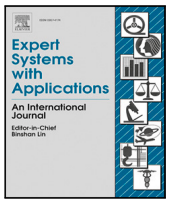
Expert Systems With Applications 265 (2025) 125767

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



On persuasion in spam email: A multi-granularity text analysis

Francisco Jáñez-Martino [a,*], Alberto Barrón-Cedeño [b], Rocío Alaiz-Rodríguez [a], Víctor González-Castro [a], Arianna Muti [b]

[a] Department of Electrical Engineering, Systems and Automation, Universidad de León, Spain
[b] DIT, Università di Bologna, Forlì, Italy

ARTICLE INFO

Keywords:
Social engineering
Persuasion detection
Spam email
Cybersecurity

ABSTRACT

Electronic mail (email) is one of the most popular communication media for direct and private communication. Being typically a free service and anonymity-friendly, massive spam email campaigns are common. Nowadays, spam email encompasses scam, phishing, malware distribution, and various other cybersecurity threats. Within these emails, recipients frequently encounter social engineering techniques aimed at persuading them to take an action, such as clicking on a hyperlink, opening an attachment or responding. In this paper, we conduct a study on supervised models to identify persuasion (binary classification) and to identify the specific persuasion techniques that are commonly used in spam email (multilabel classification). To achieve this, we develop systems capable of spotting persuasion in spam emails based on natural language processing techniques. We approach this challenging task at different levels of granularity: full email, sentences and specific text snippets (i.e. text fragments composed by one or more words, typically shorter than a sentence). We replicate and adapt two methodologies used to detect propaganda in news articles. Additionally, we build a custom spam email dataset, and fine-tune pre-trained RoBERTa-based transformer models to tackle the sentence level detection. This allows us to determine how extensively spam emails rely on persuasion to achieve their goals and, if so, to identify those techniques that would be employed for user protection and cybersecurity improvements.

1. Introduction

Electronic mail (email) has been one of the most popular communication media for organizations and citizens over the last few decades (Bhowmick & Hazarika, 2018). Despite wide use of social media, email remains a popular communication system, since it allows for a direct and private communication between peers, often through a free service, and potentially in an anonymous way. Nevertheless, *malicious* users take advantage of these characteristics to massively distribute advertisements or bothersome messages, typically known as spam. Spam email is an unsolicited, unwanted or unhelpful message. Spam appeared in the very early days of email (Ferrara, 2019) as a way of advertising products and services. It has evolved to include scams, phishing, malware or spoofing, which put users at risk of suffering cyber-attacks (Jáñez-Martino, Alaiz-Rodríguez, González-Castro, Fidalgo, & Alegre, 2022). Organizations attempt to reduce the risks by providing cybersecurity courses for employees and by deploying state-of-the-art spam filtering technology (Jáñez-Martino, Alaiz-Rodríguez, González-Castro, & Fidalgo, 2021). However, spammers continuously engineer novel and sophisticated strategies to bypass the filters and successfully make it into the user mailbox. Despite email services raising warnings, both organizations and individuals struggle to identify online frauds and scams (Wang et al., 2022).

When users open a spam email, they may run into a message, consciously created using social engineering techniques, which involve persuading them to take certain actions – such as clicking on a link, opening an attachment or responding to the message –, often putting themselves at risk of inadvertently disclosing confidential information (Dada et al., 2019; Ferreira, Coventry, & Lenzini, 2015). Persuasion is defined as the communicative process by which a source induces a change in the beliefs, attitudes, or behaviors of a target through the strategic use of communication techniques (Cialdini, 2009; Roloff & Miller, 1980). It represents a powerful tool for spammers to reach their objective of pushing users to take a given action.

Developing intelligent systems capable of spotting persuasion is a crucial tool to enhance the security and awareness of users towards spam email for three reasons: (*i*) users can be warned about spam emails with a high persuasion load, which involves an intent of pushing for a potentially undesired action; (*ii*) researchers can obtain further insights on the reason why people gets pushed to engage with spam

* Corresponding author.
E-mail addresses: francisco.janez@unileon.es (F. Jáñez-Martino), a.barron@unibo.it (A. Barrón-Cedeño), rocio.alaiz@unileon.es (R. Alaiz-Rodríguez), victor.gonzalez@unileon.es (V. González-Castro), arianna.muti2@unibo.it (A. Muti).

https://doi.org/10.1016/j.eswa.2024.125767
Received 21 December 2023; Received in revised form 28 June 2024; Accepted 10 November 2024
Available online 19 November 2024

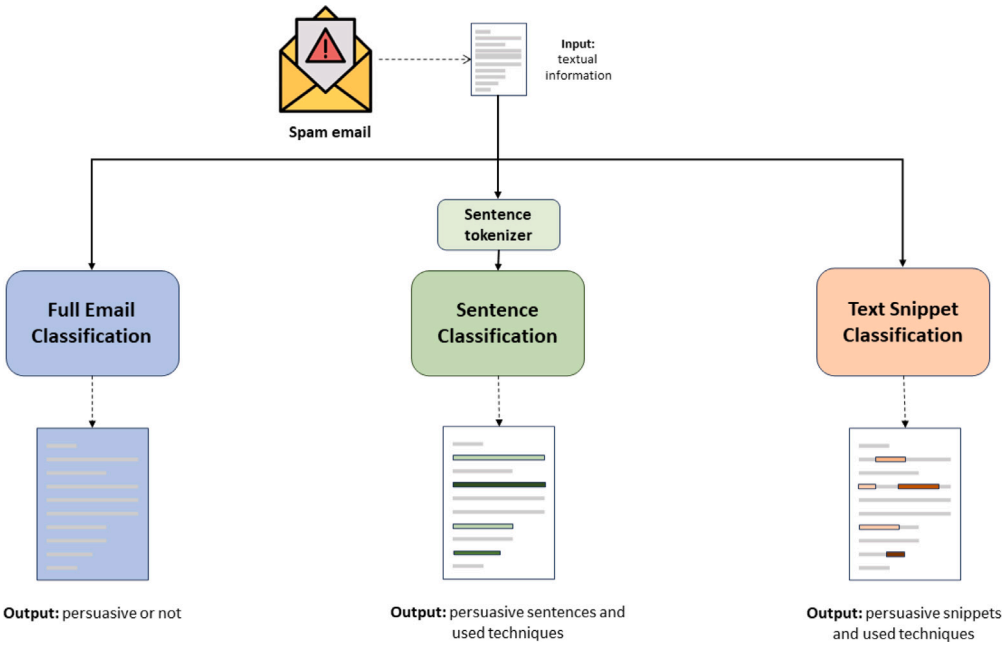


Fig. 1. The identification of persuasion at three granularity levels.

email; and (*iii*) the information about the level and kind of persuasion techniques in an email can be a factor to boost the performance of spam filtering technology.

This work seeks to respond to the following research question:

RQ: How extensively does spam email rely on persuasion to engage users and drive them to perform specific actions, and what techniques are employed in the process?

In this paper, we present a novel empirical analysis on the identification of persuasion techniques in spam email. We explore this detection at different text granularity levels: full email, sentence, and text snippet. This approach allows us to go deeper into the content and perform a thorough examination. Fig. 1 represents the approaches for persuasion identification at the three different granularity levels, for which we develop three independent models. Our work represents the first time natural language processing is employed for the identification of persuasion in spam email.

To the best of our knowledge, no empirical experimentation using natural language processing techniques has been carried out to detect and assess automatically the actual amount and type of persuasion techniques used in phishing or other kind of spam email. Most of the attention in this front comes from the analysis of persuasion in disinformation (Ali, Li, ul abdin, & Muqtadir, 2022; Chen, Xiao, & Mao, 2021) and propaganda in news articles (Barrón-Cedeño, Jaradat, Da San Martino, & Nakov, 2019; Da San Martino, Yu, Barrón-Cedeño, Petrov, & Nakov, 2019; Sony Dewantara & Budi, 2020).

We replicate and transfer the work of Barrón-Cedeño et al. (2019) to identify persuasion in spam emails at the full email level and compare it against a transformer architecture. Then, we go deeper to land at the sentence level. We create a custom dataset by manually annotating 1075 sentences from 130 spam emails following two approaches: binary – persuasive or not – and multilabel, based upon eight persuasion techniques. We then evaluate models built on top of RoBERTa (Liu et al., 2019) to address both tasks. Finally, we seek to identify which specific fragments of a spam sentence contain persuasion techniques. We follow the methodology proposed in Da San Martino, Barrón-Cedeño, Wachsmuth, Petrov, and Nakov (2020), who firstly identify the snippets, as a binary sequence labeling task, and then determine the specific persuasion technique used in each case. For both tasks, we depart from the pre-trained RoBERTa-based models using the configuration proposed by Chernyavskiy, Ilvovsky, and Nakov (2020).

Throughout our experiments, the models detected the presence of persuasive techniques, revealing that persuasion tends to be concentrated in specific sections of the spam email rather than constituting an entirely persuasive message. Our results suggest that working at the full email level presents challenges, as even the best model achieves F1 scores below 55. Our work on developing a model at the sentence level achieves a maximum F1 score of 89.40 in binary classification of persuasion. When we focused on the multilabel approach, the models perform the best in terms of F1 when detecting *Appeal to Authority* (59.30%) and *Appeal to Fear/Prejudice* (56.30%). Finally, the most frequently and accurately detected techniques at the snippet level are *Loaded Language* (30.7%), *Exaggeration or Minimization* (13.9%), and *Name Calling or Labeling* (9.9%).

The rest of the contribution is organized as follows. Section 2 reviews literature on persuasion in different genres, including email. Section 3 describes the four datasets that we work with, including the email-specific collection that we have produced. Section 4 describes our methodology and experiments. Section 5 presents and discusses our experiments and results. Section 6 sum up conclusions. Finally, Section 7 closes with the limitations.

2. Background

Due to the rise of scams in spam email, spam occurs in a variety of forms (Ferrara, 2019; Jáñez-Martino, Alaiz-Rodríguez, González-Castro, Fidalgo, & Alegre, 2023). Some authors (El Aassal, Baki, Das, & Verma, 2020; Ferreira & Teles, 2019; Magdy, Abouelseoud, & Mikhail, 2022) have adopted the term *phishing email* for a certain group of spam email. Phishing, as well as other attacks found in spam emails, aims at stealing sensitive information, extort money or cause illicit actions, usually through social engineering techniques (El Aassal et al., 2020). Such techniques take advantage of persuasive principles to manipulate people to carry out a specific action. Spammers exploit persuasion techniques to succeed in their purpose. For example, they seek users to click on malicious links masqueraded as disinformation, controversial videos and news or events, among other engaging or supposedly beneficial content for the potential victim (Frauenstein & Flowerday, 2020). Indeed, phishing has become one of the most widespread cyber-incidents (Sánchez-Paniagua, Fidalgo, González-Castro, & Alegre, 2021).

Researchers attempt to develop robust models departing from anti-spam filters, but specifically designed to spot phishing (El Aassal et al., 2020; Magdy et al., 2022; Sturman et al., 2023; Volkamer, Renaud, Reinheimer, & Kunz, 2017). While previous studies in the literature focused on phishing emails, we examine all types of spam emails in our investigation. Nowadays, anti-spam filters are mostly based on natural language processing techniques along with machine learning (El Aassal et al., 2020; Sankhwar, Pandey, & Khan, 2018) and deep learning (Alhogail & Alsabih, 2021; El Aassal et al., 2020; Halgaš, Agrafiotis, & Nurse, 2020; Lee & Verma, 2021; Magdy et al., 2022; Smadi, Aslam, & Zhang, 2018) classifiers. In addition, the process of feature selection plays a crucial role in ensuring accurate and effective filtering (Gangavarapu, Jaidhar, & Chanduka, 2020).

Despite the presence of social engineering techniques, current models have not considered persuasion as an essential factor to identify spam email yet (El Aassal et al., 2020; Sankhwar et al., 2018; Sonowal, 2020). Recent studies on phishing email detection have primarily concentrated on feature selection and extraction, focusing on aspects such as text frequency, keywords (Sonowal, 2020), readability, semantics, and lexical features (El Aassal et al., 2020), as well as link analysis (Sankhwar et al., 2018). Despite the substantial attention given to these areas, there has been a gap in research exploring the analysis of persuasive features or investigating the use of persuasion within emails through natural language processing techniques.

These persuasive factors can be potential features for developing effective systems to combat phishing attacks. Weapons of influence, such as persuasive techniques and life domains (a specific topic or aspect of an individual's life used by attackers), significantly impact the susceptibility of the user towards phishing email (Lin et al., 2019).

Some works are focused on figuring out and understanding patterns regarding the user behavior towards phishing email (Molinaro & Bolton, 2018). Hakim et al. (2021) introduced and empirically evaluated the Phishing Email Suspicion Test (PEST), which was created to understand the cognitive and neural mechanisms behind phishing susceptibility. In their experiment, participants were asked to assess the level of disguise (i.e. whether a common user would be able to identify phishing as such) of different email collections. Participants could only discriminate phishing from non-phishing emails in about 62% of the cases, which is indicative of how persuasive and harmful phishing emails can be, since they have a high potential of going unnoticed. Other authors have connected and studied the principles of persuasion in phishing email (Ferreira et al., 2015; Ferreira & Teles, 2019; Lawson, Pearson, Crowson, & Mayhorn, 2020). In particular, Ferreira et al. (2015) studied the relationship among existing principles of influence, psychological triggers and principles of scamming to identify a list of five *principles of persuasion* in phishing email:

Authority: Individuals tend to follow an expert or reference and do a great deal for someone they think is an authority. **Social poof**: Individuals often mimic the behavior of the majority. **Linking, similarity and deception**: Individuals tend to gravitate towards those who are similar, familiar, and appealing to them. **Commitment, reciprocation and consistency**: Individuals have a tendency to believe what others say and want to appear consistent in what they do; for instance, when they owe a favor. **Distraction**: Individuals often focus their attention on what they can gain or lose, for example if something will soon be unavailable, has been censored, restricted or will be more expensive later on.

Later, Ferreira and Teles (2019) set the ground for a method for the automated identification of these persuasive principles in social engineering within phishing email. They empirically observed that subject lines often provide enough information to judge whether an email is phishing. Nevertheless, they did not implement any solution or carry out experiments about it.

Most research on phishing deception and persuasion supports the notion that content and wording used in phishing messages influence susceptibility, but which persuasion tactic is the most effective remains unclear (Frank, Jaeger, & Ranft, 2022). Parsons, Butavicius, Delfabbro, and Lillie (2019) carried out an empirical experiment to determine, among others (Cialdini, 2009), the influence of these principles on a group of subjects. They concluded that both *social influence* and *impulsivity* can affect how users respond to phishing email. Phishing emails using the persuasion principles of *scarcity* and *social proof* were identified as the least successful ones in misleading the participants. Conversely, emails employing the principles of *consistency* and *reciprocity* were found the most successful.

Da San Martino et al. (2019) proposed a multilabel approach to detect 18 persuasion techniques in news articles. They also organized a SemEval-2020 shared task (Da San Martino et al., 2020) proposing two problems: snippet identification and technique classification. In the first task, given a plain-text document, a model has to identify those specific text snippets that are propagandist. The second task involves identifying the specific persuasion technique being employed, given a previously spotted text snippet and its document context.

The spammer is conscious about persuading the users to perform an action (Ferreira et al., 2015), in a similar fashion as creators of biased and hyperpartisan contents for disinformation and propaganda do. Despite the genre drift, a similar approach can be applied to identify specific persuasion mechanisms in spam email. Eight of these 18 techniques can be related to the principles of persuasion in social engineering introduced in Ferreira et al. (2015), since some techniques may be recurrent patterns in phishing email directly associated to the appearance of certain principles:

Loaded language: Using words/phrases with powerful emotional connotations (positive or negative). **Name calling or Labeling**: Labeling the object of the propaganda campaign as something to fear, hate or find undesirable. **Repetition**: employing a message over and over again. **Exaggeration minimization**: Amplifying or downplaying the representation of something by intensifying its scale, quality, or intensity. **Doubt**: Questioning the credibility of an individual or entity. **Appeal to fear/prejudice**: Trying to obtain support for an idea by inducing anxiety and/or panic within the population regarding an alternative, often relying on preconceived judgments. **Flag-waving**: Appealing to a strong national feeling (or with respect to a group, e.g., race, gender, political preference) to justify or advocate an action or idea. **Appeal to authority**: Pleading a claim as true solely based on the endorsement of a credible authority or expert in the field, without presenting any additional corroborating evidence.

Table 1 represents the association between persuasion principles in phishing email (Ferreira et al., 2015) and the persuasion techniques in news (Da San Martino et al., 2019) that could be applied to pursue them. We focus on the techniques that showed higher prevalence in Da San Martino et al. (2019). Although the technique *Appeal to Authority* tends to be less frequent (Da San Martino et al., 2019), we keep it because it remains the only technique related to the *Authority* principle. Table 2 shows instances where spammers apply some of these techniques to persuade users to perform a given action.

This association opens the door for the development, for the first time, of supervised models for the identification of persuasion in spam email.

3. Corpora

We use four datasets for our experiments. Section 3.1 describes two existing propaganda datasets: Qprop (Barrón-Cedeño et al., 2019) and PTC (Da San Martino et al., 2019). Section 3.2 introduces one existing spam email corpus, collected from the public repository Spam Archive. Finally, we present our new dataset: Persuasive Sentences in Spam Emails (PerSentSE) in Section 3.3.

Table 1
Relationship between persuasion principles in phishing email (Ferreira et al., 2015) and persuasion techniques in news (Da San Martino et al., 2019).

Persuasion Principles in Phishing emails proposed by Ferreira et al. (2015)	Persuasion techniques used in new articles detected by Da San Martino et al. (2019)
Authority	Appeal to authority
Social poof	Flag-waving Loaded language Name calling or Labeling
Linking, similarity and deception	Appeal to Authority Flag-waving Loaded language
Commitment, reciprocation and consistency	Doubt Repetition
Distraction	Appeal to fear/prejudice Exaggeration, minimization Doubt Loaded Language Name Calling or Labeling Repetition

3.1. Propaganda datasets

We rely on two corpora of news articles annotated for propaganda and persuasion at different granularity levels.

The *Qprop* corpus (Barrón-Cedeño et al., 2019) is annotated at the document level in a binary fashion: propaganda or not. The propagandist articles come from ten news outlets, including 5737 articles overall. The non-propaganda class includes 45,557 articles, coming from 94 news outlets. The data originates from the period from October 2017 till December 2018. Qprop was labeled using distant supervision, grounded on the judgments of Media Bias/Fact Check.[1]

Table 3 shows some statistics. We use Qprop for training a model at full document granularity level to detect persuasion.

To build the *Propaganda Techniques Corpus (PTC)*, Da San Martino et al. (2020) retrieved a set of news articles from a period between mid-2017 and early 2019. The annotation process consisted of manually labeling snippets along with the specific persuasion technique being applied (Da San Martino et al. (2020) originally refer to *propaganda techniques*). PTC includes 536 articles, with a total of 8981 identified snippets. Table 4 shows the PTC overall and multilabel statistics, focusing only on the eight relevant techniques for this work, resulting in 302 news articles and 4503 persuasive snippets.

Furthermore, we leveraged the PTC dataset to produce binary and multilabel annotations at the sentence level. In the binary setting, a sentence is considered *persuasive* if it contains at least one persuasive snippet. We randomly added the same number of *non-persuasive* sentences to create a balanced dataset. For the multilabel setting, a sentence is assigned the labels from all persuasion techniques occurring within it. Table 5 shows the statistics of the resulting dataset.

3.2. Spam email dataset

Since our aim is to ascertain whether spam emails contain persuasion, we need a spam email corpus to carry out our experiments. Spam Archive of Bruce Guenter[2] is a public repository that shares spam emails from his mailbox on a monthly basis since 1998. This repository represents the most up-to-date spam only corpus. We downloaded all emails from April 2021 to March 2022 and extracted the textual information following the processing pipeline in Jáñez-Martino et al. (2023). The email preprocessing method extracts the textual content from subject, body and images (through OCR) considering spammer tricks. These strategies include embedding portions or even the entire spam message within images and hiding random text within the email body, a technique known as "salting". We selected only emails in English, which resulted in 20,588 documents. We refer to this corpus as *SpamArchive2122*.

3.3. Persuasion sentences in spam email

PerSentSE stands for Persuasion Sentences in Spam Email, a corpus that consists of 130 emails randomly extracted from the *SpamArchive2122* corpus. We split emails into sentences using nltk[3] and annotate them according to the use of the eight persuasive techniques (plus a non-persuasion class). Two expert volunteers annotated twenty of the emails independently. The γ inter-annotator agreement (Mathet, Widlöcher, & Métivier, 2015), which can deal with overlapping labels, was 0.27, which is relatively low, despite γ being a pessimistic measure of agreement. Discrepancies were discussed by the two annotators until converging to a common decision. Then, they annotated twenty more emails independently, resulting in an agreement increase up to $\gamma = 0.63$, which is a fair agreement. The rest of the emails were labeled by one single annotator. Out of the 1075 sentences in the 130 emails, 216 were judged as persuasive.

Table 6 shows statistics on the PerSentSE corpus, after dividing it into training and testing partitions using a 60%–40% ratio. As for PTC, in the binary setting, a sentence is considered persuasive if it contains any of the persuasive techniques. Various persuasion techniques tend to co-occur, as stated by the statistics shown in Fig. 2. We found that the following combinations were the most common ones: *Appeal to fear/prejudice* with *Loaded Language* appear together a total of 25 times, *Exaggeration,minimization* co-occurs with *Loaded Language* 24 times and *Appeal to fear/prejudice* with *Exaggeration, minimization* 20 times. Both *Appeal to authority* and *Appeal to fear/prejudice* as well as *Exaggeration, minimization* and *Name Calling or Labeling* co-occur ten times.

In order to estimate the expected amount of persuasion in different parts of a spam message, we divided the emails according to their typical sections: *Subject* corresponds to the content of the subject header; *Greeting* is the section that serves to introduce the message to the recipient and often provides a concise explanation of the reason why the email has been sent; *Body* describes the purpose of the email at length; and *Farewell* closes the email, typically including what the recipient should do, a goodbye, and unsubscribe/contact information. Spam emails encompass a fusion of these sections, even if not all of them appear in every message. Table 7 shows the statistics of the amount of persuasion techniques appearing in each email section. The *Body* concentrates the highest amount of persuasion (82.46% of the bodies are considered persuasive), with an average of 5.51 techniques. The *Greeting* section follows, with 54.17% of the occurrences containing persuasion, with an average of 1.46 techniques.

These statistics appear to contradict the findings of Ferreira and Teles (2019), who highlighted the relevance of the subject when it comes to a persuasive email, even claiming that looking at the subject alone is enough to judge. However, it is worth noting that their study focused exclusively on phishing email, whereas our analysis encompasses all forms of spam.

4. Methodology

We design three experiments, each one focused on training models to identify persuasion at a different granularity level. These experiments provide empirical evidence to answer our research question. Experiment 1 is focused on the entire spam message, in a binary setting. Experiment 2 looks into each of the sentences of a spam email, in both binary and multilabel settings. Finally, Experiment 3 zooms into spotting specific snippets and their associated persuasion technique.

[1] https://mediabiasfactcheck.com/. Retrieved December 2023.
[2] http://untroubled.org/spam/ Retrieved December 2023.
[3] https://www.nltk.org/ Retrieved December 2023.

Table 2
Sentences extracted from spam email with persuasive text snippets to push the target user to perform an action (e.g., clicking on a link) highlighted.

	Instance	Persuasion techniques
1.	Because further research shows, **ONE of these positions: is linked to the first signs of dementia**.	**Appeal to fear** to scary the user about a health-related concern.
2.	Find out more here. This new discovery is **shocking the medical world.**	**Exaggeration** to increase the curiosity of the user.
3.	If **you're a true supporter** then you'll want to show your support for Trump 2020 in every way possible.	**Flag-waving** to call for the users's patriotism and compromise.
4.	The **$390 billion dollar diabetes industry doesn't want you** to see this video and discover the **extraordinary diabetes-fighting secret**.	**Loaded language** and **Doubt** to call for attention about the disclosure of a secret.
5.	Better yet, he'll show you how to make those same guns legally invisible to the registration system...so **snooping government goons** never come looking for them in the first place.	**Name calling/labeling** to give a negative connotation about a character.
6.	**Every Trump supporter will LOVE** this brand new, **never-before-released**, Trump 2020 commemorative golf ball.	**Flag-waving and Exaggeration** to instigate an action through patriotism and a once-in-a-time opportunity.
7.	While chemical sprays and plug-in repellents might work, they often have a pungent odour, and **may even be harmful to your family's health**.	**Appeal to fear, Doubt and Loaded language** to create a sense of danger and urgency.

Table 3
Statistics about the QProp corpus including the number of articles in each partition (Train, Development and Test) and their average length in tokens.

Label	Articles	Train	Dev	Test	Avg. length
Propaganda	5,737	4,021	575	1,141	1084.46 ± 890.59
Non-Propaganda	45,557	31,972	4564	9,021	620.31 ± 518.92
Total	51,294	35,993	5139	10,162	672.22 ± 590.98

Table 4
PTC statistics including **inst**ances per technique for the Train, Development and Test sets and their average length in terms of characters.

Persuasion technique	Train	Dev	Test	Avg. length
Loaded language	1811	304	432	23.70 ± 25.30
Name calling, labeling	931	154	209	26.10 ± 19.88
Repetition	456	115	196	16.90 ± 18.92
Exaggeration, minimization	398	81	92	45.36 ± 35.55
Doubt	423	67	72	123.21 ± 97.65
Appeal to fear/prejudice	187	52	128	93.56 ± 74.59
Flag-waving	206	34	90	61.88 ± 68.61
Appeal to authority	91	25	53	131.23 ± 123.2
Total	4503	832	1272	42.26 ± 58.34

Table 5
PTC class distribution and their average length in terms of characters in the binary (top) and in the multilabel (bottom) setting.

Class	Sentences	Avg. length
Binary		
Persuasive	1946	132.82 ± 96.24
Non-persuasive	1946	171.49 ± 79.36
Multilabel		
Appeal to authority	61	193.68 ± 121.17
Appeal to fear/prejudice	140	169.60 ± 94.87
Doubt	102	194.46 ± 134.21
Exaggeration, minimization	206	180.05 ± 98.65
Flag-waving	122	154.31 ± 101.44
Loaded Language	926	176.02 ± 102.91
Name Calling or Labeling	340	169.56 ± 99.63
Repetition	258	169.59 ± 101.33

Table 6
Binary and multilabel class distribution at the sentence level in the PerSentSE corpus. Multilabel labels come from all persuasive sentences of binary distribution.

Class	Training	Test	Avg. length
Binary			
Non-persuasive	523	336	108.86 ± 77.05
Persuasive	130	86	121.37 ± 114.03
Multilabel			
Appeal to authority	18	10	156.89 ± 88.58
Appeal to fear/prejudice	45	36	120.21 ± 74.10
Doubt	13	11	109.88 ± 63.43
Exaggeration, minimization	46	24	131.31 ± 79.36
Flag-waving	4	5	127.22 ± 76.94
Loaded Language	52	33	115.59 ± 62.25
Name Calling or Labeling	13	6	157.58 ± 97.18
Repetition	7	5	130.58 ± 89.19

Table 7
Total amount of email sections in the PersentSE corpus and extent of sections considered persuasive. Column Techniques (avg) shows the average number of techniques in each email section considered as persuasive.

Section	Total	Persuasive (%)	Techniques (avg)
Subject	59	21 (35.59)	1.62
Greeting	24	13 (54.17)	1.46
Body	57	47 (82.46)	5.51
Farewell	35	11 (31.43)	1.45

Experiment 1. We seek a binary classifier to discriminate between two classes at the document level, persuasive or not. We address it by replicating *proppy* (Barrón-Cedeño et al., 2019), a model originally intended to organize news articles according to their level of propaganda. Its pipeline includes two steps: feature extraction and classification on the basis of a maximum entropy classifier. It involves six sets of feature: (1) A *Term frequency–inverse document frequency (TF–IDF) of words* $[1,3]$-*grams* vector; (2) *lexicon features*, which include the relative frequency of words associated to specific propagandist techniques; a set of features concerning (3) *vocabulary richness* and (4) *readability*; (5) TF–IDF character 3-grams to reflect writing *style*; and, finally, (6) the *NEws LAndscape (NELA)* features (Horne, Dron, Khedr, & Adali, 2018) to measure different aspects, such as sentiment, morality and complexity.

In addition, owing to the benefits of attention mechanisms (Vaswani et al., 2017) in many natural language processing tasks for transfer learning, we have assessed the effectiveness of XLM-RoBERTa (Conneau et al., 2020). Our selection of this Transformer is informed by its broad advantages in handling diverse linguistic structures and styles in different languages, which could address challenges posed by spam structures and varying styles.

Experiment 2. Here we aim at flagging those specific sentences in a spam email that intend to persuade the target recipient. Approaching this problem with a multilabel setting allows finding out the most prevalent persuasion techniques used in spam. We built two models on top of RoBERTa (Liu et al., 2019): (*i*) a binary model to discriminate persuasive from non-persuasive sentences and (*ii*) a multilabel model to identify among the eight persuasive techniques (cf. Section 3).

Experiment 3. From the point of view of individual users, cybersecurity organizations and psychologists, it is worth highlighting those specific text snippets in which the spammer is trying to persuade an action. Experiment 3 aims at spotting specific persuasive snippets as well as the technique employed in them. Following Da San Martino et al. (2019), we address the problem in two steps: identifying snippets and then associating them to a persuasion technique. We adapt the

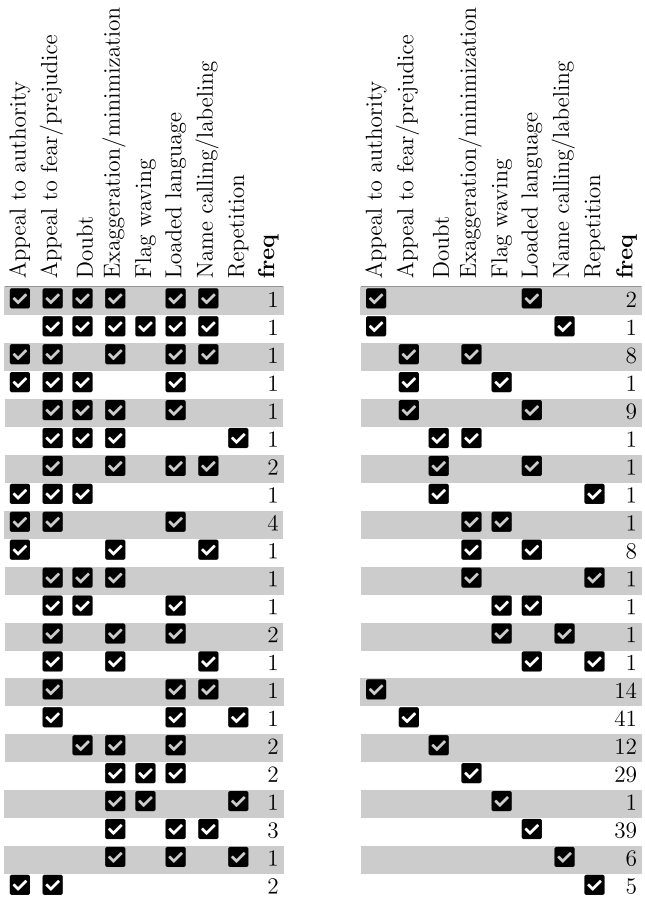


Fig. 2. Co-occurrences of the persuasion techniques within the same sentence in the PerSentSE corpus.

approach of Chernyavskiy et al. (2020), one of the top performers at the SemEval 2011 Task 11 shared task (Da San Martino et al., 2020).[4]

5. Experiments and results

5.1. Experiment 1: binary classification at the full email level

We first replicate the model of Barrón-Cedeño et al. (2019) by training and testing a document-level binary model on the *Qprop* corpus (cf. Section 3). Differently from the original approach, we subsample the word and character *n*-grams to the most frequent 15,000 elements and applied stopwording for the word *n*-grams. These heuristics reduce the feature dimension and results in a better performance. Table A.11 in Appendix shows the results on the Qprop's development and testing partitions. We took the best three configurations according to those results for our experiments on the spam email dataset: (a) a combination of word and character *n*-grams, lexicon, vocabulary richness and NELA; (b) a combination of words and character *n*-grams alone; and (c) a combination of character *n*-grams, lexicon, vocabulary richness and NELA features. Additionally, we also reported the results of the XLM-RoBERTa.

We use the full *QProp* dataset to train a Maximum Entropy classifier. Then, we apply the trained model on the *SpamArchive2122* (Section 3.2) to obtain a score per spam email. Table 8 shows the performance on 130 spam emails manually annotated from the *SpamArchive2122*. The original proppy model (Barrón-Cedeño et al., 2019) sets the decision boundary at 0.50. Due to the cross-domain, training on news articles and testing on spam, we consider various thresholds to verify the robustness of the models. We consider 0.60, 0.50, 0.40, 0.30, 0.20, 0.10, 0.05. All models performance increase as the threshold becomes lower, indicating that the differences between detected classes among the models are very small. On the other hand, the attention model barely shows differences whether the threshold is increased or decreased, suggesting that it manages to find greater distinctions between the two classes, although its F_1 score is lower. We used XLM-RoBERTa model to estimate the volume of persuasion in spam emails. Out of the 20,717 emails, the model considers 3254 as persuasive, which represents 15.71% of the emails.

This behavior is due to the cross-domain: the model trained on news articles is applied to spam emails, where the text addresses the reader directly. Moreover, rather than the usual deep coverage provided by news articles, spam email tends to be shorter in extension and more concise to reduce suspicion and capture the recipient's attention. As a result, they tend to be direct, easy to read and intend to be understood swiftly, which is not always the case for news articles. As there is not enough evidence, it is necessary to move to a lower level of granularity to reduce complexity.

5.2. Experiment 2: binary and multilabel classification at the sentence level

In this experiment, our goal is to analyze whether our models learn to distinguish persuasive from non-persuasive sentences. We consider three collections of training material: (a) the PTC alone (Section 3.1), (b) the training partition of PerSentSE alone (Section 3.2), and (c) the union of both. In all cases, we fine-tune a RoBERTa pre-trained model during 25 epochs with a batch size of 24. The number of epochs was determined through early stopping analysis. This batch size is the maximum capacity permitted by our GPUs.

Table 9 (top) shows the results on the test partition of PerSentSE. We repeat the process 10 times and report mean and standard deviation. The combination of the PerSentSE training partition with that of the PTC corpus outperforms the individual use of each of them. The boost goes from 68.80 when using PerSentSE alone to 82.20 F_1-score when using both. Adding more examples helps achieve the highest macro-average of F_1-score 89.40%. The model trained on the PTC corpus exhibits a significantly larger standard deviation, indicating a lack of robustness, which increases when both datasets are used. Such high value of standard deviation can be attributed to a cross-domain. This may be considerable between the data sources used for training, news and spam email, and testing data, consisted of only spam email. When considering different data sources, variations can arise in vocabulary, writing style, and topics. These variations can lead to differences in the model's predictions. The extent of the cross-domain in the training data plays a significant role in influencing the model's accuracy and performance. Increasing the batch size can enhance performance by generating more homogeneous batches, which helps prevent data from the same source from being isolated.

As in Experiment 1, we applied the binary model trained on the PTC and both sets of PerSentSE corpus to the *SpamArchive202122*. In this scenario, we label an email as persuasive if at least one sentence of the email is classified as persuasive. We found 7952 of 20 583 spam emails, i.e., 38.63% of them are classified as persuasive.

For the multilabel setting, we fine-tuned a multilabel RoBERTa pre-trained model, trained it for 50 epochs with a batch size of 24, considering the same settings as in the binary experiment. Table 9 (bottom) shows the results. In terms of macro-F_1, the combination of both datasets achieves the highest F_1-score, reaching 36.10%. The techniques which yielded the highest results in the PTC model also enhance the performance of PTC + PerSentSE. These techniques are *Exaggeration, minimization*, *Flag-waving* and *Name Calling, Labeling*. We may attribute this behavior to the fact that the PTC model increases the number of sentences in two minority techniques within the PerSentSE. The

[4] Their implementation is available at https://github.com/aschern/semeval2020_task11 Retrieved December 2023.

Table 8
F_1-measures (%) for Experiment 1 on 130 spam emails from the SpamArchive2122 dataset using the top-performing models on the Qprop dataset (see Appendix).

Features groups	Thresholds						
	>0.60	>0.50	>0.40	>0.30	>0.20	>0.10	>0.05
(a) word & char *n-grams*	29.33	36.59	47.73	58.95	60.61	62.39	70.09
(b) char *n-grams*+lexicon+voc. richness+nela	36.36	39.51	40.96	50.55	55.32	60.61	63.46
(c) word & char *n-grams*+lexicon+voc. richness+nela	27.78	29.73	35.90	40.00	55.56	59.41	64.81
Attention model							
XLM-RoBERTa	52.87	52.87	52.87	52.87	52.87	54.54	54.54

Table 9
Results of experiment 2 on spotting persuasive sentences when training on PTC, PerSentSE and both. The classes are **non-per**suasive and **per**suasive. The values of Precision (%), Recall (%)and F_1-Score (%) are the mean of 10 evaluations of the experiment. We also report the standard deviation (σ) of these 10 repetitions.

	PTC				PerSentSE				PTC+PerSentSE			
	P	R	F_1	σ	P	R	F_1	σ	P	R	F_1	σ
Binary setting												
non-per	87.00	92.50	89.60	1.50	91.20	94.20	92.60	0.49	94.90	97.00	95.90	1.70
per	60.50	45.60	51.80	6.98	74.40	64.20	68.80	2.40	86.40	79.50	82.80	8.07
macro	73.80	69.10	70.70	3.95	82.60	79.20	80.70	1.42	90.60	88.30	89.40	4.88
Multilabel setting												
AA	25.90	16.00	17.90	10.74	50.30	73.00	**59.30**	6.69	22.70	19.00	20.50	3.77
AF	11.60	0.90	1.80	2.75	65.40	50.70	**56.60**	6.59	68.80	30.20	41.50	5.66
D	33.10	19.30	23.50	7.92	72.50	20.00	**31.20**	1.25	55.10	21.40	30.60	2.11
EM	38.80	33.40	35.50	5.95	44.90	26.30	32.80	4.98	62.70	36.30	**45.60**	3.58
FW	55.20	34.60	41.90	7.97	30.00	4.20	7.50	11.46	82.60	30.40	**43.10**	4.74
LL	47.40	7.10	11.90	4.89	45.40	41.00	**43.00**	4.49	40.60	28.70	33.50	2.77
NCL	21.00	54.70	30.30	4.17	24.10	6.60	10.30	8.45	52.40	45.50	**47.80**	6.05
R	0.90	8.40	1.60	2.15	100.00	17.00	**29.00**	0.00	83.30	17.00	27.50	2.42
macro	29.40	21.90	20.60	1.50	54.00	29.80	33.70	2.10	58.50	28.40	**36.10**	1.70

addition of the PTC resulted in a notable decrease in the performance on techniques such as *Appeal to authority*, *Appeal to fear/prejudice*, *Doubt* and *Loaded Language*. This may indicate that the vocabulary mismatch between news and spam negatively impacts the ability to spot these techniques. The levels of standard deviation show that the combination of both datasets rises the reliability on most techniques. Fragment-level inference can provide us with more information about the keywords used by spammers for each technique.

5.3. Experiment 3: snippet identification and technique classification

Our base model[5] is RoBERTa for both tasks with an optional Conditional Random Field (CRF). The CRF layer only uses tokens that start with words and skips tokens starting with special characters or numbers, thereby, we do not use the CRF layer to avoid losing potential worthy information. The model for technique classification is built on top of RoBERTa as well. Due to our GPUs limitations, we select 256 as the maximum sequence length used. Thus, we split larger spam emails before feeding them to the model. We use the same configuration of Chernyavskiy et al. (2020): learning rate of 2e-5 and a batch size of 24 during 30 epochs for both subtasks (this configuration managed to replicate fairly well the results reported in Chernyavskiy et al. (2020)).

After training the models, we applied them on the *SpamArchive202122* dataset. Table 10 shows the performance in terms of precision. We focus on precision alone because we lack annotation at this level of granularity and we performed a manual verification of the predictions. snippets with **Loaded Language**, **Name Calling, Labeling** and **Exaggeration, minimization** were found the most, and with a precision of 55.36, 62.50 and 77.78, respectively. The model spotted eight cases of doubt, but with a low level of precision: 37.50. The rest of techniques were hardly observed, resulting also in some precision values of 0. Hence, there is still room to improve the knowledge transferring approach.

[5] https://github.com/aschern/semeval2020_task11, Retrieved December 2023.

Table 10
Results of Experiment 3 when detecting persuasion techniques in the multilabel classification setting.

Persuasive technique	Spotted snippets	Precision
Appeal to authority	1	0.00
Appeal to fear/prejudice	1	0.00
Doubt	8	37.50
Exaggeration, minimization	18	77.78
Flag-Waving	1	0.00
Loaded Language	56	55.36
Name Calling or Labeling	16	62.50
Repetition	–	–
Micro average	101	58.00

5.4. Discussion

Our results show that working at the full email level is far from convenient, since even the best model reaches F_1 levels below 55. Beside the domain shift – from news articles to email – this subpar performance gives evidence about the nature of persuasion in spam email, which appears in specific parts of the message, rather than being spread.

When we zoom into the sentence level, we have managed to produce models that identify persuasive content at a more reasonable level of performance, F_1 of 82.80 on the positive class with a Recall close to 80. Indeed, persuasion in spam tends to be applied in specific text fragments. Identifying the specific persuasion technique remains an open issue, with *Name Calling or Labeling* being the techniques that we afford to identify the best. Nevertheless, there is still room for improvement if a model is expected to identify specific techniques.

The same trend translates to the snippet level, where the models identify *Name Calling or Labeling* and *Exaggeration, minimization* the best. If we consider that these techniques are among the top-three techniques in terms of frequency, together with *Loaded Language*, we can conclude that we are close to make models to characterize persuasion in spam email operational.

Regarding our RQ, we have identified indicators that substantiate the presence of persuasion in spam emails, specially in specific sections

of the message. We can assert that spammers use persuasive content, and this is more easily detectable when the model is focused on individual sentences rather than the full email. By analyzing our dataset PersentSE, we found that 20.1% of the sentences in spam email contain persuasive content. Specifically, *Loaded Language* is used in 7.9% of the sentences, *Appeal to Fear/Prejudice* in 7.5%, and *Exaggeration or Minimization* in 6.5%. Our models tend to frequently detect *Name Calling or Labeling*, as well as instances employing *Loaded Language* and *Exaggeration or Minimization* techniques.

6. Conclusions

In this paper, we aimed at analyzing persuasion used in spam email at three granularity levels: full document, sentence and text snippet. This analysis enabled us to address our **RQ** by providing the following insights into the presence of persuasive elements in spam emails and identifying the most commonly used techniques. In order to do that, we delved on existing datasets with annotation both at the document and at the fragment level in binary and multi-label (snippet labeling) settings from the news domain.

Our experiments in binary settings – persuasive or not – show that persuasion in spam email is localized in specific text fragments and therefore working at the document level is not sensitive. Spammers try to persuade their target in specific text fragments rather than on the whole contents of the messages.

Spotting persuasive sentences within an email seems doable and is close to turn this process practical in order to both understand how spammers try to persuade their targets and to provide additional information to spam filtering models in order to better identify different kinds of spam. This could be useful not only to create better spam filters, but also to better explain the final user why a given message has been filtered out from their mailbox.

As in the domain of news, only some persuasion techniques are identified at reasonable performance levels, including *Exaggeration, minimization*, *Name Calling or Labeling* and *Loaded Language*. Indeed these techniques are the most frequent in spam email and tend to appear together with other techniques. As a result, we envision that the efforts to characterize and uncover persuasion in spam email should focus on these three techniques.

Spam emails encompass a wide range of writing styles, from unstructured advertisements to formal letters that simulate legitimate messages. However, the primary aim of all spam emails is to manipulate the recipient in some way to achieve their target. This work sheds light on the use of persuasion techniques for manipulation, enhancing our psychological understanding of how spammers can achieve their goals. This information can be valuable for cybersecurity organizations, scientists, companies, and individuals, by helping them to recognize, analyze, and be wary of the persuasive elements in emails.

Our work provides a starting point for future studies investigating how persuasion techniques affect different cybersecurity challenges, e.g., spam emails, online scams, and other related areas. It remains crucial to create spam email datasets annotated at the snippet level to train more robust models.

7. Limitations

We trained most of the models for full document and text snippet classification on news datasets annotated for persuasion. Even with the domain shift, we could carry out a preliminary assessment of persuasive content within spam email. Still, as shown for the case of sentence-level classification, training models on annotated emails results in a more robust model, evidencing the necessity of producing higher volumes of high-quality email corpora annotated for persuasion at different levels of granularity.

We focused only on the visible textual information. Following the methodology of Jáñez-Martino et al. (2023), we used Optical Character Recognition (OCR) techniques to extract text from the subject and body, as well as from text embedded in images. In future research, we can consider other components of emails, such as the images themselves or the visual layout created by HTML. This would align with studies that adopt a multi-modal perspective to detect persuasion, such as Dimitrov et al. (2021). We also focus on English alone, but spam is a global challenge and further research should be carried out in other languages, be it in a monolingual or in a multilingual fashion.

CRediT authorship contribution statement

Francisco Jáñez-Martino: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Alberto Barrón-Cedeño:** Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Supervision, Project administration. **Rocío Alaiz-Rodríguez:** Conceptualization, Validation, Investigation, Resources, Writing – review & editing Supervision. **Víctor González-Castro:** Conceptualization, Validation, Investigation, Resources, Writing – review & editing Supervision. **Arianna Muti:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix. Original results of Proppy

Table A.11 shows a comparison between the performance of Proppy in the original work (Barrón-Cedeño et al., 2019) and the present work

Table A.11
Results for Development and Test sets of the Proppy evaluation on the Qprop dataset using feature groups proposed by Barrón-Cedeño et al. (2019) in terms of F_1-Score.

Features	Original work (Barrón-Cedeño et al., 2019)		Present work	
	Dev	Test	Dev	Test
word *n-grams*	74.42	75.55	80.61	78.11
lexicon	46.55	44.87	46.55	44.87
voc. richness	29.45	29.72	29.45	29.72
readability	21.96	21.50	21.96	21.50
char *n-grams*	**82.93**	**82.13**	80.65	79.02
nela	54.60	50.98	54.60	50.98
word *n-grams* + char*n-grams*	78.37	79.01	84.05	83.78
char *n-grams* + lexicon	83.02	81.94	81.03	79.04
char *n-grams* + nela	**83.21**	82.75	81.66	79.56
readability + nela	75.34	76.83	54.50	50.98
char *n-grams* + lexicon + voc.richness + nela	83.17	**82.89**	81.60	79.82
word & char *n-grams* + lexicon + voc. richness + nela	79.04	79.50	**84.68**	**83.81**

in terms of F_1 score. The results of features lexicon, vocabulary richness, readability and NELA are similar to the original ones. Regarding word and char n-grams features, the overall results improve slightly, thereby, their concerning combinations also enhance their performance. We found the combination of word and chars n-grams, lexicon, vocabulary richness and NELA features as the highest F_1-score performance in both validation and testing sets. Followed by the combination of words and chars -grams and, in the third position, that of chars n-grams, lexicon, vocabulary richness and NELA features.

Data availability

Data will be made available on request.

References

Alhogail, A., & Alsabih, A. (2021). Applying machine learning and natural language processing to detect phishing email. *Computers & Security*, *110*, Article 102414. http://dx.doi.org/10.1016/j.cose.2021.102414.

Ali, K., Li, C., ul abdin, K. Z., & Muqtadir, S. A. (2022). The effects of emotions, individual attitudes towards vaccination, and social endorsements on perceived fake news credibility and sharing motivations. *Computers in Human Behavior*, *134*, Article 107307. http://dx.doi.org/10.1016/j.chb.2022.107307.

Barrón-Cedeño, A., Jaradat, I., Da San Martino, G., & Nakov, P. (2019). Proppy: Organizing the news based on their propagandistic content. *Information Processing & Management*, *56*(5), 1849–1864. http://dx.doi.org/10.1016/j.ipm.2019.03.005.

Bhowmick, A., & Hazarika, S. M. (2018). E-mail spam filtering: A review of techniques and trends. *Advances in Electronics, Communication and Computing*, *443*, 583–590. http://dx.doi.org/10.1007/978-981-10-4765-7_61.

Chen, S., Xiao, L., & Mao, J. (2021). Persuasion strategies of misinformation-containing posts in the social media. *Information Processing & Management*, *58*(5), Article 102665. http://dx.doi.org/10.1016/j.ipm.2021.102665.

Chernyavskiy, A., Ilvovsky, D., & Nakov, P. (2020). Aschern at SemEval-2020 task 11: It takes three to tango: RoBERTa, CRF, and transfer learning. In *Proceedings of the fourteenth workshop on semantic evaluation* (pp. 1462–1468). Barcelona (online): International Committee for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.semeval-1.191.

Cialdini, R. B. (2009). *vol. 4, Influence: Science and Practice*. Pearson education Boston, MA.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., et al. (2020). Unsupervised cross-lingual representation learning at scale.

Da San Martino, G., Barrón-Cedeño, A., Wachsmuth, H., Petrov, R., & Nakov, P. (2020). SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the fourteenth workshop on semantic evaluation* (pp. 1377–1414). Barcelona (online): International Committee for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.semeval-1.186, URL https://aclanthology.org/2020.semeval-1.186.

Da San Martino, G., Yu, S., Barrón-Cedeño, A., Petrov, R., & Nakov, P. (2019). Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 5636–5646). Hong Kong, China: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D19-1565, URL https://aclanthology.org/D19-1565.

Dada, E. G., Bassi, J. S., Chiroma, H., Abdulhamid, S. M., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon*, *5*(6), Article e01802. http://dx.doi.org/10.1016/j.heliyon.2019.e01802.

Dimitrov, D., Bin Ali, B., Shaar, S., Alam, F., Silvestri, F., Firooz, H., et al. (2021). Detecting propaganda techniques in memes. In C. Zong, F. Xia, W. Li, R. Navigli (Eds.), *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)* (pp. 6603–6617). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.acl-long.516.

El Aassal, A., Baki, S., Das, A., & Verma, R. M. (2020). An in-depth benchmarking and evaluation of phishing detection research for security needs. *IEEE Access*, *8*, 22170–22192. http://dx.doi.org/10.1109/ACCESS.2020.2969780.

Ferrara, E. (2019). The history of digital spam. *Communications of the ACM*, *62*(8), 82–91, https://www.doi.org/10.1145/3299768.

Ferreira, A., Coventry, L., & Lenzini, G. (2015). Principles of persuasion in social engineering and their use in phishing. In T. Tryfonas, & I. Askoxylakis (Eds.), *Human aspects of information security, privacy, and trust* (pp. 36–47). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-319-20376-8_4.

Ferreira, A., & Teles, S. (2019). Persuasion: How phishing emails can influence users and bypass security measures. *International Journal of Human-Computer Studies*, *125*, 19–31. http://dx.doi.org/10.1016/j.ijhcs.2018.12.004.

Frank, M., Jaeger, L., & Ranft, L. M. (2022). Contextual drivers of employees' phishing susceptibility: Insights from a field study. *Decision Support Systems*, *160*, Article 113818. http://dx.doi.org/10.1016/j.dss.2022.113818.

Frauenstein, E. D., & Flowerday, S. (2020). Susceptibility to phishing on social network sites: A personality information processing model. *Computers & Security*, *94*, Article 101862. http://dx.doi.org/10.1016/j.cose.2020.101862.

Gangavarapu, T., Jaidhar, C., & Chanduka, B. (2020). Applicability of machine learning in spam and phishing email filtering: review and approaches. *Artificial Intelligence Review*, *53*, 64. http://dx.doi.org/10.1007/s10462-020-09814-9.

Hakim, Z. M., Ebner, N. C., Oliveira, D. S., Getz, S. J., Levin, B. E., Lin, T., et al. (2021). The Phishing Email Suspicion Test (PEST) a lab-based task for evaluating the cognitive mechanisms of phishing detection. *Behavior Research Methods*, *53*, 11. http://dx.doi.org/10.3758/s13428-020-01495-0.

Halgaš, L., Agrafiotis, I., & Nurse, J. R. C. (2020). Catching the phish: Detecting phishing attacks using recurrent neural networks (RNNs). In I. You (Ed.), *Information security applications* (pp. 219–233). Cham: Springer International Publishing.

Horne, B., Dron, W., Khedr, S., & Adali, S. (2018). Sampling the news producers: A large news and feature data set for the study of the complex media landscape. *Proceedings of the International AAAI Conference on Web and Social Media*, *12*, http://dx.doi.org/10.1609/icwsm.v12i1.14982.

Jáñez-Martino, F., Alaiz-Rodríguez, R., González-Castro, V., & Fidalgo, E. (2021). Trustworthiness of spam email addresses using machine learning. In *Proceedings of the 21st ACM symposium on document engineering* (p. 4). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3469096.3475060.

Jáñez-Martino, F., Alaiz-Rodríguez, R., González-Castro, V., Fidalgo, E., & Alegre, E. (2022). A review of spam email detection: analysis of spammer strategies and the dataset shift problem. *Artificial Intelligence Review*, 29. http://dx.doi.org/10.1007/s10462-022-10195-4.

Jáñez-Martino, F., Alaiz-Rodríguez, R., González-Castro, V., Fidalgo, E., & Alegre, E. (2023). Classifying spam emails using agglomerative hierarchical clustering and a topic-based approach. *Applied Soft Computing*, *139*, Article 110226. http://dx.doi.org/10.1016/j.asoc.2023.110226.

Lawson, P., Pearson, C. J., Crowson, A., & Mayhorn, C. B. (2020). Email phishing and signal detection: How persuasion principles and personality influence response patterns and accuracy. *Applied Ergonomics*, *86*, Article 103084. http://dx.doi.org/10.1016/j.apergo.2020.103084.

Lee, D., & Verma, R. M. (2021). Adversarial machine learning in text: A case study of phishing email detection with RCNN model. In *Adversary-aware learning techniques and trends in cybersecurity* (pp. 61–83). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-55692-1_4.

Lin, T., Capecci, D. E., Ellis, D. M., Rocha, H. A., Dommaraju, S., Oliveira, D. S., et al. (2019). Susceptibility to spear-phishing emails: Effects of internet user demographics and email content. *ACM Transactions on Computer-Human Interaction*, *26*(5), http://dx.doi.org/10.1145/3336141.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. CoRR abs/1907.11692. arXiv:1907.11692. URL http://arxiv.org/abs/1907.11692.

Magdy, S., Abouelseoud, Y., & Mikhail, M. (2022). Efficient spam and phishing emails filtering based on deep learning. *Computer Networks*, *206*, Article 108826. http://dx.doi.org/10.1016/j.comnet.2022.108826.

Mathet, Y., Widlöcher, A., & Métivier, J.-P. (2015). The unified and holistic method Gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, *41*(3), 437–479. http://dx.doi.org/10.1162/COLI_a_00227, URL https://aclanthology.org/J15-3003.

Molinaro, K. A., & Bolton, M. L. (2018). Evaluating the applicability of the double system lens model to the analysis of phishing email judgments. *Computers & Security*, *77*, 128–137. http://dx.doi.org/10.1016/j.cose.2018.03.012.

Parsons, K., Butavicius, M., Delfabbro, P., & Lillie, M. (2019). Predicting susceptibility to social influence in phishing emails. *International Journal of Human-Computer Studies*, *128*, 17–26. http://dx.doi.org/10.1016/j.ijhcs.2019.02.007.

Roloff, M. E., & Miller, G. R. (Eds.), (1980). *SAGE series in communication research, Persuasion: New Directions in Theory and Research*. Thousand Oaks, CA: SAGE Publications, Inc.

Sánchez-Paniagua, M., Fidalgo, E., González-Castro, V., & Alegre, E. (2021). Impact of current phishing strategies in machine learning models for phishing detection. In A. Herrero, C. Cambra, D. Urda, J. Sedano, H. Quintián, & E. Corchado (Eds.), *13th international conference on computational intelligence in security for information systems (CISIS 2020)* (pp. 87–96). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-57805-3_9.

Sankhwar, S., Pandey, D., & Khan, P. R. (2018). Email phishing: An enhanced classification model to detect malicious URLs. *ICST Transactions on Scalable Information Systems*, *6*, Article 158529. http://dx.doi.org/10.4108/eai.13-7-2018.158529.

Smadi, S., Aslam, N., & Zhang, L. (2018). Detection of online phishing email using dynamic evolving neural network based on reinforcement learning. *Decision Support Systems*, *107*, 88–102. http://dx.doi.org/10.1016/j.dss.2018.01.001.

Sonowal, G. (2020). Phishing email detection based on binary search feature selection. *SN Computer Science*, *1*, 14. http://dx.doi.org/10.1007/s42979-020-00194-z.

Sony Dewantara, D., & Budi, I. (2020). Combination of LSTM and CNN for article-level propaganda detection in news articles. In *2020 fifth international conference on informatics and computing* (pp. 1–4). http://dx.doi.org/10.1109/ICIC50835.2020.9288532.

Sturman, D., Valenzuela, C., Plate, O., Tanvir, T., Auton, J. C., Bayl-Smith, P., et al. (2023). The role of cue utilization in the detection of phishing emails. *Applied Ergonomics*, *106*, Article 103887. http://dx.doi.org/10.1016/j.apergo.2022.103887.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. CoRR abs/1706.03762. arXiv:1706.03762.

Volkamer, M., Renaud, K., Reinheimer, B., & Kunz, A. (2017). User experiences of TORPEDO: Tooltip-powered phishing email detection. *Computers & Security*, *71*, 100–113. http://dx.doi.org/10.1016/j.cose.2017.02.004.

Wang, J., Shi, J., Wen, X., Xu, L., Zhao, K., Tao, F., et al. (2022). The effect of signal icon and persuasion strategy on warning design in online fraud. *Computers & Security*, Article 102839. http://dx.doi.org/10.1016/j.cose.2022.102839.