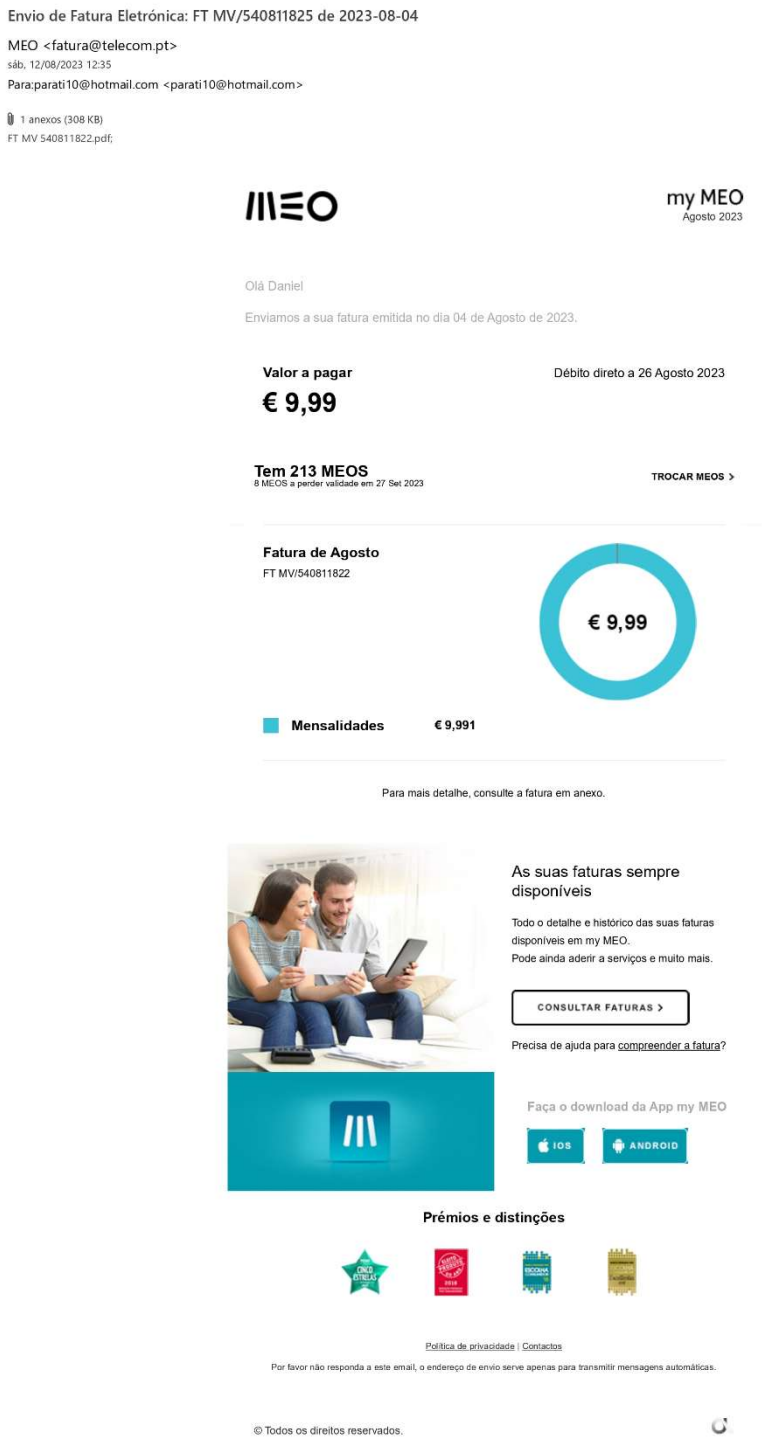
Apoio ao Cliente: Dias úteis das 9h às 20h
T: 217 948 800 - Chamada para rede fixa nacional.

Figure 8 Legitimate Email Fidelidade

Figure 9 Legitimate Email Flixbus displays a promotional email from Flixbus, which is a global mobility provider. This email highlights various travel destinations, promotes the download of the app, and features a prominent button at the top of the page encouraging users to make reservations.

Figure 9 Legitimate Email Flixbus

Figure 10 presents an email received from MEO, a telecommunications company, which includes an attached invoice. Within the email, you can find details such as the invoice amount and date. Additionally, there is a clickable button provided to easily access and review the invoices.


Envio de Fatura Eletrónica: FT MV/540811825 de 2023-08-04

MEO <fatura@telecom.pt>

sáb, 12/08/2023 12:35

Para:parati10@hotmail.com <parati10@hotmail.com>

1 anexos (308 KB)

FT MV 540811822.pdf;

MEO

my MEO
Agosto 2023

Olá Daniel

Enviamos a sua fatura emitida no dia 04 de Agosto de 2023.

Valor a pagar

€ 9,99

Débito direto a 26 Agosto 2023

Tem 213 MEOS

8 MEOS a perder validade em 27 Set 2023

TROCAR MEOS >

Fatura de Agosto

FT MV/540811822

Mensalidades € 9,991

Para mais detalhe, consulte a fatura em anexo.

As suas faturas sempre disponíveis

Todo o detalhe e histórico das suas faturas disponíveis em my MEO.
Pode ainda aderir a serviços e muito mais.

CONSULTAR FATURAS >

Precisa de ajuda para compreender a fatura?

Faça o download da App my MEO

iOS ANDROID

Prémios e distinções

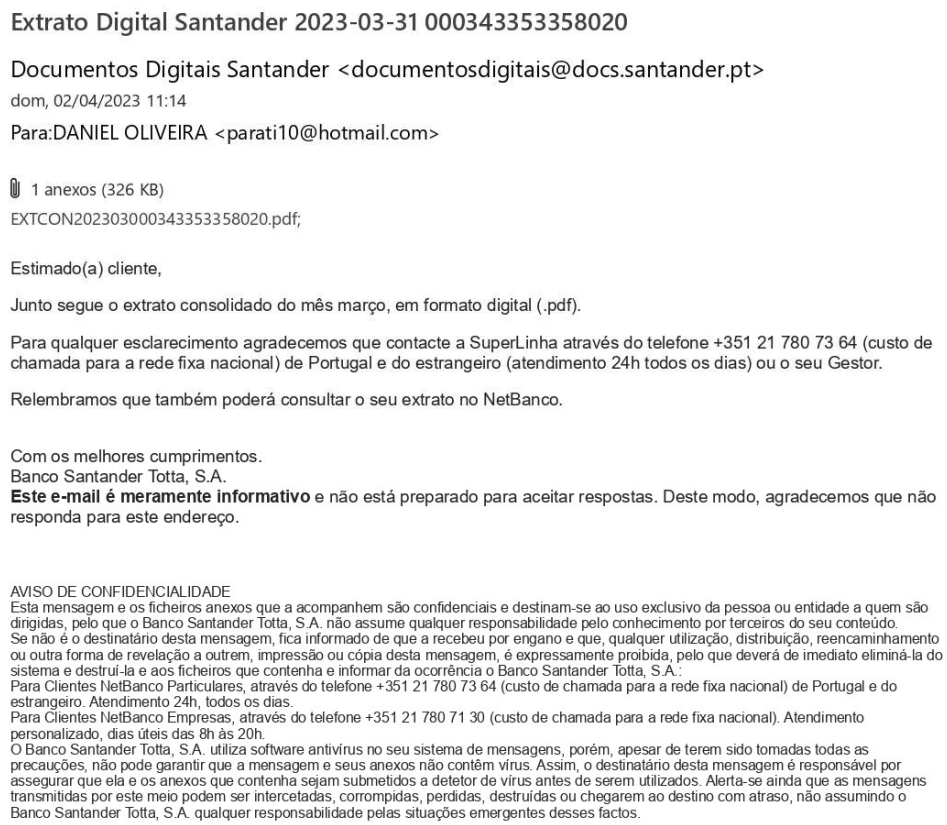
Política de privacidade | Contactos

Por favor não responda a este email, o endereço de envio serve apenas para transmitir mensagens automáticas.

© Todos os direitos reservados.

Figure 10 Legitimate email MEO

The email shown in Figure 11 is a straightforward message from Santander, a bank, containing the monthly account statement as an attachment. This email does not contain any hyperlinks or buttons; it solely includes the PDF attachment.



Extrato Digital Santander 2023-03-31 000343353358020

Documentos Digitais Santander <documentosdigitais@docs.santander.pt>

dom, 02/04/2023 11:14

Para:DANIEL OLIVEIRA <parati10@hotmail.com>

1 anexos (326 KB)

EXTCON20230300034335335802O.pdf;

Estimado(a) cliente,

Junto segue o extrato consolidado do mês março, em formato digital (.pdf).

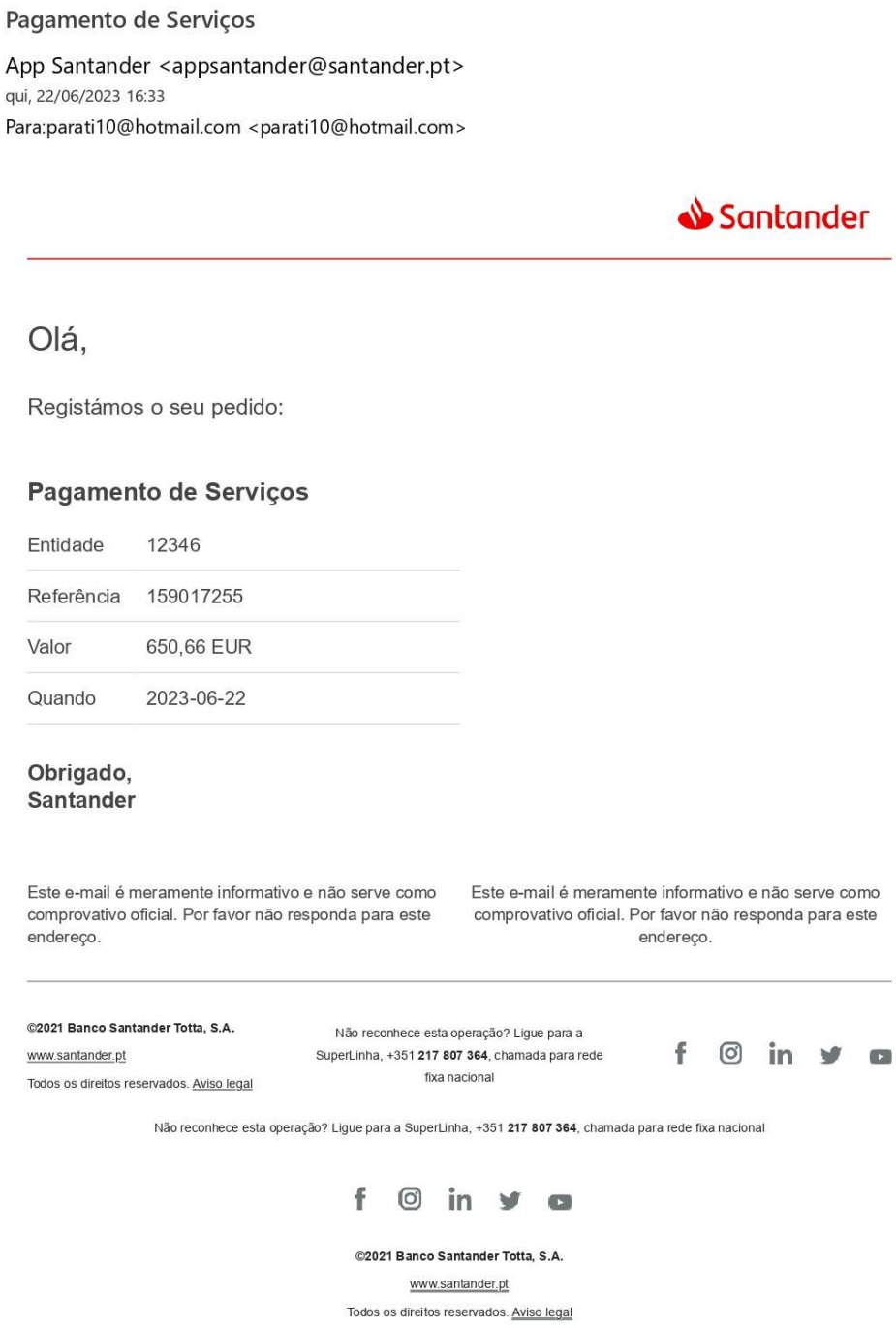
Para qualquer esclarecimento agradecemos que contacte a SuperLinha através do telefone +351 21 780 73 64 (custo de chamada para a rede fixa nacional) de Portugal e do estrangeiro (atendimento 24h todos os dias) ou o seu Gestor.

Relembramos que também poderá consultar o seu extrato no NetBanco.

Com os melhores cumprimentos.
Banco Santander Totta, S.A.
Este e-mail é meramente informativo e não está preparado para aceitar respostas. Deste modo, agradecemos que não responda para este endereço.

Figure 11 Legitimate email Santander

The email on Figure 12 features a confirmation email from Santander, a bank, regarding a service payment of €650.66. While this amount may be substantial for many individuals, it is a legitimate transaction. Notably, this email does not include any clickable buttons, and it repeats the same message twice, which, although may seem unusual, is not an indication of phishing; it is a genuine communication.



Pagamento de Serviços

App Santander <appsantander@santander.pt>
qui, 22/06/2023 16:33
Para:parati10@hotmail.com <parati10@hotmail.com>

Santander

Olá,

Registámos o seu pedido:

Pagamento de Serviços

Entidade	12346
Referência	159017255
Valor	650,66 EUR
Quando	2023-06-22

Obrigado,
Santander

Este e-mail é meramente informativo e não serve como comprovativo oficial. Por favor não responda para este endereço.

Este e-mail é meramente informativo e não serve como comprovativo oficial. Por favor não responda para este endereço.

©2021 Banco Santander Totta, S.A.
www.santander.pt
Todos os direitos reservados. Aviso legal

Não reconhece esta operação? Ligue para a SuperLinha, +351 **217 807 364**, chamada para rede fixa nacional

Não reconhece esta operação? Ligue para a SuperLinha, +351 **217 807 364**, chamada para rede fixa nacional

©2021 Banco Santander Totta, S.A.
www.santander.pt
Todos os direitos reservados. Aviso legal

Figure 12 Legitimate Email Santander

This email presents a special offer from Galp, an integrated energy provider, aimed at users of their mobile app. Figure 13 illustrates this email, which includes several clickable buttons that direct users to the app.

Figure 13 Legitimate Email Galp

This email from Meu Presente, a platform from Deco proteste, appearing to resemble a phishing attempt and featured in Figure 14, offers a gift, specifically JBL headphones. It also displays the personal information of the recipient, including their name, address, and country. The email includes a button for users to track their request and several links.

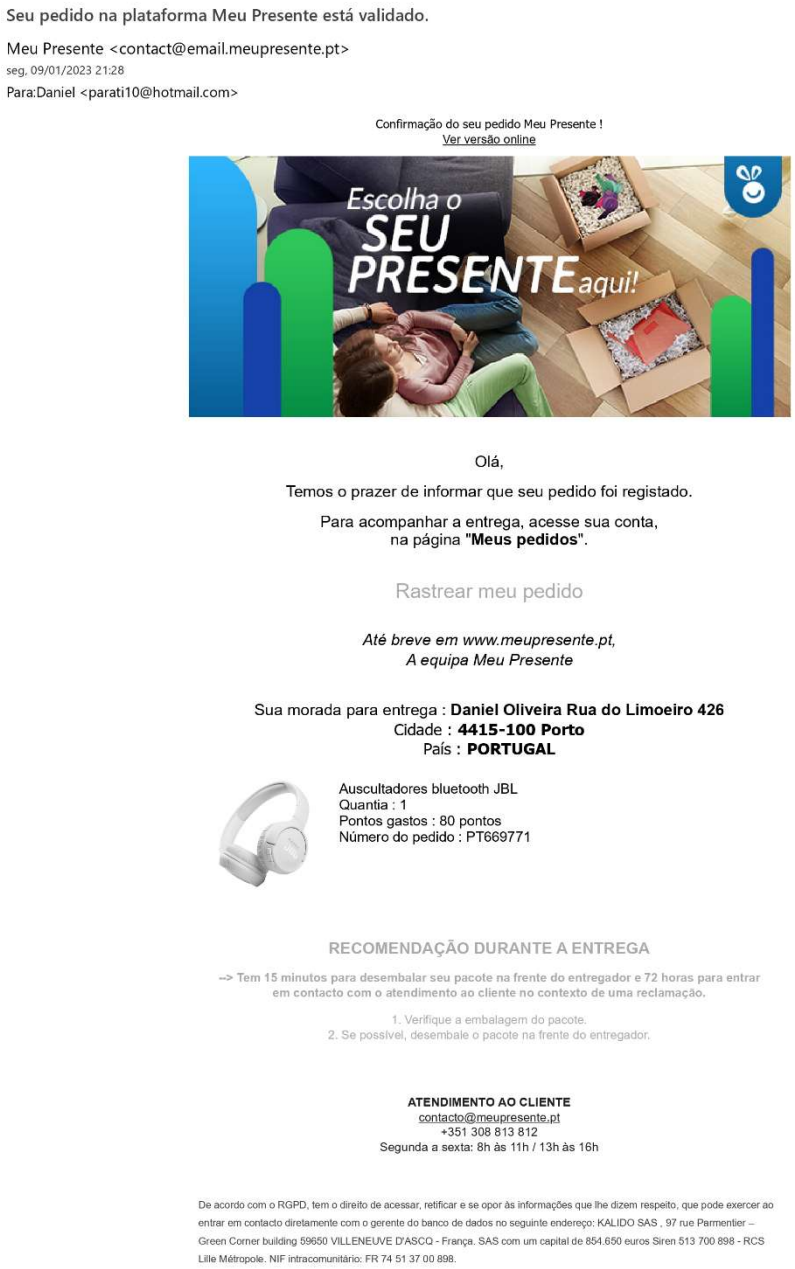


Figure 14 Legitimate Email Meu Presente

3.1.2. Phishing Emails

The phishing email, as shown in Figure 15, imitates an email from CTT (Portuguese Postal Service). The sender email address mentions dental, and the email content claims that a package has been delivered. However, it instructs the recipient to click a button to make a payment for alleged delivery costs. The email also emphasizes a sense of urgency, stating that this fee must be paid within the next three days.

Figure 15 Phishing Email CTT

This email, on Figure 16, imitates a communication from Nescafé. The email subject features two languages, Portuguese and Spanish. The sender email address includes terms like "correos" (Spanish for mail) and "caso clínico" (Spanish for clinical case). The email content urges the recipient to click a button to validate a request. It displays an alleged purchase of a coffee machine valued at 379.90€ for a mere 1.95€. The email concludes with an extended text in the German language, referencing a football story.

Su pedido #eb4e55eb9ba821b foi feito! 🎉

Nespresso.pt™ <correos@casoclinico.net>
sex, 25/11/2022 08:10
Para:parati10 <parati10@hotmail.com>

Seu pedido C40337 está esperando por você

Valide seu pedido

Ponto de retirada
ARMAZÉM DHL MAIS PRÓXIMO

Daniel Oliveira,

O seu pedido foi submetido no dia 24.11.2022 às 12:50:31.(ordem por conversão de pontos).

Resumo da encomenda:

Valor dos produtos: 379.90 €
Custo de envio: 0.00 €
Total dos pontos: 324 pontos
Desconto: 377.05 €
Valor total da encomenda: 1.95 €

Produtos encomendados:

Sophisticated Darks De'Longhi Gran Lattissima

Valide seu pedido

 # Email do pedido
parati10@hotmail.com

Rastreamento
indisponível

WM 2022: Thomas Hitzlsperger kontert Kommentar von Eden Hazard über "Mund-zu"-Aktion der DFB-Elf
Eden Hazard konnte sich einen bissigen Kommentar zur "Mund-zu"-Aktion der deutschen Nationalmannschaft vor dem Anpfiff des WM-Gruppenspiels gegen Japan nicht verkneifen. "Es wäre besser gewesen, wenn sie es nicht getan und ihr Spiel gewonnen hätten", stichelte Belgiens Kapitän bei "TV2". Ein gefundenes Fressen für Thomas Hitzlsperger. Der Ex-Nationalspieler knöpfte sich Hazard auf Twitter vor.
Lusail/München - Brasilien zauberte - aber Neymar weinte. Dem Superstar schossen auf der Bank plötzlich Tränen in die Augen, sein rechtes Fußgelenk musste nach der späten Auswechslung mit Eis gekühlt werden: Der Rekordweltmeister hat die Jagd auf den sechsten WM-Titel zwar glanzvoll eröffnet, das 2:0 (0:0) gegen Serbien aber womöglich teuer bezahlt.

Ob Neymar bei insgesamt neun Fouls und in etlichen harten Zweikämpfen mehr abbekommen hat als einen dicken Knöchel, wird das Thema der kommenden Tage sein.

Vom Ergebnis zumindest ist der Start gelungen. Denn als Neymar noch auf dem Rasen hockte, wurde der Matchwinner unter tosendem Beifall ausgewechselt: Richarlison hatte ihm mit einem Abstauber (62.) und einem spektakulären Seitfallzieher (73.) die Show gestohlen.

"Vor vier Wochen lag ich selbst noch weinend auf dem Rasen, ich war verletzt", sagte Richarlison, der in England für Tottenham Hotspur stürmt. "Ich hatte sogar Angst, überhaupt nicht zur WM fliegen zu können. Aber alle Anstrengungen haben sich gelohnt. Ich bin super glücklich."

Brasilien jubelte dort, wo es am 18. Dezember unbedingt den "Hexacampeonato" feiern will: im Finalstadion von Lusail. Am Ende drehte sich das Passkarussell schon schwindelerregend - doch Neymar vergrub sein Gesicht traurig im Trikot.

WM 2022: Pele fordert WM-Titel
"Bringt den Pokal nach Hause!", hatte die nationale Fußball-Ikone Pele in den sozialen Medien gefordert. Die Selecao gehorchte: Sie verdiente sich ihren 17. Sieg im 22. WM-Auftaktspiel und setzte sich an die Spitze der Gruppe G mit der Schweiz und Kamerun.

Neymar nimmt seinen dritten Anlauf auf den WM-Titel - für ein tief gespaltenes Land. Das gelb-grüne Nationaltrikot, zu Zehntausenden im Stadion zu sehen, war in den vergangenen Wochen vom Lager des Rechtspopulisten Jair Bolsonaro politisch vereinnahmt worden. Auch der Superstar trommelte bei der Präsidenten-Stichwahl für den Amtsinhaber, vergeblich.

Einigender ist der große Pele, der einzige Spieler der Geschichte mit drei WM-Titeln. Der 82-Jährige schrieb bei Instagram: "Unsere 200 Millionen Herzen werden schlagen wie eines. Lasst uns jedes Spiel wie ein Finale angehen!"

Figure 16 Phishing Email Nespresso

The email shown in Figure 17 represents a phishing attempt, allegedly from Worten Contest. Within the email, it falsely claims that the recipient has won a "brand new" iPhone. Additionally, it includes a message stating that to claim this prize, the user must respond to a series of questions and click a button to initiate the survey.

Figure 17 Phishing Email Worten

Figure 18 shows an attempt to obtain user information while claiming that the user account is blocked. The notification also states that the user mobile device has not yet been verified in the records and that to unlock the account, they must click the "continue" option. The message warns that if the user does not unlock the account in the next 24 hours, it can only be done at a bank counter– but this bank has none; it is entirely virtual.

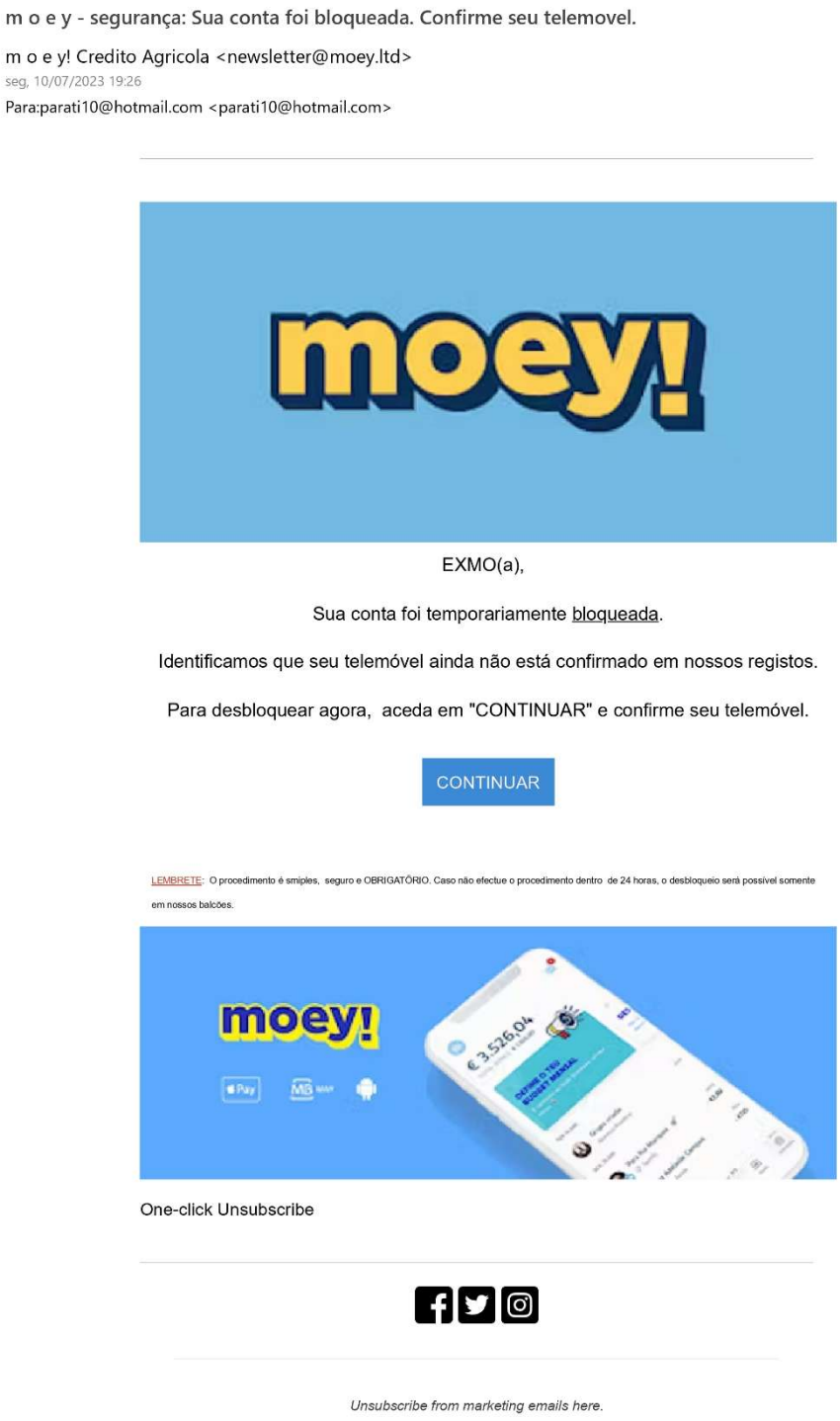


Figure 18 Phishing Email Moey!

An email that purports to be from the insurance company MAPFRE is showed in Figure 19. The email invites the recipient to create a simulation with several buttons. Words like "mail" and "use makeups," which are not in the original context of the email, are used in the text.

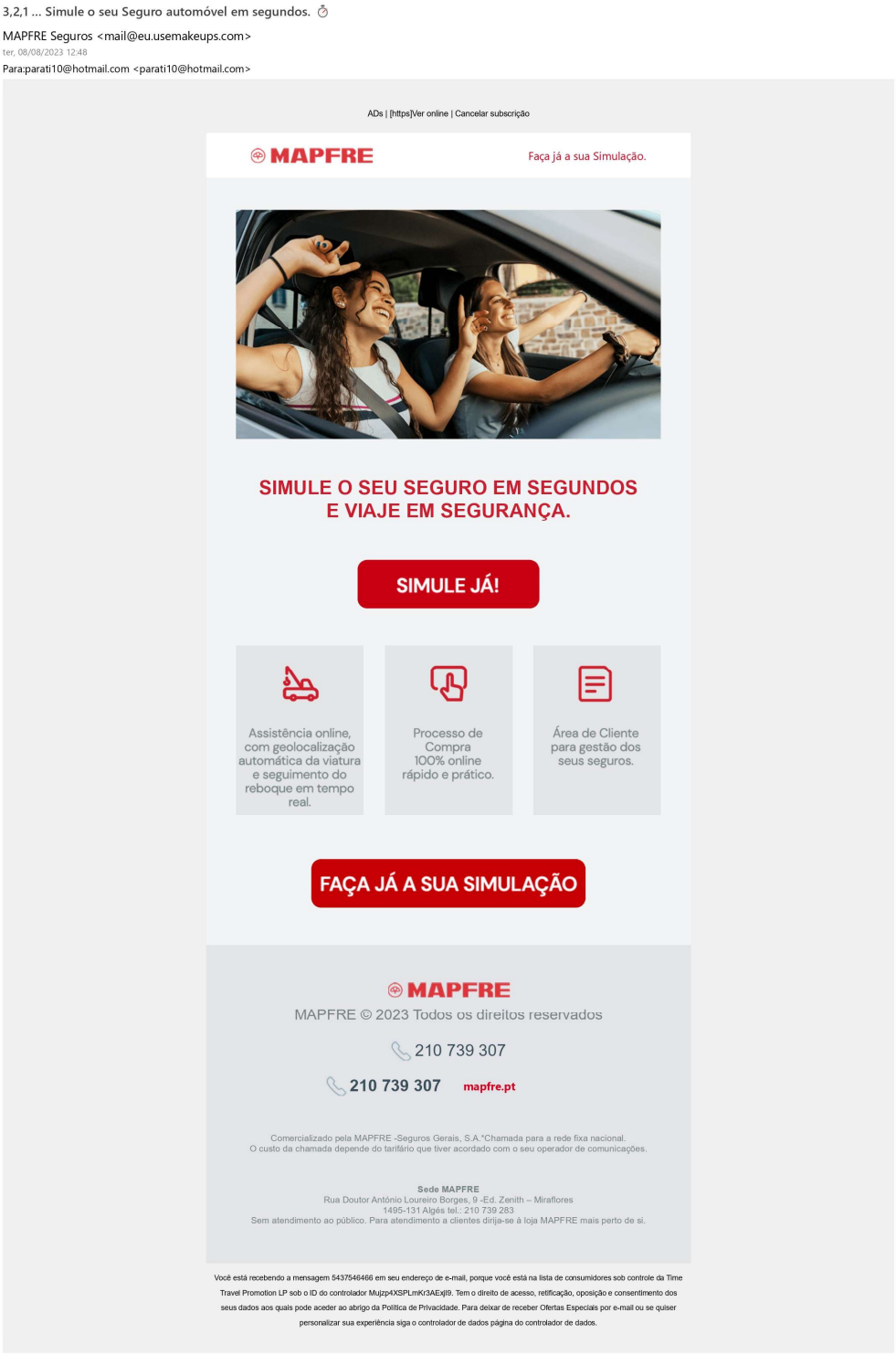


Figure 19 Phishing Email MAPFRE insurances

According to the email from Figure 20, the user will receive a receipt for the payment he has made to InfoS. This email contains a fake link that was supposed to allow the user to confirm receipt.


Recibo

InfoS <recibosbatia32ka883kksaf@recibosbatia.com.mx>
ter, 29/08/2023 10:08
Para:parati10@hotmail.com <parati10@hotmail.com>

Boa tarde

Agradecemos o V. pagamento - Online AQUI.

Cumprimentos

Rua Gonçalves Zarco, nº 176
4440-659 Valongo
Portugal
tlf. 351 224 229 037
fax. 351 969 004 727

Figure 20 Phishing Email Info S

A fraudulent email from CTT claiming that the address is uncomplete is presented in Figure 21. The email advises the user to click the button of the email within the next two days to resolve this issue.


Endereço incompleto - ação necessária

CTT-Expresso <bkdxgvfesds@truvaturizm.com>
seg, 04/09/2023 17:25
Para:parati10@hotmail.com <parati10@hotmail.com>

ctt

Endereço incompleto - ação necessária

Estimado cliente,

Escrevemos para informar que não podemos entregar seu pacote recente devido a um endereço incompleto.

Tentamos entrar em contato com você pelo número de telefone para obter as informações em falta, mas infelizmente não conseguimos contatá-lo.

Para garantir que seu pacote seja entregue no destino correto e no prazo, pedimos que atualize suas informações de endereço e detalhes de perfil com o CTTGroup nos próximos 2 dias, clicando no botão abaixo.

Atualizar Agora !

Figure 21 Phishing Email CTT

This email, showed in Figure 22, states that they are looking for testers for Nivea products. The email urges the recipient to click the link and enter their personal information in the form that opens.

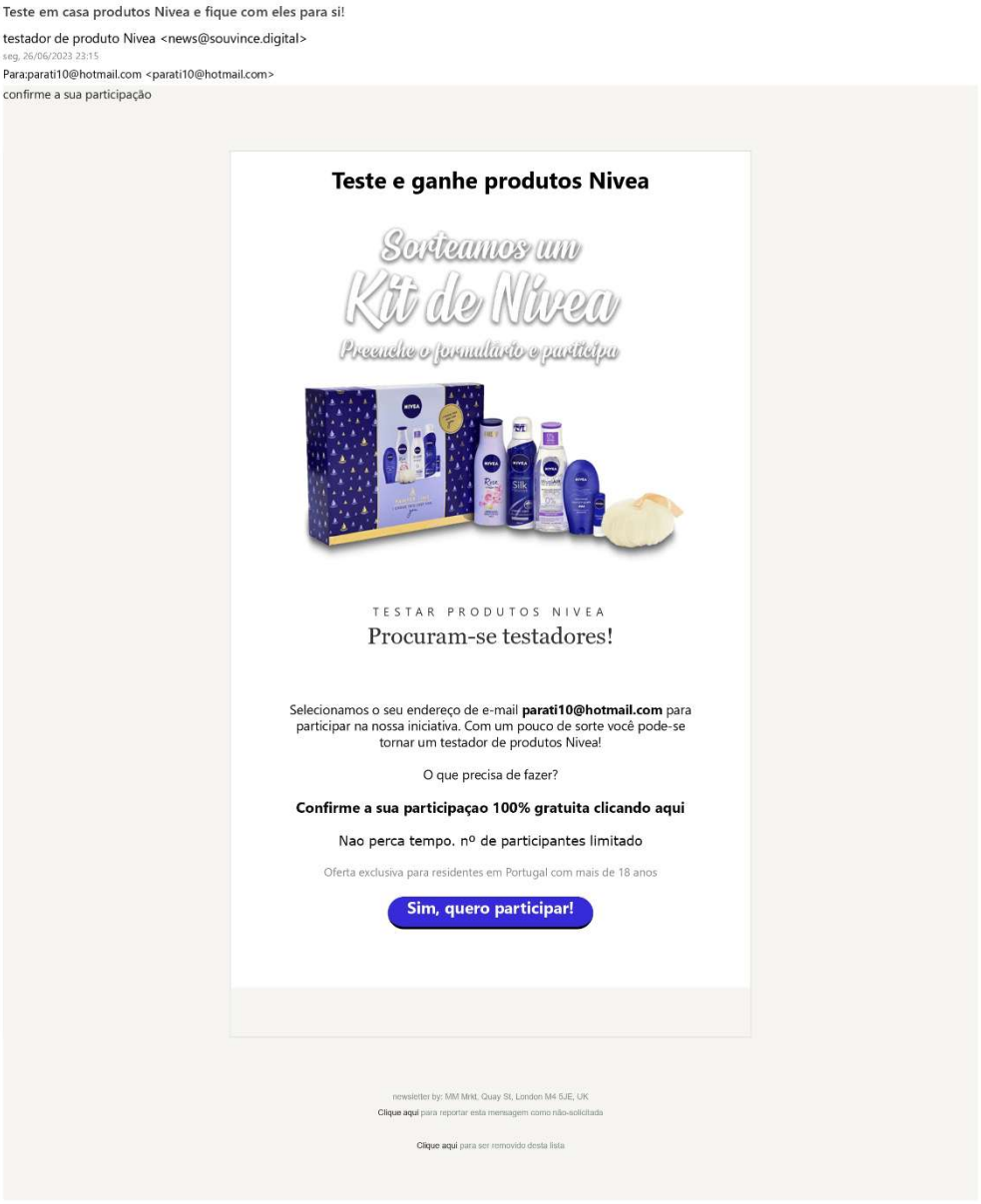
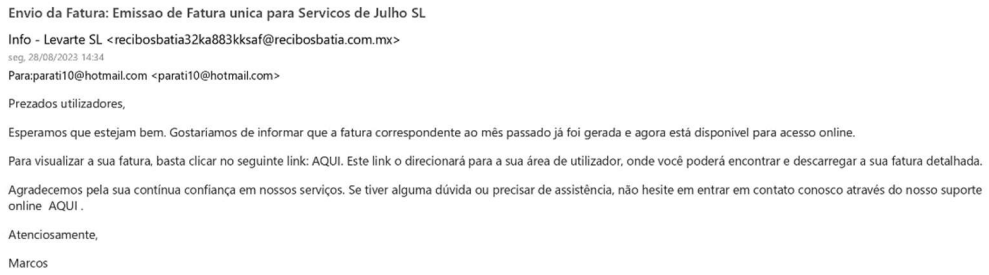


Figure 22 Phishing Email Nivea

The email shown in Figure 23 persuades the user to accept this big promotion by warning that there are only a few days left to do so. Multiple buttons in the email encourage the user to simulate the pricing and there are many references to quality of the product. We can observe from the sender email that it is quite dubious.

Figure 23 Phishing Email Grupo8

The email shown in Figure 24 encourages the user to click a link to view their monthly invoice. The email instructs the user to click the link, which will take him to his user panel and allow him to download the comprehensive invoice.



Envio da Fatura: Emissao de Fatura unica para Servicos de Julho SL

Info - Levarte SL <recibosbatia32ka883kksaf@recibosbatia.com.mx>

seg, 28/08/2023 14:34

Para:parati10@hotmail.com <parati10@hotmail.com>

Prezados utilizadores,

Esperamos que estejam bem. Gostariamos de informar que a fatura correspondente ao mês passado já foi gerada e agora está disponivel para acesso online.

Para visualizar a sua fatura, basta clicar no seguinte link: AQUI. Este link o direcionará para a sua área de utilizador, onde você poderá encontrar e descarregar a sua fatura detalhada.

Agradecemos pela sua contínua confiança em nossos serviços. Se tiver alguma dúvida ou precisar de assistência, não hesite em entrar em contato conosco através do nosso suporte online AQUI .

Atenciosamente,

Marcos

Figure 24 Phishing Email Info- Levarte SL

3.2. Research inquiry

To investigate the correlation between personality, emotions, and responses to phishing emails, a survey was conducted using the LimeSurvey platform. LimeSurvey was chosen due to its availability, features, and ease of use.

The survey was structured into four main sections: demographic data, the personality Inventory NEO- FFI-20, the Melbourne Decision Making Questionnaire (Portuguese version) and by the task involving the classification of 20 phishing and non-phishing emails. Figure 25 illustrates the inquiry workflow.

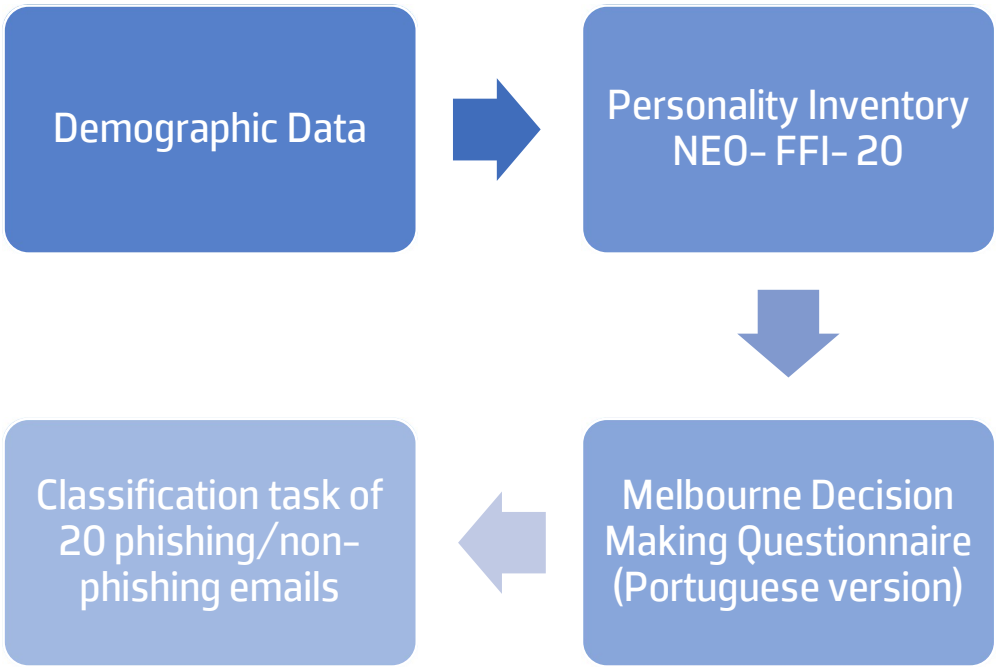


Figure 25 Inquiry methodology workflow

The demographic data section included information such as educational level, sex, profession, native language, monthly income, age, and two rating questions. The rating questions ask participants to rate their frequency of working with online platforms and handling sensitive information on a scale from 1 to 5. Educational level, sex, profession, and monthly income provide background data for profiling the participants. Age was collected to cross-reference results and assess user accuracy. The questions about online platform usage and handling sensitive information aimed to gauge the participants' comfort with technology, especially online platforms, and their experience in dealing with sensitive information. This helped evaluate the correlation between these factors and participants' phishing detection accuracy. In Figure 26 the inquiry demographic data group workflow is represented.

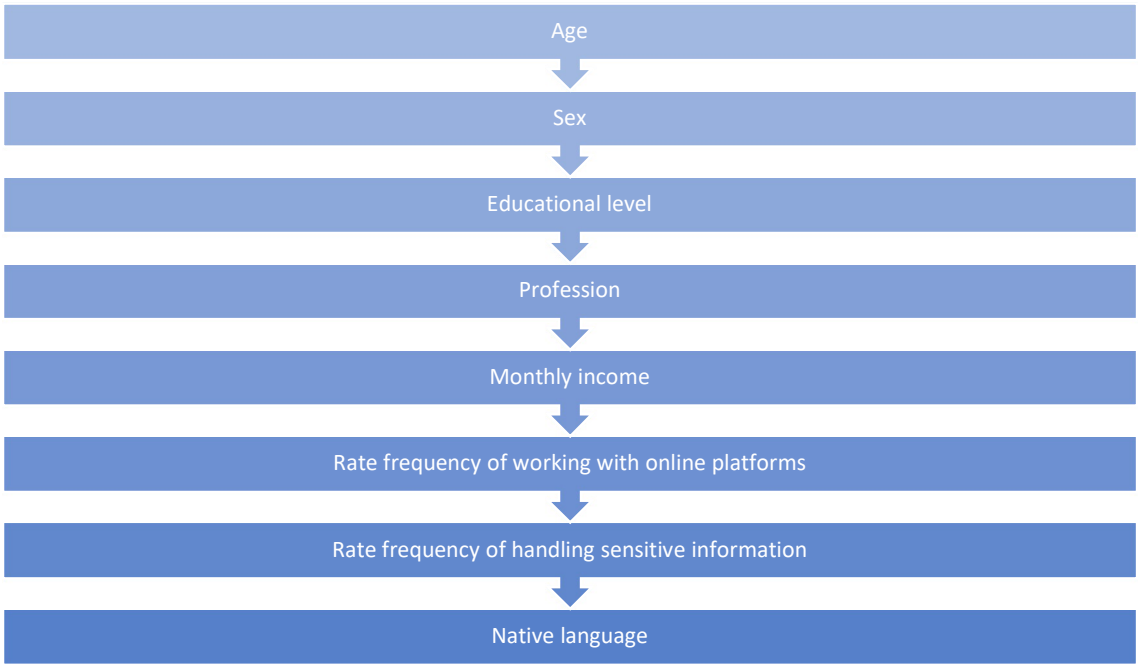


Figure 26 Inquiry demographic data section workflow

In the classification task, participants were presented with an email and given a specific time to read it, determined by the average time taken by the author and the orienteer or co-orienteer. After reading, participants were asked to report the emotions triggered by the email. Participants are then asked to classify the email as either "relaxing" or "activating" on a scale from 1 (relaxing) to 5 (activating).

Similarly, participants were asked to classify the email as "pleasant" or "unpleasant" on a scale from 1 (pleasant) to 5 (unpleasant). The final question regarding the email is whether participants believe it to be a phishing email or not. This question was positioned last to avoid influencing previous responses.

The classification task was repeated twenty times, with ten phishing emails and ten non-phishing emails presented in random order to each participant.

Following the demographic data questions, participants were invited to complete a personality inventory NEO-FFI-20 and the Melbourne Decision Making Questionnaire (Portuguese version) which was followed by the classification task.

To maintain data anonymization, all of these inquiries were integrated into a single LimeSurvey questionnaire, which also featured the NEO-FFI-20 personality inventory and the Melbourne Decision Making Questionnaire (Portuguese version).

The "Questionário de Melbourne de Tomada de Decisão" was generously supplied by its author, Luis Filipe, who granted permission for its use in the survey (see Appendix 3). Additionally, the NEO-FFI-20

personality inventory was kindly provided by its author, José Luís Pais Ribeiro, who also granted authorization for its inclusion in the survey as presented in Appendix 4.

The NEO-FFI-20 Personality Inventory is a shortened instrument validated for the Portuguese population by [46], designed to assess personality traits. This tool operationalizes the five-factor model: Extroversion, Openness to Experience, Agreeableness, Conscientiousness, and Neuroticism, utilizing a five-point Likert scale (0 - strongly disagree to 4 - strongly agree). Scoring is computed per dimension, ranging from 0 to 16 points. Higher scores in a dimension indicate a greater presence of that trait in an individual personality. The instrument has exhibited strong psychometric properties in prior studies (α > 0.70) [47].

The Melbourne Decision Making Questionnaire (MDMQ), adapted and validated by [44], assesses a personal pattern of an individual for dealing with decision-related stress, delineating how individuals make choices and judgments. This instrument evaluates four primary strategies: vigilance (e.g., "When making decisions I like to collect lots of information"), buck-passing (e.g., "I do not like to take responsibility for making decisions"), procrastination (e.g., "Even after I have made a decision I delay acting upon it"), and hypervigilance (e.g., "I feel as if I'm under tremendous time pressure when making decisions"). Respondents rate 22 items on a 3-point Likert scale (0-strongly disagree to 2-strongly agree). The total scores for each scale range from 0 to 10 (for procrastination and hypervigilance) or 0 to 12 (for vigilance and buck-passing) [48].

The questionnaire was divided into two phases: a pilot run and a trial run. During the pilot run, three participants were asked to complete the questionnaire. Following the suggested improvements, data collection for the questionnaire started.

3.2.2. Sample

The survey incorporated responses from a total of 106 participants. Of these, 54 provided complete responses, while 52 submitted incomplete ones.

The inclusion criteria for participation in this study were as follows: (1) being of legal age, (2) of any sex, male or female, (3) having the ability to use a computer as an end-user, and (4) proficiency in the Portuguese language. The exclusion criteria primarily involved illiteracy and a lack of computer skills.

The participants' mean age was 35 years, with a standard deviation of 14.25 and a range of 50 years, spanning from a minimum age of 17 to a maximum of 67.

Regarding the highest level of completed education among participants, "Graduation" was the most common, with 41% of respondents, totaling 22 participants. "Secondary education aimed at continuing higher studies" constituted 24% of the sample, representing 13 participants. "Secondary education obtained through double certification or with an internship" accounted for 13% of the participants,

comprising 7 individuals. Both "Masters degree" and "2nd cycle of basic education" were each represented by 7% of the sample, which equates to 4 individuals. A total of 6% (3 persons) had completed "3rd cycle of basic education obtained in basic education or through double certification." Only one person, or 2%, possessed a "Non-tertiary post-secondary qualification (CTESP)." No respondents held a Doctorate.

In terms of occupation, none of the respondents worked as "Representatives of the legislative branch and executive bodies, directors, directors, and executive managers," and only one person was employed as a "Skilled worker in industry, construction, and craftsmanship," constituting 2%. Two individuals worked as "Administrative staff," while two others were "Farmers and skilled workers in agriculture, fishing, and forestry," each group representing 4% of the sample. "Plant and machine operators and assembly workers" and "Unskilled workers" both accounted for 9% of the respondents, each comprising 5 individuals. "Specialists in intellectual and scientific activities" and "Personal, safety, and security service workers and salespeople" each constituted 13% of the sample with seven individuals. The category of "Intermediate-level technicians and professions" made up 19% of the sample, comprising 10 individuals. The most substantial group identified as "students," representing 28% of the sample, or 15 people.

Among the respondents, 20% (11 individuals) reported personal incomes falling between €500 and €700. A total of 24% (13 individuals) earned more than €1000, while 26% (14 individuals) reported earning between €0 and €500. The remaining 30% (16 individuals) earned between €700 and €1000.

Regarding the frequency of working with online platforms, none of the respondents indicated a rating of 1, which corresponds to "Not frequent at all." Eight participants (15%) selected 2, 26% (14 individuals) chose 3, and 28% (15 individuals) opted for 4. The majority, 31% (17 individuals), chose 5, indicating "very frequent" usage.

In terms of handling sensitive information, 26% (14 individuals) rated their frequency as 1, indicating "Not frequent at all." A total of 22% (12 individuals) selected 2, 19% (10 individuals) chose 3, 20% (11 individuals) picked 4, and 13% (7 individuals) rated their frequency as 5, which denotes "very frequent" handling of sensitive information.

The sociodemographic data results are covered in Table 2 above.

	Sociodemographic results **% (n=54)**
Age (Years)	35,28 (14,25) (17 - 67)
Sex % (n)	
Male	46% (25)

Female	54% (29)
Completed education % (n)	
2nd cycle of basic education	7% (4)
3rd cycle of basic education obtained in basic education or through double certification	6% (3)
Secondary education aimed at continuing higher studies	24% (13)
Secondary education obtained through double certification or with an internship	13% (7)
Non-tertiary post-secondary qualification (CTESP)	2% (1)
Graduation	41% (22)
Master's degree	7% (4)
Doctorate	0% (0)
Profession % (n)	
Representatives of the legislative branch and executive bodies, directors, directors and executive managers	0% (0)
Specialists in intellectual and scientific activities	13% (7)
Intermediate-level technicians and professions	19% (10)
Administrative staff	4% (2)
Personal, safety and security service workers and salespeople	13% (7)
Farmers and skilled workers in agriculture, fishing and forestry	4% (2)
Skilled workers in industry, construction and craftsmen	2% (1)
Plant and machine operators and assembly workers	9% (5)
Unskilled workers	9% (5)
Student	28% (15)
Personal Income % (n)	
Between €0 and €500	26% (14)
Between €500 and €700	20% (11)
Between €700 and €1000	30% (16)
More than €1000	24% (13)

Frequency of working with online platforms	
1	0 %
2	15% (8)
3	26% (14)
4	28% (15)
5	31% (17)
Frequency of handling sensitive information	
1	26% (14)
2	22% (12)
3	19% (10)
4	20% (11)
5	13% (7)

Table 2 - Inquiry Results - Sociodemographic data

3.3. Tools Research

In this study, data was gathered from two AI tools: IBM Watson and Chat GPT. The methodology for each tool was as follows.

3.3.1. IBM Watson

The email sentiments were analyzed using IBM Watson Natural Language Understanding Text Analysis page, which can be found at this link: https://www.ibm.com/demos/live/natural-language-understanding/self-service/home.

For each of the selected emails, it was extracted the full text of the email body, recognizes by the PDF file of the email and applied into the AI on August 29th, 2023.

3.3.2. Chat GPT

This research was performed from August 16th till September 1st on https://chat.openai.com/ using to the GPT-3.5 version of this software. On a Blank Chat, it was asked the AI Tool: "Analyze the emotions based on 6 Types of Basic Emotions (fear, disgust, anger, joy, sadness, and surprise) of the following email:" and the given text of the email was given. This process was repeated for all 10 legitimate emails and 10 phishing emails.

The results are presented in the next chapter.

Chapter 4 – Results

This chapter contains two subchapters: the results from various AI Tools and the results from the inquiry.

4.1. AI Tools Research

This subsection presents the outcomes of the sentiment analysis conducted using IBM Watson and Chat GPT.

4.1.1. IBM Watson

The IBM Watson Tool categorizes text sentiments based on their polarity. No previous classification was present for any of the emails that were examined.,

The emails overall polarity for the Legit 1 message from CTT Tolls, in Figure 27 sender was positive with a punctuation of 0.92. Figure 27 illustrates how the words "detail" and "confidence" categorize as positive what reflects the final result.

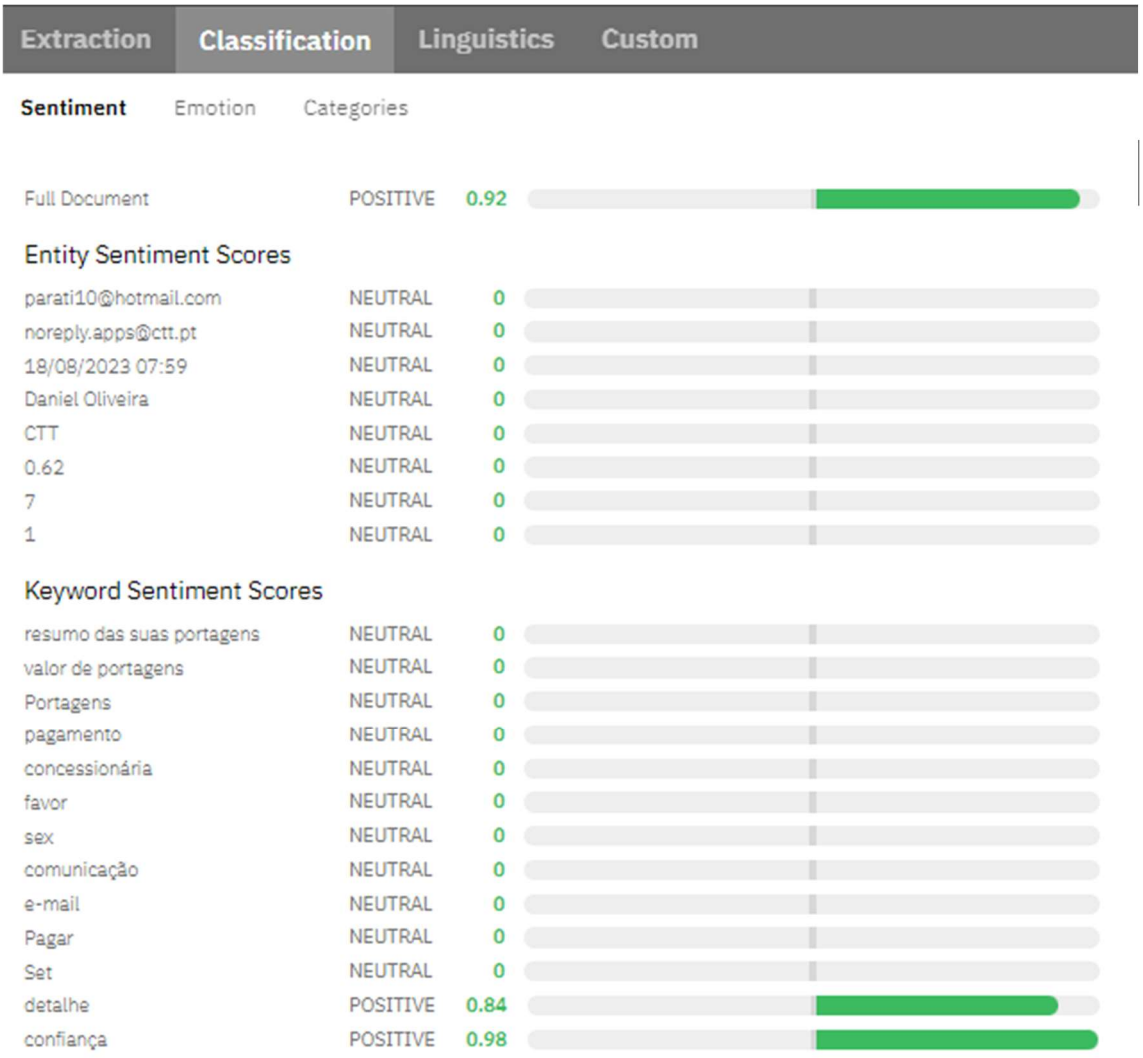


Figure 27 IBM Watson Classification results – Sentiment analysis for CTT Tolls email

With a punctuation score of 0.28, the Legit Email 2 from Meu Presente shown in Figure 28 is categorized as favorable. Positive connotations are assigned to the URL (www.meupresente.pt), the phrases "PT," and "team meu Presente." "Contact," "context of complaint," and "request" are considered negative terms.

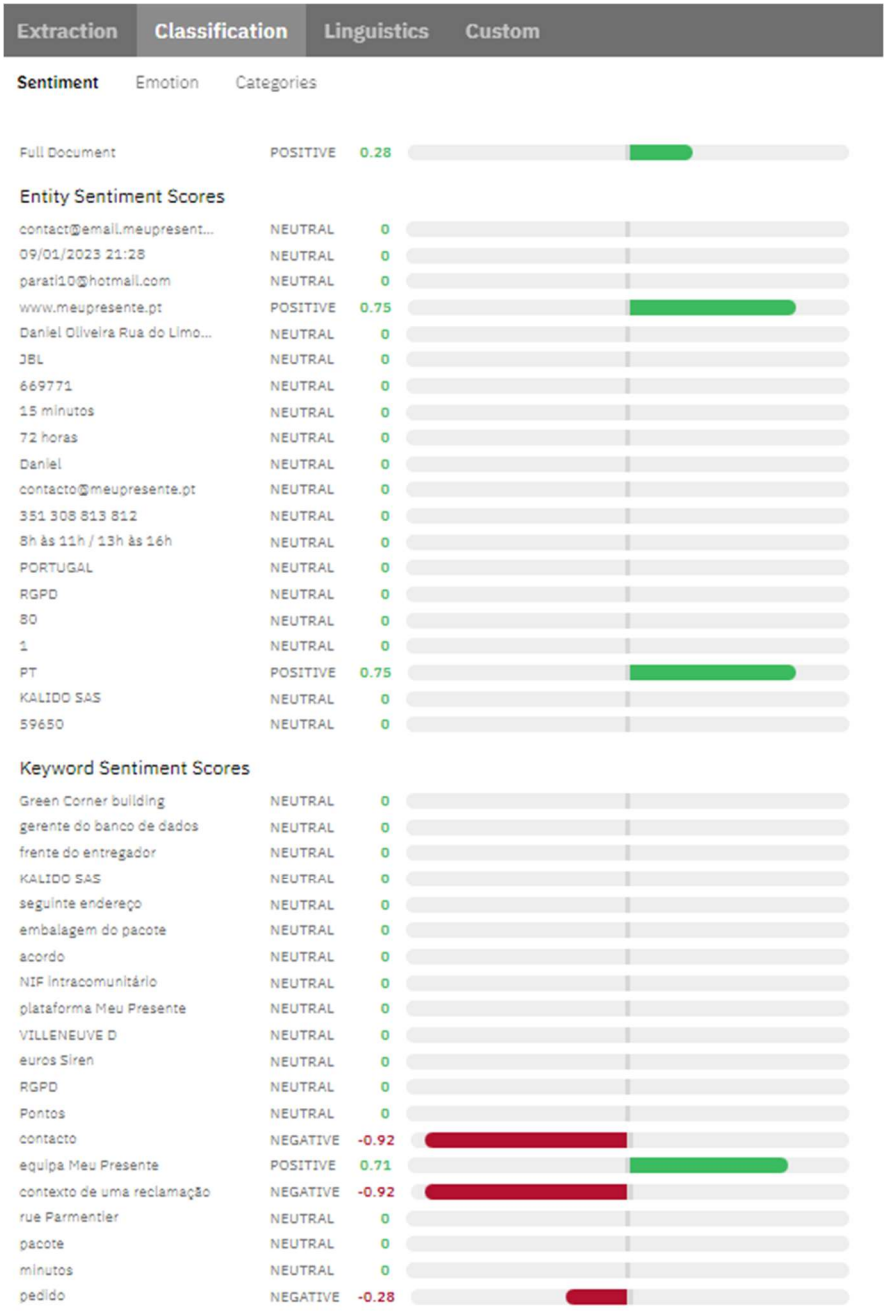


Figure 28 IBM Watson Classification results – Sentiment analysis for Meu Presente

The phrase "best regards" in the email Legit 4 from techonline, which is seen in Figure 29, is what gives it a categorization of positive by 0.82 points.

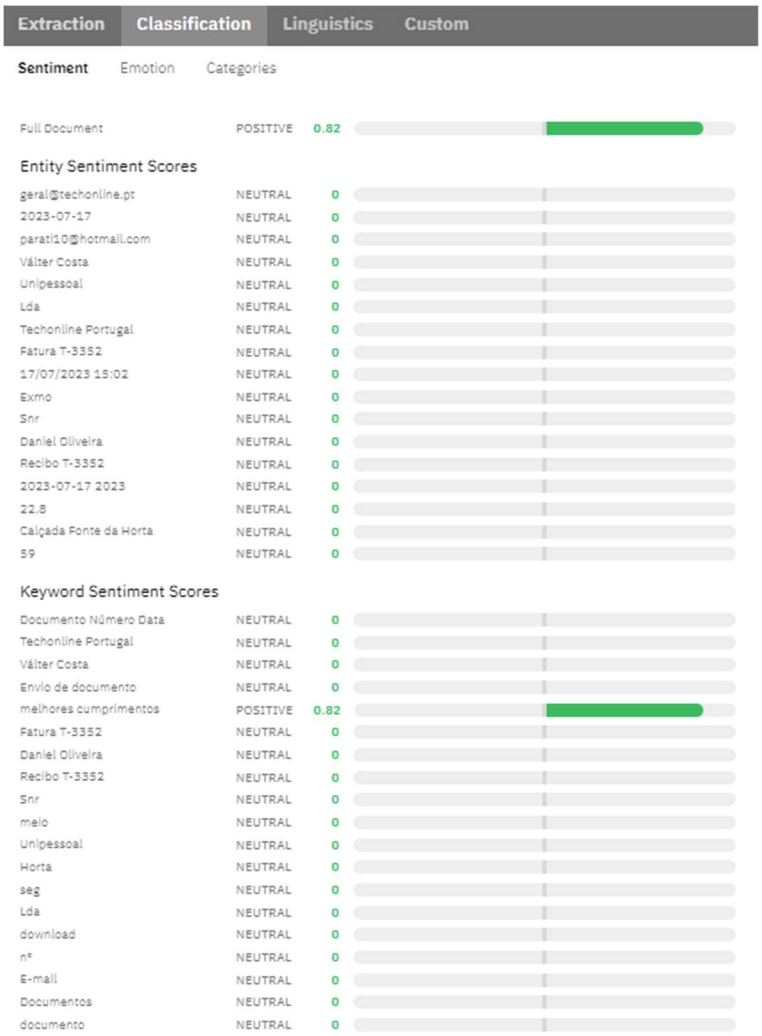


Figure 29 IBM Watson Classification results – Sentiment analysis for Techonline

The email from Moey! on Figure 30, with the ID Legit 3, is rated negatively (-0.26 points), according to the data. The terms "Daniel," "access code," and "app" all contributed to this outcome. The word "moey" and the internet URL "https://moey.pt" both contribute to a favorable classification.

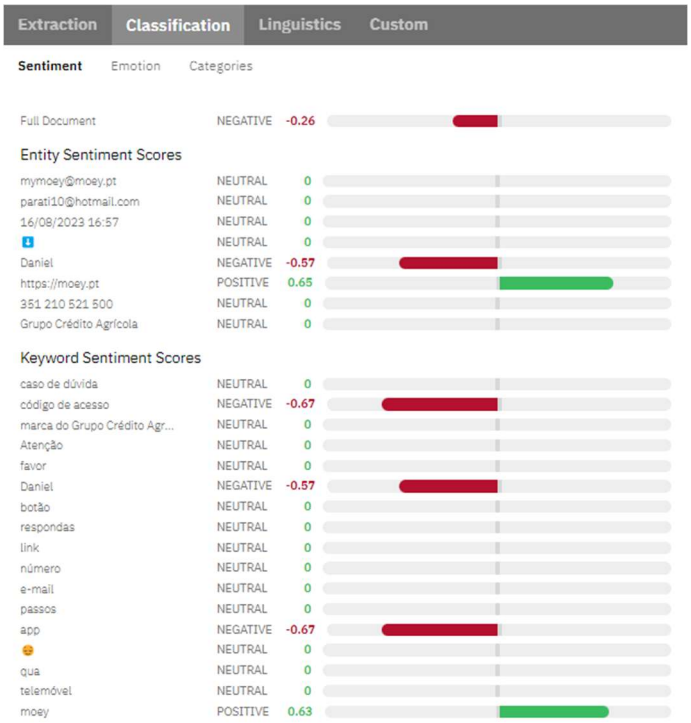


Figure 30 IBM Watson Classification results – Sentiment analysis for Moey!

The overall punctuation score for the Legit email 5 from Fidelidade is -0.41 points, although there is no information about this punctuation, as can be shown in Figure 31.

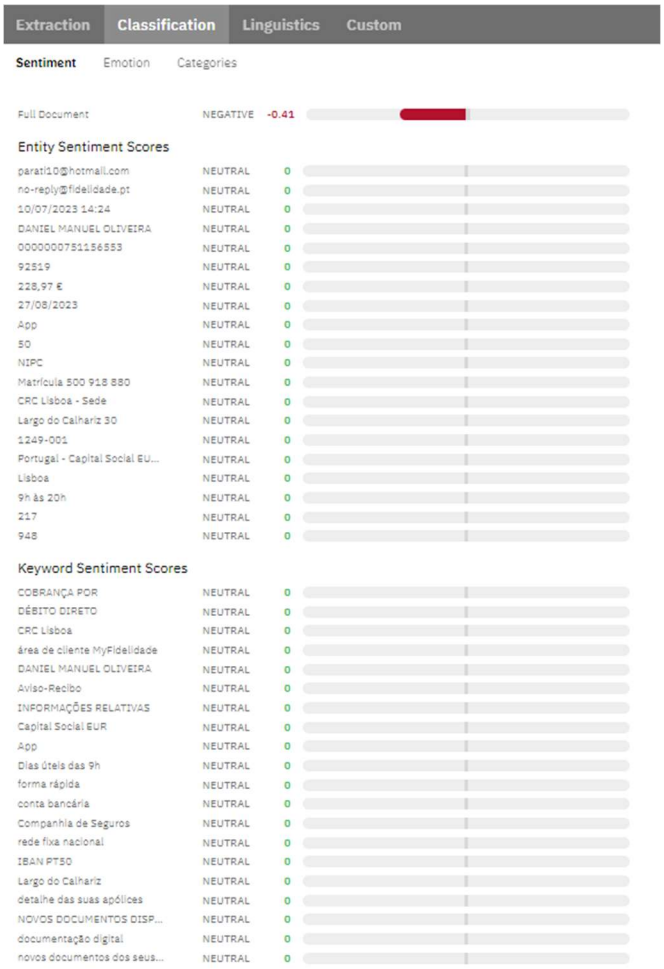


Figure 31 IBM Watson Classification results – Sentiment analysis for Fidelidade

According to Figure 32, the legitimate email 7 from MEO scores negatively with -0.28 points. The terms "Daniel," "Olá Daniel," and "detail" are classified as positive polarity, while "MEOS," "2023," "27 Set 2023," "MEO," "App my MEO," "my MEO," "validity," "Set," and "download" are classified as negative polarity.

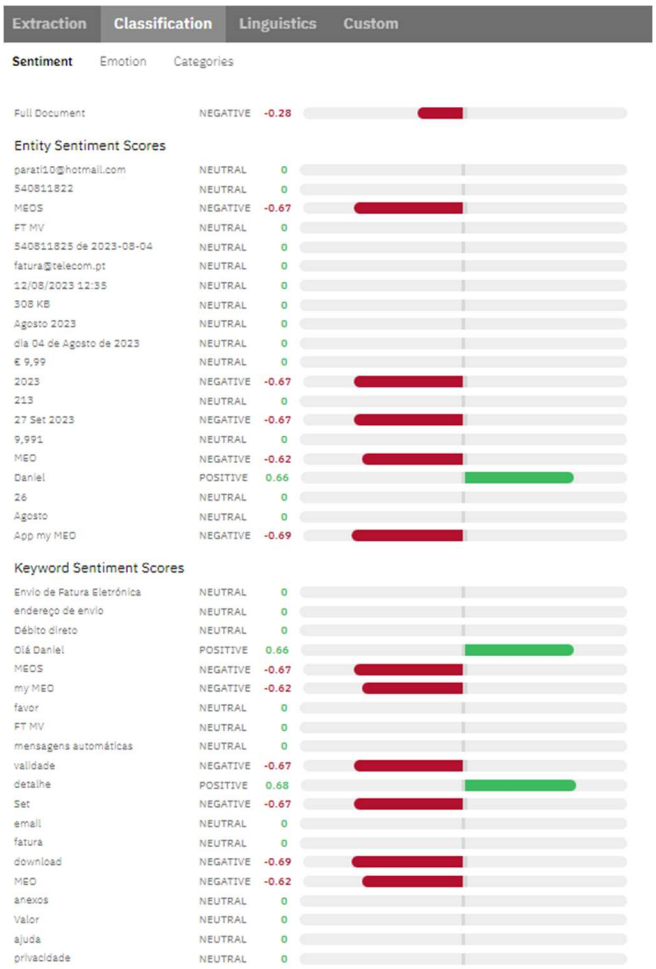


Figure 32 IBM Watson Classification results – Sentiment analysis for MEO

IBM Watson classified the legitimate email 6 from FlixBus, which is shown in Figure 33 as positive with 0.53 points. The terms "actor," "big screen," "best scenarios," "main actor," "time to be," "adventurous spirit," "best destinations," and "adventure" all played a part in this classification. The sole word classified as negative was "app".

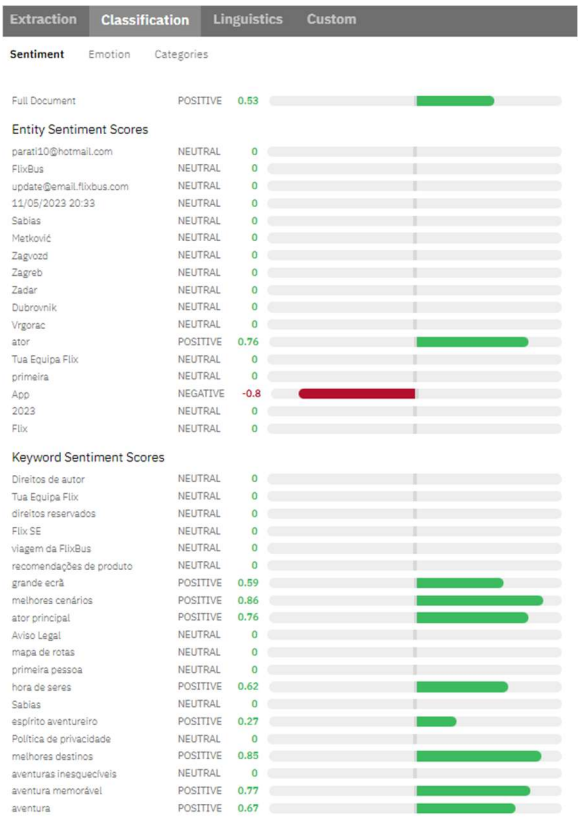


Figure 33 IBM Watson Classification results – Sentiment analysis for FlixBus

The Santander Legit email 8, which is depicted in Figure 34, is rated negatively (-0.27 points). The words "message" and "best regards" are categorized as positive and negative, respectively.

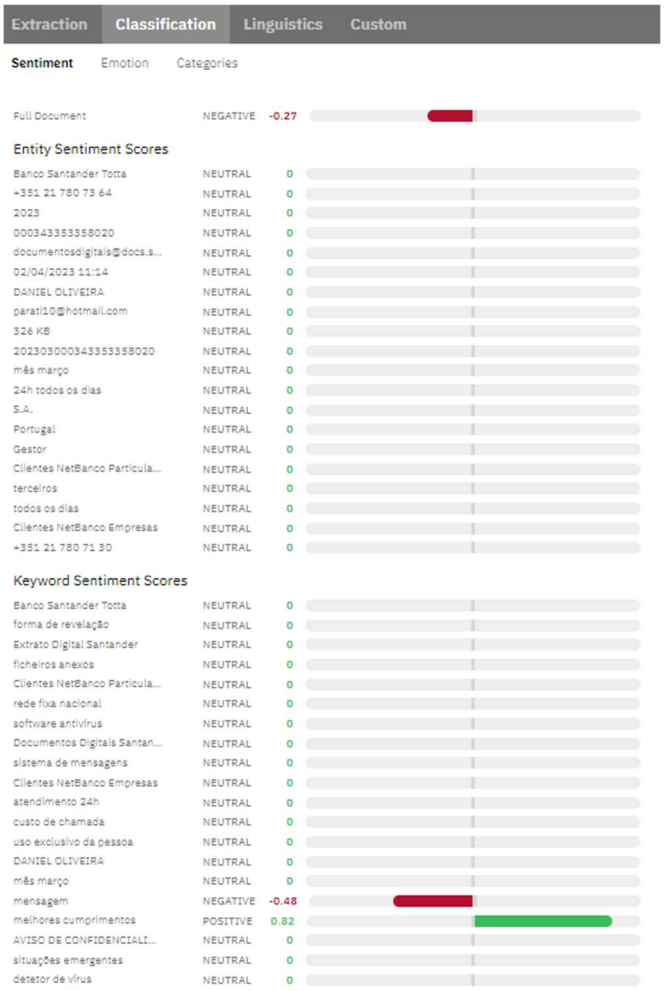


Figure 34 IBM Watson Classification results – Sentiment analysis for Santander extract

As seen in Figure 35, the Santander payment email Legit 9 is rated negatively with a score of -0.34. Positive words include "Santander," "www.Santander.pt," and " legal notice," whereas negative ones include "e-mail," "favor," and "operation."

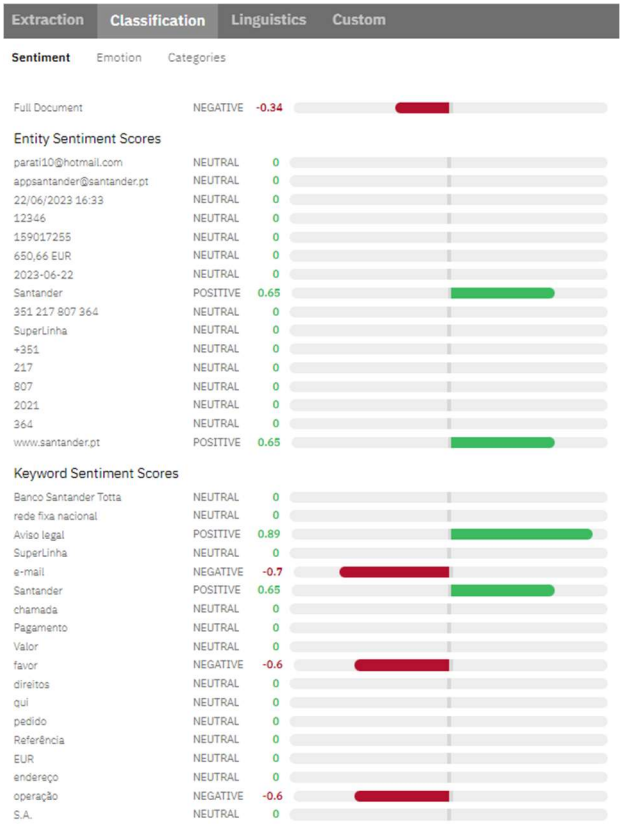


Figure 35 IBM Watson Classification results – Sentiment analysis for Santander payment

Since the IBM Watson website returned a blank page with no results, it was unable to extract any information from the legitimate email 10 from Galp.

IBM Watson rated the phishing email 1 from MAPFRE insurance as negative (-0.82 points). The word "service," which is featured on Figure 36, was used to create this classification.

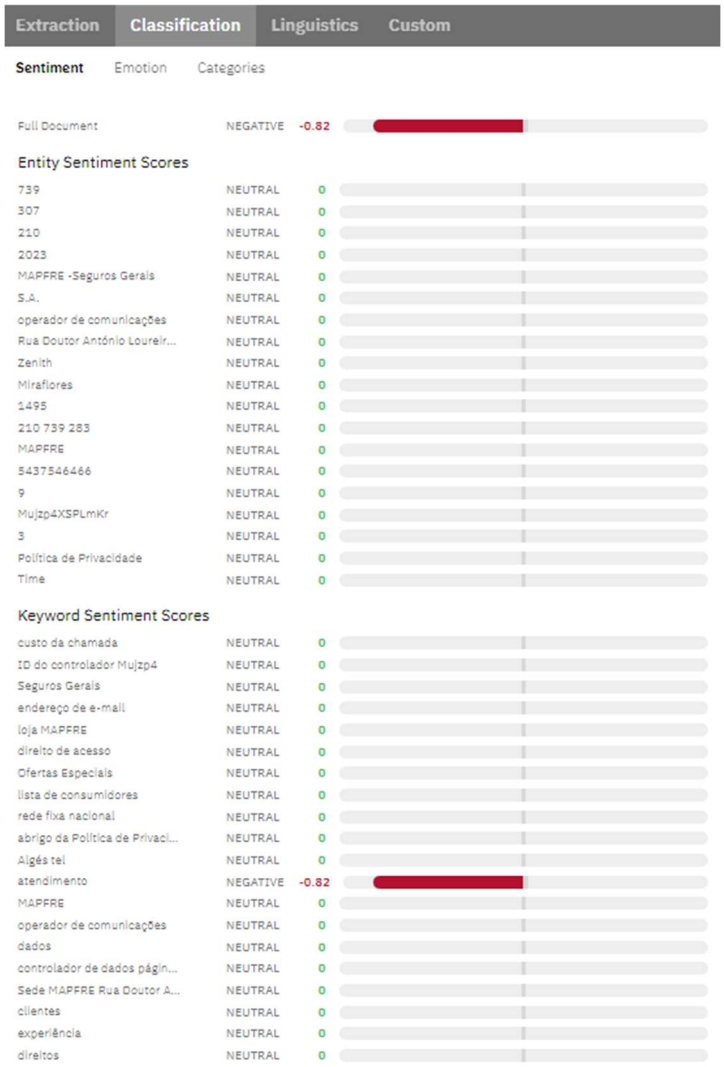


Figure 36 IBM Watson Classification results – Sentiment analysis for MAPFRE

According to Figure 37, the phishing email 2 from CTT is categorized as negative with -0.42 points. The following words were used to determine the outcome: "Correio - CTT", "corre@imagendental.com" and "CTT" were classed as positives, while "package" was classified as a negative.

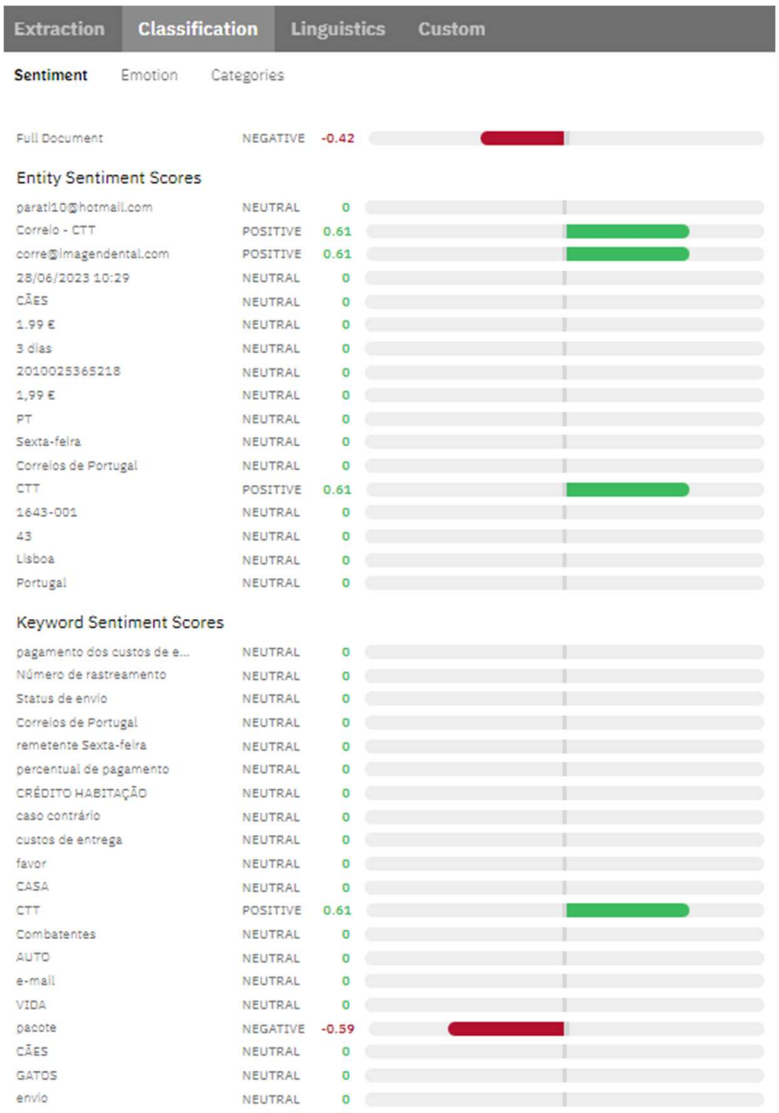


Figure 37 IBM Watson Classification results – Sentiment analysis for CTT Package

The Nescafé phishing email 3, Figure 38 receives a negative rating of -0.4 points. Even though this email had content in both German and Spanish, IBM Watson was able to execute the analysis because these languages were supported. The terms "Eden Hazard," "Neymar," " Brazil," "past weeks," "yellow and green national jersey," "World Cup group game," "tough duels," "snappy comment," "Action by the German national team," "Tottenham Hotspur," and "Bank suddenly" all played a part in the production of these results.

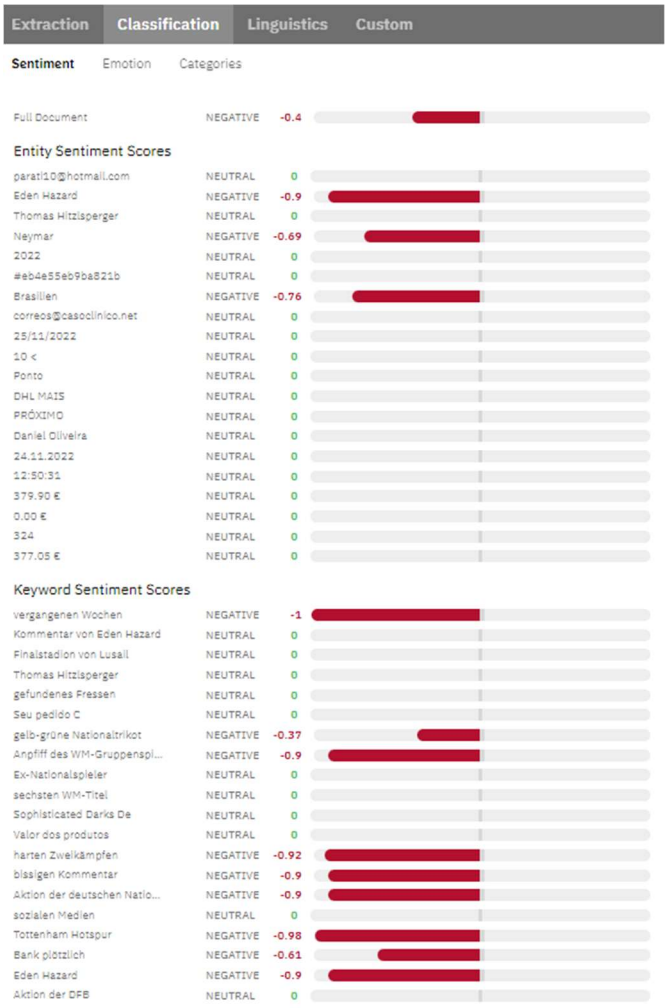


Figure 38 IBM Watson Classification results – Sentiment analysis for Nescafé

Analysis of Worten phishing email 4, on Figure 39, had a positive 0.69 rating. "You were," "monthly winner," "winner," "opportunity," and "congratulations" were the terms that led to this good rating. The terms "offer" and "enrollment" were negative.

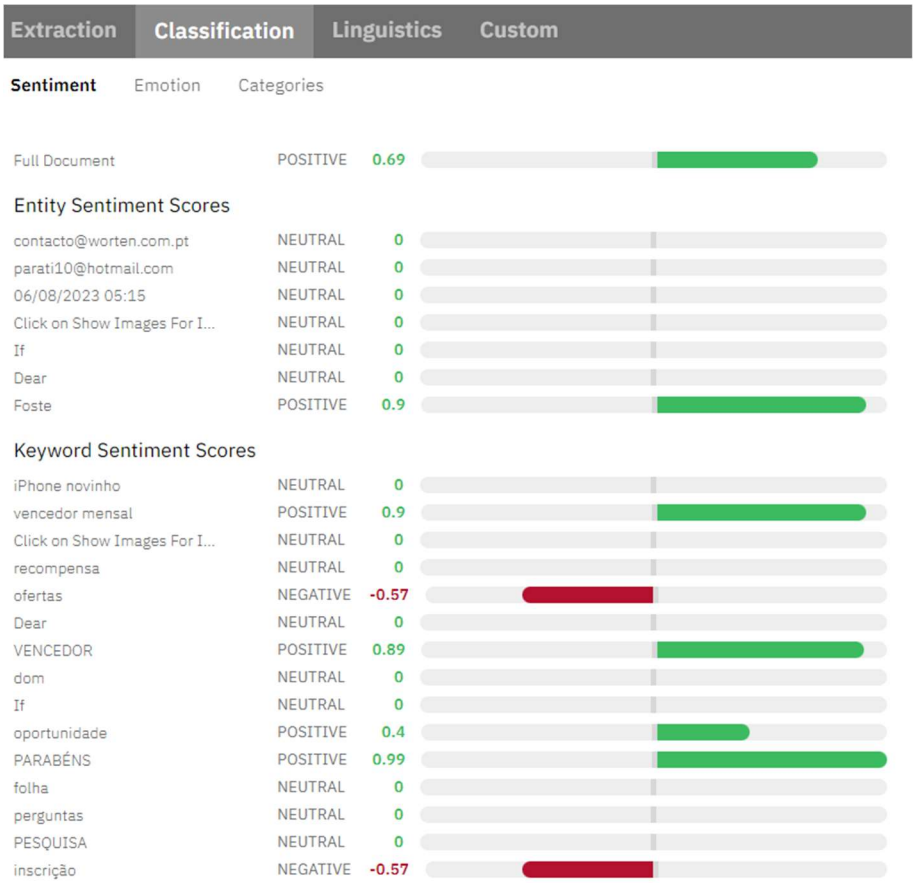


Figure 39 IBM Watson Classification results – Sentiment analysis for Worten

The phishing email 5 from Moey! was rated as positive on Figure 40**Erro! A origem da referência não foi encontrada.** with 0.32 points. "Counters" was a helpful contribution. The word "account" had a negative contribution.

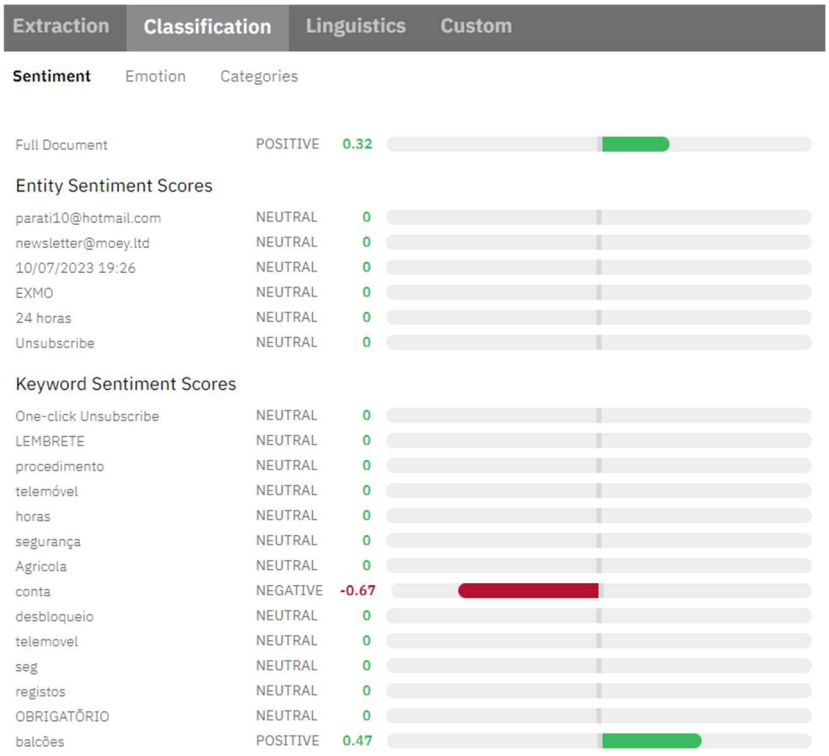


Figure 40 IBM Watson Classification results – Sentiment analysis for Moey!

With 0.45 points, the email from infoS (Figure 41) was rated as positive. The phrase "good afternoon" served as the classification for phishing email number 6 and was rated as favorable. As all the other keywords were neutral, this led to a classification that was overall positive.

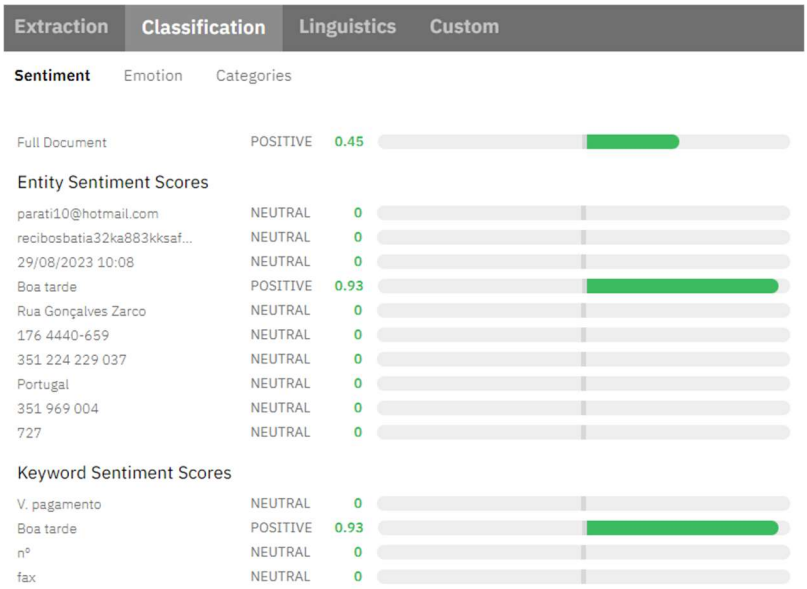


Figure 41 IBM Watson Classification results – Sentiment analysis for InfoS

The total punctuation score for the phishing email number 7 from the CTT incomplete address is -0.83, which is considered negative. The phrases "recent package," "Incomplete address," "phone number," "information," "contact," "package," and "missing" were responsible for this outcome, as seen in Figure 42 Figure 42.

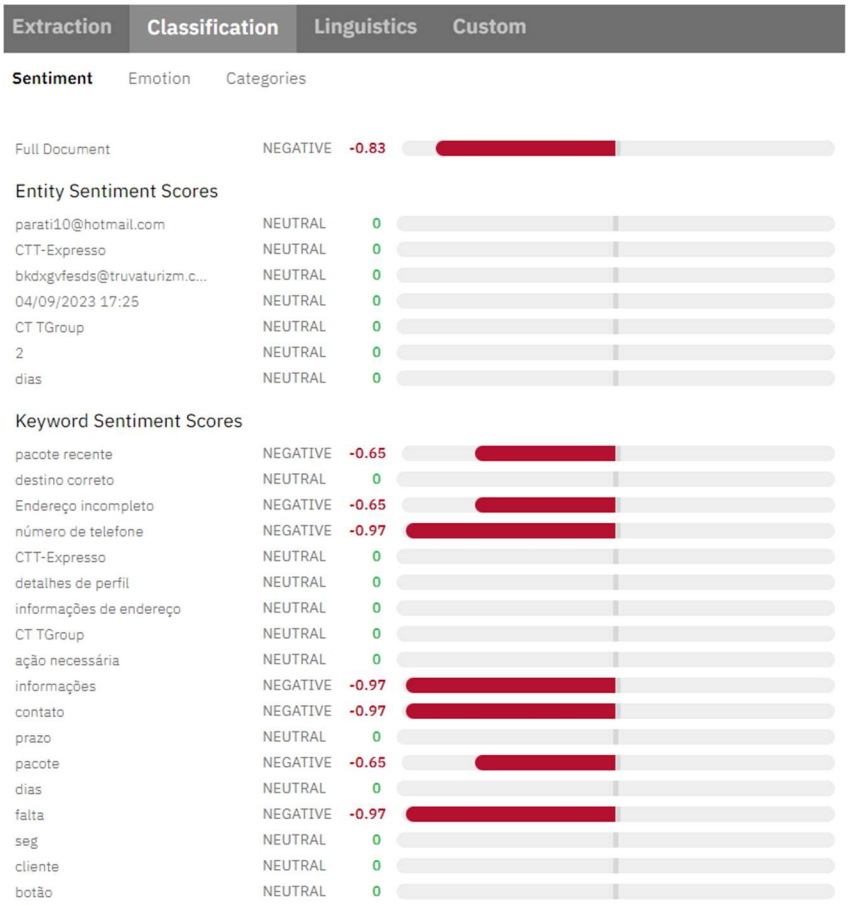


Figure 42 IBM Watson Classification results – Sentiment analysis for CTT Incomplete address

The IBM Watson website gave a blank page with no results, making it impossible to examine the email phishing number 8 from Nivea.

The phrases "continued trust" and "services" in phishing email number 9 from Info-Levarte SL were scored as positive with 0.61 points, as we can see on Figure 43.

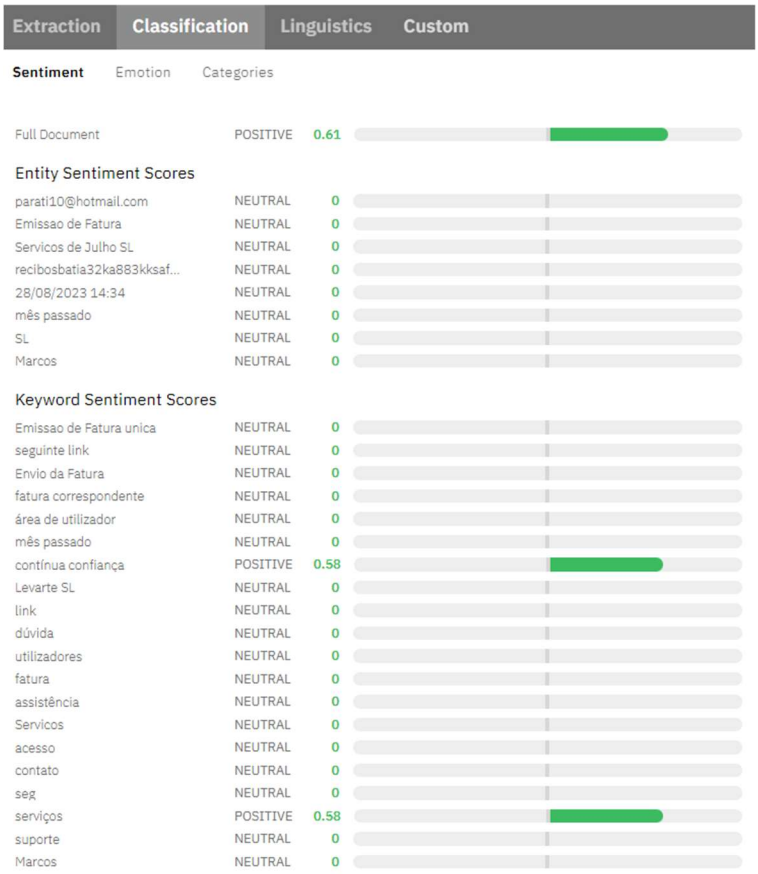


Figure 43 IBM Watson Classification results – Sentiment analysis for Info - Levarte SL

The IBM Watson algorithm classified the email from Grupo8—the phishing email 10—as negative with -0.37 points. The phrase "last days" was used as a positive statement, and the phrases "most innovative alert" and "advertiser" were used negatively, as shown on Figure 44.

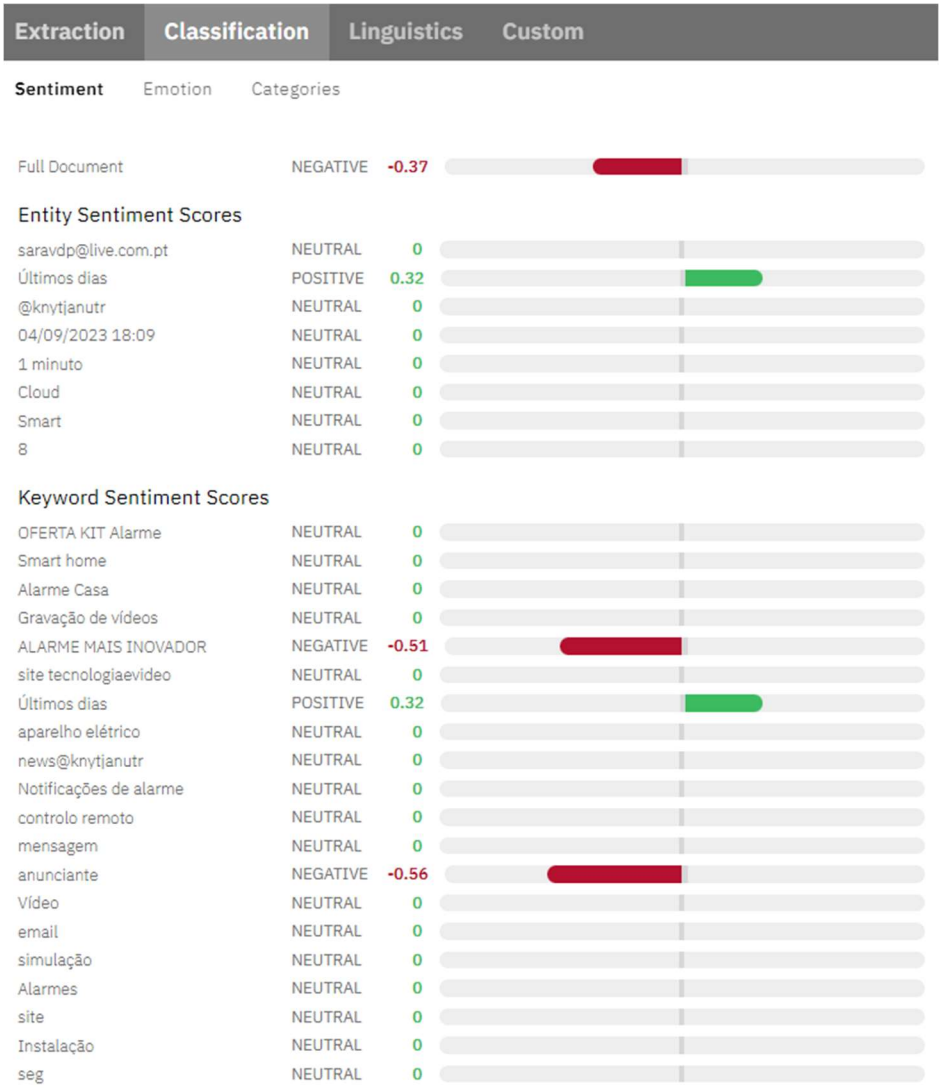


Figure 44 IBM Watson Classification results – Sentiment analysis for Grupo8

The polarity values are measured from -1 to 1. For legitimate emails, the average polarity is positive, standing at 0.11. In contrast, phishing emails exhibit a markedly negative average polarity, with a score of -0.77. Table 3 shows a summary of IBM Watson Analysis of Emails Polarity and Points.

Email ID	Sender	IBM Watson	
		Polarity	Points
Legit 1	CTT Tolls	Positive	0.92
Legit 2	Meu Presente	Positive	0.28
Legit 3	Moey! Password	Negative	-0.26
Legit 4	Techonline	Positive	0.82
Legit 5	Fidelidade	Negative	-0.41
Legit 6	Flixbus	Positive	0.53
Legit 7	Meo	Negative	-0.28
Legit 8	Santander Documents	Negative	-0.27
Legit 9	Santander Payment	Negative	-0.34
Legit 10	Galp	N/A	N/A
Phishing 1	MAPFRE	Negative	-0.82
Phishing 2	CTT	Negative	-0.42
Phishing 3	Nescafé	Negative	-0.4
Phishing 4	Worten	Positive	0.69
Phishing 5	Moey!	Positive	0.32
Phishing 6	Info S	Positive	0.45
Phishing 7	CTT Incomplete Address	Negative	-0.83
Phishing 8	Nivea	N/A	N/A
Phishing 9	Info Levarte SL	Positive	0.61
Phishing 10	Grupo8	Negative	-0.37

Table 3 - IBM Watson Analysis of Email Polarity and Points

4.1.2. Chat GPT

Analyzing the email **Legit 1, from CTT tolls** based on the 6 types of basic emotions, chat GPT can identify the following sentiments and emotions:

1. Fear: The email might evoke a mild sense of fear or concern. The subject line "Portagens a pagamento" (Toll payments) and the mention of "Tem portagens a pagamento no veículo" (You have toll payments on your vehicle) could potentially create a sense of worry about unpaid tolls.
2. Disgust: There is no content in the email that would typically trigger feelings of disgust.

3. Anger: The email could potentially trigger a mild feeling of annoyance or frustration if the recipient is not expecting toll payments or if there are any issues with the toll payment system. The mention of payments and the need to make payments might lead to a subtle sense of irritation.
4. Joy: There is no content in the email that would evoke feelings of joy.
5. Sadness: The email is unlikely to evoke feelings of sadness, as it is more focused on practical information about toll payments.
6. Surprise: The email could potentially evoke surprise if the recipient wasn't aware of the toll payments mentioned. The unexpected nature of such notifications might lead to a moment of surprise.

Overall, the primary emotions in this email appear to be a mix of fear, mild annoyance, and potential surprise. The email informs the recipient about toll payments that need to be made and encourages them to take action, which could lead to varying emotional responses based on individual circumstances.

Studying the **Legit email 2 from Meu Presente** , and analyzing the email based on the 6 types of basic emotions, chat GPT can identify the following sentiments and emotions:

1. Fear: There doesn't seem to be any explicit content or context in the email that would evoke fear.
2. Disgust: There is no content in the email that would typically trigger feelings of disgust.
3. Anger: The email does not contain any language or context that would likely provoke anger.
4. Joy: The email seems to convey a positive and reassuring tone. The phrases like "Temos o prazer de informar" (We are pleased to inform) and "Até breve" (See you soon) suggest a sense of positivity and excitement. The confirmation of the order being registered and the mention of "A equipa Meu Presente" (The Meu Presente team) further contribute to a joyful sentiment.
5. Sadness: There is no content in the email that would evoke feelings of sadness.
6. Surprise: While the email is informative and contains details about the order, there is no element that would elicit a strong sense of surprise.

Overall, the predominant emotions in this email seem to be joy and perhaps a mild sense of reassurance. The tone is positive and informative, focusing on the confirmation of the order and providing instructions for tracking and receiving the package.

When analyzing the **email legit 3 from Moey!** based on the 6 types of basic emotions, chat GPT can identify the following emotions:

1. Fear: The email is not likely to evoke fear. It is a routine reminder about a forgotten access code.
2. Disgust: The email content is not likely to trigger feelings of disgust.
3. Anger: The email is presented in a calm and helpful manner, so it is unlikely to provoke anger.

4. Joy: The email is not intended to evoke joy and does not contain content that would typically elicit joy.
5. Sadness: The email is not expected to trigger feelings of sadness.
6. Surprise: The email might evoke a mild sense of surprise due to its subject line and content. The phrase "Parece que te esqueceste do teu código de acesso à app" (It seems you forgot your access code to the app) could elicit a moment of surprise or realization for the recipient. The use of the 😌 emoji also suggests a mild, reassuring surprise.

Overall, the primary emotion conveyed in this email is surprise, along with a sense of reassurance. The email informs the recipient about the forgotten access code and provides steps to rectify the situation, which might lead to a moment of surprise followed by a practical response to the issue.

Analyzing the **email legit 4 from Techonline** based on the 6 types of basic emotions, chat GPT can identify the following emotions:

1. Fear: The email content is not likely to evoke fear. It appears to be a routine document notification.
2. Disgust: The email content is not expected to trigger feelings of disgust.
3. Anger: The email is informational and formal, so it is unlikely to provoke anger.
4. Joy: The email does not contain content that would typically elicit joy.
5. Sadness: The email is not likely to trigger feelings of sadness.
6. Surprise: The email might evoke a mild sense of surprise, especially if the recipient was not expecting to receive this specific document. The fact that a document is being sent could lead to a small moment of surprise, but this is likely to be a fleeting reaction.

Overall, the primary emotion conveyed in this email is a subtle sense of surprise, mostly due to the unexpected arrival of the document. The tone is neutral and formal, focusing on providing necessary information about the document for download.

Analyzing the **email legit 5 from Fidelidade** based on the 6 types of basic emotions, chat GPT can identify the following emotions:

1. Fear: The email content could potentially evoke a mild sense of fear or concern, especially due to the mention of financial information and "cobrança por débito direto" (direct debit collection). The recipient might worry about the amount to be debited from their bank account.
2. Disgust: The email content is not expected to trigger feelings of disgust.
3. Anger: The email could potentially trigger a mild feeling of frustration or irritation if the recipient has concerns about the mentioned debit or if they find the communication unclear. The financial nature of the content might lead to a subtle sense of annoyance.
4. Joy: The email is not likely to evoke feelings of joy. It contains practical and informational content.

5. Sadness: The email is not likely to trigger feelings of sadness.
6. Surprise: The email might evoke a mild sense of surprise due to the unexpected arrival of the notification about new insurance documents. The mention of the "Débito na conta bancária" (Debit from the bank account) and the specific amount might also lead to a brief moment of surprise.

Overall, the primary emotions conveyed in this email are a mix of potential mild fear, slight surprise, and perhaps a hint of irritation or concern due to the financial details mentioned. The tone is informative and functional, focusing on notifying the recipient about new insurance documents and relevant financial information.

Analyzing the email **Legit 6, from Flixbus,** based on the 6 types of basic emotions, chat GPT can identify the following emotions:

1. Fear: The email is not likely to evoke fear. It has a positive and encouraging tone.
2. Disgust: The email content is not expected to trigger feelings of disgust.
3. Anger: The email is presented in a light and motivational manner, so it is unlikely to provoke anger.
4. Joy: The email is likely to evoke feelings of joy and excitement. Sentences like "Deixa a aventura saltar do grande ecrã!" (Let the adventure jump off the big screen!) and "É hora de seres tu o ator principal na tua estória" (It is time for you to be the main actor in your story) suggest a sense of adventure and empowerment. The idea of creating one own stories and exploring new destinations can bring about feelings of joy and enthusiasm.
5. Sadness: The email is not likely to trigger feelings of sadness.
6. Surprise: The email could evoke a mild sense of surprise due to its unexpected nature, especially if the recipient hasn't interacted with FlixBus before. The encouragement to switch from cinema trips to personal adventures might also lead to a moment of surprise or intrigue.

Overall, the primary emotion conveyed in this email is joy. The tone is motivational and inspiring, encouraging the recipient to embrace new adventures and experiences. The idea of being the protagonist of one own story and the promise of memorable adventures are likely to resonate positively with the reader.

Studying the email **Legit 7, from MEO** , based on the 6 types of basic emotions, chat GPT can identify the following emotions:

1. Fear: The email content is not likely to evoke fear. It is a routine notification about a bill.
2. Disgust: The email content is not expected to trigger feelings of disgust.
3. Anger: The email is factual and informative, so it is unlikely to provoke anger.
4. Joy: The email is not intended to evoke joy and does not contain content that would typically elicit joy.

5. Sadness: The email is not likely to trigger feelings of sadness
6. Surprise: The email could potentially evoke a mild sense of surprise due to the unexpected arrival of the notification or if the recipient was not anticipating this specific bill. Additionally, the mention of "213 MEOS" and "8 MEOS a perder validade" could lead to a brief moment of surprise or intrigue, especially if the recipient is not familiar with these terms.

Overall, the primary emotion conveyed in this email is a mix of potential mild surprise and informational content. The tone is neutral and focused on providing necessary information about the electronic invoice and its details.

In the analysis of the email **legit 8 from Santander with documents** , based on the 6 types of basic emotions, chat GPT can identify the following emotions:

1. Fear: The email content is not likely to evoke fear. It is a routine communication about a financial statement.
2. Disgust: The email content is not expected to trigger feelings of disgust.
3. Anger: The email is informative and formal, so it is unlikely to provoke anger.
4. Joy: The email is not intended to evoke joy and does not contain content that would typically elicit joy.
5. Sadness: The email is not likely to trigger feelings of sadness.
6. Surprise: The email might evoke a mild sense of surprise due to the unexpected arrival of the financial statement or if the recipient was not anticipating it. Additionally, the inclusion of specific information such as the document title "Extrato Digital Santander," the reference number "2023-03-31 000343353358020," and the attached PDF could lead to a moment of surprise.

Overall, the primary emotion conveyed in this email is informational. The tone is formal and focused on providing the recipient with their financial statement and the necessary contact information for assistance.

Analyzing the **email legit 9 from Santander, related to a payment,** based on the 6 types of basic emotions, chat GPT can identify the following emotions:

1. Fear: The email content might evoke a mild sense of fear or concern, especially due to the mention of a payment of a significant amount ("Valor 650,66 EUR") and the possibility of an unrecognized operation. The recipient might worry about the legitimacy of the payment.
2. Disgust: The email content is not expected to trigger feelings of disgust.
3. Anger: The email could potentially trigger a mild feeling of frustration or irritation if the recipient did not initiate the payment or if there are concerns about the operation accuracy. The mention of the payment and the need to contact SuperLinha might lead to a subtle sense of annoyance.

4. Joy: The email is not intended to evoke joy and does not contain content that would typically elicit joy.
5. Sadness: The email is not likely to trigger feelings of sadness.
6. Surprise: The email could evoke a mild sense of surprise due to the unexpected arrival of the payment notification. Additionally, the mention of the specific payment details, such as the amount and reference numbers, could lead to a brief moment of surprise or intrigue.

Overall, the primary emotions conveyed in this email are a mix of potential mild fear, slight surprise, and perhaps a hint of irritation or concern due to the financial transaction and the mention of potential unrecognized operations. The tone is informative and functional, focusing on notifying the recipient about the payment and providing necessary contact information for clarification.

Analyzing the **email legit 10, from Galp** , based on the 6 types of basic emotions, chat GPT can identify the following emotions:

1. Fear: The email content is not likely to evoke fear. It is an advertising email about a discount offer.
2. Disgust: The email content is not expected to trigger feelings of disgust.
3. Anger: The email is promotional and informative, so it is unlikely to provoke anger.
4. Joy: The email is likely to evoke feelings of joy and excitement. The phrases like "Aproveite o cupão de desconto," (Take advantage of the discount coupon) and "vá de férias com combustível aditivado Evologic 2.0 ao preço do combustível simples" (go on vacation with Evologic 2.0 fuel at the price of regular fuel) suggest a sense of saving money and getting a good deal. The offer of a discount and the chance to save on fuel costs could lead to feelings of joy and enthusiasm.
5. Sadness: The email is not likely to trigger feelings of sadness.
6. Surprise: The email could evoke a mild sense of surprise due to the unexpected nature of the offer or if the recipient was not anticipating such a promotion. The mention of exclusive discounts and the new Simulador de Poupança Evologic 2.0 could lead to a moment of intrigue and interest.

Overall, the primary emotion conveyed in this email is joy. The tone is promotional and inviting, focusing on a discount offer for fuel and encouraging the recipient to take advantage of the deal. The offer to save money and the potential benefits of using Evologic 2.0 fuel are likely to resonate positively with the reader.

Analyzing the **phishing email 1, from MAPFRE** , based on the 6 types of basic emotions, chat GPT can identify the following emotions:

1. Fear: The email content could potentially evoke a mild sense of fear or concern, especially due to the mention of insurance and security ("VIAJE EM SEGURANÇA"). Additionally, the mention of "recebendo a mensagem" and "seus dados" might lead to slight apprehension for privacy or data security.
2. Disgust: The email content is not expected to trigger feelings of disgust.
3. Anger: The email is promotional and informational, so it is unlikely to provoke anger.
4. Joy: The email is not intended to evoke joy and does not contain content that would typically elicit joy.
5. Sadness: The email is not likely to trigger feelings of sadness.
6. Surprise: The email could evoke a mild sense of surprise due to the unexpected arrival of the email and its mention of a "Seguro automóvel em segundos" (car insurance in seconds). The mention of receiving a message based on being in a consumer list might also lead to a moment of surprise or curiosity.

Overall, the primary emotions conveyed in this email are a mix of potential mild fear, slight surprise, and perhaps a hint of curiosity or concern regarding data privacy. The tone is promotional, focused on encouraging the recipient to simulate their car insurance and emphasizing safety. The additional note about data control and privacy rights is also present in the email.

Analyzing the **phishing email 2, allegedly from CTT,** based on the 6 types of basic emotions, chat GPT can identify the following emotions:

1. Fear: The email content could evoke a mild sense of fear or concern. The mention of a package awaiting delivery, a required payment, and a time limit for confirmation might create a sense of urgency and worry that the recipient could lose their package if they don't take action.
2. Disgust: The email content is not likely to trigger feelings of disgust.
3. Anger: The email could potentially trigger a mild feeling of frustration or annoyance if the recipient is unsure about the mentioned package, payment, or if they find the communication unclear. The mention of payment and potential consequences could lead to a subtle sense of irritation.
4. Joy: The email is not intended to evoke joy and does not contain content that would typically elicit joy.
5. Sadness: The email is not likely to trigger feelings of sadness.

6. Surprise: The email could evoke a mild sense of surprise due to the unexpected arrival of the notification about a package. The mention of a required payment and the need to confirm it within a specified time frame might also lead to a moment of surprise or intrigue.

Overall, the primary emotions conveyed in this email are a mix of potential mild fear, slight surprise, and perhaps a hint of irritation or concern due to the financial transaction and the time-sensitive nature of the communication. The tone is informative and functional, focusing on notifying the recipient about the package and its associated payment requirements.

Analyzing the **phishing email 3, apparently from Nescafé** , based on the 6 types of basic emotions, chat GPT can identify the following emotions:

1. Fear: The email content is not likely to evoke fear.
2. Disgust: The email content is not expected to trigger feelings of disgust.
3. Anger: The email is not likely to provoke anger.
4. Joy: The email is likely to evoke feelings of joy and excitement. The phrases like "Seu pedido #eb4e55eb9ba821b foi feito!" (Your order #eb4e55eb9ba821b is made!) and "Resumo da encomenda" (Order summary) suggest a successful purchase and the acquisition of products through a point-based system. The mention of discounted prices ordered products, and a successful order could lead to feelings of satisfaction and joy.
5. Sadness: The email is not likely to trigger feelings of sadness.
6. Surprise: The email could evoke a mild sense of surprise due to the unexpected arrival of the email and its content about the order confirmation and details. The user might not have been anticipating the email or the order update.

Overall, the primary emotion conveyed in this email is joy. The tone is informative and celebratory, focusing on the successful order placement and the details of the purchased products. The information about the order and its status is presented in a straightforward manner.

Analyzing the **phishing email 4, from Worten** , based on the 6 types of basic emotions, chat GPT can identify the following emotions:

1. Fear: The email content could potentially evoke a mild sense of fear or concern, especially considering the context of online scams and phishing attempts. The unexpected announcement of being a "vencedor" (winner) and the request to respond to questions might raise suspicion about the legitimacy of the email.

2. Disgust: The email content is likely to trigger feelings of disgust, especially if the recipient views it as a potential scam or deceptive marketing strategy. The mention of winning a prize and being asked to respond to questions might be seen as manipulative or misleading.
3. Anger: The email could potentially provoke a mild feeling of annoyance or anger if the recipient perceives it as spam, a scam, or an attempt to deceive them. The mention of being a "vencedor" (winner) might lead to frustration if the recipient believes it to be false or misleading.
4. Joy: The email is designed to evoke feelings of joy and excitement by claiming that the recipient is a "vencedor" (winner) and has been chosen for a prize, specifically a new iPhone. However, given the context of online scams, recipients might be cautious about feeling genuine joy based on this email.
5. Sadness: The email is not likely to trigger feelings of sadness.
6. Surprise: The email could evoke a sense of mild surprise due to the unexpected announcement of being a winner and the offer of a new iPhone. However, given the prevalence of online scams, recipients might approach this surprise with caution and skepticism.

Overall, the primary emotions conveyed in this email are a mix of potential mild fear, skepticism, and perhaps some curiosity or intrigue regarding the unexpected claim of being a winner. The tone is celebratory and enticing, aiming to encourage the recipient to engage with the content, but the potential for skepticism is also present.

Studying the **phishing email 5, from Moey!,** based on the 6 types of basic emotions, chat GPT can identify the following emotions:

1. Fear: The email content is likely to evoke a sense of fear or alarm due to the message indicating that the recipient account has been temporarily blocked. The urgency of the situation, along with the potential consequences of not confirming the mobile phone, could cause the recipient to feel anxious or fearful.
2. Disgust: The email content is not likely to trigger feelings of disgust.
3. Anger: The email content could provoke feelings of mild anger or frustration if the recipient perceives the account blockage and the request for mobile phone confirmation as an inconvenience or a hassle.
4. Joy: The email is not likely to evoke feelings of joy.

5. Sadness: The email is not likely to trigger feelings of sadness.
6. Surprise: The email might evoke mild surprise due to the unexpected claim that the recipient account has been temporarily blocked and the urgency to confirm their mobile phone. However, the recipient might also approach this surprise with skepticism given the prevalence of phishing and scam attempts.

Overall, the primary emotions conveyed in this email are potential fear and caution due to the urgency of the situation and the request to confirm the mobile phone to unblock the account. The emails tone is serious and urgent, aiming to prompt the recipient to take immediate action. Recipients might also approach the email with a degree of skepticism, given the prevalence of phishing and scam emails that use similar urgency tactics.

Analyzing the emotions based on the 6 types of basic emotions, the **phishing email 6 from infoS** doesn't strongly evoke any of these emotions:

1. Fear: The emails content is not likely to trigger fear, as it appears to be a simple receipt or confirmation of payment. There are no elements that suggest a threat or danger.
2. Disgust: The emails content is not likely to trigger feelings of disgust. It contains standard payment confirmation information.
3. Anger: The emails content is not likely to provoke anger, as it is a routine payment confirmation message.
4. Joy: The email does not convey a sense of joy, but it is not intended to do so. It is a formal and straightforward message.
5. Sadness: The email does not evoke feelings of sadness. It is a factual and neutral communication.
6. Surprise: The email does not contain elements that would trigger surprise. It provides standard payment information and contact details.

In summary, this email is informational and transactional in nature, lacking emotional elements. It is a standard payment receipt with no emotional cues.

Studying the **phishing email 7, allegedly from CTT,** based on the 6 types of basic emotions, chat GPT can identify the following:

1. Fear: The emails content may evoke a sense of fear or concern as it informs the recipient that their recent package cannot be delivered due to an incomplete address. This could make the recipient anxious about the potential consequences of not receiving the package.
2. Disgust: The emails content is not likely to trigger feelings of disgust.
3. Anger: The emails content could potentially provoke mild frustration or anger if the recipient perceives the inconvenience caused by the incomplete address issue. The request for action may be seen as an annoyance.
4. Joy: The email is not likely to evoke feelings of joy. It addresses a problem rather than offering a positive experience.
5. Sadness: The emails content is not likely to trigger feelings of sadness. It is more of an informational and practical nature.
6. Surprise: The email might evoke mild surprise due to the unexpected message informing the recipient about the incomplete address issue. However, this surprise is unlikely to be a positive one, as it presents a problem to be resolved.

Overall, the primary emotions conveyed in this email are related to potential fear, caution, and mild frustration due to the urgency of the situation and the request to update address information to ensure timely package delivery. The email tone is practical and action-oriented, aimed at prompting the recipient to take immediate steps to resolve the issue.

Studying the **phishing email 8 from Nivea** "Teste em casa produtos Nivea e fique com eles para si!" (Test Nivea products at home and keep them for yourself!) based on the 6 types of basic emotions, chat GPT can identify the following:

1. Fear: The email content is not likely to evoke feelings of fear. It appears to offer a testing opportunity rather than posing a threat.
2. Disgust: The email content is not likely to trigger feelings of disgust.
3. Anger: The email content is not likely to provoke feelings of anger. It offers a chance to test Nivea products, which is generally a positive opportunity.
4. Joy: The email content is more likely to evoke feelings of joy. It presents the opportunity for the recipient to test and potentially receive Nivea products, which could be seen as a rewarding and enjoyable experience.

5. Sadness: The email content is not likely to trigger feelings of sadness. It offers an exciting opportunity rather than conveying sad news.
6. Surprise: The email might evoke mild surprise due to the unexpected offer of testing and potentially receiving Nivea products. However, this surprise is more likely to be seen as a pleasant and positive surprise.

Overall, the primary emotion conveyed in this email is joy, as it presents an opportunity for the recipient to test and potentially receive Nivea products. The email aims to generate interest and enthusiasm about the testing opportunity, and its tone is generally positive and inviting.

Studying the **phishing email 9 from Info Levarte SL** "Envio da Fatura: Emissão de Fatura única para Serviços de Julho SL" (Invoice Delivery: Issuance of a Single Invoice for July Services SL) based on the 6 types of basic emotions, chat GPT can identify the following:

1. Fear: The email content is not likely to evoke feelings of fear. It is a routine message about an invoice and does not contain any alarming or threatening information.
2. Disgust: The email content is not likely to trigger feelings of disgust.
3. Anger: The email content is not likely to provoke feelings of anger. It is a standard notification about an invoice, and its tone is neutral.
4. Joy: The email is not intended to evoke feelings of joy. It conveys straightforward information about accessing an invoice.
5. Sadness: The email content is not likely to trigger feelings of sadness. It provides practical information and does not convey any sad news.
6. Surprise: The email is not likely to evoke feelings of surprise. It follows a typical format for notifying users about their invoices, and its content is expected.

Overall, the primary emotion conveyed in this email is neutrality. It serves as a routine notification about the availability of an invoice for services rendered in July. The tone is professional and straightforward, aiming to inform users rather than evoke strong emotional responses.

Studying the **phishing email 10 from GRUPO8** "OFERTA KIT Alarme + Instalação. Últimos dias! Saiba mais." (OFFER Alarm Kit + Installation. Last days! Learn more.) based on the 6 types of basic emotions, chat GPT can identify the following:

1. Fear: The email content is likely to evoke a sense of fear or concern because it promotes the use of an alarm system, implying a potential threat or danger that users might want protection from. The urgency expressed with "Últimos dias!" (Last days!) may also create a sense of unease.
2. Disgust: The email content is not likely to trigger feelings of disgust. It provides information about a product or service without any offensive or repulsive elements.
3. Anger: The email content is not likely to provoke feelings of anger. It is primarily promotional and informational, and its tone is not aggressive.
4. Joy: The email is not designed to evoke feelings of joy. It is a promotional message focused on security solutions and may not be associated with joyful emotions.
5. Sadness: The email content is not likely to trigger feelings of sadness. It provides information about an alarm system and installation without conveying any sad news.
6. Surprise: The email might evoke feelings of mild surprise due to the offer of an alarm kit and installation. The phrase "Últimos dias!" (Last days!) might also create a sense of urgency and surprise for some recipients.

Overall, the primary emotions conveyed in this email are related to security and potentially fear or concern due to the promotion of an alarm system. The email tone is promotional and focused on encouraging recipients to learn more about the offered product.

In sum, the Figure 45 displays the frequency of different emotions as retrieved by ChatGPT for both phishing and legitimate emails. Notably, phishing emails evoke slightly higher counts of emotions like fear, anger, and surprise compared to legitimate emails. However, legitimate emails exhibit more instances of joy. Interestingly, the AI model rarely detected sadness in either category, and it classified some phishing emails as having "no emotion." These findings highlight the variations in emotional content between phishing and legitimate emails as interpreted by an AI model, underscoring the nuances in the valence of email communication.

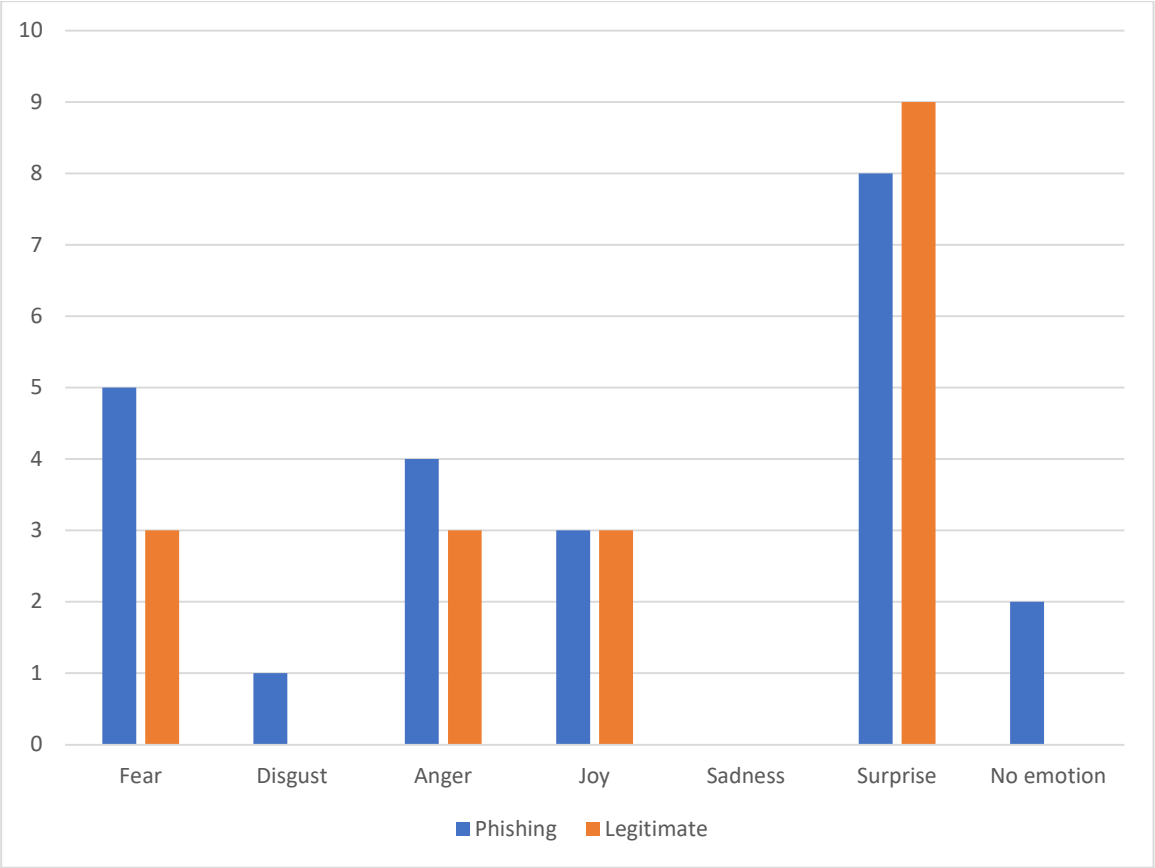


Figure 45 Frequency of emotions in retrieved by ChatGPT for Phishing emails vs Legitimate emails

IBM Watson focuses on evaluating whether the words within an email convey a positive, negative, or neutral sentiment, which makes it less suitable for determining if an email is phishing or not. Analyzing complete phrases could potentially lead to an improvement, and this avenue deserves further investigation. Regarding ChatGPT, it appears to be more precise in determining emotions and linking these emotions to the classification of emails as phishing or legitimate. In this case, it becomes necessary to conduct a more extensive experiment to ascertain whether ChatGPT can be a valuable tool in deciding whether an email is phishing or not.

4.2. Inquiry results

This chapter delves into the outcomes of the conducted survey, unveiling findings pertaining to the alignment in the recognition of legitimate and phishing emails, the associations between personality traits, emotional responses as well as the evaluation of valence, activation, emotions, and decision patterns.

4.2.1. Legitimate identification concordance

The Kappa statistic, on Table 4 which quantifies the level of agreement among 10 evaluators when assessing 54 effective subjects, is found to be 0.2517. This value suggests a fair degree of consensus (Cohen interpretation) among the evaluators in their evaluations of legitimate emails.

The standard error (0.0203) associated with the Kappa statistic provides an estimate of the variability or precision in the agreement measurement. Smaller standard errors typically indicate a more stable and reliable measure.

The Z-statistic, which stands at 12.4052, indicates how many standard errors the Kappa value is from the expected value under the null hypothesis. This high Z-statistic is a strong indicator that the observed agreement is significantly different from what one would expect by random chance.

The "SIG." value is reported as 0.0E0, signifying an extremely low p-value. This implies that the observed global agreement is statistically significant and is unlikely to occur due to random chance alone.

The 95% asymptotic confidence interval, ranging from 0.2119 to 0.2914, provides a measure of the precision of the Kappa estimate. Importantly, this interval does not encompass zero, which reinforces the statistical significance of the observed global agreement.

In summary, the Kappa statistic demonstrates a statistically fair level of global agreement among the 10 evaluators in assessing legitimate emails. The high Z-statistic, low p-value, and the confidence interval that excludes zero collectively underscore the strength of this consensus, which is a valuable finding, particularly in scenarios involving multiple evaluators and assessments.

Global agreement [a]						
		Asymptotic			95% asymptotic confidence interval	
	Kappa	**Standard error**	**z**	**SIG.**	**Inferior limit**	**Upper limit**
Global agreement	0.251652	0.020286	12.405201	0.0E0	0.211892	0.291412
[a]The sample data contains 54 effective subjects and 10 evaluators.						

Table 4 - Analysis of Global Agreement Legitimate emails

4.2.2. Phishing Identification concordance

The Kappa statistic, in Table 5 a measure of agreement among 10 evaluators in the assessment of 54 effective subjects in legitimate emails, produced a value of 0.2279. This value indicates a fair degree of agreement among the evaluators.

The standard error, at 0.0203, offers an estimate of the precision or variability in the agreement measurement. Smaller standard errors are associated with more stable and reliable measures.

The Z-statistic, with a value of 11.2361, highlights how many standard errors the Kappa value deviates from the expected value under the null hypothesis. This relatively high Z-statistic underscores that the observed agreement is significantly different from what would be anticipated by random chance.

The "SIG." value, reported as 0.0E0, reflects an exceedingly low p-value. This signifies that the observed global agreement is statistically significant and unlikely to have occurred randomly.

The 95% asymptotic confidence interval, ranging from 0.1882 to 0.2677, provides an estimate of the precision of the Kappa statistic. Importantly, this interval does not include zero, further emphasizing the statistical significance of the observed global agreement.

In summary, the Kappa statistic reveals a statistically fair global agreement among the 10 evaluators when assessing legitimate emails. The high Z-statistic, low p-value, and the confidence interval that excludes zero collectively underscore the strength of this consensus, a valuable finding, especially when involving multiple evaluators and assessments.

Global agreement [a]						
		Asymptotic			95% asymptotic confidence interval	
	Kappa	**Standard error**	**z**	**SIG.**	**Inferior limit**	**Upper limit**
Global agreement	0.227935	0.020286	11.236082	0.0E0	0.188176	0.267695
[a]The sample data contains 54 effective subjects and 10 evaluators.						

Table 5 - Analysis of Global Agreement Phishing emails

4.2.3. Valence evaluation

Table 6 displays the results of a statistical test known as a Student T test to understand the relationship between email valence (Valencia_Legitima and Valencia_Phishing) and phishing classification.

- Valencia_Legitima (Legitimate Email Valence): The mean valence score for legitimate emails was 3.08, with a standard deviation of 0.35. The test result indicates a significant relationship between legitimate email valence and phishing classification (p = 0.001).
- Valencia_Phishing (Phishing Email Valence): The mean valence score for phishing emails was 3.27, with a standard deviation of 0.40.

	Mean	n	Standard deviation	t	gl	Student T test
Valencia_Legitima	3,077778	54	0,353242	-3,368257	53	0,001416
Valencia_Phishing	3,270370	54	0,401708			

Table 6 - Student T test of Email Valence and Phishing Classification

4.2.4. Activation Evaluation

Table 7 displays the results of a statistical test, specifically a Student T test, to explore the relationship between email activation in legitimate emails and phishing emails.

- Ativaçao_Legitima (Legitimate Email Activation): The mean activation score for legitimate emails was 3.06, with a standard deviation of 0.27. The test result indicates a significant association between the activation level of legitimate emails and phishing classification (p = 0.002).
- Ativaçao_Phishing (Phishing Email Activation): The mean activation score for phishing emails was 3.17, with a standard deviation of 0.30.

The statistical test suggests that there is a significant difference in activation levels between legitimate and phishing emails. This implies that the level of activation in emails might have an impact on their classification, which will be further discussed in the subsequent sections.

	Mean	n	Standard deviation	t	gl	Student T test
Ativaçao_Legitima	3,057407	54	0,272377	-3,246041	53	0,002031
Ativaçao_Phishing	3,168519	54	0,295133			

Table 7 - Email Activation and Phishing Classification

4.2.5. Emotion Evaluation

This graphic on Figure 46 displays the frequency of different emotions in both phishing and legitimate emails.

- Fear: Fear was observed in 43 phishing emails and 33 legitimate emails.

- Disgust: Disgust was way more frequent in phishing emails, with 31 occurrences, compared to only 4 instances in legitimate emails.
- Anger: Anger was present in 38 phishing emails and 12 legitimate emails.
- Joy: Joy was more prevalent in legitimate emails, with 46 instances, while 29 phishing emails show this emotion.
- Sadness: Sadness was identified in 15 phishing emails and 21 legitimate emails.
- Surprise: Surprise was observed in 46 phishing emails and 40 legitimate emails.
- No emotion: The majority of both phishing and legitimate emails did not express a specific emotion, with 338 and 384 occurrences, respectively.

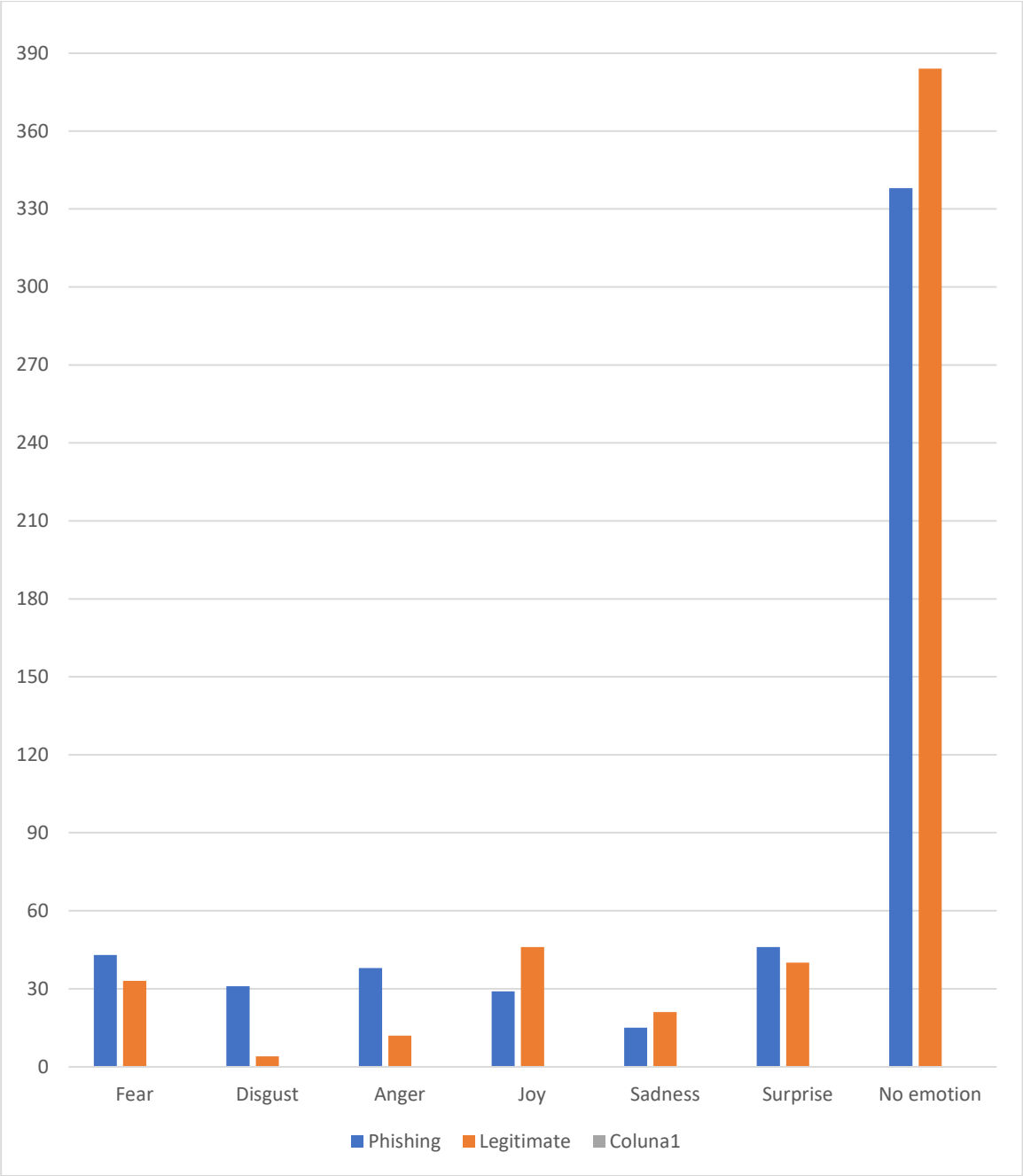


Figure 46 Emotional Frequency Comparison: Phishing vs. Legitimate Emails

4.2.6. Personality traits Evaluation

The table below summarizes the Point biserial correlation and their significance levels for various personality traits and the "Tipo_Email_Phishing" variable. The significance levels are denoted as follows: ** for a significant correlation at the 0.01 level (two-sided) and * for a significant correlation at the 0.05 level (two-sided). Table 8 provides an overview of the correlations between different personality traits and the likelihood of an email being classified as phishing. The sample size for each correlation was 54 participants. The following results were observed, with "r" representing the Pearson correlation coefficient:

• Neuroticism (N): These results were not statistically significant with the classification of emails as phishing (r = -0.047, p > 0.005).

• Extroversion (E): The correlation with extroversion and email classification was not statistically significant (r = -0.105, p > 0.005).

• Openness to Experience (O): No significant correlation was found between openness to experience and email classification (r = 0.190, p > 0.005).

• Agreeableness (A): No significant correlation was found between agreeableness and email classification (r = -0.071, p > 0.005).

• Conscientiousness (C): Similar to agreeableness, conscientiousness did not show a significant correlation with email classification (r = -0.071, p > 0.005).

• Vigilance (V): Vigilance also did not exhibit a significant correlation with email classification (r = 0.153, p > 0.005).

		Neuroticism (N)	Extroversion (E)	Openness to Experience (O)	Agreeableness (A)	Conscientiousness (C)	Vigilance(V)	Tipo_Email_Phishing
Neuroticism (N)	**Pearson correlation**	1	-0,422234 **	0,291549 *	-0,105913	-0,105913	0,012811	0,046571
	Sig. (bilateral)		0,001471	0,032434	0,445919	0,445919	0,926743	0,738077
	N	54	54	54	54	54	54	54
Extroversion (E)	**Pearson correlation**	-0,422234**	1	0,114211	0,169388	0,169388	0,102007	-0,105421
	Sig. (bilateral)	0,001471		0,410887	0,220768	0,220768	0,462966	0,448049
	N	54	54	54	54	54	54	54
	Pearson	0,291549 *	0,114211	1	0,038400	0,038400	0,168744	0,190097

Openness to Experience (O)	**correlation**							
	Sig. (bilateral)	0,032434	0,410887		0,782792	0,782792	0,222554	0,168566
	N	54	54	54	54	54	54	54
Agreeableness (A)	**Pearson correlation**	-0,105913	0,169388	0,038400	1	1,000000 **	0,026085	-0,071482
	Sig. (bilateral)	0,445919	0,220768	0,782792		0,0E0	0,851481	0,607498
	N	54	54	54	54	54	54	54
Conscientiousness (C)	**Pearson correlation**	-0,105913	0,169388	0,038400	1,000000 **	1	0,026085	-0,071482
	Sig. (bilateral)	0,445919	0,220768	0,782792	0,0E0		0,851481	0,607498
	N	54	54	54	54	54	54	54
Vigilance (V)	**Pearson correlation**	0,012811	0,102007	0,168744	0,026085	0,026085	1	0,153341
	Sig. (bilateral)	0,926743	0,462966	0,222554	0,851481	0,851481		0,268285
	N	54	54	54	54	54	54	54
Tipo_Email_Phishing	**Pearson correlation**	0,046571	-0,105421	0,190097	-0,071482	-0,071482	0,153341	1
	Sig. (bilateral)	0,738077	0,448049	0,168566	0,607498	0,607498	0,268285	
	N	54	54	54	54	54	54	54

**** The correlation is significant at the 0.01 level (two-sided).**

*** The correlation is significant at the 0.05 level (two-sided).**

Table 8 - Correlation Between Personality Traits and Email Classification

4.2.7. Decisional patterns evaluation

The Table 9 below summarizes the Pearson correlation coefficients and their significance levels for different decisional patterns and the phishing classification.

- Tipo_Email_Phishing: The "Tipo_Email_Phishing" variable itself shows no significant correlation with the other decisional patterns.
- Vigilance (V): Vigilance exhibits a low positive correlation with "Tipo_Email_Phishing," which is not statistically significant ($r = 0.153$, $p > 0.005$).
- Buck-passing (D): Buck-passing, displays a low positive correlation with "Tipo_Email_Phishing" ($r = 0.048$, $p > 0.005$).
- Procrastination (P): Procrastination reveals a low positive correlation with "Tipo_Email_Phishing" ($r = -0.188$, $p > 0.005$).
- Hypervigilance (Hv): Hypervigilance also exhibits a low positive correlation with "Tipo_Email_Phishing" ($r = -0.225$, $p > 0.005$).

		Tipo_Email_Phishing	Vigilance (V)	Buck-passing (D)	Procrastination (P)	Hypervigilance (Hv)
Tipo_Email_Phishing	Pearson correlation	1	0,153341	0,048315	-0,187671	-0,225395
	Sig. (bilateral)		0,268285	0,728639	0,174167	0,101272
	N	54	54	54	54	54
Vigilance (V)	Pearson correlation	0,153341	1	-0,140771	-0,089949	0,100607
	Sig. (bilateral)	0,268285		0,309956	0,517738	0,469160
	N	54	54	54	54	54
Buck-passing (D)	Pearson correlation	0,048315	-0,140771	1	0,681112 **	0,498215 **
	Sig. (bilateral)	0,728639	0,309956		1,4451E-8	0,000126
	N	54	54	54	54	54

Procrastination (P)	Pearson correlation	-0,187671	-0,089949	0,681112**	1	0,718096**
	Sig. (bilateral)	0,174167	0,517738	1,4451E-8		9,8233E-10
	N	54	54	54	54	54
Hypervigilance (Hv)	Pearson correlation	-0,225395	0,100607	0,498215**	0,718096**	1
	Sig. (bilateral)	0,101272	0,469160	0,000126	9,8233E-10	
	N	54	54	54	54	54

Table 9 - Correlation Between Personal Decisional Patterns and Email Phishing Classification

4.3. Result Analysis

The classification of emotions by ChatGPT diverges in several ways from the inquiry results, as depicted in the Figure 47 comparing ChatGPT's classifications to those obtained through inquiry for emotions like fear, disgust, anger, joy, sadness, surprise, and the absence of emotion. These findings will be further examined and discussed in the next chapter.

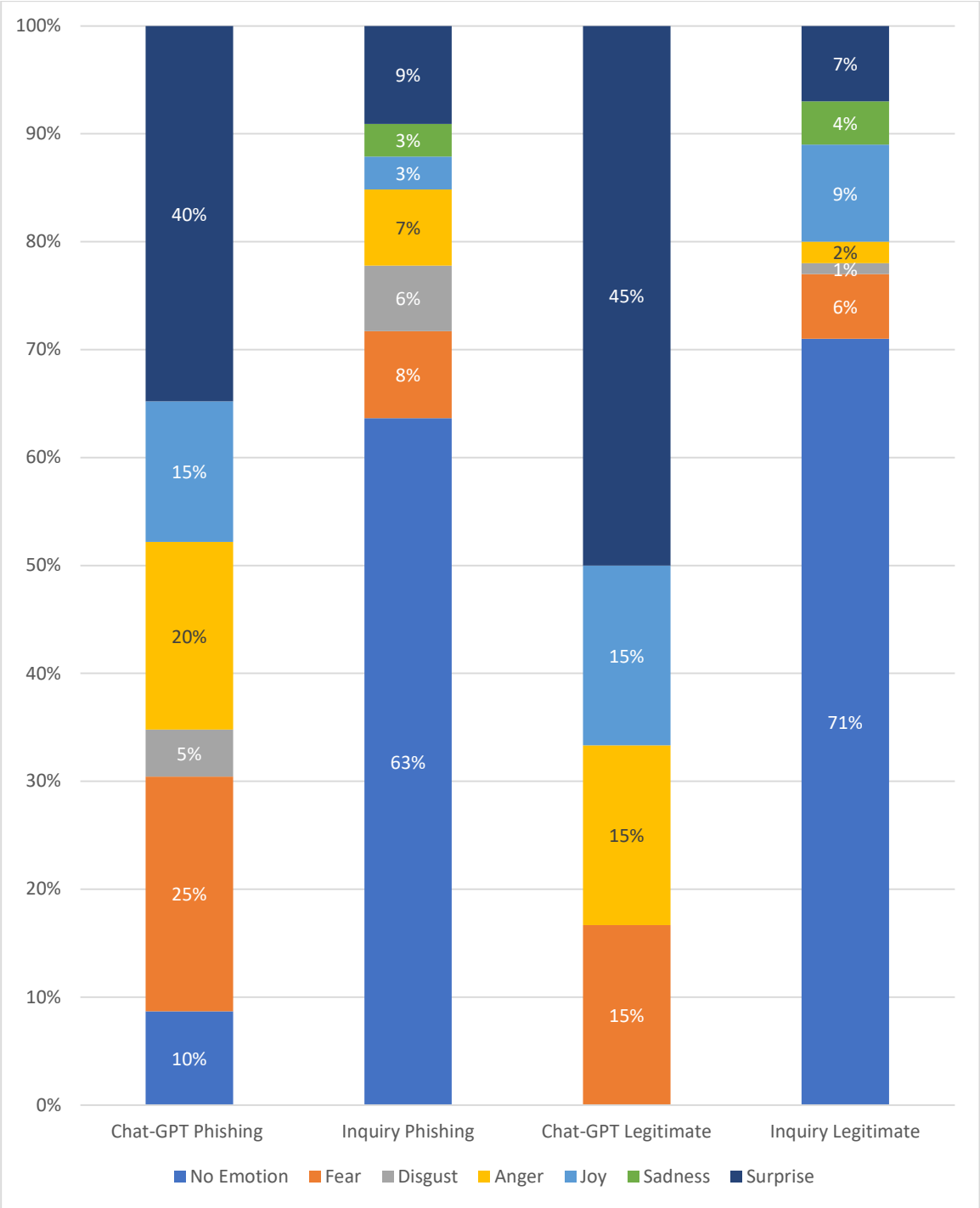


Figure 47 Summary of Chat-GPT vs Inquiry emotion classification

Chapter 5 – Discussion

This project seeks to undertake a comprehensive investigation into the realm of Phishing Email Classification, drawing insights from both Psychology and Machine Learning. To this end, our research objectives encompass several key facets, including participants' ability to discern legitimate emails from phishing ones, an analysis of emotional valence in these emails, examination of activation levels, identification of emotional responses, exploration of personality traits' potential influence, and investigation of decision-making patterns. By addressing these objectives, this study endeavors to advance our comprehension of the multifaceted factors that affect phishing email detection, ultimately contributing to the development of more robust mitigation strategies.

5.1. Polarity

The positive polarity on legitimate emails suggests that, on average, legitimate emails tend to have a slightly positive valence, as indicated by the positive polarity scores of 5 out of 10 emails. This may imply that legitimate emails often convey information or messages that elicit positive or neutral emotional responses from recipients.

Despite an equal frequency of 5 out of 10 emails with negative polarity, the magnitude of negativity in phishing emails is notably higher than in legitimate ones. This striking difference underscores the potential use of strongly negative emotional cues in phishing emails as a strategy to evoke specific reactions from recipients.

The divergence in polarity between legitimate emails and phishing emails highlights the emotional manipulation techniques employed in phishing attempts. While legitimate emails generally maintain a more neutral to positive valence, phishing emails utilize stronger negative emotional cues to trigger desired responses.

5.2. Concordance

For legitimate emails, a low level of agreement among the 10 evaluators indicates that their ratings are not arbitrary but rather reflective of a significant consensus. The small standard error further strengthens the results, signifying the stability and reliability of this agreement measurement. Essentially, this suggests that the observed consensus is not a result of random chance, thus enhancing the validity of the findings.

Conversely, when assessing phishing emails, the Kappa statistic reveals a significantly lower level of agreement. The Kappa value points to a minimal consensus among the 10 evaluators. Similar to legitimate emails, the small standard error confirms the reliability and stability of the agreement measurement. This implies that the observed agreement is not subject to random fluctuations, thereby reinforcing the findings.

In light of these results, one can interpret these weak agreement levels as indicative of challenges faced by participants when classifying both types of emails. This observation raises questions regarding the samples age composition, as it consisted of both young and adult individuals who might be expected to distinguish these emails due to their familiarity with electronic communication. However, it appears that there is still a need to raise awareness about such email types and provide education for individuals to become proficient at detecting them.

5.3. Valence

The assessment of valence in emails plays a critical role in understanding the factors influencing email classification. The Student T test analysis aims to elucidate the relationship between email valence (Valencia_Legitima and Valencia_Phishing) and phishing classification.

In examining the emotional valence of emails, the results suggest distinct patterns between legitimate and phishing emails. Phishing emails tend to elicit more negative emotional responses when compared to their legitimate counterparts. This indicates that when individuals interact with phishing emails, they often experience discomfort and negative emotions. In contrast, legitimate emails are perceived as emotionally neutral, neither pleasant nor unpleasant.

In conclusion, the exploration of valence in the context of email classification has unveiled intriguing insights. The significant relationship between valence and email classification, along with the linear association, implies that the valence tone of emails is an influential factor in their categorization as legitimate or phishing. These findings contribute to a more refined understanding of the difficult interplay between valence and email classification.

5.4. Activation

Examining Ativaçao_Legitima (Legitimate Email Activation), we find a mean activation. The statistical test results reveal a notable revelation—an association between the activation level of legitimate emails and phishing classification, denoted by a low p-value of 0.002. This discovery underlines that the degree of activation in legitimate emails holds significance in their categorization, potentially affecting the perception of email authenticity.

In the case of Ativaçao_Phishing (Phishing Email Activation), the mean activation score is slightly higher, at 3.17, with a standard deviation of 0.30. This difference in activation levels between legitimate and phishing emails is an intriguing point of inquiry that warrants further exploration.

In conclusion, the exploration of email activation uncovers an intriguing aspect of email classification.

5.5. Emotions

In contrast, the Chat-GPT analysis of emotional expressions in emails presents some distinct findings. Chat-GPT consistently attributes some emotional categorization to the analyzed emails. Phishing emails tend to exhibit higher instances of emotions such as fear and anger, whereas legitimate emails show a greater prevalence of surprise, which contrasts with the human analysis where surprise was more prominent in phishing emails. Intriguingly, the AI analysis did not classify any email as expressing sadness.

These disparities between human and AI-generated results highlight the intricate nature of interpreting emotional nuances within emails. Human evaluators can grasp subtle contextual cues and nuances, enabling them to provide comprehensive emotional assessments. AI models, on the other hand, might struggle to capture the full spectrum of emotional intricacies, leading to variations in the outcomes.

The significance of emotions in phishing attacks has been highlighted in previous literature [6], [15], [12], [13], indicating the complexity of the relationship between emotions and responses. Prior studies have demonstrated that emotions like fear, anticipation, and trust are commonly associated with phishing attacks [6], while high-risk scenarios tend to evoke more negative emotions [15]. Additionally, incorporating sentiment-based emotion analysis, such as Affect Intensities [12], has shown performance improvement in phishing email detection algorithms. Moreover, the correlation between participants' ability to identify phishing emails and persuasion principles like authority, commitment, liking, and scarcity has been explored, suggesting the potential benefits of integrating such principles into anti-phishing training programs [13]. In our study, we identified various emotions in both phishing and legitimate emails, with fear and anger being more pronounced in phishing emails, while surprise was prevalent in legitimate emails, which contrasts with human analysis [6]. Notably, our AI analysis did not classify any email as expressing sadness, indicating variations in emotional interpretation.

In summary, the analysis of emotional expressions in phishing and legitimate emails yields valuable insights into the psychological and emotional dimensions of email content. The divergences between human and AI-generated results underscore the importance of continued research aimed at enhancing AI models' accuracy in detecting and interpreting emotions. Understanding how emotional content influences email perception is vital, particularly in the realm of email security and communication analysis.

Due to the use of different emotional expressions in various studies, makes it difficult to draw direct comparisons with previous literature.

5.6. Personality traits

The examination of the Point biserial correlation between various personality traits and email classification as phishing revealed intriguing results. None of the personality traits, including Neuroticism (N), Extroversion (E), Openness to Experience (O), Agreeableness (A), Conscientiousness (C), and Vigilance (V), demonstrated a statistically significant correlation with the classification of emails as phishing. These findings imply that an individual's level of Neuroticism, Extroversion, Openness to Experience, Agreeableness, Conscientiousness, or Vigilance does not strongly influence their ability to identify emails as phishing or non-phishing.

While these results offer a comprehensive overview of the relationships between personality traits and email classification, they suggest a poor correlation between the two. Certain personality traits exhibit weak correlations with phishing detection, with Openness to Experience (O) and Vigilance (V) appearing to be associated with an increased likelihood of detecting phishing attempts. Conversely, the personality trait Extroversion (E) displays an inverse relationship, indicating that higher scores in this trait reduce the probability of successfully identifying phishing emails.

The absence of a clear link between personality traits and the ability to identify phishing emails underscores the complexity of email categorization and the potential influence of factors beyond these specific personality dimensions.

Due to the use of different personality traits in various studies, makes it difficult to draw direct comparisons with previous literature ([10], [3], [5], [15], [14], [17]).

5.7. Decision patterns

The analysis of the correlations between different decisional patterns and phishing classification, as represented by the phishing emails variable, provides valuable insights into the relationship between decision-making strategies and email categorization. Surprisingly, the phishing emails variable itself shows no significant correlation with the various decisional patterns, suggesting that the ability to identify phishing emails is not inherently linked to distinct decision-making approaches.

When we delve into the individual decisional patterns, a subtle pattern emerges. Vigilance, buck-passing, procrastination, and hypervigilance all exhibit low correlations with phishing emails albeit without reaching statistical significance. This implies that the tendencies to classify emails as phishing or not are weakly

associated with these decision-making strategies. These findings highlight the complexity of the human decision-making process and suggest that the ability to detect phishing emails might be influenced by factors other than traditional decision-making patterns.

In essence, this study underscores that identifying phishing emails is a multifaceted task, potentially characterized by active affective regulation. This emphasizes the importance of considering emotional and cognitive factors in email categorization, as they contribute to an individual overall well-being and life satisfaction.

Chapter 6 – Limitations and recommendations

It is essential to acknowledge the limitations of this research, such as the relatively small sample size, scalability constraints, subjective emotion categorization, and the absence of a pilot test. These limitations serve as a catalyst for future research endeavors, highlighting the need for more extensive, diverse, and controlled studies to refine our understanding of email content evaluation and its implications.

Future studies in this domain should explore several topics. Firstly, expanding the sample size to include a more diverse and extensive population of email evaluators would enhance the generalizability of findings, offering a broader perspective on email security and content evaluation. Secondly, investigating email content analysis in the context of larger-scale email systems and corporate environments would provide insights into the unique challenges and dynamics faced by organizations. Additionally, future research can delve deeper into emotion categorization by developing standardized methods that reduce subjectivity and bias. Conducting pilot tests and calibration exercises to enhance the reliability and consistency of email classification could be a valuable step. These future studies hold the potential to refine email security practices and contribute to more effective communication analysis.

Chapter 7 – Conclusions

This dissertation conducted a comprehensive investigation into Phishing Email Classification, utilizing insights from Psychology and Machine Learning. The research objectives encompassed different aspects, including participants' ability to differentiate between legitimate and phishing emails, emotional valence and activation levels in these emails, emotional responses, connections between personality traits and phishing email detection, and the impact of decision-making patterns. This study aimed to improve our understanding of the complex factors influencing phishing email detection and contribute to the development of more effective mitigation strategies. The findings provide critical insights into email content evaluation, decisional patterns, and their implications for email security and communication, shedding light on the role of emotions, personality traits, and decision-making in this context.

The study has highlighted the significant role of emotional polarity in the categorization of emails. Legitimate emails tend to exhibit a predominantly neutral valence, while phishing emails incorporate stronger negative cues aimed at eliciting specific responses. The discernible disparities in emotional valence between phishing and legitimate emails serve as valuable indicators for individuals in identifying potential phishing attempts. A deeper comprehension of the emotional intricacies within emails contributes to more effective email classification and detection, ultimately enhancing email security protocols and reducing susceptibility to phishing attacks.

Phishing emails tend to be emotionally rated as unpleasant and induce higher emotional activation compared to legitimate emails. This emotional discrepancy presents challenges in the detection and differentiation of phishing emails from legitimate ones. Some personality traits, such as openness to experience and vigilance, seem to relate to the detection of phishing emails, as do various decision-making styles, particularly vigilance and hypervigilance. On the other hand, the personality trait of extroversion appears to be less likely to detect phishing emails, while decision-making characterized by procrastination is also less effective in identifying phishing emails.

The examination of personality traits and decisional patterns has unveiled intriguing insights into the potential impact of individual characteristics on email security. The correlations between these traits and email classification underscore the relevance of understanding the human element in email assessment.

This dissertation holds potential for several practical applications, primarily with regard to enhancing email security measures and mitigating the susceptibility of both individuals and organizations to phishing attacks. By investigating the influential factors in phishing email detection, such as emotional valence, personality traits, and decision-making patterns, this research can contribute to the advancement of more

effective email classification and filtering systems. This, in turn, can fortify cybersecurity efforts by enhancing the precision of phishing email identification and reducing instances of false positives and negatives. Furthermore, the findings of this study may serve as valuable input for awareness and training programs, educating individuals in recognizing phishing emails and thus empowering them to better safeguard sensitive information and assets from the ever-present threat of cyberattacks.

References

[1] P. S. Bogawar and K. K. Bhoyar, "Soft Computing Approaches to Classification of Emails for Sentiment Analysis.," *ACM International Conference Proceeding Series*, pp. 1–7, Feb. 2016, [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=edb&AN=119697132&lang=pt-pt&site=eds-live&scope=site

[2] C. Sur, "Ensemble one-vs-all learning technique with emphatic & rehearsal training for phishing email classification using psychology.," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 30, no. 6, pp. 733–762, Feb. 2018, [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=133380109&lang=pt-pt&site=eds-live&scope=site

[3] R. Fatima, A. Yasin, L. Liu, and J. Wang, "How persuasive is a phishing email? A phishing game for phishing awareness.," *J Comput Secur*, vol. 27, no. 6, pp. 581–612, Feb. 2019, [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=139099709&lang=pt-pt&site=eds-live&scope=site

[4] D. E. O'Leary, "What Phishing E-mails Reveal: An Exploratory Analysis of Phishing Attempts Using Text Analysis.," *Journal of Information Systems*, vol. 33, no. 3, pp. 285–307, 2019, [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=139861153&lang=pt-pt&site=eds-live&scope=site

[5] Y. Li, K. Xiong, and X. Li, "Applying Machine Learning Techniques to Understand User Behaviors When Phishing Attacks Occur.," *EAI Endorsed Transactions on Security & Safety*, vol. 6, no. 21, pp. 1–27, Feb. 2019, [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=edb&AN=142116275&lang=pt-pt&site=eds-live&scope=site

[6] T. Sharma, M. Bashir, C. I., N. E., and A. T., "An Analysis of Phishing Emails and How the Human Vulnerabilities are Exploited," *AHFE Virtual Conference on Human Factors in Cybersecurity, 2020*, vol. 1219, pp. 49–55, 2020, doi: 10.1007/978-3-030-52581-1_7.

[7] A. Alhogail and A. Alsabih, "Applying machine learning and natural language processing to detect phishing email," *Comput Secur*, vol. 110, 2021, doi: 10.1016/j.cose.2021.102414.

[8] H. Abroshan, J. Devos, G. Poels, and E. Laermans, "COVID-19 and Phishing: Effects of Human Emotions, Behavior, and Demographics on the Success of Phishing Attempts during the Pandemic," *IEEE Access*, vol. 9, pp. 121916–121929, 2021, doi: 10.1109/ACCESS.2021.3109091.

[9] X. Lu, M. Head, J. Jiang, and J. Yang, "AN INVESTIGATION OF THE INTERPLAY BETWEEN EMOTIONAL CONGRUENCE AND REGULATORY FIT IN PHISHING SUSCEPTIBILITY.," *Proceedings of the European Conference on Information Systems (ECIS)*, pp. 1–11, Feb. 2021, [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=edb&AN=155093521&lang=pt-pt&site=eds-live&scope=site

[10] F. K.-Y. Chou, A. P.-S. Chen, and V. C.-L. Lo, "Mindless Response or Mindful Interpretation: Examining the Effect of Message Influence on Phishing Susceptibility.," *Sustainability (2071-1050)*, vol. 13, no. 4, p. 1651, Feb. 2021, [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=edb&AN=152671745&lang=pt-pt&site=eds-live&scope=site

[11] A. Pyke, E. Rovira, S. Murray, J. Pritts, harlotte L. Carp, and R. Thomson, "Predicting Individual Differences to Cyber Attacks: Knowledge, Arousal, Emotional and Trust Responses.," *Cyberpsychology (Brno)*, vol. 15, no. 4, pp. 1–24, Feb. 2021, [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=153724629&lang=pt-pt&site=eds-live&scope=site

[12] N. A. Farahisya and F. A. Bachtiar, "Spam Email Detection with Affect Intensities using Recurrent Neural Network Algorithm," *2022 2nd International Conference on Information Technology and Education (ICIT&E), Information Technology and Education (ICIT&E), 2022 2nd International Conference on*, pp. 206–211, Feb. 2022, doi: 10.1109/ICITE54466.2022.9759865.

[13] F. Sharevski and P. Jachim, "'Alexa, What's a Phishing Email?': Training users to spot phishing emails using a voice assistant.," *EURASIP J Inf Secur*, vol. 2022, no. 1, pp. 1–13, Feb. 2022, [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=edb&AN=160349252&lang=pt-pt&site=eds-live&scope=site

[14] S. Shahriar, A. Mukherjee, and O. Gnawali, "IMPROVING PHISHING DETECTION VIA PSYCHOLOGICAL TRAIT SCORING.," *International Conference on ICT, Society & Human Beings*, pp. 131–139, Feb. 2022, [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=edb&AN=159129514&lang=pt-pt&site=eds-live&scope=site

[15] A. Pyke, E. Rovira, S. Murray, J. Pritts, harlotte L. Carp, and R. Thomson, "Predicting Individual Differences to Cyber Attacks: Knowledge, Arousal, Emotional and Trust Responses.," *Cyberpsychology (Brno)*, vol. 15, no. 4, pp. 1–24, 2021, [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=153724629&amp

[16] A. Alhogail and A. Alsabih, "Applying machine learning and natural language processing to detect phishing email," *Comput Secur*, vol. 110, 2021, doi: 10.1016/j.cose.2021.102414.

[17] C. Sur, "Ensemble one-vs-all learning technique with emphatic & rehearsal training for phishing email classification using psychology.," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 30, no. 6, pp. 733–762, Dec. 2018, [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=133380109&amp

[18] E. O. Yeboah-Boateng and P. M. Amanor, "Journal of Emerging Trends in Computing and Information Sciences Phishing, SMiShing & Vishing: An Assessment of Threats against Mobile Devices Campus and a database administrator," vol. 5, no. 4, 2014, [Online]. Available: http://www.cisjournal.org

[19] R. Heartfield and G. Loukas, "A taxonomy of attacks and a survey of defence mechanisms for semantic social engineering attacks," *ACM Computing Surveys*, vol. 48, no. 3. Association for Computing Machinery, Dec. 01, 2015. doi: 10.1145/2835375.

[20] Adhiparasakthi Engineering College. Department of Electronics and Communication Engineering, Institute of Electrical and Electronics Engineers. Madras Section, and Institute of Electrical and Electronics Engineers, *Proceedings of the 2018 IEEE International Conference on Communication and Signal Processing (ICCSP) : 3rd - 5th April 2018, Melmaruvathur, India.*

[21] N. Saud Al-Musib, F. Mohammad Al-Serhani, M. Humayun, and N. Z. Jhanjhi, "Business email compromise (BEC) attacks," *Mater Today Proc*, vol. 81, pp. 497–503, Jan. 2023, doi: 10.1016/J.MATPR.2021.03.647.

[22] S. Davis and J. Chaudhry, "Don't Take the Bait! Protect Yourself from the Latest Email Scams," 2018, Accessed: Jul. 15, 2023. [Online]. Available: https://www.consumer.ftc.gov/articles/0003-

[23] E. R. S. and R. Ravi, "A performance analysis of Software Defined Network based prevention on phishing attack in cyberspace using a deep machine learning with CANTINA approach (DMLCA)," *Comput Commun*, vol. 153, pp. 375–381, Mar. 2020, doi: 10.1016/J.COMCOM.2019.11.047.

[24] K. Chetioui, B. Bah, A. O. Alami, and A. Bahnasse, "Overview of Social Engineering Attacks on Social Networks," *Procedia Comput Sci*, vol. 198, pp. 656–661, 2022, doi: 10.1016/j.procs.2021.12.302.

[25] "Honeytraps, A Network Forensic Tool (Paper Draft)".

[26] N. Odeh, D. Eleyan, A. Eleyan, and N. A. Odeh, "A SURVEY OF SOCIAL ENGINEERING ATTACKS: DETECTION AND PREVENTION TOOLS The ant lion optimizer (ALO): theories, variants, and applications View project Software engineering View project A SURVEY OF SOCIAL ENGINEERING ATTACKS: DETECTION AND PREVENTION TOOLS," *J Theor Appl Inf Technol*, vol. 99, no. 18, 2021, Accessed: Jul. 15, 2023. [Online]. Available: https://www.researchgate.net/publication/355410947

[27] S. Ray, "A Quick Review of Machine Learning Algorithms," pp. 14–16, 2019.

[28] A. A. Soofi and A. Awan, "Classification Techniques in Machine Learning: Applications and Issues," *Journal of Basic & Applied Sciences*, vol. 13, pp. 459–465, 2017.

[29] A. Chhabra, K. Masalkovait ˙ E, and P. Mohapatra, "An Overview of Fairness in Clustering", doi: 10.1109/ACCESS.2021.3114099.

[30] J. Alzubi, A. Nayyar, and A. Kumar, "Machine Learning from Theory to Algorithms: An Overview," *J. Phys*, p. 12012, 2018, doi: 10.1088/1742-6596/1142/1/012012.

[31] M. Huljanah, Z. Rustam, S. Utama, and T. Siswantining, "Feature Selection using Random Forest Classifier for Predicting Prostate Cancer", doi: 10.1088/1757-899X/546/5/052031.

[32] A. Ganatra and D. Panchal, "Behaviour Analysis of Multilayer Perceptronswith Multiple Hidden Neurons and Hidden Layers," *Article in International Journal of Computer Theory and Engineering*, vol. 3, no. 2, 2011, doi: 10.7763/IJCTE.2011.V3.328.

[33] R. M. Schmidt, "Recurrent Neural Networks (RNNs): A gentle Introduction and Overview".

[34] M. K. Eckstein, L. Wilbrecht, and A. G. Collins, "What do reinforcement learning models measure? Interpreting model parameters in cognition and neuroscience," *Curr Opin Behav Sci*, vol. 41, pp. 128–137, Oct. 2021, doi: 10.1016/J.COBEHA.2021.06.004.

[35] E. D. Liddy, "Natural Language Processing," *Encyclopedia of Library and Information Science, 2nd Ed.*, 2001, Accessed: Jul. 15, 2023. [Online]. Available: https://surface.syr.edu/istpub

[36] J. Li, A. Sun, J. Han, and C. Li, "A Survey on Deep Learning for Named Entity Recognition," *IEEE Trans Knowl Data Eng*, vol. 34, no. 1, pp. 50–70, Jan. 2022, doi: 10.1109/TKDE.2020.2981314.

[37] J. Deng and Y. Lin, "Frontiers in Computing and Intelligent Systems The Benefits and Challenges of ChatGPT: An Overview".

[38] J. Chung, "What Should We Do About Artificial Intelligence in Health Care?," *NYSBA Health Law Journal*, vol. 22, no. 3. Jan. 30, 2018. Accessed: Jul. 15, 2023. [Online]. Available: https://papers.ssrn.com/abstract=3113655

[39] "Universal Emotions | What are Emotions? | Paul Ekman Group." Accessed: Sep. 25, 2023. [Online]. Available: https://www.paulekman.com/universal-emotions/

[40] "What Are Basic Emotions? | Psychology Today." Accessed: Sep. 25, 2023. [Online]. Available: https://www.psychologytoday.com/us/blog/hide-and-seek/201601/what-are-basic-emotions

[41] "The 6 Types of Basic Emotions." Accessed: Sep. 25, 2023. [Online]. Available: https://www.verywellmind.com/an-overview-of-the-types-of-emotions-4163976

[42] M. Pedroso-Lima *et al.*, "Revista da Associação Portuguesa de Psicologia A versão portuguesa do NEO-FFI: Caracterização em função da idade, género e escolaridade," *Revista PSICOLOGIA*, vol. 28, no. 2, pp. 1–10, 2014, Accessed: Oct. 11, 2023. [Online]. Available: https://revista.appsicologia.org/index.php/rpsicologia/article/download/534/702/2309

[43] U. Isaksson, S. Hajdarević, L. Jutterström, and Å. Hörnsten, "Validity and reliability testing of the Swedish version of Melbourne Decision Making Questionnaire," *Scand J Caring Sci*, vol. 28, no. 2, pp. 405–412, 2014, doi: 10.1111/SCS.12052.

[44] L. Filipe, M. J. Alvarez, M. S. Roberto, and J. A. Ferreira, "Validation and invariance across age and gender for the Melbourne Decision-Making Questionnaire in a sample of Portuguese adults," *Judgm Decis Mak*, vol. 15, no. 1, pp. 135–148, Jan. 2020, doi: 10.1017/S1930297500006951.

[45] C. Cotrena, L. D. Branco, and R. P. Fonseca, "Adaptation and validation of the Melbourne Decision Making Questionnaire to Brazilian Portuguese," *Trends Psychiatry Psychother*, vol. 40, no. 1, pp. 29–37, Dec. 2017, doi: 10.1590/2237-6089-2017-0062.

[46] "(PDF) Estudo de formas muito reduzidas do Modelo dos Cinco Factores da Personalidade." Accessed: Dec. 28, 2023. [Online]. Available: https://www.researchgate.net/publication/310458188_Estudo_de_formas_muito_reduzidas_do_Modelo_dos_Cinco_Factores_da_Personalidade

[47] J. Pais-Ribeiro, "Estudo de formas muito reduzidas do Modelo dos Cinco Factores da Personalidade," 2006. [Online]. Available: https://www.researchgate.net/publication/310458188

[48] L. Filipe, M.-J. Alvarez, M. S. Roberto, and J. A. Ferreira, "Validation and invariance across age and gender for the Melbourne Decision-Making Questionnaire in a sample of Portuguese adults," *Judgm Decis Mak*, vol. 15, no. 1, pp. 135–148, 1930, doi: 10.1017/S1930297500006951.

[49] "Regression vs Classification. In machine learning Regression and... | by Ali Reza Kohani | Medium." Accessed: Jun. 18, 2023. [Online]. Available: https://medium.com/@ali_88273/regression-vs-classification-87c224350d69

[50] "How Paul Ekman's Theory of Emotion Can Help You Better Understand Others - Envision your Evolution." Accessed: Sep. 25, 2023. [Online]. Available: https://www.envisionyourevolution.com/evolution-emotion/paul-ekman-theory-of-emotion/2149/

Appendix I – Inventário de Personalidade NEO-FFI-20

Leia cada afirmação com atenção. Para cada afirmação, nas páginas seguintes, marque com um ☒ apenas a coluna que melhor corresponde à sua opinião.
Se mudar de opinião ou se se enganar apague completamente a resposta ou, no caso de isso não ser possível, preencha o ▩ errado e assinale com um ☒ a resposta correcta.
Não existem respostas certas nem erradas. Descreva as suas opiniões rápida, espontânea e honestamente. Responda a todas as questões.

Assinale **Discordo Fortemente** se a afirmação for definitivamente falsa ou se discordar fortemente dela. • Assinale **Discordo** se a afirmação for, na maior parte das vezes, falsa ou se discordar dela. • Assinale **Neutro** se a afirmação for igualmente falsa e verdadeira, se não se decidir ou se a sua posição perante o que foi dito é completamente neutra. • Assinale **Concordo** se a frase for, na maior parte das vezes, verdadeira ou se concordar com ela. • Assinale **Concordo Fortemente** se a frase for definitivamente verdadeira ou se concordar fortemente com ela.

	Discordo Fortemente	Discordo	Neutro	Concordo	Concordo Fortemente
1. Raramente estou triste ou deprimido(a).	☐	☐	☐	☐	☐
2. Sou uma pessoa alegre e bem disposta.	☐	☐	☐	☐	☐
3. A poesia pouco ou nada me diz.	☐	☐	☐	☐	☐
4. Tendo a pensar o melhor acerca das pessoas.	☐	☐	☐	☐	☐
5. Sou eficiente e eficaz no meu trabalho.	☐	☐	☐	☐	☐
6. Sinto-me, muitas vezes, desamparado(a), desejando que alguém resolva os meus problemas por mim.	☐	☐	☐	☐	☐
7. Muitas vezes, sinto-me a rebentar de energia.	☐	☐	☐	☐	☐
8. Às vezes ao ler poesia e ao olhar para uma obra de arte sinto um arrepio ou uma onda de emoção.	☐	☐	☐	☐	☐
9. A minha primeira reacção é confiar nas pessoas.	☐	☐	☐	☐	☐
10. Sou uma pessoa muito competente.	☐	☐	☐	☐	☐
11. Raramente me sinto só ou abatido(a).	☐	☐	☐	☐	☐
12. Sou uma pessoa muito activa.	☐	☐	☐	☐	☐
13. Acho as discussões filosóficas aborrecidas.	☐	☐	☐	☐	☐
14. Algumas pessoas consideram-me frio(a) e calculista.	☐	☐	☐	☐	☐
15. Esforço-me por ser excelente em tudo o que faço.	☐	☐	☐	☐	☐
16. Houve alturas em que experimentei ressentimento e amargura.	☐	☐	☐	☐	☐
17. Sou dominador(a), cheio(a) de força e combativo(a).	☐	☐	☐	☐	☐
18. Não dou grande importância às coisas da arte e da beleza.	☐	☐	☐	☐	☐
19. Tendo a ser descrente ou a duvidar das boas intenções dos outros.	☐	☐	☐	☐	☐
20. Sou uma pessoa aplicada, conseguindo sempre realizar o meu trabalho.	☐	☐	☐	☐	☐

NEO-FFI-20 desenvolvido por Bertoquini, V. & Pais Ribeiro, L. (2005).

Cotação

1. q71*	
2. q177	
3. q128*	**Neuroticismo = 1 + 6 + 11 + 16**
4. q184	
5. q215	
6. q26	
7. q107	**Extroversão = 2 + 7 + 12 + 17**
8. q188	
9. q154	
10. q185	
11. q11*	**Abertura à Experiência = 3 + 8 + 13 + 18**
12. q227	
13. q53*	
14. q74*	
15. q200	**Amabilidade = 4 + 9 + 14 + 19**
16. q186	
17. q12	
18. q8*	
19. q4*	**Conscienciosidade = 5 + 10 + 15 + 20**
20. q85	

Nota. Os números a seguir do q (q4; p8; p11, etc.) correspondem aos itens do NEO-PI-R.
A cotação é feita de 0 a 4.
*itens com cotação invertida

NEO-FFI-20 desenvolvido por Bertoquini, V. & Pais Ribeiro, L. (2005).

Appendix II – Melbourne Decision Making Questionnaire (Portuguese Version)

Versão Portuguesa

Questionário de Melbourne de Tomada de Decisão

Instruções

As pessoas diferem na forma como tomam decisões. Por favor indique como toma decisões indicando, em cada questão, a resposta que melhor representa o seu estilo habitual de decidir.

		Escala	Verdade em relação a mim	Por vezes verdade	Falso em relação a mim
1	Sinto como se estivesse sobre grande pressão de tempo quando estou a tomar decisões.	Hv	2	1	0
2	Gosto de considerar todas as alternativas.	V	2	1	0
3	Prefiro deixar as decisões para outras pessoas.	D	2	1	0
4	Tento descobrir as desvantagens de todas as alternativas.	V	2	1	0
5	Gasto muito tempo em assuntos triviais antes de me dedicar à decisão final.	P	2	1	0
6	Analiso a melhor forma de levar a cabo uma decisão.	V	2	1	0
7	Mesmo depois de tomar uma decisão demoro a agir de acordo com ela.	P	2	1	0
8	Quando tomo decisões gosto de recolher muita informação.	V	2	1	0
9	Evito tomar decisões.	D	2	1	0
10	Quando tenho que tomar uma decisão espero muito tempo antes de começar a pensar no assunto.	P	2	1	0
11	Não gosto de assumir a responsabilidade de tomar decisões.	D	2	1	0
12	Tento ser claro(a) sobre os meus objetivos antes de escolher.	V	2	1	0
13	A possibilidade que alguma pequena coisa possa correr mal leva-me a mudar rapidamente as minhas preferências.	Hv	2	1	0
14	Se uma decisão pode ser tomada por mim ou por outra pessoa eu deixo que outra pessoa a faça.	D	2	1	0
15	Sempre que enfrento uma decisão difícil sinto-me pessimista sobre encontrar uma boa solução.	Hv	2	1	0
16	Tomo muito cuidado antes de escolher.	V	2	1	0
17	Não tomo decisões a não ser que tenha mesmo que ser.	D	2	1	0
18	Adio tomar decisões até que seja demasiado tarde.	P	2	1	0

19	Prefiro que pessoas que estão melhor informadas que decidam por mim.	D	2	1	0
20	Depois de uma decisão ser tomada gasto muito tempo convencendo-me que era a correta.	Hv	2	1	0
21	Adio ter que tomar decisões.	P	2	1	0
22	Não consigo pensar adequadamente se tenho que decidir rapidamente.	Hv	2	1	0

Escalas

(V) Vigilância : 6 itens: 2,4,6,8,12,16 (média=9.41, dp=2.22, α=0.80)

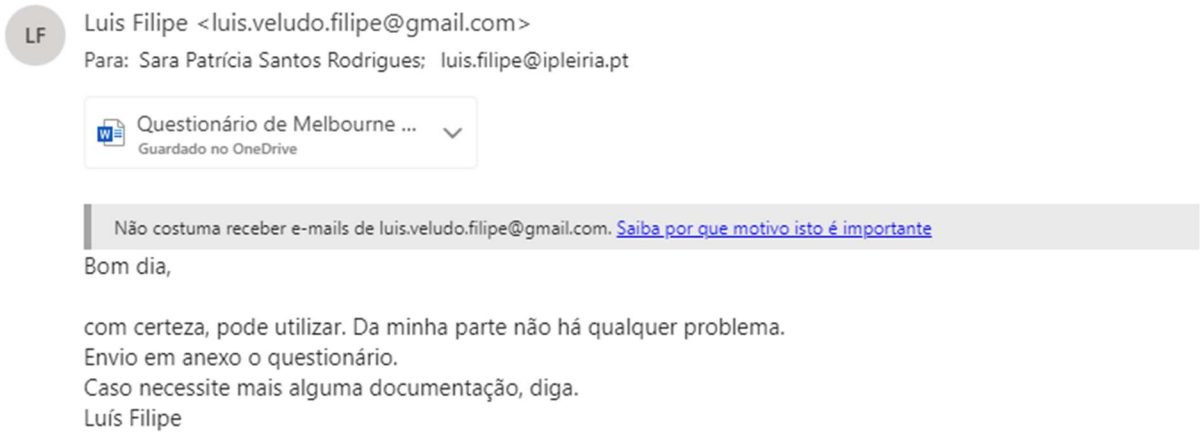
(D) Desresponsabilização : 6 itens: 3,9,11,14,17,19 (média=4.85, dp=2.93, α=0.87)

(P) Procrastinação: 5 itens: 5,7,10,18,21 (média=3.88, dp=2.39, α=0.81)

(Hv) Hipervigilância: 5 itens: 1,13,15,20,22 (média=4.61, dp=2.26, α=0.74)

Appendix III – Questionário de Melbourne de Tomada de Decisão

Authorization



LF

Luis Filipe <luis.veludo.filipe@gmail.com>

Para: Sara Patrícia Santos Rodrigues; luis.filipe@ipleiria.pt

Questionário de Melbourne ...
Guardado no OneDrive

Não costuma receber e-mails de luis.veludo.filipe@gmail.com. Saiba por que motivo isto é importante

Bom dia,

com certeza, pode utilizar. Da minha parte não há qualquer problema.
Envio em anexo o questionário.
Caso necessite mais alguma documentação, diga.
Luís Filipe

Appendix IV -Inventário de Personalidade NEO-FFI-20 Authorization

JR

José Pais Ribeiro <jlpr@fpce.up.pt>
Para: Sara Patrícia Santos Rodrigues

Cara colega
Autorizamos o uso do Inventário de Personalidade NEO-FFI-20 que estudámos e desenvolvemos para uno com a população portuguesa
Cordialmente

José Luís Pais Ribeiro
jlpr@fpce.up.pt
mobile phone: (351) 965045590

web page: http://sites.google.com/site/jpaisribeiro/

ORCID iD: http://orcid.org/0000-0003-2882-8056

ResearchGate- https://www.researchgate.net/profile/Jose_Pais-Ribeiro/publications

https://scholar.google.pt/citations?user=8OnOW5MAAAAJ&hl=en