

## Towards Automatic Principles of Persuasion Detection Using Machine Learning Approach

Bustio-Martínez, Lázaro; Herrera-Semenets, Vitali; García-Mendoza, Juan-Luis; González-Ordiano, Jorge Ángel; Zúñiga-Morales, Luis; Sánchez Rivero, Rubén; Quiróz-Ibarra, José Emilio; Santander-Molina, Pedro Antonio; van den Berg, Jan; Buscaldi, Davide

**DOI**

[10.1007/978-3-031-49552-6\\_14](https://doi.org/10.1007/978-3-031-49552-6_14)

**Publication date**

2024

**Document Version**

Final published version

**Published in**

Progress in Artificial Intelligence and Pattern Recognition - 8th International Congress on Artificial Intelligence and Pattern Recognition, IWAIPR 2023, Proceedings

**Citation (APA)**

Bustio-Martínez, L., Herrera-Semenets, V., García-Mendoza, J.-L., González-Ordiano, J. Á., Zúñiga-Morales, L., Sánchez Rivero, R., Quiróz-Ibarra, J. E., Santander-Molina, P. A., van den Berg, J., & Buscaldi, D. (2024). Towards Automatic Principles of Persuasion Detection Using Machine Learning Approach. In Y. Hernández Heredia, V. Milián Núñez, & J. Ruiz Shulcloper (Eds.), *Progress in Artificial Intelligence and Pattern Recognition - 8th International Congress on Artificial Intelligence and Pattern Recognition, IWAIPR 2023, Proceedings* (pp. 155-166). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 14335 LNCS). Springer.  
[https://doi.org/10.1007/978-3-031-49552-6\\_14](https://doi.org/10.1007/978-3-031-49552-6_14)

**Important note**

To cite this publication, please use the final published version (if applicable).  
Please check the document version above.

**Copyright**

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

**Takedown policy**

Please contact us and provide details if you believe this document breaches copyrights.  
We will remove access to the work immediately and investigate your claim.

***Green Open Access added to TU Delft Institutional Repository***







***'You share, we take care!' - Taverne project***

**<https://www.openaccess.nl/en/you-share-we-take-care>**

Otherwise as indicated in the copyright section: the publisher is the copyright holder of this work and the author uses the Dutch legislation to make this work public.



# Towards Automatic Principles of Persuasion Detection Using Machine Learning Approach

Lázaro Bustio-Martínez<sup>1</sup> , Vitali Herrera-Semenets<sup>2</sup> ,  
Juan-Luis García-Mendoza<sup>3</sup> , Jorge Ángel González-Ordiano<sup>1</sup> ,  
Luis Zúñiga-Morales<sup>1</sup>, Rubén Sánchez Rivero<sup>2</sup>, José Emilio Quiróz-Ibarra<sup>1</sup>,  
Pedro Antonio Santander-Molina<sup>5</sup>, Jan van den Berg<sup>4</sup> ,  
and Davide Buscaldi<sup>3</sup> 

- <sup>1</sup> Universidad Iberoamericana, Ciudad de México, Prolongación Paseo de Reforma  
880, 01219 Mexico City, Mexico  
{lazaro.bustio, jorge.gonzalez, jose.quiróz}@ibero.mx,  
luis.zuniga@correo.uia.mx
- <sup>2</sup> Centro de Aplicaciones de Tecnologías de Avanzada (CENATAV), 7a # 21406,  
Playa, 12200 Havana, La Habana, Cuba  
{vherrera, rsanchez}@cenatav.co.cu
- <sup>3</sup> Université Sorbonne Paris Nord, LIPN, Villetaneuse, France  
{garciamendoza, davide.buscaldi}@lipn.univ-paris13.fr
- <sup>4</sup> Intelligent Systems Department, Delft University of Technology, Mekelweg 4,  
2628CD Delft, The Netherlands  
j.vandenberg@tudelft.nl
- <sup>5</sup> Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile  
pedro.santander@pucv.cl

**Abstract.** Persuasion is a human activity of influence. In marketing, persuasion can help customers find solutions to their problems, make informed choices, or convince someone to buy a useful (or useless) product or service. In computer crimes, persuasion can trick users into revealing sensitive information, or even performing actions that benefit attackers. Phishing is one of the most common and dangerous forms of persuasion-based attacks, as it exploits human vulnerabilities rather than technical ones. Therefore, an intelligent system capable of detecting and classifying persuasion attempts might be useful in protecting users. In this work, an approach that uses Machine Learning to analyze messages based on principles of persuasion and different data representations is presented. The aim of this research is to detect which data representation and which classification algorithm obtain the best results in detecting each principle of persuasion as a prior step to detecting phishing attacks. The results obtained indicate that among the combinations tested, there is one combination of data representation and classification algorithm that performs best. The related classification models obtained can detect the principles of persuasion at a rate that varies between 0.78 and 0.86 of AUC-ROC.

**Keywords:** Principles of Persuasion · Machine Learning · Artificial Intelligence · Data representation · Phishing detection

# 1 Introduction

There is a branch of psychology where the persuasion is studied. This concerns studying the reasons that cause a person to change his/her behavior due to an external influence [6]. In marketing, persuasion aims to create a positive image of a product or service to influence the customer's decision making process. In computer crimes such as phishing, persuasion is used to deceive and to seduce people into revealing sensitive information or performing harmful actions. Phishing is a serious threat that exploits the human factor, which is often the most vulnerable element in a security system. Detecting phishing emails is not an easy task, as they vary greatly in sophistication and appearance. Therefore, a tool that can assist human users in identifying and avoiding phishing emails is needed and would be highly valued [8]. Although phishing has been used for a long time, there are still no completely effective ways to prevent it or to make people aware that they are exposed to it. There is always a risk of falling victim to some type of phishing attack. In this sense, identifying persuasion attempts would be valuable in identifying and preventing a phishing attack. Persuasion can be grouped into some patterns called the "principles of persuasion". So, the principles of persuasion are patterns that can be used to influence the reasoning process by promoting certain opinions, beliefs, and moods. Robert Cialdini was the first to study these principles of persuasion in his book "Influence: The Psychology of Persuasion" [1].

The prevailing view in literature is that phishing messages use principles of persuasion to increase their effectiveness. Thus, the identification of such principles might improve the phishing detection tools. A tool based on Machine Learning (ML) techniques can be an effective solution to support human judgment about phishing attacks. Based on this assumption, this work aims to evaluate the performance of ML techniques in detecting persuasion attempts in emails using an optimized set of principles of persuasion and applying different data representations. Based on Cialdini's work, Ana Ferreria et al. proposed an optimized set of principles of persuasion specially focused on phishing attacks [4], which are used in this research. The data representations investigated in this research are based on different ways of encoding the textual information of messages into numerical vectors that can be processed by ML algorithms. These include bag-of-words, term frequency-inverse document frequency (TF-IDF), and sentence embeddings. Furthermore, this work evaluates the performance of different ML classification algorithms such as Support Vector Machines (SVM), Random Forests (RF), k-Nearest Neighbors (KNN), Naïve Bayes (NB) and pretrained language models.

The main contribution of this research is to identify both effective data representation and data classification algorithms for detecting persuasion principles in messages using an optimized set of persuasion principles. The knowledge obtained will be used to propose an approach to phishing detection based on principles of persuasion.

The remainder of this paper is structured as follows. First, the related work about persuasion detection using ML is analyzed in Sect. 2. Second, the proposed

methodology for persuasion detection, and the data used, are presented in Sect. 3. Third, in Sect. 4, the experimental results using different settings are presented and discussed. Finally, the conclusions and future work are outlined in Sect. 5.

## 2 Related Work

Since 2015, Ana Ferreira et al. [3, 4] have focused on comprehending and identifying how the principles of persuasion can be employed in phishing attacks. Their work searched for a comprehensive and unified list of principles by integrating three different perspectives (i.e., those found in [1, 5] and [12]). The principles of persuasion identified by Ferreira et al. when looking at phishing attacks are: i)- “authority”: people are trained to follow authority without questioning; ii)- “social proof”: people mimic majority to share responsibility; iii)- “liking, similarity and deception”: people follow familiar individuals but can be manipulated; iv)- “distraction”: emotions can cloud decision-making; and v)- “commitment, integrity and reciprocation”: reciprocation and trust can influence behavior.

According to [4], there is a close relationship between these principles of persuasion and the content of phishing messages. In 2019, Van der Heijden et al. [13] identified cognitive vulnerabilities in email content. The authors use a supervised method based on labeled Latent Dirichlet Allocation (LDA). Their solution treats each incoming email as a mixture of topics derived from the labeled input data and estimates email-label distributions, where labels correspond to principles of persuasion.

Only three principles of persuasion proposed by Cialdini (“authority”, “reciprocation” and “scarcity”, which are equivalent to Ferreira’s “distraction” principle), were used by Li et al. in [9] to label a dataset. However, the reason why they only use these three principles is not substantiated in their report. Applying the TF-IDF data representation, the authors associated a set of words with each of these principles: if one of those words appears in an email content, it is labeled with the corresponding principle of persuasion. To detect phishing emails, they trained and evaluated several ML classifiers on the labeled data set, being the Nearest Neighbor the one that achieved the best results. This approach may be inadequate since several principles of persuasion can be associated with the same words, and therefore non-present principles would be identified. The authors do not provide details on this point that could offer a better understanding.

In [11], emotion recognition was performed on spam emails. Their dataset consists of 343 sentences of a random sample sentence from emails labeled under six classes, each one associated with Cialdini’s principles of persuasion. The basic idea is to identify principles of persuasion and associate them with emotions. For this, a transformer-based pretrained model called “Bidirectional Encoder Representations from Transformers” (BERT), and some other variants of BERT, were used. This work is based on a flawed premise: assuming that spam messages are the same as phishing messages. This premise is incorrect because, although both types of messages are morphologically similar, they do not express the same intentions semantically.

The work proposed by Karki et al. in [7] also evaluates ML models based on transformer networks. This work focuses on email classification using principles of persuasion. The goal is to find out if these principles are used in the construction of phishing emails and if it is possible to detect them automatically using Natural Language Processing (NLP) techniques. The models used to classify emails, into categories defined according to Cialdini’s principles. Additionally, just as in [13], they use LDA for automated topic modeling to label the given emails. However, the results obtained show that LDA is not very effective for email classification. The topic modeling offered by LDA was too broad and generic, so it did not improved the classification results.

From the reviewed literature it has been observed that there are almost no labeled datasets for phishing detection available, and none for the classification of principles of persuasion. Since 2018, Rakesh Verma leded the International Workshop on Security and Privacy Analytics Anti-Phishing Shared Task (IWSPA-AP). This workshop focused on identifying phishing emails. In addition to the basic contributions of this workshop, another contribution was the provision of a dataset composed of legitimate and phishing emails [14], which is also used in this work.

The discussed approaches have contributed to the comprehension and identification of principles of persuasion, but they entails some limitations. Manual extraction of persuasion principles is a time-consuming and subjective process that lacks automation. Li et al.’s method of associating words for labeling may lead to inaccuracies, as multiple principles can be linked to the same words. Pepe et al.’s assumption that spam and phishing messages are the same is flawed, as their semantic intentions differ. The ineffectiveness of LDA topic modeling, as observed by Karki et al., emphasizes the need for more robust techniques, which are explored in this research. Also, the literature review reveals that the detection of principles of persuasion in phishing messages has not been fully explored or thoroughly studied. Consequently, this research investigates various data representations and Machine Learning algorithms to determine the most effective combination for identifying persuasion principles as a main step for detecting phishing messages.

## 3 Data and Methodology

### 3.1 Principles of Persuasion Dataset

Given the absence of datasets specifically designed for the detection of principles of persuasion, it was necessary to develop one. To facilitate this endeavor, a comprehensive phishing dataset was required. The IWSPA-AP dataset, proposed by Rakesh Verma et al. [14], was selected for its inclusion of both phishing and legitimate emails<sup>1</sup>.

---

<sup>1</sup> This dataset is available upon request to Rakesh Verma in the following link: <https://www2.cs.uh.edu/~rmverma/>.

**Table 1.** Details of PoP dataset.

PoP dataset	Positive	Negative	% of Positive
authority	681	432	61.18
commitment, integrity and reciprocaton	141	972	12.66
distraction	201	912	18.05
liking, similarity and deception	39	1074	3.50
social proof	61	1052	5.48

Principles of persuasion are used in all kinds of communication, not only in phishing attacks, but considering that the success of such attacks largely depends on the use of Social Engineering, it can be concluded that the principles of persuasion are used more intensely in phishing attacks than in normal communication. In consequence, all the phishing emails from the IWSPA-AP dataset were considered for the creation of the Principles of Persuasion dataset (hereafter referred to as the “PoP dataset”). This resulted in a collection of 1113 confirmed phishing emails. Afterwards, the principles of persuasion of each data sample were labeled manually. To facilitate this process, 3 referees were instructed in the detection of principles of persuasion as proposed in [4]. The labels used corresponded to the principles of persuasion proposed by Ferreira et al. A “blind” methodology was employed during labeling, in which none of the referees were aware of the labels assigned by their colleagues. At the conclusion of the labeling process, the level of agreement between the 3 referees was 94.75%. A “majority vote” label assignment strategy was utilized, in which labels receiving the highest number of votes from referees were assigned. In cases where no consensus was reached among referees, the label was assigned by the authors. This occurred in 5.25% of cases. The obtained set is presented in Table 1. The resulting PoP dataset contains the text of the messages and five columns indicating the presence or absence of each five principle of persuasion within the messages.

### 3.2 Learning Phase

Once the PoP dataset was created, the next phase was to train a classifier that learns the patterns for detecting the principles of persuasion in each data sample. To accomplish this phase, a crucial issue is the selection of the data representation. A comprehensive literature review revealed that no single data representation method demonstrates superiority over others in detecting principles of persuasion within texts. Similarly, an examination of classification algorithms reported in the state of the art for detecting principles of persuasion in texts yielded comparable results. Therefore, the goal of this research is to identify a highly effective combination of data representation techniques and classifiers for accurately detecting principles of persuasion in texts. The strategy delineated in Pseudocode 1 endeavors to achieve this aim.

Given a dataset  $D$  of phishing messages (which were pre-processed in order to remove stop words, removed non alpha-numeric symbols and down-cased all characters), and given a set of principles of persuasion  $P$  composed of {“authority”, “commitment, integrity and reciprocation”, “distraction”, “liking, similarity and deception”, “social proof”}, a set of features extraction algorithms  $F$  that includes {Universal Sentence Encoder<sup>2</sup>, LASER, RoBERTa, TF-IDF, Words Unigrams, Bigrams, Trigrams} was used to train a set of classification algorithms  $C$  composed of {Naïve Bayes, K-Nearest Neighbors, Random Forest, Support Vector Machines, BERT\_base [2], RoBERTa [10]}. For storing the classification results obtained, a list  $L$  is used.

For each principle of persuasion  $p \in P$ , all messages in  $D$  are obtained and stored in  $d_p$ .  $D$  is composed of 6 columns: one column labeled “txt” that stores the text of the messages and 5 additional columns that store the voting of each message according the principle of persuasion  $p$ . Subsequently, for  $d_p$ , its corresponding data representation  $d_{p_f}$  is computed using each feature extraction algorithm  $f \in F$  except for BERT\_base and RoBERTa, which include their own feature extraction method. Each  $d_{p_f}$  is then used to train each classifier  $c \in C$  using a 10-fold cross-validation process. Then the principle of persuasion  $p$ , the feature extraction algorithm  $f$  and the resulting classification model  $model$  are stored as a tuple in a list  $L$ . All of this processing is performed in parallel using Spark. Finally, the combination  $\langle p, f, model \rangle$  that achieves the best accuracy according to some pre-established metric (AUC-ROC in this research) will be returned as output of the proposed approach. After the processing, the combination of data representation and classification model that achieves the best results in detecting that principle of persuasion is determined.

## 4 Experimental Work

As explained above, this research focuses on obtaining a ML model capable of detecting the principles of persuasion contained in messages. To achieve this goal, a processing scheme is proposed in which each principle of persuasion is detected independently of the others. Considering the fact that principles of persuasion detection is a crucial stage for automatically detection of phishing attacks, three research questions arise:

- RQ1: Given the chosen set of data representations, which is best suited for detecting principles of persuasion regardless of classification algorithms?
- RQ2: Given the chosen set of classification algorithms, which is best suited for detecting principles of persuasion regardless of data representation?
- RQ3: Given the chosen set of data representations and classification algorithms, which combination of them is best suited for improving the detection rate of each principle of persuasion?

---

<sup>2</sup> Universal Sentence Encoder includes two feature extractor algorithms based on Deep Averaging Networks (DAN) and Transformers (TRANSF).



**Algorithm 1:** Principles of persuasion extraction method.

---

**Input:**  $D$ : PoP dataset of phishing messages.  
**Output:**  $\langle p, f, model \rangle$ : for each principle of persuasion  $p$ , this list contains the features extraction algorithm  $f$  and the model  $model$  that obtains the best results in detecting principles of persuasion.

```

1 Procedure Train_Models( $D, P, F, C$ )
2    $L = list()$ 
3   do in parallel
4     foreach  $p \in P$  do
5        $d_p = \langle D[txt], D[p] \rangle$ 
6       foreach  $f \in F$  do
7         if ( $f \in \{\text{BERT\_base, RoBERTa}\}$ ) then
8            $d_{p_f} = d_p$ 
9         else
10           $d_{p_f} = f(d_p)$ 
11          foreach  $c \in C$  do
12             $cross\_val = True$ 
13             $folds = 10$ 
14             $model = c(d_{p_f}, cross\_val, folds)$ 
15             $L.append([p, f, model])$ 
16   return  $L$ 

1 Procedure Get_Model_by_Principle( $L$ )
2    $result = list()$ 
3   do in parallel
4     foreach  $p \in P$  do
5        $model = \underset{i=1}{\text{argmax}}^{|L|} (\text{AUC-ROC}(\forall L_i.model \in$ 
6          $L, \text{ if } L_i.model \text{ has been trained for } p))$ 
7        $f = model_f$ 
8        $result.append([p, f, model])$ 
9   return  $result$ 

```

---

The platform used for conduct the experiment was an Intel(R) Xeon(R) Gold 6248 CPU @ 2.50 GHz equipped with 2 sockets, 20 cores per socket, 80 CPUs and 256 GB of RAM. Additionally, 8 GPUs Tesla V100-SXM2 with 32 GB of RAM was used.

#### 4.1 Experiments Results

Several experiments were designed and conducted to answer the 3 proposed research questions. To do so, the messages in  $D$  were represented as a matrix of feature vectors using each feature extraction algorithm in  $F$ , resulting in eight different representations of  $D$ . Each data representation was then used to train each of the six classification algorithms in  $C$  using 10-fold cross-validation to mitigate over-fitting issues. This led to the training of 240 classification models using AUC-ROC as the performance evaluation metric, and the results obtained for each model were stored in  $L$ . Tables 2, 3, and 4 shows the obtained result. Furthermore,  $F$  is the set of features extractor algorithms used. Although AUC-ROC was selected as main evaluation metric, due to their well-known performance on unbalanced classification problems, other evaluation metrics were included such as the Macro Precision (Pr), the Macro Recall (Re), and the Macro F1 Score (F1 Score). These evaluation metrics were reported since they measure the behavior of the obtained classification models using other particular perspective. This allows to obtain other specific points of view.

**Table 2.** Best features extractor algorithm for detecting each principle of persuasion regardless of classification algorithm used.

Principle of Persuasion	$F$	Pr	Re	F1 Score	AUC-ROC
authority	DAN	$0.76 \pm 0.07$	$0.75 \pm 0.09$	$0.75 \pm 0.09$	$0.82 \pm 0.08$
commitment, integrity and reciprocation	RoBERTa	$0.56 \pm 0.09$	$0.62 \pm 0.14$	$0.53 \pm 0.07$	$0.76 \pm 0.07$
distraction	DAN	$0.74 \pm 0.11$	$0.68 \pm 0.06$	$0.66 \pm 0.08$	$0.80 \pm 0.07$
liking, similarity and deception	DAN	$0.55 \pm 0.15$	$0.58 \pm 0.17$	$0.55 \pm 0.14$	$0.75 \pm 0.19$
social proof	DAN	$0.53 \pm 0.14$	$0.55 \pm 0.12$	$0.52 \pm 0.1$	$0.77 \pm 0.13$

**Best Features Extraction Algorithm for Each Principle of Persuasion.**

Table 2 describes the performance metrics obtained when each feature extractor in  $F$  is used to determine which of them is best suited to detect principles of persuasion regardless the classification algorithm used. In this experiment, the performance results obtained for classifiers in  $C$  were averaged for each feature extractor.

The obtained results indicate that there is no single data representation that consistently yields superior detection results for all principles of persuasion. Furthermore, the detection rate for the principles of persuasion varies between 0.75 and 0.82 for AUC-ROC. The principle ‘authority’ was most effectively detected using the DAN feature extractor, achieving an AUC-ROC of 0.82 with 0.08 of standard deviation across all classifiers. The second most effectively detected principle was ‘distraction’, with an AUC-ROC of 0.80 and a standard deviation of 0.07, also using the DAN feature extractor. ‘social proof’ was the third most effectively detected principle with an AUC-ROC of 0.77 with 0.13 of standard deviation, while ‘commitment, integrity and reciprocation’ was detected with AUC-ROC of 0.76 and a standard deviation of 0.07 using RoBERTa. Finally ‘liking, similarity and deception’ was detected with 0.75 of AUC-ROC and a standard deviation of 0.19 using DAN.

Irrespective of the classification algorithm employed, DAN and RoBERTa were found to be the most effective feature extractors. In response to research question RQ1, the evidence collected suggests that DAN and RoBERTa are the two feature extractors that yield superior classification rates. Specifically, RoBERTa is recommended for detecting the principle of ‘commitment, integrity and reciprocation,’ while DAN is recommended for detecting the remaining principles.

**Best Classifier for Each Principle of Persuasion.** In this experiment, each classifier in  $C$  was used to determine which of them is best suited to determine the principles of persuasion regardless the data representation used. Similarly to the

**Table 3.** Best classifier for detecting each principle of persuasion regardless the feature extractor.

Principle of Persuasion	$C$	Pr	Re	F1 Score	AUC-ROC
authority	BERT_base	$0.78 \pm 0.08$	$0.77 \pm 0.07$	$0.77 \pm 0.08$	$0.83 \pm 0.06$
commitment, integrity and reciprocity	SVM	$0.56 \pm 0.15$	$0.58 \pm 0.14$	$0.54 \pm 0.1$	$0.72 \pm 0.15$
distraction	BERT_base	$0.73 \pm 0.08$	$0.71 \pm 0.07$	$0.72 \pm 0.07$	$0.80 \pm 0.06$
liking, similarity and deception	SVM	$0.55 \pm 0.15$	$0.56 \pm 0.15$	$0.54 \pm 0.12$	$0.75 \pm 0.15$
social proof	SVM	$0.56 \pm 0.14$	$0.57 \pm 0.15$	$0.55 \pm 0.12$	$0.75 \pm 0.16$

former experiment, the performance results obtained for the features extractors in  $F$  were averaged for each classifier. The obtained results are expressed in Table 3.

The best classification results concerning AUC-ROC were obtained for “authority”, with 0.83 and a standard deviation of 0.06. This result was obtained using BERT\_base. “distraction” was the second-highest rated principle of persuasion detected, with an AUC-ROC of 0.8 with a standard deviation of 0.06, also obtained using BERT\_base. The third-highest value of AUC-ROC at 0.75 with a standard deviation of 0.15 was for “liking, similarity and deception”, which was obtained using SVM. The “social proof” principle was detected also with 0.75 of AUC-ROC and 0.16 of standard deviation using SVM; while “commitment, integrity and reciprocity” was detected also using SVM with 0.72 of AUC-ROC and 0.15 of standard deviation.

The findings of this experiment indicate that BERT\_base and SVM can be effectively utilized for the detection of principles of persuasion, irrespective of the feature extraction technique employed. Specifically, BERT\_base is recommended for detecting the principles of “authority” and “distraction”, while SVM is recommended for detecting the remaining principles. Consequently, research question RQ2 is addressed based on the results obtained in this experiment.

**Best Combination of Features Extractor and Classifier for Each Principle of Persuasion.** As result of this experiment, it can be noticed that there is no a single combination of feature extractor and classifier that detects all principles of persuasion. In consequence, each principle of persuasion should be detected using their own combination of feature extractor and classifier. Opposite to the previous experiments, Table 4 shows the detection results obtained without averaging any value to show the classification metrics obtained for each combination in each principle of persuasion. From this table it can be observed that the principle of “authority” achieves the highest AUC-ROC value with a value of 0.86 with an standard deviation of 0.07 when using DAN as the fea-

ture extractor and Random Forest as the classifier. This is obtained employing the optimal data representation (DAN) which was determined as conclusion of experiment described in Sect. 4.1. For the principle “authority”, and considering the best data classification results described in Table 3, BERT\_base was found to be the best classifier. This classifier employs its own feature extractor approach, and using it, an AUC-ROC value of 0.84 with a standard deviation of 0.07 was obtained. Considering the AUC-ROC values for “authority”, the best combination of features extractor and data classifier is DAN with Random Forest.

A similar behavior is found for “distraction”, which achieves an AUC-ROC value of 0.82 and a standard deviation of 0.08 when DAN and SVM are used. This result was obtained from Table 2, where DAN was determined to be the best data representation for detecting “distraction”. Subsequently, the best classification result using DAN was achieved using SVM. With regard to the best classifier, according with Table 3, BERT\_base achieves the best classification result. BERT\_base employs its own feature extractor approach, and using this combination of feature extractor and classifier, an AUC-ROC value of 0.80 with a standard deviation of 0.07 is obtained. Then, it can be concluded that for detecting the “distraction”, the best classification results are achieved using DAN and SVM.

The third principle of persuasion that is most effectively detected is “social proof”, with an AUC-ROC value of 0.83 and a standard deviation of 0.11. These results were obtained when TRANSF and SVM were used. For detecting this principle of persuasion, the best results regarding the data representation (see Table 2) were achieved using DAN. Considering DAN, the best classification results were achieved using SVM with an AUC-ROC value of 0.79 and a standard deviation of 0.13 (see Table 3). Also considering Table 3, the best results were reported for SVM, but this time the best results for SVM were achieved using TRANSF as the data representation. Using this combination, the detection was 0.83 for AUC-ROC and 0.11 for standard deviation. As a conclusion, and concerning this principle of persuasion, the best classification results were obtained for the combination of TRANSF and SVM.

A different behavior is observed for the principle of “liking, similarity and deception”. For this principle, according the Table 2, the best discrimination result is achieved using DAN. Considering DAN and according to Table 3, the best classification result was achieved using Naïve Bayes, with an AUC-ROC value of 0.80 and a standard deviation of 0.16. Also for “liking, similarity and deception”, the best classification results with regard to the classifier were obtained for SVM, and these results were achieved using DAN as feature extractor, with an AUC-ROC value of 0.80 and a standard deviation of 0.16. Subsequently, the best detection rate according to the AUC-ROC values obtained for identifying “liking, similarity and deception” was achieved using LASER as features extractor and SVM as classifier, which was 0.82 with a standard deviation of 0.14. It was expected that the best classification results for “liking, similarity and deception” would be obtained using the best data representation (DAN) and the best classifier (SVM), but this was not the case.

**Table 4.** Best combination of features extractor algorithm and classifier for each principle of persuasion.

Principle of Persuasion	<i>F</i>	<i>C</i>	Pr	Re	F1 Score	AUC-ROC
<i>authority</i> *	<b>DAN</b>	<b>RF</b>	0.79 ± 0.01	0.78 ± 0.08	0.78 ± 0.09	<b>0.86 ± 0.07</b>
<i>authority</i> <sup>+</sup>	–	BERT_base	0.78 ± 0.08	0.77 ± 0.07	0.77 ± 0.08	0.84 ± 0.07
<i>commitment, integrity and reciprocation</i> *	<b>RoBERTa</b>	<b>NB</b>	0.58 ± 0.04	0.72 ± 0.09	0.53 ± 0.07	<b>0.78 ± 0.09</b>
<i>commitment, integrity and reciprocation</i> <sup>+</sup>	DAN	SVM	0.48 ± 0.11	0.51 ± 0.04	0.48 ± 0.06	0.77 ± 0.12
<i>commitment, integrity and reciprocation</i> <sup>#</sup>	RoBERTa	SVM	0.57 ± 0.04	0.65 ± 0.16	0.56 ± 0.04	0.73 ± 0.13
<i>distraction</i> *	<b>DAN</b>	<b>SVM</b>	0.70 ± 0.10	0.67 ± 0.08	0.69 ± 0.09	<b>0.82 ± 0.08</b>
<i>distraction</i> <sup>+</sup>	–	BERT_base	0.74 ± 0.08	0.71 ± 0.07	0.72 ± 0.07	0.80 ± 0.07
<i>liking, similarity and deception</i> *	DAN	NB	0.60 ± 0.12	0.71 ± 0.22	0.60 ± 0.12	0.80 ± 0.16
<i>liking, similarity and deception</i> <sup>+</sup>	<b>LASER</b>	<b>SVM</b>	0.48 ± 0.01	0.50 ± 0.0	0.49 ± 0.01	<b>0.82 ± 0.14</b>
<i>liking, similarity and deception</i> <sup>#</sup>	DAN	SVM	0.53 ± 0.16	0.55 ± 0.16	0.54 ± 0.16	0.80 ± 0.16
<i>social proof</i> *	DAN	SVM	0.57 ± 0.22	0.56 ± 0.16	0.56 ± 0.17	0.79 ± 0.13
<i>social proof</i> <sup>+</sup>	<b>TRANSF</b>	<b>SVM</b>	0.62 ± 0.20	0.58 ± 0.17	0.57 ± 0.17	<b>0.83 ± 0.11</b>

\* Combination of the best feature extractor algorithm and the best classifier associated with it.

+ Combination of the best classifier and the best feature extractor algorithm associated with it.

# Combination of the best individual feature extractor and classifier.

In **bold** text, the results of the combination of feature extractor and classifier that obtains the best overall detection results for each principle of persuasion

Finally, for detecting “commitment, integrity and reciprocation”, the best detection rate concerning data representation (see Table 2), was achieved using RoBERTa. For RoBERTa, the best detection AUC-ROC value was achieved using Naïve bayes according to Table 3, with 0.78 and 0.09 standard deviation. Concerning the classifier, and according to Table 3, the best detection rate for “commitment, integrity and reciprocation” was achieved using SVM as classifier and, and were obtained using DAN as features extractor. With this combination it was obtained an AUC-ROC of 0.77 with 0.12 of standard deviation. At this point the best classification results using the best feature extractor (RoBERTa) and the best classifier (SVM) were used, was 0.73 of AUC-ROC and 0.13 of standard deviation. Similarly to “liking, similarity and deception”, the best classification result is not achieved using the best data representation and the best classifier, but was achieved using RoBERTa and Naïve Bayes with 0.78 of AUC-ROC and 0.09 of standard deviation. Considering these results, the research question RQ3 is answered.

## 5 Conclusions

Phishing is a highly profitable and effective scam that exploits the human factor in information systems. In such attacks, messages are delivered with the intent of provoking emotions such as urgency, greed, and curiosity in their victims. In the literature reviewed, the majority of proposed approaches focused on detecting *what* is communicated in a phishing message rather than *how* the message is communicated. One approach to detecting *how* a phishing message is communicated is by identifying the principles of persuasion included in the messages. This article presents a study aimed at determining the data representation and classifier that improve the detection rate of each principle of persuasion, both

independently and in combination. This approach is novel in that the Machine Learning models obtained in this research are specifically tailored for detecting principles of persuasion most commonly used in phishing attacks, rather than broader principles of persuasion addressed in the literature. The detection rate, as measured by AUC-ROC, ranges between 0.78 and 0.86.

**Acknowledgement.** This research was supported by the IBERO and InIAT through the project “*Detección de ataques de phishing en mensajes electrónicos mediante técnicas de Inteligencia Artificial*”. Additionally, the authors thank CONACYT for the computer resources provided through the INAOE Supercomputing Laboratory’s Deep Learning Platform for Language Technologies.

## References

1. Cialdini, R.B.: *Influence: The Psychology of Persuasion*, vol. 55. Collins New York (2007)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. [arXiv:1810.04805v1](https://arxiv.org/abs/1810.04805v1) [cs.CL] (2018)
3. Ferreira, A., Coventry, L., Lenzini, G.: Principles of persuasion in social engineering and their use in phishing. In: Tryfonas, T., Askoxylakis, I. (eds.) HAS 2015. LNCS, vol. 9190, pp. 36–47. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-20376-8\\_4](https://doi.org/10.1007/978-3-319-20376-8_4)
4. Ferreira, A., Teles, S.: Persuasion: how phishing emails can influence users and bypass security measures. *Int. J. Hum.-Comput Stud.* **125**, 19–31 (2019)
5. Gragg, D.: A multi-level defense against social engineering. *SANS Reading Room* **13**, 1–21 (2003)
6. Hogan, K.: *The Psychology of Persuasion: How to Persuade Others to Your Way of Thinking*. Pelican Publishing (2010)
7. Karki, B., Abri, F., Namin, A.S., Jones, K.S.: Using transformers for identification of persuasion principles in phishing emails. In: 2022 IEEE International Conference on Big Data (Big Data), pp. 2841–2848. IEEE (2022)
8. Koddebusch, M.: Exposing the phish: the effect of persuasion techniques in phishing e-mails. In: DG. O 2022: The 23rd Annual International Conference on Digital Government Research, pp. 78–87 (2022)
9. Li, X., Zhang, D., Wu, B.: Detection method of phishing email based on persuasion principle. In: 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), vol. 1, pp. 571–574. IEEE (2020)
10. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. [arXiv:1907.11692](https://arxiv.org/abs/1907.11692) (2019)
11. Pepe, E.: Human-centric approach to emails phishing detection. Ph.D. thesis, Dublin, National College of Ireland (2022)
12. Stajano, F., Wilson, P.: Understanding scam victims: seven principles for systems security. *Commun. ACM* **54**(3), 70–75 (2011)
13. Van Der Heijden, A., Allodi, L.: Cognitive triaging of phishing attacks. In: SEC 2019, pp. 1309–1326. USENIX Association (2019)
14. Verma, R.M., Zeng, V., Faridi, H.: Data quality for security challenges: case studies of phishing, malware and intrusion detection datasets. In: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS 2019, pp. 2605–2607. Association for Computing Machinery, New York (2019)