



A review of spam email detection: analysis of spammer strategies and the dataset shift problem

Francisco Jáñez-Martino^{1,2}  · Rocío Alaiz-Rodríguez^{1,2} · Víctor González-Castro^{1,2} · Eduardo Fidalgo^{1,2} · Enrique Alegre^{1,2}

© The Author(s) 2022

Abstract

Spam emails have been traditionally seen as just annoying and unsolicited emails containing advertisements, but they increasingly include scams, malware or phishing. In order to ensure the security and integrity for the users, organisations and researchers aim to develop robust filters for spam email detection. Recently, most spam filters based on machine learning algorithms published in academic journals report very high performance, but users are still reporting a rising number of frauds and attacks via spam emails. Two main challenges can be found in this field: (a) it is a very dynamic environment prone to the dataset shift problem and (b) it suffers from the presence of an adversarial figure, i.e. the spammer. Unlike classical spam email reviews, this one is particularly focused on the problems that this constantly changing environment poses. Moreover, we analyse the different spammer strategies used for contaminating the emails, and we review the state-of-the-art techniques to develop filters based on machine learning. Finally, we empirically evaluate and present the consequences of ignoring the matter of dataset shift in this practical field. Experimental results show that this shift may lead to severe degradation in the estimated generalisation performance, with error rates reaching values up to 48.81%.

Keywords Spam email detection · Dataset shift · Adversarial machine learning · Spammer strategies · Feature selection

1 Introduction

Communication media is an essential tool for society and a considerable vector for fraudulent content, like fake rewards, identity fraud, extortion, phishing or malware transmission. Many cybercriminals design harmful scam messages daily sent to millions of people worldwide taking advantage of the technology advances. Email services provide a free, possibly anonymous, and quick way of propagating the scams via the Internet (Ferrara 2019).

✉ Francisco Jáñez-Martino
francisco.janez@unileon.es

¹ Department of Electrical, Systems and Automation, University of León, León, Spain

² Researcher at INCIBE (Spanish National Cybersecurity Institute), León, Spain

Although email users traditionally see spam just as annoying, unsolicited advertisements or a loss of time, it is increasingly associated with a tricky and potential risk for their security, integrity and reliability on the web (Gangavarapu et al. 2020). Additionally, Kaspersky Lab¹ and Cisco Talos² place spam emails between 50% and 85% of total worldwide emails sent in a day, above 200 billion, which turns it into a big scale problem.

Since spam email has been a problem during the last few decades, organisations and researchers aim to build robust and efficient filters to stop it. Nowadays, in the literature, many models based on machine learning algorithms (Dedeturk and Akay 2020; Saidani et al. 2020) show excellent performance with accuracies over 90% to detect whether an email is spam or legitimate (often referred to as *ham*). However, despite remarkable performance results and enhancements of filters, users are still reporting scams and attacks whose roots are spam emails.

Spammers—i.e., any people or organisation sending unwanted emails—obtain a benefit using scams included in emails and, thereby, seek to keep invisible for the filters. To accomplish their purpose, they continuously apply new strategies to bypass the spam filters (Redmiles et al. 2018), taking advantage of filters weaknesses. They manipulate the emails in different ways, like inserting the spam message into an image to avoid being detected by textual filters. Thus, spammers may be considered as the adversarial figure in this field. From a forensic perspective, investigating spammer strategies in emails may help to find out these kinds of disguises in other fields that suffer both an adversarial figure and digital crimes (Yu 2015). First, Wang et al. (2013) and, then, Bhowmick and Hazarika (2018) already analysed spam trends pointing out the dynamic nature of spam content. The authors (Wang et al. 2013) warned that the spam email was not dying, but becoming to be more fancy and sophisticated.

This artificial change in data joined with the natural evolution of email data over time means that researchers should develop tools for detecting spam emails in a non-stationary environment (Mohammad 2020). Approaches using supervised learning are based on the assumption that training and test data come from the same distribution. However, effective strategies able to tackle the dataset shift and adversarial manipulation problems are necessary in order to handle security attacks and detect spammer corruption in data (Gangavarapu et al. 2020).

In this paper, unlike classical spam email reviews (Bhowmick and Hazarika 2018; Dada et al. 2019; Gangavarapu et al. 2020; Karim et al. 2019), we present a literature revision focused on both the analysis of the increasingly sophisticated spammer tricks and the dataset shift that appears in this practical application. Our goal is to highlight the importance of both challenges to build more robust spam filters over time. We pay attention to the main spammer strategies, their aim, evolution over the years, properties and presence in the previous decade emails using datasets provided by Spam Archive of Bruce Guenter.³ We also review how researchers deal with detecting and reducing the spammer manipulation effects in filters.

Besides, we hypothesise that ignoring the fact that the spam email field is a changing environment may lead to severe degradation of generalisation and performance of any model. A similar idea was previously used by Pérez-Díaz et al. (2012), proposing an evaluation methodology to anticipate possible filter problems and avoid a drop-in performance

¹ <https://www.statista.com/statistics/420391/spam-email-traffic-share/> Retrieved June 2021.

² https://talosintelligence.com/reputation_center/email_rep Retrieved June 2021.

³ <http://untroubled.org/spam/> Retrieved June 2021.

during their operation. In our study, we increased the number of datasets and considered the temporal evolution, though. Therefore, we assess the consequences of assuming that training samples follow the same distribution as the email samples in the operational environment. The spam classifiers we train result from the combination between two frequency text encoders: Term Frequency-Inverse Document Frequency (TF-IDF) and Bag of Words (BOW) alongside two machine learning algorithms: Naïve Bayes (NB) and Support Vector Machine (SVM), both widely used as spam email filters. We use five datasets from 2000 to 2010: Ling-Spam, SpamAssassin, Enron-Spam, TREC07 and CSDMC, for calibrating different spam filters and, later on, we apply them to categorise email from different scenarios using data from 2000 to 2018.

The rest of the paper is organised as follows: Sect. 2 introduces the background of the dataset shift problem and the adversarial machine learning approach. We review recent spammer strategies in Sect. 3 and spam email filters in Sect. 4. Section 5 includes the datasets description, experimentation setup, results and discussion. Finally, in Sect. 6, we present our conclusions and future highlights.

2 Background

2.1 The problem of dataset shift

A fundamental assumption of supervised learning is that training and test data remained the same (although unknown) distribution (Hand 2006). However, it may be some mismatches that are likely to appear in real-world scenarios, and as a result, this assumption is often violated.

This problem has been referred to with different terms, i.e., dataset shift, concept shift, concept drift or just drift (Moreno-Torres et al. 2012; Quionero-Candela et al. 2009) and it has attracted increasing attention over the last decade (Biggio and Roli 2018; Gama et al. 2014; González-Castro et al. 2013; Liu et al. 2020; Simester et al. 2020; Webb et al. 2016).

It is important to highlight that in the presence of dataset shift, classification models often fail to generalise in the deployment environment and their performance may deteriorate significantly (Alaiz-Rodríguez and Japkowicz 2008; Kull and Flach 2014). A comprehensive analysis of the effects of dataset shift on probability distributions has been presented in Moreno-Torres et al. (2012) and Quionero-Candela et al. (2009) as well as a categorisation of the different types of dataset shift: covariate shift (distribution shift in features), prior probability shift (shift in classes), concept shift (shift in the relationship between features and classes) and other types of shift.

In order to mitigate the problem of dataset shift, different strategies have been proposed (Gama et al. 2014; Kadwe and Suryawanshi 2015; Yu et al. 2019): (i) first, to detect the presence of dataset shift and categorise it into different types and (ii) to choose the most suitable classifier from a pool of calibrated classifiers according to the shift detected.

Spam email filtering has been tackled using different machine learning approaches including NB, SVM, Random Forest (RF) or Neural Networks (NN), among others. Some of these proposals have reported high performance, i.e. around 90% (Bhowmick and Hazarika 2018; Dada et al. 2019; Ferrara 2019). Thus, a recent work in 2020, Dedetürk and Akay (2020) developed a spam filter model which achieved an accuracy of 98.70%. However, the spam and ham messages used for evaluation in that work were extracted from a dataset of email examples generated during the 2000–2010 decade. The same applies to

(Bahgat et al. 2018; Dedetürk and Akay 2020; Diale et al. 2019; Faris et al. 2019; Gibson et al. 2020; Naem et al. 2018; Saidani et al. 2020). It is important to consider that spam email has a changing nature due to the evolution of the topics through time and the techniques used by spammers wishing to elude spam filters, leading to shifts in the dataset. The presence of dataset shift in this domain suggests that the anti-spam filters presented above are likely to fail more than expected on new unseen examples.

Some of the first approaches to handle the dataset shift, also known as concept drift, inherent in email spam data, has relied on lazy learners (Delany et al. 2005; Fdez-Riverola et al. 2007). Basically, the proposal in Delany et al. (2005) was based on (i) daily taking the misclassified by the system cases to update the case-base at the end of each day, (ii) periodically re-training the system and re-selecting features using the most recent cases. Other two techniques were presented in Fdez-Riverola et al. (2007) for tracking concept drift in this domain and using a lazy learner. Firstly, the RTI (Relevant Term Identification) technique, which performed a selection of representative terms based on the information contained in each email. Secondly, RMS (Representative Message Selection), that selected those emails more applicable given the actual context implementation.

A study (Ruano-Ordas et al. 2018a) has shown different weaknesses of several spam filtering alternatives. In particular, the authors have provided a detailed analysis of the real impact of different types of concept drift (i.e., sudden drift, re-occurring drift, gradual drift, and incremental drift) on the spam-filtering domain. Their study highlighted many issues caused by concept drift on this problem: concept drift in ham messages, different kinds of concept drift in both ham and spam messages, or topics with multiple concept drift types. In addition, they identified the inner causes of concept drift, e.g., changes in business activities, the variation in marketing interests along time, communication, linguistic aspect or economy. In this domain, the Dynamic Weighted Majority Concept Drift Detection (DWM-CCD) algorithm (Nosrati and Pour 2011) was able to deal with sudden and gradual concept drift, but was unsuitable to tackle more complex scenarios of dataset shift.

The dynamic characteristics of the spam email domain have also been studied in Mohammad (2020). The authors considered a cyclical concept drift appears in this field because the list of characteristics used for spam emails may disappear and reappear every certain period of time. This paper addressed the concept drift as well as other catastrophic forgetting issues, i.e. past strategies from spammers, in order to get a lifelong classification model based on the ensemble learning strategy. Their proposal relied on the Early Drift Detection Method (EDDM) (Baena-García et al. 2006) to confirm whether a concept drift was actually happening and in that case an Ensemble based Lifelong Classification using Adjustable Dataset Partitioning (ELCADP) attempted to adapt the spam filter to any change in the class distribution. The performance of ELCADP has not been examined with a virtual concept drift where the input features are unchanged, however, a new class value might appear.

2.2 Adversarial machine learning

Machine learning algorithms have been applied in many fields with efficient performances (Al Nabki et al. 2017; Riesco et al. 2019). However, there is a set of them, such as phishing detection (Sánchez-Paniagua et al. 2021), spam detection (Lam and Yeung 2008; Dedetürk and Akay 2020) or botnet detection (Velasco-Mata et al. 2019) that continuously require updating the models due to an adversarial figure. Nevertheless, organizations and

researchers should tackle this problem considering the specific nature of each field. For instance, phishing has different properties from spam, such as imitating branch logos, asking for sensitive information or transmitting urgency to users.

The adversary takes advantage of the vulnerability caused by dataset shift and consciously alters the data to mislead the classifiers. Dalvi et al. (2004) defined the adversarial figure as the one who introduces malicious data to defeat the classifiers. Barreno et al. (2006) created a taxonomy of adversarial attacks through three criteria to identify the attack and how to defend a classification model from it. Huang et al. (2011) extended the study of Barreno et al. (2006) introducing a deeper analysis of the adversarial features, attack taxonomy and adversarial capabilities. Recently, Wang et al. (2019) presented an overview of this field emphasising three challenges for next few years: security in deep learning models, effective and efficient data encryption to ensure the model privacy and new evaluation mechanisms.

The adversarial classification has been mainly studied from two different perspectives. The first one seeks to measure the classifier stability against adversarial attacks (Biggio et al. 2013; Goodfellow et al. 2015; Laskov and Kloft 2009; Lu et al. 2020; Nelson et al. 2011; Paudice et al. 2018). Following this approach, Nelson et al. (2011) quantified the classifier stability under adversarial training data contamination by introducing a metric to classify its robustness. Laskov and Kloft (2009) first, and Biggio et al. (2013) later, proposed frameworks for security analysis and evaluation of classification algorithms which simulated attack scenarios. Goodfellow et al. (2015), focused on non-linearity and overfitting problems, inspected examples of adversarial data to find out the weaknesses in the NN classifiers. In order to mitigate the effects of poisoning attacks, Paudice et al. (2018) created an algorithm for pre-training. Lu et al. (2020) uncovered the remarkable weaknesses of quantum machine learning algorithms against the adversarial settings.

Other authors approach this field attempting to evaluate the attack efficiency (Apruzzese et al. 2019; Papernot et al. 2015a, 2017; Shi et al. 2019). For instance, Papernot et al. (2015a) formalised the space of attacks against deep NNs and presented an algorithm capable of crafting adversarial samples based on a precise understanding of the mapping between inputs and outputs. Papernot et al. (2017) carried out an attack using a black box adversary to a real-world deep learning application and demonstrated the viability of evading the defence strategies. Apruzzese et al. (2019) explored the possible damages that an attack, based on poisoning and evasion strategies, can cause to a cyber-detector. They highlighted the need to develop more robust machine learning techniques, in cybersecurity terms. Shi et al. (2019) assessed an effective poisoning attack to spectrum recognition applications. The adversarial classification seems to be a never-ending game between adversary and defender, who attempts to alleviate the adversarial attacks.

Generally, the studies that deal with understanding machine learning security in adversarial strategies focus on spam email detection (Chen et al. 2018), whose adversarial figure is known as spammer. Spammers aim to evade the classifier without affecting the readability of the email content, for instance, introducing specific misspelling or legitimate words into the message (Biggio and Roli 2018). Hence, spam emails may contain malicious data properly injected by spammers to harm the information used for training the classifiers and, therefore, subvert their normal operation filter (Xiao et al. 2018). Nelson et al. (2008) made visible the vulnerabilities of SpamBayes filter⁴ using a dictionary attack by contaminating

⁴ <http://spambayes.sourceforge.net/> Retrieved June 2021.

only a little number of training set emails. Despite having successfully explored two defences against dictionary attacks, they observed that an attack with extra knowledge would be challenging to defend. To design machine learning models with more robust and effective security defences, works like (Dasgupta and Collins 2019; Rota Bulo et al. 2017) built models taking its basis on game theory, e.g. automatically simulating attacks from adversaries, and validated them using spam email datasets. Naveiro et al. (2019) provided an alternative framework based on adversarial risk analysis and evaluated it on spam email datasets, in contrast to following a game theory approach.

3 Spammer tricks

Spammers continuously contaminate emails with smart and creative strategies to bypass the anti-spam filters (Wittel and Wu 2004). On the opposite side, organisations and researchers develop new techniques to mitigate the effects of these strategies or tricks on spam filters. Given the dynamic and vulnerable nature (Bhowmick and Hazarika 2018) of this domain due to this adversarial activity, its concept drift does not only appear as a natural change, which can be fixed by updating the model, but also it is designed to avoid detection by traditional concept drift techniques (Sethi and Kantardzic 2018).

The concept drift with these characteristics is commonly known as adversarial drift. This drift always seeks to subvert the most recently used classifiers, gaining information about them previously and manipulating data accordingly to make them produce false negatives (Dalvi et al. 2004). Dada et al. (2019) stated that recent researchers and main providers of email services, like Gmail or Outlook, employ a combination of different machine and deep learning techniques to build their filters generally involving text classification. This approach has caused the spammers to focus their efforts in outmatching the textual filters by disturbing the data extracted from the email body and legible headers like the subject. Reflecting these considerations, Ruano-Ordas et al. (2018b) used evolutionary computation to automatically generate regular expressions as an aid to filter spam emails and to detect patterns.

Spammers adapt emails to surpass anti-spam filters and also attempt to cheat the receiver by imitating the appearance of legitimate emails, creating confusing situations like false forwards or taking advantage of spam campaigns (Oliveira et al. 2019; Redmiles et al. 2018).

In this study, we review the tricks designed to evade spam filters, normally based on machine learning algorithms, such as poisoning text, obfuscated words or hidden text salting. For this reason, we only include these spammer strategies without considering a final user perspective, such as cyber-attacks using backscatter spam (Hijawi et al. 2021, 2017). We refer to social engineering as a set of techniques to seek to steal personal or sensitive confidential information from users using fake online companies, providers or people. Next, we will explore the main strategies followed by spammers during the last few decades and the respective research advances to mitigate their influence on the spam filters.

3.1 Poisoning text and obfuscated words

A spam email is typically transmitted in textual format and analysed by rule-based filters and text classifiers based on machine learning algorithms (Biggio and Roli 2018) such as NB, Logistic Regression (LR) or SVM. Since anti-spam filters based on machine learning

classifiers are widely used for detecting spam emails (Pitropakis et al. 2019), spammers often attempt to mislead them through contaminating textual information. Spammers use textual manipulation techniques, known as poisoning text and obfuscated words, on the entire emails body or certain words within it, e.g. misspellings or adding random or ham-legitimate words to a spam message (Wang et al. 2019).

The poisoning text technique depends on the spammer's knowledge level about the anti-spam filter and the receiver. Thereby, there are non-personalised and personalised poisoning attacks. The former may insert into the email body random words or popular ham-legitimate words, namely "word salad", and avoid using well-known spam words (Kuchipudi et al. 2020). Nevertheless, due to their high knowledge level, personalised attacks are more harmful and hard to detect, since they include specific words targeted to the anti-spam filter and the victim.

The use of obfuscated words attempt to modify words keeping them readable, e.g. embedding special character, using HTML comments (Bhowmick and Hazarika 2018), or leetspeak (Peng et al. 2018). Table 1 shows the main obfuscated word techniques with their respective examples. The actual text is "free spam message" for all examples.

Several works have addressed poisoning text in the spam email field (Kuchipudi et al. 2020; Peng et al. 2018; Shams and Mercer 2016) considering linguistic attributes for training supervised classifiers. They used: (1) word-level attributes, such as counting the number of spam words, alphanumeric words or function words, (2) error attributes, both misspelling words and grammatical mistakes, (3) readability attributes analysed by simple and complex words, TF-IDF and different reading scores, and (4) HTML attributes. Thus, Peng et al. (2018) proposed an enhancement for the NB classifier capable of detecting text modifications, in particular leetspeak and diacritics.


Recently, Kuchipudi et al. (2020) demonstrated the vulnerability of a spam message filter based on a Bayesian classifier attacked by three invasive techniques: synonym replacement, ham-word injection and spam-word spacing. The classifier was easily bypassed in the three scenarios analysed. Chan et al. (2021) designed a transfer learning based on a countermeasure to deal with the flipping label poisoning by extracting the benign knowledge from contaminated data in adversarial environments. Feature selection has also been explored in this context (Méndez et al. 2019) introducing a new semantic-based feature selection method to spam models and highlighting that it is worth exploring obfuscated tricks and poisoning attacks to enhance the feature selection process.

3.2 Hidden text salting

The hidden text salting is a technique to disturb the proper behaviour of textual filtering by introducing random text opportunely hidden in the email background. Spammers started to use this strategy around the middle of the decade of the 2000s in multiple communication platforms, and some works (Bergholz et al. 2008; Lioma et al. 2008; Moens et al. 2010) highlighted that it could be commonly found in phishing emails.

Since the spammers often insert this unseen text into HTML tags (Moens et al. 2010), some studies were focused on analysing and using the font colour, font size or glyph as features (Bergholz et al. 2008; Lioma et al. 2008) and Optical Character Recognition (OCR) approaches (Bergholz et al. 2008) to find out emails containing text hidden intentionally. These works implemented an SVM classifier to detect the hidden text, achieving high accuracy. Moens et al. (2010) identified statistics of salting detection in spam email datasets. Despite seeming a past approach, we still found it in several emails provided by recent

Table 1 Most popular obfuscated word techniques with examples for each one

Obfuscated word technique	Behaviour	Example
Embedding characters	Words are divided by special characters, letters, number or spaces, visually readable for human eye, but complicated to process for machines	F-r-e-e s p a m m e l s l s l a l g l e
Pattern recognition	Consist in varying the order of the words' letters remaining the first and last letter or using leetspeak and diacritics changes. Despite these modifications, the human eye recognises the words, but machines do not	Fr33 späm message
HTML encoding	Use of HTML entity encoding to code entire words or characters of spam message which is decoded for the users	freespad message
HTML tag	Create custom HTML tags to split words taking advance of HTML rendering engines which do not recognize up this tag. Thus, the message is readable for user	Free <customtag> random text </customtag> spam message
HTML code	Change the font size or colour to embed specific words and hide with the background, making the message to look "random" for machines	Free spam <p style="color:white;"> D </p> cccc message
HTML tables	Introduce the spam message into HTML tables which allows to be legible for user, but it looks like random letter for textual filters due to the vertical understanding	fsm rpe eas ems a g e
ASCII mural	The spammer draws the message using a characters combination	

The original text for all the examples is "free spam message"

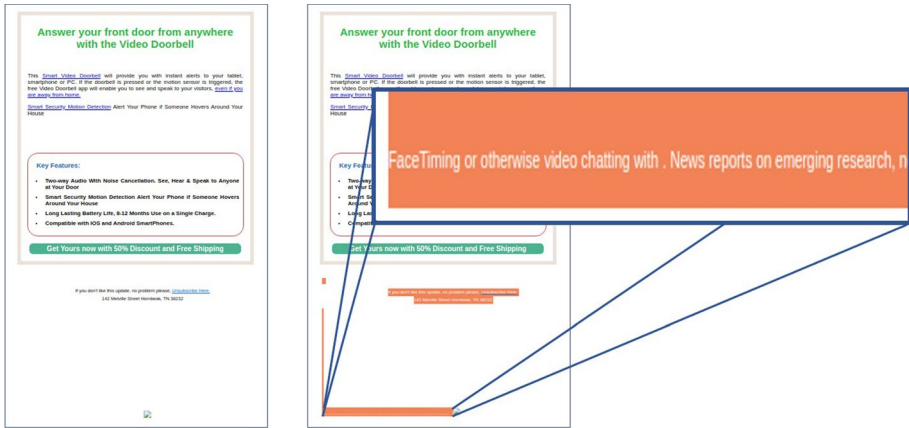


Fig. 1 Example of Hidden text in an email. Example extracted from Bruce Guenter 2020

datasets, like *Bruce Guenter* from 2020, as we show in the example that appears in Fig. 1. *Bruce Guenter* dataset, also known as Spam Archive, is a publicly available dataset used in many works (Metsis et al. 2006; Ruano-Ordas et al. 2018a; Méndez et al. 2019), in which the author has uploaded sets of spam emails from personal honey pots since 1998 every month. This fact indicates that spammers still make use of the hidden text salting technique nowadays as a strategy to evade anti-spam filters.

3.3 Image-based spam email

In the mid 2000s, spammers started to introduce the spam message into images, instead of writing it in the email body. Image-based spam made the textual processing ineffective (Biggio et al. 2011).

During that period, several works that presented a binary classification of spam images into ham or spam were published (Byun et al. 2007; Mehta et al. 2008; Wang et al. 2007). To evaluate the proposed classifiers, researchers built and made publicly available several datasets of ham and spam images extracted from emails. The most popular image datasets are Image Spam Dataset (Dredze et al. 2007), Standard Dataset or Image Spam Hunter (Gao et al. 2008) and Princeton Spam Image Benchmark (Wang et al. 2007). Moreover, Biggio et al. (2007) faced spammer tricks on images, such as using content obscuring techniques to defeat OCR tools.

More recent models have used machine learning algorithms and image properties, like the metadata or colour as features (Aiwan and Zhaofeng 2018; Chavda et al. 2018; Zamil et al. 2019). Authors trained and evaluated their models on the above-mentioned datasets obtaining high performance. However, techniques to generate the image-based spam have evolved, and spammers have modified the message content format and the image appearance, as shown in Fig. 2. Annadatha and Stamp (2016) evaluated Principal Component Analysis (PCA), and SVM approaches on Standard Dataset (Gao et al. 2008) and their own private and improved dataset gathered in 2016, obtaining a much higher performance on Standard Dataset than on the private one. In order to overcome this degradation against

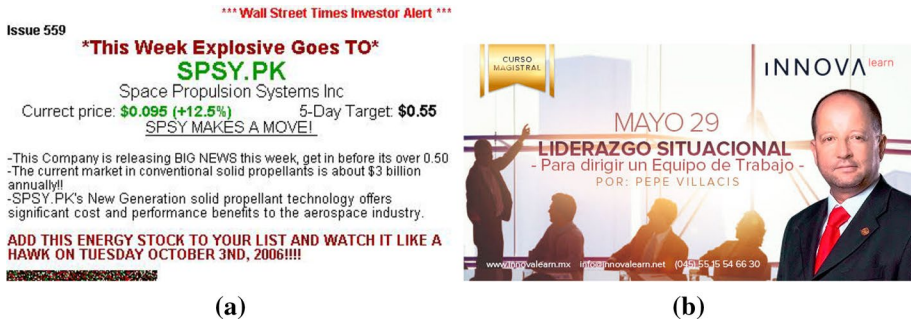


Fig. 2 Examples of image-based spam. Image a) from the 2007 dataset Image Spam Dataset, whereas b) is an example from a 2019 private collection of spam emails

new and unseen datasets, Kim et al. (2020) built a CNN-XGBoost model along with data augmentation based techniques.

To sum up, the lack of models assessed on recent datasets, containing current spam images, may make us question whether the high performance, demonstrated on the datasets from the late 00s, degrades on recent datasets or not. Additionally, Naiemi et al. (2019) handled image-based spam from an OCR perspective to extract the letters and words from the attached images. Dhah et al. (2019) achieved an improvement in the effectiveness of the classification by processing both text and image features rather than using each one separately.

Experimental results have shown that the combination of both image and text features improves the effectiveness of the classification with regard to the case in which only image or text features are used.

3.4 Other emerging strategies

Since spam email is an open arms-race, spammers continuously look for alternative strategies to bypass whatever advance against them. By observing previous strategies, we may affirm spammers hide their tricks from the user, taking advantage of any breach in email formats and attachments.

According to Alazab and Broadhurst (2016), Ferrara (2019) and Tran et al. (2013), attachments and URLs contained into emails are one of the main vectors of malicious files. Some works (Arivudainambi et al. 2019; Cohen et al. 2018) proposed a model to detect malware in web-mail files. However, attached files and URLs may be a hook for sharing malware and a means of transmitting a spam message via the email. File extensions, like PDF or docx, allow to show a message on the mail client interface and, at the same time, evade the textual filter focused on the email body or image.

An apparently legitimate email may include URLs which load images containing a spam message, or link to the true spam website. Current spam emails tend to avoid attaching images and directly load them through links inside HTML code. URLs are widely used on phishing emails that becomes an open problem in this field and deserve further research (El Aassal et al. 2020; Gupta et al. 2017). We depict in Fig. 4 the evolution of the percentage of spam email containing images or other files, like PDF or docx, as an attachment or inline forms. We analysed the spam emails from the last ten years extracted from Spam Archive of Bruce Guenter.

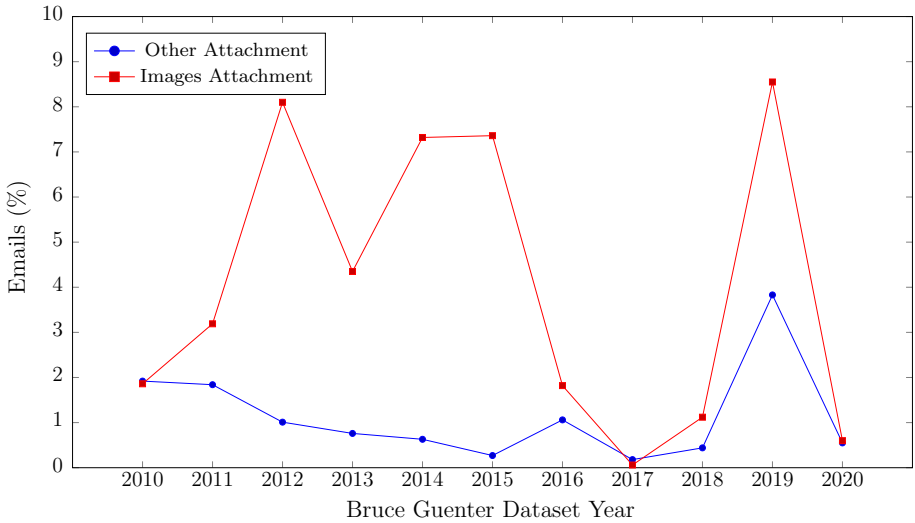


Fig. 3 The graphic depicts the percentage of spam emails (axis Y) which contains images attached or other files over the last decade (axis X). We used spam emails provided by Spam Archive of Bruce Guenter and considered spam emails which include an image or other file in both inline and attachment form

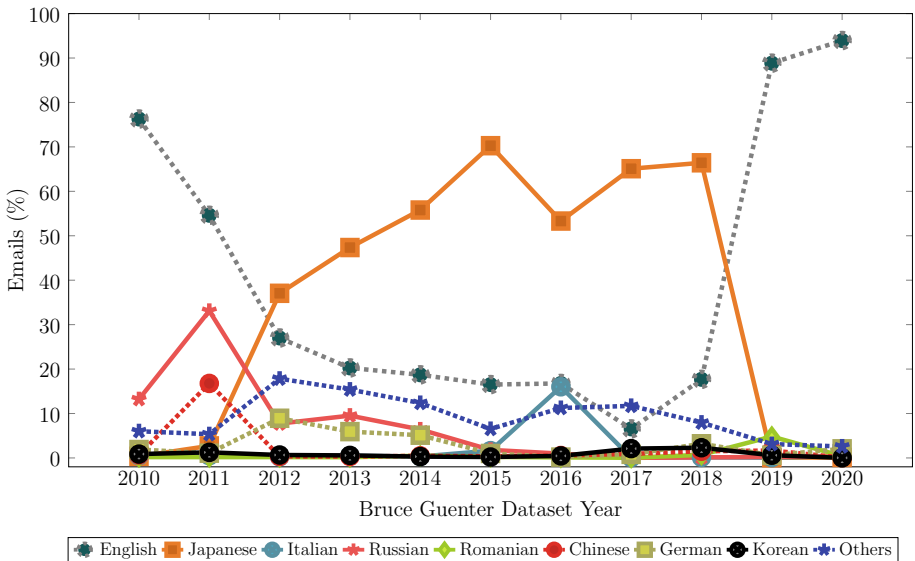


Fig. 4 The graphic depicts the percentage of spam emails (axis Y) written in most used languages, including minor languages in “Other” legend, over the last decade (axis X). We used spam emails provided by Spam Archive of Bruce Guenter and the Python3 library langdetect (<https://pypi.org/project/langdetect/>) to implement the language classifier

We depict in Fig. 3 the evolution of the percentage of spam email containing at least an image and other files apart from emails as an attachment or inline form. In Fig. 3, we can see a cyclical behaviour, images attached in emails increased from 2010 to 2012, decreased

from 2015 to 2017, reaching its minimum level in 2017 (0.06% of emails with attached images and 0.18% with other attachments), and, finally, reappeared from 2017 to 2019. Mohammad (2020) already pointed out that the concept drift in spam email is cyclical, i.e. some characteristic disappear at certain period of time, but return later on.

Including parts of the message in a different language represents an emerging textual strategy. English has been the most representative language in the spam email field for the last few decades. Nevertheless, during the last few years, spam emails are written in increasingly numerous languages (Fig. 3). There may be several reasons, such as increasing the number of spam emails inside each country or a conscious use of the languages to mislead the users and filters. In general, emails in a foreign language, depending on the user settings, are categorised as spam. However, a mix of languages may trick some filters by detecting an expected language among them.

Merging languages plays with the user's confusion and, additionally, may include social engineering techniques. All these new strategies open new challenges to cyber-security and NLP research. In Fig. 3 we show the language evolution in the last decade, taking the spam emails from Spam Archive as reference. In April of 2018, Bruce Guenter let his main domain of receiving spam expired; thereby, he lost his most source of spam email. In 2017, Bruce Guenter uploaded 1,400,401 of spam emails, reducing this number up to 23,859 in 2019, which may explain the sudden decrease of Japanese spam emails from 2019.

4 Binary spam email classification

4.1 Spam filters

Motivated by the dynamic nature of the adversarial environment created by spammers, many Artificial Intelligence approaches, mainly based on Natural Language Processing techniques, have addressed the problem of filtering spam email, i.e., categorising email into two classes: legitimate or unwanted email, popularly known as ham and spam, respectively (Bassiouni et al. 2018). More recently, some works have also tackled the problem of spam email from a different perspective, introducing the classification of spam emails into multiple classes depending on their topic (Jáñez-Martino et al. 2020; Murugavel and Santhi 2020).

To feed the filters, researches make use of headers, body or both from emails (Mohammad 2020). Recent works are focused on analysing body parts, textual message and attachments (Dada et al. 2019)—usually with text classification techniques—on or extracting features from the email like semantic-based properties (Saidani et al. 2020). In this section, we will explore the proposals based on feature weight, feature extraction or feature selection to enhance email filters based on machine learning classifiers.

Spam filters are the main tool to deal with the spam problem (Bhowmick and Hazarika 2018). Initially, the filters were exclusively based on rule-user settings, black and white lists and keywords detection, i.e., solutions provided from a knowledge engineering perspective (Sanghani and Kotecha 2019). However, these approaches tend to be inefficient and quickly outdated, which requires a manual, continuous and rigorous maintenance and upgrade, which is time and resource consuming (Gibson et al. 2020). During the last decade, machine learning algorithms have proved to overcome knowledge engineering drawbacks (Dada et al. 2019). Recent studies in this field rely on traditional classifiers

Table 2 The most used classifiers in spam email filtering with their paper references

Classifiers	References
Naïve Bayes	2018), Gibson et al. (2020), Méndez et al. (2019), Peng et al. (2018), Saidani et al. (2020), Shams and Mercer (2016) and Zavvar et al. (2016)
Support Vector Machine	2018), Diale et al. (2019, 2016), Gibson et al. (2020), M. et al. (2012), Méndez et al. (2019), Sanghani and Kotecha (2019), Shams and Mercer (2016), Sumathi and Pugalendhi (2020) and Zavvar et al. (2016)
Random Forest	2018), Bassiouni et al. (2018), Diale et al. (2019), Gibson et al. (2020), Saidani et al. (2020) and Shams and Mercer (2016)
Decision Tree	2019), Gibson et al. (2020), Méndez et al. (2019) and Saidani et al. (2020)
Logistic Regression	2018), Dedetürk and Akay (2020), Méndez et al. (2019) and Zavvar et al. (2016)
Boosting	2018), Saidani et al. (2020) and Sumathi and Pugalendhi (2020)
K-Nearest Neighbour	2018), Saidani et al. (2020) and Sumathi and Pugalendhi (2020)
Neural Network	2016), Faris et al. (2019), Awad and Foqaha (2016), Bahgat et al. (2018) and Bassiouni et al. (2018)
Deep Learning	2018), Srinivasan et al. (2021) and Sumathi and Pugalendhi (2020)

(Bhowmick and Hazarika 2018; Dada et al. 2019). Thus, the most common machine learning algorithms recently used in the literature are listed in Table 2.

Despite the rise of deep learning methods, traditional algorithms (Dedetürk and Akay 2020; Gibson et al. 2020; Méndez et al. 2019; Saidani et al. 2020) are still leading the spam email filtering field (Faris et al. 2019; Sumathi and Pugalendhi 2020). One of the reasons behind this may be the high performance already achieved with more simple models based on traditional classifiers (Ferrara 2019). Additionally, this is a practical application where the model weight and capability of adapting to any environment play an important role. So far, deep learning models tend to be heavier and require more computational resources, which benefits the continuity of traditional machine learning algorithms for this practical application (Barushka and Hajek 2018).

Table 3 summarises the accuracy and F1-score metrics reported in these studies as well as the classification model proposed. It also highlights the performance for each method evaluated on the most widely used datasets in the literature.

In general, most of the models report an accuracy above 90%, when they are evaluated on well-know and relevant public spam email datasets. Nevertheless, these encouraging studies ignore the issue of dataset shift and the adversarial data manipulations from spammers. The models analysed were published between 2018 and 2020, but they were calibrated and evaluated using emails recollected in the period from 2000 to 2010. Therefore, they overlook the issue that spam email is a changing environment. These datasets may not be representative because they do not totally cover the current range of spammer strategies, that rapidly evolve and go back over time.

The common assumption that the joint distribution of inputs and outputs is stationary definitely leads to an important loss of generalisation in applications heavily affected by dataset shift, like spam filtering. To address this problem, a lifelong model is presented in Mohammad (2020) to automatically adjust the number of dataset partitions to create a new classification model.

Finally, Ferrara (2019) stated that state-of-the-art research of spam detection lies behind a close curtain. Although companies like Microsoft and Google do not often publish studies about spam filtering to avoid revealing their system, spammers seem to obtain enough

Table 3 Results in terms of Accuracy (Acc) and F1-Score (F1) of three-year spam emails filters published in the literature and evaluated on dataset from 2000s decade

Paper	Year	Dataset year	Dataset	Model	Acc (%)	F1
Barusha and Hajek	2018	2006*	SpamAssassin	DBB-RDNN-ReL	99.89	–
		2006	Enron 1		98.76	–
Bahgat et al.	2018	2006	Enron-Spam	Semantic measures-LR	95.00	0.950
Naem et al.	2018	2006*	SpamAssassin	Antlion	98.91	0.999
		2010	CSDMC	Optimization Boosting	99.80	0.995
Bassiouni et al.	2018	1999	SpamBase Dataset	Random Forest	95.45	0.954
Diale et al.	2019	2000	Enron	Unsupervised learning: SVM	–	0.978
		2007	TREC07	Unsupervised learning: SVM	–	0.984
Sanghani and Kotecha	2019	2006	ENRON1	TFDCR: SVM	96.69	0.960
		2006	ENRON2		97.38	0.965
		2006	ENRON3		96.74	0.959
		2006	ENRON4		99.20	0.988
		2006	ENRON5		97.25	0.967
		2006	ENRON6		97.63	0.966
		2000	PU1		97.97	0.979
		2003	PU2		96.18	0.935
		2003	PU3		97.57	0.975
Faris et al.	2019	2006*	SpamAssassin	Auto GA: RWN	96.80	0.875**
		2000	Ling-Spam		93.30	0.640**
		2010	CSDMC		91.10	0.876**
Sumathi and Pugalendhi	2020	1999	SpamBase Dataset	Random Forest – Deep Neural Network	88.59	–
Saidini et al.	2020	2010	CSDMC	Semantic Features-AdaBoost	98.53	0.988
Dedeturk et al.	2020	2010	CSDMC	Artificial bee colony	98.70	–
		2006	Enron1	– LR	98.91	–

*Emails from SpamAssassin dataset correspond to the period between 2002 and 2006. **Original works did not report the F1-Score metric, which has been calculated using reported precision and recall values

information about state-of-the-art spam filtering from literature and the behaviour of the practical filters deployed in current applications.

4.2 Text encoders

Sometimes, as we indicated in Sect. 3.1, spammers introduce random, legitimate, tricky or useless words to poison the spam textual message to confuse the filters. This manipulation increases the number of textual and semantic features that may negatively affect the predictive

performance of filters based on text encoders, the computational time for processing as well as the memory resources. Both preprocessing and feature extraction play an important role in execution time and classification accuracy has been the goal of several research papers (Bahgat et al. 2018; Diale et al. 2019; Saidani et al. 2020).

The preprocessing phase attempts to remove unnecessary text, like stop-words, duplicated content or special characters. Depending on the application field, it may include stemming and lemmatization techniques and detecting entities such as web pages, email addresses or bitcoin wallets (Sanghani and Kotecha 2019). This step should be adapted depending on the application field properties (Al Nabki et al. 2020), e.g. in spam environment; it may have a possible adversarial manipulation (Kuchipudi et al. 2020).

After the preprocessing stage, the text is converted into a numerical vector which later feeds the classifier. Most anti-spam filters use term count and frequency-based feature techniques to represent the semantic meaning of an email message into a vector (Diale et al. 2019) following an n-gram approach, such as BOW (Bhowmick and Hazarika 2018; Saidani et al. 2020) or TF-IDF (Barushka and Hajek 2018; Bhowmick and Hazarika 2018; Dedetürk and Akay 2020; Diale et al. 2016; Gibson et al. 2020; Sumathi and Pugalethi 2020).

However, TF-IDF or BOW do not capture the word order or context and tend to generate high dimensional vectors. To ensure the lowest possible dimensional feature space with a fixed-length numerical vector for every email, Diale et al. (2019) generate a vector space model by means of distribution BOW and distributed Memory. To decrease the space and time complexity of feature vectors, Bahgat et al. (2018) applied semantic-based methods and similarity measures using WordNet ontology, reducing the number of extracted textual features. Méndez et al. (2019) employed a semantic-based approach and WordNet ontology to generate a reduced feature space of grouping message knowledge in topics rather than only words. A method based on semantic analysis was proposed by Saidani et al. (2020) to detect spam emails. Their goal was to categorise both ham and spam emails into predefined domains and then, extracting semantic features for each domain.

Unlike frequency encoders, word embedding techniques are based on the fact that words with similar context tend to be closer and related in the vector space. The models are trained on natural language vocabulary and their relationship between words. Word embedding has evolved from precursors like Word2vec (Mikolov et al. 2013a, b) to contextual language embedding, e.g. EIMo (Peters et al. 2018) and then, to transformers such as Devlin et al. (2018), RoBERTa (Liu et al. 2019) or GPT-3 (Brown et al. 2020).

The number of research papers that successfully apply word embedding techniques to spam email filtering is quite limited (Saidani et al. 2020; Srinivasan et al. 2021). Some reasons behind this fact are: (i) the number of words that appear on spam environments is larger than in regular natural language vocabulary due to the use of obfuscated or misspelling words by spammers and (ii) frequency and semantic-based techniques already show apparently high performance. However, word embedding approaches are a powerful tool to detect spammer strategies related to language quality and ambiguous information and extremely sentimental messages that spammers tend to send to mislead the users.

4.3 Feature selection

Feature selection is an essential phase in many classification problems, specially, in those with high dimensional datasets that may contain a large number of irrelevant, noisy and redundant features (Vinitha and Renuka 2020).

It is a popular statement that the size of the training dataset grows exponentially with the number of dimensions (Méndez et al. 2019).

Feature selection techniques may provide many advantages for spam filtering, such as Cai et al. (2018): (a) the improvement of the classification performance by selecting an optimum subset of features and (b) the obtention of faster and more cost-effective spam filters. Both efficiency and computational cost are essential requirements for spam email filtering applications.

Several works have addressed the problem of dimensionality reduction for email spam classification (see Vinita and Renuka 2020 and references therein). They are mainly focused on text classification approaches, i.e. text encoded into numeric vectors.

Feature selection methods assess the relevance of a feature or a set of features according to a given measure. A traditional perspective to select features in the spam environment is to use classic statistical methods like Information Gain or Chi-Square (Diale et al. 2016; Rehman et al. 2017) or more recent ones such as Infinite Latent Feature Selection (Bassiouni et al. 2018). Other approaches involve heuristic algorithms such as Genetic Algorithm (Gibson et al. 2020; Hong et al. 2015), Artificial Bee Colony (Dedetürk and Akay 2020) or Particle Swarm Optimisation (Gibson et al. 2020; Zavvar et al. 2016). There are also hybrid models based on joining a heuristic approach with machine learning algorithms (M. et al. 2012) like NNs, e.g. Genetic Algorithm along with Random Weighted Network (Faris et al. 2019) or Particle Swarm Optimisation and Radial Basis Function NNs combination (Awad and Foqaha 2016; Sumathi and Pugalendhi 2020).

Apart from statistical, heuristic or hybrid methods, other semantic-based approaches that consider the semantic similarity among words have been proposed (Bahgat et al. 2018; Méndez et al. 2019). Thus, Bahgat et al. (2018) compressed features in a reducing dimensional space by considering the semantic relationship and semantic similarity measures. Méndez et al. (2019) developed a semantic topic-based feature selector to filter spam using topics instead of words that decreased remarkably the number of features.

The presence of an adversary, i.e., the spammer, implies that features extracted from the email may be contaminated. Hence, feature selection can also enhance the security of spam email classifier against evasion attacks, such as random and common-ham words, by incorporating specific assumptions of data manipulation strategies (Zhang et al. 2016). Several efforts have been made to readjust the feature sets continuously (Diale et al. 2019; Sanghani and Kotecha 2019). Thus, Sanghani and Kotecha (2019) presented an incremental learning mechanism for a feature selector able to update a discriminate function and heuristic function to identify new relevant features automatically, making more robust the personalised email spam filters. In Diale et al. (2019), the authors developed an unsupervised feature learning based on Autoencoders.

5 Experimental study of binary spam email classification

In this section, we assess four spam filters calibrated with five email datasets that were gathered from different sources in different periods of time. Previous studies report high performance for the spam filters on these datasets. Our aim is to find out whether or not the spam filters maintain their generalisation performance when applied to categorise email from a dataset different to the one used for learning. Our initial hypothesis is that filter performance deteriorates as a result of the dataset shift and spammer strategies inherent to the spam email datasets.

This section is organised as follows: Sect. 5.1 presents the datasets used in the experimentation. The experimental settings are provided in Sect. 5.2 and finally, the performance of the different scenarios is discussed in Sect. 5.3.

5.1 Spam email datasets

Most of the publicly available datasets used for training and testing novel spam email filtering models presented in the literature correspond to the period between 2000 and 2010. To develop our experimentation, we have selected five well-known datasets, containing both spam and ham emails, allowing us to cover epochs with several years of difference. These datasets are: Ling-Spam Androutsopoulos et al. (2000), SpamAssassin⁵ (Project 2005), Enron-Spam (Metsis et al. 2006), TREC07 (Cormack (2007)) and CSDMS 2010.⁶

Ling-Spam recollects 2893 emails (without attachments, HTML tags and duplicate spam emails) from the Linguist List whose major linguistic interests involve job postings, research opportunities, software availability announcements and flame-like responses.

SpamAssassin, created by Justin Mason, from Network Associates, contains 6047 emails published in public fora or donated by users, being thereby less topic-specific than a single-user dataset.

Enron-Spam joins ham emails from six Enron employees and spam emails obtained from four different sources, i.e. the SpamAssassin corpus, the Honeypot Project,⁷ Bruce Guenter collection and the personal mailbox of one of the dataset creators. The creators focused on building a personalised spam dataset and published six sub-datasets for emulating different situations faced by real users.

TREC07 contains all emails of a particular server, which had many accounts, and honeypots accounts published on the web from April 8 to July 6, 2007. This corpora was distributed to participants in a competition for developing a spam filter.

CSDMC is a dataset used for the data mining competition associated with ICONIP 2010 and all messages were published on public fora and received from non-spam-trap sources, i.e. corresponding to non-personalised email datasets as SpamAssassin and Ling-Spam.

Additionally, for the evaluation of the models, we have also used two datasets extracted from the Spam Archive of Bruce Guenter,⁸ a repository which publicly shares spam emails from their mailbox on a monthly basis since 1998. We have taken the 2010 and 2018 folders for this experimentation, in order to determine the current performance on recent spam emails of models trained on last-decade datasets and their generalisation against spam environments.

⁵ <https://spamassassin.apache.org/old/publiccorpus/> Retrieved June 2021.

⁶ <http://csmine.org/index.php/spam-email-datasets-.html> Retrieved June 2021.

⁷ <https://www.projecthoneypot.org/> Retrieved June 2021.

⁸ <http://untroubled.org/spam/> Retrieved June 2021.

Table 4 Main characteristic of the publicly available datasets selected for our experiments

Dataset	Year	Total	Ham	Spam	Spam Rate (%)	Image Att (%)	Other Att (%)	English (%)
LingSpam	2000	2893	2412	481	16.63	0	0	100
SpamAssassin	2002–2006	6047	4150	1897	31.37	0.63	0.63	95.31
EnronSpam	2006	29,694	16,921	12,773	43.02	3.34	1.53	96.94
TREC07	2007	75,419	25,220	50,199	66.56	18.33	0.72	97.53
CSDMC	2010	4327	2949	1378	31.85	0.94	0.21	81.70
BG 2010	2010	649,974	0	649,974	100.00	1.86	1.92	76.28
BG 2018	2018	77261	0	77261	100.00	3.83	0.44	17.69

We include the total number of emails, the number of ham and spam emails, spam rate (%), percentage of emails containing image (Image Att.) and other files (Others Att.) attached and percentage of emails written in English

Table 4 shows the main characteristics of each dataset, including the total number of emails, the number of ham and spam emails, spam rate, percentage of emails containing image and other files attached and percentage of emails written in English.

5.2 Experimental setup

We carried out our experiments on an Intel(R) Core(TM) i7-7th Gen with 16G of RAM, under Ubuntu 18.04 OS and Python 3.

We designed four spam email filters for each dataset, based on a text classification pipeline which comprises two machine learning algorithms widely used for detecting spam, NB and SVM, and two text encoders, TF-IDF and BOW.

We used Ling-Spam, SpamAssassin, TREC07 and CSDMC without any modification. Since Enron-Spam Dataset contained a folder with emails from SpamAssassin, we removed the folder to avoid interference among both datasets. To deal with the huge size of Bruce Guenter 2010 we only selected 50K emails randomly. Finally, we only took into account English emails of both Bruce Guenter datasets 2010 and 2018.

To implement the pipelines, we used Scikit-Learn and NLTK⁹ to remove the English stopwords. The preprocessing, text representation and classification were set after the evaluation of different configurations as follows.

After we tried out our models using different configurations, we considered the following setting to be the most suitable. Firstly, for the preprocessing phase, we removed single URLs, characters, numbers, single letters, stopwords, duplicated words and we tokenized the message from emails. For the text encoders, BOW and TF-IDF, we selected a vocabulary size of 9000 words and 3 minimum appearances per word. Regarding the classification step, we show below the parameter tuning per model, and the rest of the model parameters are left with their default values. We chose a linear kernel for the SVM model, and the C value was set 1000. C parameter is an optimiser for classifiers; a low value looks for a higher margin of hyperplane separation. For NB, we used a Multinomial distribution.

⁹ <https://www.nltk.org/> Retrieved June 2021.

Table 5 Predictive performance for the spam filters based on the combination of TF-IDF and NB

Test set	Train set									
	Ling-Spam		SpamAssassin		EnronSpam		TREC07		CSDMC	
	2000		2002–2006		2006		2007		2010	
	Acc	FPR	Acc	FPR	Acc	FPR	Acc	FPR	Acc	FPR
Ling-Spam	99.14	0.00	86.52	13.76	91.98	8.95	94.05	5.97	51.19	56.38
SpamAssassin	74.54	30.75	97.41	1.04	64.28	46.79	84.09	17.11	93.86	4.30
EnronSpam	81.83	11.54	75.64	18.44	97.55	1.49	83.50	23.72	72.22	28.29
TREC07	75.45	20.76	75.89	6.74	81.68	34.13	96.65	1.41	77.37	20.23
CSDMC	74.81	27.05	88.84	0.48	70.81	39.71	85.66	9.29	95.51	2.96

Models are trained with different datasets (columns) and accuracy and false positive rate (FPR) metrics are reported when they are applied to different scenarios (rows). Values highlighted in grey refer to performance estimated with 10-fold cross-validation using a specific dataset

Table 6 Predictive performance for the spam filters based on the combination of TF-IDF and SVM

Test set	Train set									
	Ling-Spam		SpamAssassin		EnronSpam		TREC07		CSDMC	
	2000		2002–2006		2006		2007		2010	
	Acc	FPR	Acc	FPR	Acc	FPR	Acc	FPR	Acc	FPR
Ling-Spam	99.52	0.04	56.10	50.70	63.64	41.46	37.37	74.38	54.72	52.86
SpamAssassin	69.11	39.17	99.33	0.31	61.11	50.36	63.87	45.24	98.64	0.85
EnronSpam	80.92	12.95	63.48	59.23	97.97	2.64	56.90	73.27	72.76	36.69
TREC07	69.52	26.49	85.17	11.01	78.10	47.56	99.38	0.83	83.24	14.55
CSDMC	70.00	29.36	92.24	0.88	66.74	45.15	76.45	27.42	99.04	0.58

Models are trained with different datasets (columns) and accuracy and false positive rate (FPR) metrics are reported when they are applied to different scenarios (rows). Values highlighted in grey refer to performance estimated with 10-fold cross-validation using a specific dataset

The classifier performance has been reported in terms of accuracy and false positive rate (FPR). Since our aim is to ascertain whether a filter calibrated on a dataset preserves their capability of generalisation on other datasets collected over different periods of time and spam emails sources, we divided our evaluation procedure into two parts. On the one hand, we defined estimated generalisation values as the results of calibrating and assessing a filter on the same dataset, using the 10-fold cross-validation technique. On the other hand, the estimated values were compared with the results of training a filter with a specific dataset and testing it on another dataset, called evaluation values.

Table 7 Predictive performance for the spam filters based on the combination of BOW and NB

Test set	Train set									
	Ling-Spam		SpamAssassin		EnronSpam		TREC07		CSDMC	
	2000		2002–2006		2006		2007		2010	
	Acc	FPR	Acc	FPR	Acc	FPR	Acc	FPR	Acc	FPR
Ling-Spam	98.93	1.00	87.80	13.06	92.84	7.75	94.50	4.52	33.08	79.27
SpamAssassin	37.07	83.25	97.78	0.91	68.31	41.41	86.61	11.22	92.88	6.37
EnronSpam	62.65	59.68	73.74	27.46	97.24	1.28	85.21	17.39	66.87	44.99
TREC07	68.72	72.55	78.52	7.90	82.18	29.85	96.05	0.94	81.63	21.30
CSDMC	36.67	86.83	92.24	0.54	74.28	33.68	84.89	5.95	95.58	3.98

Models are trained with different datasets (columns) and accuracy and false positive rate (FPR) metrics are reported when they are applied to different scenarios (rows). Values highlighted in grey refer to performance estimated with 10-fold cross-validation using a specific dataset

Table 8 Predictive performance for the spam filters based on the combination of BOW and SVM

Test set	Train set									
	Ling-Spam		SpamAssassin		EnronSpam		TREC07		CSDMC	
	2000		2002–2006		2006		2007		2010	
	Acc	FPR	Acc	FPR	Acc	FPR	Acc	FPR	Acc	FPR
Ling-Spam	99.38	0.17	52.89	54.06	71.07	31.34	39.13	71.56	51.54	56.34
SpamAssassin	48.55	66.96	99.16	0.45	64.83	44.13	65.95	42.56	97.94	1.68
EnronSpam	76.12	17.64	63.06	59.20	96.86	3.38	59.74	67.12	61.69	60.19
TREC07	55.25	55.60	86.11	7.37	74.89	45.24	98.78	1.45	86.29	16.34
CSDMC	50.58	60.77	93.36	1.80	62.44	50.36	77.62	26.13	98.36	1.16

Models are trained with different datasets (columns) and accuracy and false positive rate (FPR) metrics are reported when they are applied to different scenarios (rows). Values highlighted in grey refer to performance estimated with 10-fold cross-validation using a specific dataset

5.3 Results and discussion

Tables 5 and 6 show the performance of the spam filters using the TF-IDF pipeline and the classifiers NB and SVM, respectively. On the other hand, Tables 7 and 8 present the BOW pipelines results achieved by the classifiers NB and SVM combinations, respectively.

All of them show the predictive performance of the above-mentioned models trained with five different datasets (shown in the columns) when they are applied to different scenarios, i.e., datasets. Each of these datasets reflects emails gathered on different years, as explained in Sect. 5.1.

In light of the overall results, we can confirm our initial hypothesis that anti-spam filters deteriorate due to the dataset shift and changes in spammer strategies reflected in the different spam email datasets. Firstly, the filters have shown a performance similar to models published in the literature, above 95% of accuracy by using cross-validation to assess them, i.e. estimated generalised values. These are the values highlighted in grey in the tables that

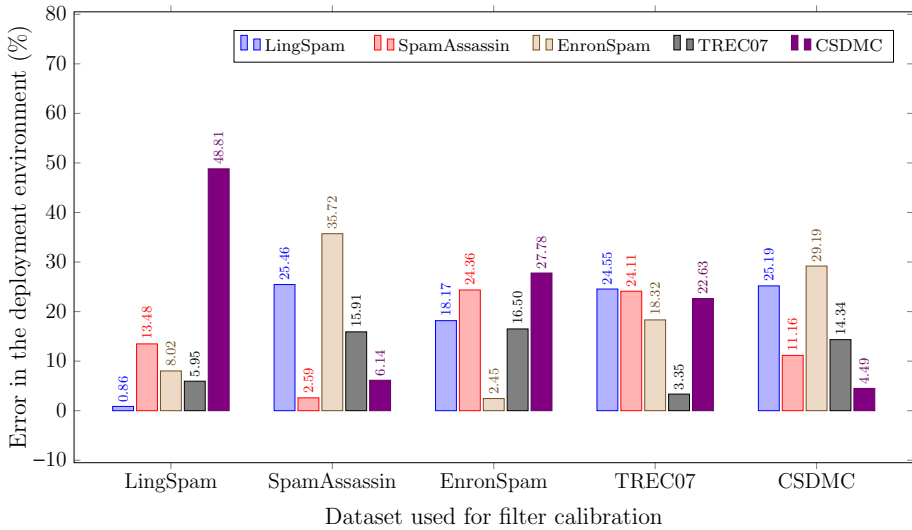


Fig. 5 Predictive performance for the spam filters based on the combination of TF-IDF and NB. Models are trained with different datasets (axis X) and Error Rate metric is reported (Axis Y) when they are applied to different scenarios (legend). Our evaluation includes the following datasets: Ling-Spam (2000), SpamAssassin(2002-2006), Enron-Spam (2006), TREC07 (2007) and CSDMC (2010). The legend order also indicates the order of bars on the chart

refer to performance estimated with 10-fold cross-validation using each specific dataset. In addition, the FPR represents the percentage of legitimate or ham emails with a wrong prediction, i.e., labeled as spam. Despite obtaining a low value when the performance is evaluated with 10-fold cross-validation using a specific dataset, the degradation of FPR may reach up to 83.25% by assessing models trained and tested in different scenarios.

However, when the filters calibrated with a given dataset are applied to a different dataset (both back and forth in time), the performance decreases. In some cases, a severe degradation in performance is seen, such as a filter calibrated with the oldest dataset, Ling-Spam. Filters calibrated with Enron-Spam, SpamAssassin and CSDMC also suffer a degradation in their results.

In Fig. 5 we show this situation in a specific case (i.e., NB classifier with TF-IDF) to show this situation graphically. We can see that the filter error rate from the estimated values increases. For instance, when the filter is calibrated with LingSpam (estimated error rate 0.86%), this error increases from 5.95 to 48.81% or with CSDMC (estimated error rate 4.49), this error increases from 11.16 to 29.19%

In general, performance is no longer as good as the estimated value, with error rates reaching even values of 48.81%, 35, 72% or 29, 19% (they vary from 5.95 to 48.81%).

According to the results, the deterioration of filters may be affected by the dataset shift problem and how personalised the anti-spam filters are, i.e. spam email resource, or the spammer strategies included. Similar sources of some datasets, e.g. Ling-Spam and Enron-Spam or SpamAssassin and CSDMC, may allow a slight deterioration in the results respect to their estimated values.

Regarding the machine learning and text encoders models, we found out that SVM filters decreased their performance on models trained on TREC07 and tested on the other ones. A possible reason may be the well-known problems of strength and efficacy of SVM

Table 9 Accuracy metric of twenty calibrated filters (based on BOW and TF-IDF text encoders along with NB and SVM machine learning algorithms) when applied to two Bruce Guenter spam (only) dataset, from the year 2010 and 2018

	Test set	Train set				
		Ling-Spam 2000	SpamAssassin 2002–2006	EnronSpam 2006	TREC07 2007	CSDMC 2010
BOW	NB	BG 2010 82.35	57.51	89.64	67.72	82.41
		BG 2018 80.44	60.55	81.35	77.71	72.13
	SVM	BG 2010 63.22	80.81	89.16	88.73	93.30
TF-IDF		BG 2018 52.11	81.82	88.08	95.58	90.01
	NB	BG 2010 55.69	47.81	91.02	73.28	81.55
		BG 2018 52.90	50.39	83.72	88.88	61.70
	SVM	BG 2010 64.32	75.45	90.74	86.38	88.33
		BG 2018 52.15	80.43	90.73	97.23	75.75

with high dimensional datasets or number of parameters (Dada et al. 2019) apart from memory cost and execution time. NB filters improve their performance on a large dataset, and the accuracy remains above 83.50% in all cases. BOW filters trained on Ling-Spam suffer a remarkable drop in their performance. The size of the dataset and, thereby, the sparse vocabulary size may explain this behaviour.

Finally, we evaluated the filters calibrated with the previous datasets on only-spam environments to be aware of their generalisation on recent and spam datasets. This gives us further information to interpret the results and verify our initial hypothesis. Table 9 shows the accuracy of each of the twenty classification models we have assessed so far—i.e., SVM and NB classifiers using TF-IDF or BOW calibrated with five different datasets—in the task of identifying the spam from Bruce Guenter dataset, specifically, those from years 2010 and 2018 (let us recall that only spam data is available in Bruce Guenter dataset). In general terms, the generalisation performance of the filters reflected a large variability, far away from the estimated values. Most filters trained on Ling-spam obtained between 52.11 and 64.32% of accuracy. Similarly, most SpamAssassin-calibrated filters achieved from 47.81 to 75.45% of accuracy.

Nevertheless, filters calibrated with Enron-spam attained the most balanced and highest results, between 83.72 and 91.02%. This may be because the Enron-Spam dataset contains a spam emails subset from the Bruce Guenter dataset (2004–2005), which confirms our hypothesis about the relevance of spam source to calibrate a filter.

6 Conclusions

This paper presents a review on spam email detection, focusing on the analysis of spammer strategies and the changing nature of the data in this field.

The spammer, or the adversarial figure in this environment, follows sophisticated strategies to bypass the filters. In our study on spam datasets over the last few decades, we identified the use of poisoning text, obfuscated words, hidden text salting and image-based spam as the most popular spammer tricks. More recent strategies include multi-language emails to poison the text, attachments like PDF or “.docx” files to introduce a spam message, and URLs to connect to potentially harmful sites. We also reviewed the studies involved in detecting these tricks and minimising their effect.

We explored the most recent works on filtering spam emails emphasising on the phases of feature extraction and feature selection to mitigate overfitting by reducing the dimensionality of the input space. Despite the rise of deep learning in other application fields, traditional machine learning algorithms are still the most popular approaches in the literature for the spam email filtering. Their high performance as well as their simplicity when compared with deep learning models may be the reasons.

We also evaluated the impact on the performance when there is a mismatch between training and operational data distributions. Our initial hypothesis assumes the dataset shift and spammer strategies presented in the spam email datasets are the main factors of the deterioration of an anti-spam filter. We selected five popular email datasets from different years (Ling-Spam, SpamAssassin, Enron-Spam, TREC07 and CSDMC) to calibrate different spam filters. For each dataset, four classifiers have been trained as a result of combining two text representation techniques, TF-IDF and BOW, along with two machine learning algorithms, NB and SVM. We used the Bruce Guenter spam email datasets from 2010 and 2018 only for testing the models. Experimental results show an overall deterioration of

the estimated generalisation accuracy when the models are applied on unseen future data coming from datasets different from the ones used for training. We found that the datasets which were collected from different sources and, thereby, more or less personalised-specific fields, tend to show larger differences in their performance, and are also affected by the changing nature of spam over time due to dataset shift and spammer strategies. Finally, we evaluated our calibrated filters on only-spam Bruce Guenter datasets. Generally, the results show a wide range of accuracy which differ from the estimated values and certain dependency of the spam source.

To sum up, we encourage to unify the criteria to assess anti-spam filters based on machine learning, keeping in mind the dataset shift problem. Identifying the spammer strategies used recently, the topics that appear and their current relevance would help to assign specific filters in spam email detection. In addition, spammer—adversarial—strategies are a challenge for this field. Analysing these strategies in depth could help to detect and anticipate new popular tricks or develop models that allow to assess the filter robustness against attacks. Gaining insight into the data by knowing the most relevant features could also help to increase the classification model robustness.

Acknowledgements This work was supported by the Framework Agreement between the Universidad de León and INCIBE (Spanish National Cybersecurity Institute) under Addendum 01.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Aiwan F, Zhaofeng Y (2018) Image spam filtering using convolutional neural networks. *Pers Ubiquitous Comput* 22:1029–1037. <https://doi.org/10.1007/s00779-018-1168-8>
- Al Nabki MW, Fidalgo E, Alegre E, de Paz Centeno I (2017) Classifying illegal activities on Tor network based on web textual contents. In: Proceedings of the 15th conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Valencia, pp 35–43. <https://doi.org/10.18653/v1/E17-1004>
- Al Nabki W, Fidalgo E, Alegre E, Alaiz R (2020) File name classification approach to identify child sexual abuse. In: Conference: 9th international conference on pattern recognition applications and methods, pp 228–234. <https://doi.org/10.5220/0009154802280234>
- Alaiz-Rodríguez R, Japkowicz N (2008) Assessing the impact of changing environments on classifier performance. In: Conference of the Canadian Society for Computational Studies of Intelligence. Springer, pp 13–24. https://doi.org/10.1007/978-3-540-68825-9_2
- Alazab M, Broadhurst R (2016) Spam and criminal activity. In: Trends and issues in crime and criminal justice pp 1–20. <https://doi.org/10.2139/ssrn.2467423>
- Androutsopoulos I, Paliouras G, Karkaletsis V, Sakkis G, Spyropoulos C, Stamatopoulos P (2000) Learning to filter spam e-mail: a comparison of a naive Bayesian and a memory-based approach. *ArXiv* pp 1–12
- Annadatha A, Stamp M (2016) Image spam analysis and detection. *J Comput Virol Hacking Tech* 14(1):39–52. <https://doi.org/10.1007/s11416-016-0287-x>

- Apruzzese G, Colajanni M, Ferretti L, Marchetti M (2019) Addressing adversarial attacks against security systems based on machine learning. In: 2019 11th International conference on cyber conflict (CyCon), pp 1–18. <https://doi.org/10.23919/CYCON.2019.8756865>
- Arivudainambi D, Kumar KV, Chakkaravarthy SS, Visu P (2019) Malware traffic classification using principal component analysis and artificial neural network for extreme surveillance. *Comput Commun* 147:50–57. <https://doi.org/10.1016/j.comcom.2019.08.003>
- Awad M, Foqaha M (2016) Email spam classification using hybrid approach of RBF neural network and particle swarm optimization. *Int J Netw Secur Appl* 8:17–28. <https://doi.org/10.5121/ijnsa.2016.8402>
- Baena-García M, del Campo-Ávila J, Fidalgo R, Bifet A, Gavalda R, Morales-Bueno R (2006) Early drift detection method. In: Fourth international workshop on knowledge discovery from data streams, vol 6, pp 77–86. https://doi.org/10.1007/978-3-642-23857-4_12
- Bahgat EM, Rady S, Gad W, Moawad IF (2018) Efficient email classification approach based on semantic methods. *Ain Shams Eng J* 9(4):3259–3269. <https://doi.org/10.1016/j.asej.2018.06.001>
- Barreno M, Nelson B, Sears R, Joseph AD, Tygar JD (2006) Can machine learning be secure? In: Proceedings of the 2006 ACM symposium on information, computer and communications security, ASIACCS '06. Association for Computing Machinery, New York, pp 16–25. <https://doi.org/10.1145/1128817.1128824>
- Barushka A, Hajek P (2018) Spam filtering using integrated distribution-based balancing approach and regularized deep neural networks. *Appl Intell* 48(10):3538–3556. <https://doi.org/10.1007/s10489-018-1161-y>
- Bassiouni M, Shafaey M, El-Dahshan ES (2018) Ham and spam e-mails classification using machine learning techniques. *J Appl Secur Res* 13:315–331. <https://doi.org/10.1080/19361610.2018.1463136>
- Bergholz A, Paass G, Reichartz F, Strobel S, Iais F, Birlinghoven S, Moens MF, Witten B (2008) Detecting known and new salting tricks in unwanted emails. In: CEAS, p 9
- Bhowmick A, Hazarika SM (2018) E-mail spam filtering: a review of techniques and trends. *Adv Electron Commun Comput* 443:583–590. https://doi.org/10.1007/978-981-10-4765-7_61
- Biggio B, Roli F (2018) Wild patterns: ten years after the rise of adversarial machine learning. *Pattern Recognit* 84:317–331. <https://doi.org/10.1016/j.patcog.2018.07.023>
- Biggio B, Fumera G, Pillai I, Roli F (2007) Image spam filtering by content obscuring detection. In: Conference: CEAS 2007—the fourth conference on email and anti-spam, p 6
- Biggio B, Fumera G, Pillai I, Roli F (2011) A survey and experimental evaluation of image spam filtering techniques. *Pattern Recognit Lett* 32(10):1436–1446. <https://doi.org/10.1016/j.patrec.2011.03.022>
- Biggio B, Corona I, Maiorca D, Nelson B, Šrđić N, Laskov P, Giacinto G, Roli F (2013) Evasion attacks against machine learning at test time. *Lecture notes in computer science*, pp 387–402. https://doi.org/10.1007/978-3-642-40994-3_25
- Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D (2020) Language models are few-shot learners. 2005.14165
- Byun B, Lee CH, Webb S, Pu C (2007) A discriminative classifier learning approach to image modeling and spam image identification. In: Conference: CEAS 2007—the fourth conference on email and anti-spam, p 9
- Cai J, Luo J, Wang S, Yang S (2018) Feature selection in machine learning: a new perspective. *Neurocomputing* 300:70–79. <https://doi.org/10.1016/j.neucom.2017.11.077>
- Chan PP, Luo F, Chen Z, Shu Y, Yeung DS (2021) Transfer learning based countermeasure against label flipping poisoning attack. *Inf Sci* 548:450–460. <https://doi.org/10.1016/j.ins.2020.10.016>
- Chavda A, Potika K, Troia FD, Stamp M (2018) Support vector machines for image spam analysis. In: ICETE, pp 597–607. <https://doi.org/10.5220/0006921404310441>
- Chen S, Xue M, Fan L, Hao S, Xu L, Zhu H, Li B (2018) Automated poisoning attacks and defenses in malware detection systems: an adversarial machine learning approach. *Comput Secur* 73:326–344. <https://doi.org/10.1016/j.cose.2017.11.007>
- Cohen Y, Hendler D, Rubin A (2018) Detection of malicious webmail attachments based on propagation patterns. *Knowl Based Syst* 141:67–79. <https://doi.org/10.1016/j.knsys.2017.11.011>
- Cormack GV (2007) TREC 2007 spam track overview. In: The sixteenth Text REtrieval Conference (TREC 2007) proceedings, pp 1–9
- Dada EG, Bassi JS, Chiroma H, Abdulhamid SM, Adetunmbi AO, Ajibuwa OE (2019) Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon* 5(6):e01802. <https://doi.org/10.1016/j.heliyon.2019.e01802>
- Dalvi N, Domingos P, Mausam, Sanghai S, Verma D (2004) Adversarial classification. In: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining,

- KDD '04. Association for Computing Machinery, New York, pp 99–108. <https://doi.org/10.1145/1014052.1014066>
- Dasgupta P, Collins J (2019) A survey of game theoretic approaches for adversarial machine learning in cybersecurity tasks. *AI Mag* 40:31–43. <https://doi.org/10.1609/aimag.v40i2.2847>
- Dedeturk BK, Akay B (2020) Spam filtering using a logistic regression model trained by an artificial bee colony algorithm. *Appl Soft Comput* 91:106229. <https://doi.org/10.1016/j.asoc.2020.106229>
- Delany SJ, Cunningham P, Tsybmal A, Coyle L (2005) A case-based technique for tracking concept drift in spam filtering. *Knowl Based Syst* 18(4):187–195. <https://doi.org/10.1016/j.knosys.2004.10.002> (AI-2004, Cambridge, England, 13th–15th December 2004)
- Devlin J, Chang M, Lee K, Toutanova K (2018) BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR abs/1810.04805*:1–16. [arXiv: 1810.04805](https://arxiv.org/abs/1810.04805)
- Dhah EH, Naser MA, Ali SA (2019) Spam email image classification based on text and image features. In: 2019 First international conference of computer and applied sciences (CAS), pp 148–153. <https://doi.org/10.1109/CAS47993.2019.9075725>
- Diale M, Van Der Walt C, Celik T, Modupe A (2016) Feature selection and support vector machine hyper-parameter optimisation for spam detection. In: 2016 Pattern Recognition Association of South Africa and robotics and mechatronics international conference (PRASA-RobMech), pp 1–7. <https://doi.org/10.1109/RoboMech.2016.7813162>
- Diale M, Celik T, Van Der Walt C (2019) Unsupervised feature learning for spam email filtering. *Comput Electr Eng* 74:89–104. <https://doi.org/10.1016/j.compeleceng.2019.01.004>
- Dredze M, Gevaryahu R, Elias-Bachrach A (2007) Learning fast classifiers for image spam. In: 4th Conference on email and anti-spam, CEAS 2007
- El Aassal A, Baki S, Das A, Verma R (2020) An in-depth benchmarking and evaluation of phishing detection research for security needs. *IEEE Access* 8:1. <https://doi.org/10.1109/ACCESS.2020.2969780>
- Faris H, Al-Zoubi AM, Heidari AA, Aljarah I, Mafarja M, Hassonah MA, Fujita H (2019) An intelligent system for spam detection and identification of the most relevant features based on evolutionary random weight networks. *Inf Fusion* 48:67–83. <https://doi.org/10.1016/j.inffus.2018.08.002>
- Fdez-Riverola F, Iglesias EL, Díaz F, Méndez JR, Corchado JM (2007) Applying lazy learning algorithms to tackle concept drift in spam filtering. *Expert Syst Appl* 33(1):36–48. <https://doi.org/10.1016/j.eswa.2006.04.011>
- Ferrara E (2019) The history of digital spam. *Commun ACM* 62(8):82–91. <https://doi.org/10.1145/3299768>
- Gama J, Žliobaitė I, Bifet A, Pechenizkiy M, Bouchachia A (2014) A survey on concept drift adaptation. *ACM Comput Surv* 46(4):1–37. <https://doi.org/10.1145/2523813>
- Gangavarapu T, Jaidhar C, Chanduka B (2020) Applicability of machine learning in spam and phishing email filtering: review and approaches. *Artif Intell Rev* 53:64. <https://doi.org/10.1007/s10462-020-09814-9>
- Gao Y, Yang M, Zhao X, Pardo B, Wu Y, Pappas T, Choudhary A (2008) Image spam hunter. In: IEEE international conference on acoustics, speech and signal processing, 2008, ICASSP 2008, pp 1765–1768. <https://doi.org/10.1109/ICASSP.2008.4517972>
- Gibson S, Issac B, Zhang L, Jacob SM (2020) Detecting spam email with machine learning optimized with bio-inspired metaheuristic algorithms. *IEEE Access* 8:187914–187932. <https://doi.org/10.1109/ACCESS.2020.3030751>
- González-Castro V, Alaiz-Rodríguez R, Alegre E (2013) Class distribution estimation based on the Hellinger distance. *Inf Sci* 218:146–164. <https://doi.org/10.1016/j.ins.2012.05.028>
- Goodfellow IJ, Shlens J, Szegedy C (2015) Explaining and harnessing adversarial examples. [arXiv: 1412.6572](https://arxiv.org/abs/1412.6572)
- Gupta BB, Arachchilage N, Psannis K (2017) Defending against phishing attacks: taxonomy of methods, current issues and future directions. *Telecommun Syst*. <https://doi.org/10.1007/s11235-017-0334-z>
- Hand DJ (2006) Classifier technology and the illusion of progress. *Stat Sci*. <https://doi.org/10.1214/088342306000000060>
- Hijawi W, Faris H, Alqatawna J, Al-Zoubi A, Aljarah I (2017) Improving email spam detection using content based feature engineering approach. In: Conference: IEEE Jordan conference on applied electrical engineering and computing technologies (AEECT 2017), p 6. <https://doi.org/10.1109/AEECT.2017.8257764>
- Hijawi W, Alqatawna J, Al-Zoubi AM, Hassonah MA, Faris H (2021) Android botnet detection using machine learning models based on a comprehensive static analysis approach. *J Inf Secur Appl* 58:102735. <https://doi.org/10.1016/j.jisa.2020.102735>
- Hong SS, Lee W, Han MM (2015) The feature selection method based on genetic algorithm for efficient of text clustering and text classification. *Int J Adv Soft Comput Appl* 7:22–40

- Huang L, Joseph AD, Nelson B, Rubinstein BI, Tygar JD (2011) Adversarial machine learning. In: Proceedings of the 4th ACM workshop on security and artificial intelligence, AISec '11. Association for Computing Machinery, New York, pp 43–58. <https://doi.org/10.1145/2046684.2046692>
- Jáñez-Martino F, Fidalgo E, González-Martínez S, Velasco-Mata J (2020) Classification of spam emails through hierarchical clustering and supervised learning. [arXiv: 2005.08773](https://arxiv.org/abs/2005.08773)
- Kadwe Y, Suryawanshi V (2015) A review on concept drift. *IOSR J Comput Eng* 17(1):20–26. <https://doi.org/10.9790/0661-17122026>
- Karim A, Azam S, Shanmugam B, Kannoopatti K, Alazab M (2019) A comprehensive survey for intelligent spam email detection. *IEEE Access* 7:168261–168295. <https://doi.org/10.1016/j.aci.2020.01.002>
- Kim B, Abuadba S, Kim H (2020) DeepCapture: image spam detection using deep learning and data augmentation. In: Liu JK, Cui H (eds) *Information security and privacy*. Springer, Cham, pp 461–475
- Kuchipudi B, Nannapaneni RT, Liao Q (2020) Adversarial machine learning for spam filters. In: Proceedings of the 15th international conference on availability, reliability and security, ARES '20. Association for Computing Machinery, New York, pp 1–6. <https://doi.org/10.1145/3407023.3407079>
- Kull M, Flach P (2014) Patterns of dataset shift. In: *First international workshop on learning over multiple contexts (LMCE) at ECML-PKDD*, pp 1–10
- Lam HY, Yeung DY (2008) A learning approach to spam detection based on social networks. In: *Conference: CEAS 2007—the fourth conference on email and anti-spam*, p 10
- Laskov P, Kloft M (2009) A framework for quantitative security analysis of machine learning. In: *Conference: proceedings of the 2nd ACM workshop on security and artificial intelligence*, pp 1–4. <https://doi.org/10.1145/1654988.1654990>
- Lioma C, Moens MF, Gomez JC, Beer J, Bergholz A, Paass G, Horkan P (2008) Anticipating hidden text salting in emails. In: *11th International symposium on recent advances in intrusion detection*, pp 396–397. https://doi.org/10.1007/978-3-540-87403-4_24
- Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) RoBERTa: a robustly optimized BERT pretraining approach. [arXiv: 1907.11692](https://arxiv.org/abs/1907.11692)
- Liu A, Lu J, Zhang G (2020) Diverse instance-weighting ensemble based on region drift disagreement for concept drift adaptation. *IEEE Trans Neural Netw Learn Syst* 32(1):293–307. <https://doi.org/10.1109/tnnls.2020.2978523>
- Lu S, Duan LM, Deng DL (2020) Quantum adversarial machine learning. *Phys Rev Res* 2(3):22. <https://doi.org/10.1103/physrevresearch.2.033212>
- Mehta B, Nangia S, Gupta M, Nejdil W (2008) Detecting image spam using visual features and near duplicate detection. In: *Proceedings of the 17th international conference on World Wide Web*. Association for Computing Machinery, New York, pp 497–506. <https://doi.org/10.1145/1367497.1367565>
- Méndez JR, Cotos-Yañez TR, Ruano-Ordás D (2019) A new semantic-based feature selection method for spam filtering. *Appl Soft Comput* 76:89–104. <https://doi.org/10.1016/j.asoc.2018.12.008>
- Metsis V, Androutsopoulos I, Paliouras G (2006) Spam filtering with Naive Bayes—which Naive Bayes? In: *3rd Conference on email and anti-spam—proceedings, CEAS 2006*
- Mikolov T, Chen K, Corrado G, Dean J (2013a) Efficient estimation of word representations in vector space. [arXiv: 1301.3781](https://arxiv.org/abs/1301.3781)
- Mikolov T, Sutskever I, Chen K, Corrado G, Dean J (2013b) Distributed representations of words and phrases and their compositionality. [arXiv: 1310.4546](https://arxiv.org/abs/1310.4546)
- Moens M, De Beer J, Boiy E, Gomez JC (2010) Identifying and resolving hidden text salting. *IEEE Trans Inf Forensics Secur* 5(4):837–847. <https://doi.org/10.1109/TIFS.2010.2063024>
- Mohammad RMA (2020) A lifelong spam emails classification model. *Appl Comput Inform*. <https://doi.org/10.1016/j.aci.2020.01.002>
- Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, Chawla NV, Herrera F (2012) A unifying view on dataset shift in classification. *Pattern Recognit* 45(1):521–530. <https://doi.org/10.1016/j.patcog.2011.06.019>
- Murugavel U, Santhi R (2020) Detection of spam and threads identification in e-mail spam corpus using content based text analytics method. *Mater Today Proc*. <https://doi.org/10.1016/j.matpr.2020.04.742>
- Naem AA, Ghali NI, Saleh AA (2018) Antlion optimization and boosting classifier for spam email detection. *Future Comput Inform J* 3(2):436–442. <https://doi.org/10.1016/j.fcij.2018.11.006>
- Naiemi F, Ghods V, Khalesi H (2019) An efficient character recognition method using enhanced hog for spam image detection. *Soft Comput* 23:11759–11774. <https://doi.org/10.1007/s00500-018-03728-z>
- Naveiro R, Redondo A, Ríos Insua D, Ruggeri F (2019) Adversarial classification: an adversarial risk analysis approach. *Int J Approx Reason* 113:133–148. <https://doi.org/10.1016/j.ijar.2019.07.003>
- Nelson B, Barreno M, Chi FJ, Joseph A, Rubinstein BIP, Saini U, Sutton C, Tygar J, Xia K (2008) Exploiting machine learning to subvert your spam filter. In: *LEET*, pp 1–10. <https://doi.org/10.5555/1387709.1387716>

- Nelson B, Biggio B, Laskov P (2011) Understanding the risk factors of learning in adversarial environments. In: AISEC '11, pp 87–92. <https://doi.org/10.1145/2046684.2046698>
- Nosrati L, Pour AN (2011) DWM-CDD: dynamic weighted majority concept drift detection for spam mail filtering. *Int J Comput Electr Autom Control Inf Eng* 5:291–295. <https://doi.org/10.5281/zenodo.1082750>
- Oliveira DS, Lin T, Rocha H, Ellis D, Dommaraju S, Yang H, Weir D, Marin S, Ebner NC (2019) Empirical analysis of weapons of influence, life domains, and demographic-targeting in modern spam: an age-comparative perspective. *Crime Sci* 8(1):3. <https://doi.org/10.1186/s40163-019-0098-8>
- Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A (2015) The limitations of deep learning in adversarial settings. [arXiv: 1511.07528](https://arxiv.org/abs/1511.07528)
- Papernot N, McDaniel P, Goodfellow I, Jha S, Celik ZB, Swami A (2017) Practical black-box attacks against machine learning. [arXiv: 1602.02697](https://arxiv.org/abs/1602.02697)
- Paudice A, Muñoz-González L, Gyorgy A, Lupu EC (2018) Detection of adversarial training examples in poisoning attacks through anomaly detection. [arXiv: 1802.03041](https://arxiv.org/abs/1802.03041)
- Peng W, Huang L, Jia J, Ingram E (2018) Enhancing the naive Bayes spam filter through intelligent text modification detection. In: 2018 17th IEEE international conference on trust, security and privacy in computing and communications/12th IEEE international conference on big data science and engineering (TrustCom/BigDataSE), pp 849–854. <https://doi.org/10.1109/TrustCom/BigDataSE.2018.00122>
- Pérez-Díaz N, Ruano-Ordás D, Fdez-Riverola F, Méndez JR (2012) SDAI: an integral evaluation methodology for content-based spam filtering models. *Expert Syst Appl* 39(16):12487–12500. <https://doi.org/10.1016/j.eswa.2012.04.064>
- Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. [arXiv: 1802.05365](https://arxiv.org/abs/1802.05365)
- Pitropakis N, Panaousis E, Giannetos T, Anastasiadis E, Loukas G (2019) A taxonomy and survey of attacks against machine learning. *Comput Sci Rev* 34:100199. <https://doi.org/10.1016/j.cosrev.2019.100199>
- Project AS (2005) Apache SpamAssassin project. <https://spamassassin.apache.org/old/>. Accessed Dec 2020
- Quionero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND (2009) Dataset shift in machine learning. The MIT Press. <https://doi.org/10.7551/mitpress/9780262170055.001.0001>
- Redmiles EM, Chachra N, Waismeyer B (2018) Examining the demand for spam: who clicks? In: Proceedings of the 2018 CHI conference on human factors in computing systems, CHI '18. ACM, pp 212:1–212:10. <https://doi.org/10.1145/3173574.3173786>
- Rehman A, Javed K, Babri HA (2017) Feature selection based on a normalized difference measure for text classification. *Inf Process Manag* 53(2):473–489. <https://doi.org/10.1016/j.ipm.2016.12.004>
- Riesco A, Fidalgo E, Al-Nabkib MW, Jáñez-Martino F, Alegre E (2019) Classifying Pastebin content through the generation of PasteCC labeled dataset. In: 14th International conference on hybrid artificial intelligent systems (HAIS), pp 1–12. https://doi.org/10.1007/978-3-030-29859-3_39
- Rota Buló S, Biggio B, Pillai I, Pelillo M, Roli F (2017) Randomized prediction games for adversarial machine learning. *IEEE Trans Neural Netw Learn Syst* 28(11):2466–2478. <https://doi.org/10.1109/tnnls.2016.2593488>
- Ruano-Ordas D, Fdez-Riverola F, Mendez JR (2018a) Concept drift in e-mail datasets: an empirical study with practical implications. *Inf Sci* 428:120–135. <https://doi.org/10.1016/j.ins.2017.10.049>
- Ruano-Ordás D, Fdez-Riverola F, Méndez JR (2018b) Using evolutionary computation for discovering spam patterns from e-mail samples. *Inf Process Manag* 54(2):303–317. <https://doi.org/10.1016/j.ipm.2017.12.001>
- Saidani N, Adi K, Allili MS (2020) A semantic-based classification approach for an enhanced spam detection. *Comput Secur* 94:101716. <https://doi.org/10.1016/j.cose.2020.101716>
- Sánchez-Paniagua M, Fidalgo E, González-Castro V, Alegre E (2021) Impact of current phishing strategies in machine learning models for phishing detection. In: Herrero Á, Cambra C, Urda D, Sedano J, Quintián H, Corchado E (eds) 13th International conference on computational intelligence in security for information systems (CISIS 2020). Springer, Cham, pp 87–96. https://doi.org/10.1007/978-3-030-57805-3_9
- Sanghani G, Kotecha K (2019) Incremental personalized e-mail spam filter using novel TFDCR feature selection with dynamic feature update. *Expert Syst Appl* 115:287–299. <https://doi.org/10.1016/j.eswa.2018.07.049>
- Sethi TS, Kantardzic M (2018) Handling adversarial concept drift in streaming data. *Expert Syst Appl* 97:18–40. <https://doi.org/10.1016/j.eswa.2017.12.022>

- Shams R, Mercer RE (2016) Supervised classification of spam emails with natural language stylometry. *Neural Comput Appl* 27(8):2315–2331. <https://doi.org/10.1007/s00521-015-2069-7>
- Shi Y, Erpek T, Sagduyu YE, Li JH (2019) Spectrum data poisoning with adversarial deep learning. *arXiv: 1901.09247*
- Simester D, Timoshenko A, Zoumpoulis S (2020) Targeting prospective customers: robustness of machine-learning methods to typical data challenges. *Manag Sci* 66:2495–2522. <https://doi.org/10.1287/mnsc.2019.3308>
- Srinivasan S, Ravi V, Alazab M, Ketha S, Al-Zoubi AM, Kotti Padannayil S (2021) Spam emails detection based on distributed word embedding with deep learning. In: Maleh Y, Shojafar M, Alazab M, Baddi Y (eds) *Machine intelligence and big data analytics for cybersecurity applications*. Springer, Cham, pp 161–189. https://doi.org/10.1007/978-3-030-57024-8_7
- Sumathi S, Pugalandhi G (2020) Cognition based spam mail text analysis using combined approach of deep neural network classifier and random forest. *J Ambient Intell Humaniz Comput* 1:12. <https://doi.org/10.1007/s12652-020-02087-8>
- Temitayo M, Olabiyisi S, Baale A (2012) Hybrid GA-SVM for efficient feature selection in e-mail classification. *Comput Eng Intell Syst* 3:17–28
- Tran KN, Alazab M, Broadhurst R (2013) Towards a feature rich model for predicting spam emails containing malicious attachments and URLs. In: *Conference: proceedings of the 11th Australasian data mining conference (AusDM)*, pp 1–11
- Velasco-Mata J, Fidalgo E, González-Castro V, Alegre E, Blanco-Medina P (2019) Botnet detection on TCP traffic using supervised machine learning. In: *14th International conference on hybrid artificial intelligence systems (HAIS)*, pp 1–12. https://doi.org/10.1007/978-3-030-29859-3_38
- Vinitha VS, Renuka DK (2020) Feature selection techniques for email spam classification: a survey. In: Kumar LA, Jayashree LS, Manimegalai R (eds) *Proceedings of international conference on artificial intelligence, smart grid and smart city applications*. Springer, Cham, pp 925–935. https://doi.org/10.1007/978-3-030-24051-6_86
- Wang Z, Josephson W, Lv Q, Charikar M, Li K (2007) Filtering image spam with near-duplicate detection. In: *Conference: CEAS 2007—the fourth conference on email and anti-spam*, p 10
- Wang D, Irani D, Pu C (2013) A study on evolution of email spam over fifteen years. In: *9th IEEE international conference on collaborative computing: networking, applications and worksharing*, pp 1–10. <https://doi.org/10.4108/icst.collaboratecom.2013.254082>
- Wang X, Li J, Kuang X, Tan Y, Li J (2019) The security of machine learning in an adversarial setting: a survey. *J Parallel Distrib Comput* 130:12–23. <https://doi.org/10.1016/j.jpdc.2019.03.003>
- Webb GI, Hyde R, Cao H, Nguyen HL, Petitjean F (2016) Characterizing concept drift. *Data Min Knowl Discov* 30(4):964–994. <https://doi.org/10.1007/s10618-015-0448-4>
- Wittel G, Wu S (2004) On attacking statistical spam filters. In: *Conference: CEAS 2004—the fourth conference on email and anti-spam*, p 7
- Xiao H, Biggio B, Brown G, Fumera G, Eckert C, Roli F (2018) Is feature selection secure against training data poisoning? *CoRR abs/1804.07933*. [arXiv: 1804.07933](https://arxiv.org/abs/1804.07933)
- Yu S (2015) Covert communication by means of email spam: a challenge for digital investigation. *Digit Investig* 13:72–79. <https://doi.org/10.1016/j.diin.2015.04.003>
- Yu S, Abraham Z, Wang H, Shah M, Wei Y, Príncipe JC (2019) Concept drift detection and adaptation with hierarchical hypothesis testing. *J Frankl Inst* 356(5):3187–3215. <https://doi.org/10.1016/j.jfranklin.2019.01.043>
- Zamil YK, Ali SA, Naser MA (2019) Spam image email filtering using K-NN and SVM. *Int J Electr Comput Eng* 9(1):245. <https://doi.org/10.11591/ijece.v9i1.pp245-254>
- Zavvar M, Rezaei M, Garavand S (2016) Email spam detection using combination of particle swarm optimization and artificial neural network and support vector machine. *Int J Mod Educ Comput Sci* 8:68–74. <https://doi.org/10.5815/ijmecs.2016.07.08>
- Zhang F, Chan PPK, Biggio B, Yeung DS, Roli F (2016) Adversarial feature selection against evasion attacks. *IEEE Trans Cybern* 46(3):766–777. <https://doi.org/10.1109/tycb.2015.2415032>