Identifying threats in a large company's inbox

Luigi Gallo
Cyber Security Lab TIM S.p.A. *
University of Napoli Federico II +
*Torino,+Napoli, Italy
luigi.gallo3@unina.it

Alessio Botta
University of Napoli Federico II
Napoli, Italy
alessio.botta@unina.it

Giorgio Ventre
University of Napoli Federico II
Napoli, Italy
giorgio.ventre@unina.it

ABSTRACT

Cyber threats in emails continue to grow. Anti-spam filters have achieved good performance, but several spam emails still pass through them. Some of them are particularly dangerous as they represent attempts to breach the security policy of the company (e.g. inducing a manager to authorize a payment towards a fraudulent bank account). In this paper we propose an automated system to detect such emails, passing through antispam filter and potentially very dangerous. Our dataset is composed of real spam emails reported, collected, and labelled as critical or not by human analysts during each day of the last year in a large company's inbox. We firstly study the characteristics of dangerous emails and then train and use different supervised machine learning classifiers to detect them. Our results highlight the main distinguishing characteristics of such emails and that (a) Support Vector Machine and Random Forest classifiers achieve the best performance; (b) the full feature set considered allows to obtain up to 97% of recall and up to 92% of precision with supervised approaches; (c) highly dangerous spam emails can be easily detected with only 21 features.

CCS CONCEPTS

• **Security and privacy**;

KEYWORDS

Security, Spam, Phishing, Machine Learning.

ACM Reference Format:
Luigi Gallo, Alessio Botta, and Giorgio Ventre. 2019. Identifying threats in a large company's inbox. In *Big-DAMA '19: ACM CoNEXT Workshop on Big DAta, Machine Learning and Artificial Intelligence for Data Communication Networks, December 9, 2019, Orlando, FL, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3359992.3366637

1 INTRODUCTION AND MOTIVATION

Despite being a necessity in the work life of every business, email remains a risk and exposes to frequent attacks. Email is currently one of the most used channels for making cyber attacks and very often companies fall victim of financial fraud. Perpetrators use most commonly social engineering techniques and, in particular, spear phishing. Some top managers have even been fired for losing large amounts of money (up to $50 million) due to cyber frauds. In 2016, the FBI raised the alarm over this important problem as attacks increased in number and malignance [6]. Within the category of spam emails either fall innocuous attempts to market and to sell products, or messages that contain real threats, such as a phishing attempts or malware for espionage and theft of sensitive data. Techniques for building effective spam emails are various, using advanced techniques to escape spam filter blocks and social engineering techniques to trick people. Generally, employees of big companies are trained not to be fooled, but this is not always guaranteed for various reasons: large companies have employees of all age ranges, with varied education and technology literacy; at any rate it is not impossible for a malicious email to mislead anyone during a moment of distraction. Every single employee can represent a point of entry for spammers and attackers. According to the Internet Security Threat Report by Symantec [15], spam levels continued to increase in 2018, as they have done every year since 2015, with 55 percent of emails received in 2018 being categorized as spam. In the context of a company with tens of thousands of employees, millions of emails are received per day; although 95% of these are blocked by sophisticated spam filters, the remaining 5% is still a potentially dangerous portion of emails too large to monitor and control. In this work a spam email is simply an unwanted email, and we are not interested to most of them. We rather want to understand if any of them has created a *security incident*: "a security-relevant system event in which the system's security policy is disobeyed or otherwise breached" [11]. If an employee browses a malicious website or downloads a malicious attachment (i.e. ransomware, trojans etc.), a security incident can occur. Security incidents can have different potential, depending on the number and role of the employees involved, the nature of the threat, and how effective the security systems (i.e. corporate antivirus) are against them. We call emails that have the potential to generate a security incident *critical spam*. As the number of unsolicited e-mails received by large companies is huge and constantly increasing, their manual analysis is not feasible. An automatic mechanism to detect critical spam becomes therefore necessary.

In this paper we aim at devising an automated system to detect critical spam. We firstly analyze the composition of spam emails that passed through the spam filter of a large company. On average, 30 million e-mails per month reach the mailboxes, most of which are filtered by the filter; the dataset is composed of about 12 thousand e-mails reported in the last year by about 100 thousand users. We show what are the characteristics of critical spam emails and how the rest of them can be filtered. Through these observations interesting considerations have been reached both on the most used ways to make critical spam and on the victims of

Big-DAMA '19, December 9, 2019, Orlando, FL, USA

ACM ISBN 978-1-4503-6999-2/19/12…$15.00
https://doi.org/10.1145/3359992.3366637

it. We then train several machine learning algorithms to perform a binary classification: critical or not relevant spam. In order to find the best one, main classification algorithms based on machine learning have been tested and compared: Gaussian Naive Bayes, Decision Trees, Support Vector Machines, Neural Networks, Random Forest. We show that Support Vector Machines and Random Forest achieve the best performance, with 92% maximum Precision, 97% maximum Recall and 89% maximum F-measure. We then show the classification time of these two algorithms, analyzing the impact of different feature sets on such time. Results show that out of the 50 features considered, already the 21 best ones allow to obtain excellent performance; moreover, using just 8 well-selected features to considerably reduce the classification times, the performance degrades by (only) 5%.

2 STATE OF THE ART

The importance of the problem of spam, and more specifically spam fraud, has been recognized worldwide for many years. Spam detection and filtering technologies still can't completely solve the problem. *Blanzieri et al.* [1] explains the context in detail, highlighting the importance of a collaborative approach to reduce the impact of this problem. Automatic learning is crucial to prevent the defence process from becoming obsolete, in a scenario where attack techniques constantly evolve and adapt to defence techniques. An important differentiation between the approaches used is given by the type of features used: they can be extracted from the header, body and text content of the message. *Gansterer et al.* [7] proposes new features extracted from the header along with more traditional ones, experimenting with a classification to detect spam phishing. *Nizamani et al.* [10] shows that there is a specific set of words to recognize fraudulent spam emails. *Basavaraju et al.* [9] focuses exclusively on the text of the message, calculating for each word a metric that estimates the importance: so you get the model tf-idf, which performs especially well with deep learning approaches.
A second important differentiation is given by the type of learning used: supervised, unsupervised or semi-supervised. As *Crawford et al.* [3] explained in the specific case of spam review, the preferred approach is the supervised one, which is often impractical because a labeled dataset is not always available. A case of supervised learning has been studied by *Dai et al.* [4], who shows the classification performance of four different algorithms, of which SVM is the best. Sub-optimal results are obtained with non-supervised approaches, which, however, considering the high costs of obtaining labelled datasets, are the most common.
Note that all this work focuses on the classification and detection of spam, or a specific type of spam, regardless of the actual impact it may have on the target; they are studies that can be used in the creation of anti-spam filters. Since there are still many lacks in spam filtering, as shown by *Dada et al.* [5], the risk that some threats will succeed in their goal is real. Only recently, studies have been presented that aim to prioritize phishing emails, deepening the cognitive aspect of the message contained [16] and any attempts to spoof the identity of colleagues in the company [2]. In this work we focus on the most used methodologies to attack a real company through a simple email, identifying what are the characteristics of those that have been successful. The knowledge acquired can be used to create an automatic detection system that would mitigate the problem of the large number of unwanted emails that a large company receives every day.

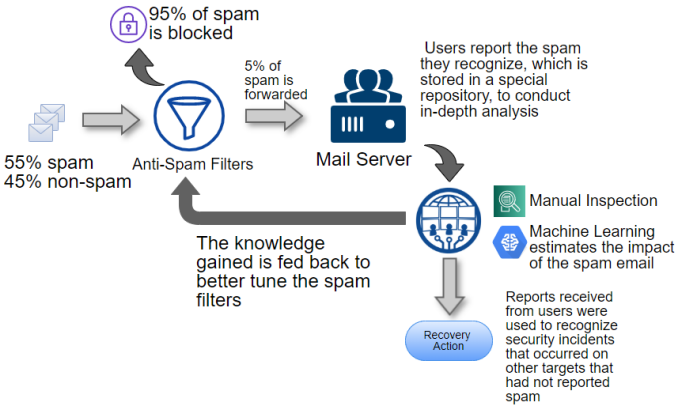


Figure 1: Ecosystem of spam defense

3 COLLABORATIVE FRAMEWORK

In the scenario of this work, the defence against attempts at spam fraud follows a collaborative approach: in addition to generic email filtering systems, there is a system of collection of reports that allows the computer emergency response team (as defined in [11]) to protect the affected users thanks to the recognition of a threat by another user. This distributed approach is important for detecting security incidents that would otherwise be undetected.

When a security incident occurs, it has to be solved with a recovery action that can be, in increasing order of relevance, one or more of the following:

- Sending notification to the users involved about the recognition of a malicious email;
- Adding a filter in the navigation proxies to block navigation or downloading from malicious or otherwise unknown sources;
- Rehabilitating of nodes and networks reached by any malware. Resetting of accounts and credentials that may have been violated;
- In-depth technical analyzing of attachments and links, in order to understand the degree of danger of them and adequately protect affected users;
- Investigating of perpetrators and possible legal actions taken.

The purpose of a collaborative framework is both to recognize and resolve security incidents that have occurred, and to intercept them before they occur. This predictive task, on which of the spam emails will actually generate a security incident if not previously reported to the recipients, can be solved with machine learning techniques. It is useful to give a level of risk, so as to prioritize the reports and allowing analysts to deepen the investigation only on the most relevant ones, and not to get lost in the huge number of them.
The architecture structured as in figure 1, has allowed to collect over time the spam that reached users and to memorize which of them has led the recipients to download an attachment or to

browse a link. In this way we have acquired a deep knowledge of the main features that the most critical spam emails have. The estimate of the risk score of the email could be made upstream for all emails (or for those for which the filter is in doubt), even before a user signals it, provided however that the process is short time consuming, because of the huge amount of emails to analyze. For this reason, the impact of feature reduction on classification performance and execution times has been studied (section 4.2).

3.1 Data collection and feature extraction

Since the beginning of 2018, all spam emails received and reported by users have been collected in a repository. More than 12,000 samples are involved. Experienced analysts have manually inspected them day by day: based on the content of the email, the navigation data of the company's proxies, the online reputation of links and attachments, they have decided for each of them whether to perform an in-depth analysis or whether to discard them. On those analyzed in depth, if a security incident was detected, this was reported.

Field	Description
General	General information, mostly extracted from the smtp headers: if any smtp server is blacklisted, size of the mail, number of recipients, role in the company of recipients etc.
Content	Features extracted from the text in the content of the email: language, number of words, number of disguised or misleading words, indices of readability, simplicity and correctness of the text etc.
Subject	Features extracted from the subject of the email: number of words, number of characters, if there are non-ASCII characters, if the email is forwarded or answered
Links	Features about the links in the email: number, number of link domains, information from url analysis service etc.
Attachments	Features about the email attachments: number, types, size, information from sandboxes and antivirus etc.
Other	Other types of information not in the previous fields: number of images, number of known entities in TIP etc.

Table 1: Features extracted from the raw data; the most important are described in detail in the table 4

This way, based on the 10 years experience of several analysts, we obtained a dataset of spam reports labelled as:

- **Critical spam - Label 1, Positive**: spam emails that have created a security incident or at least recovery action had to be taken to prevent it;
- **Not relevant spam - Label 0, Negative**: spam emails with low or no degree of danger, and have not led to any of the recovery actions listed above.

This dataset can be used to perform supervised machine learning, and obtain a classifier that allows you to immediately recognize the threats contained in the mail. The emails in the dataset are by definition all spam emails.
The full set of features extracted from the samples is described in table 1 and comprises 50 features. In this work we define "disguised word" as a word which has an edit distance of 1 from the name of the company, the names of its subsidiaries and the names of its main partners. Very often, addresses or domains similar to those normally used by the company are crafted to deceive employees. We also define the words in table 2 as "misleading word".

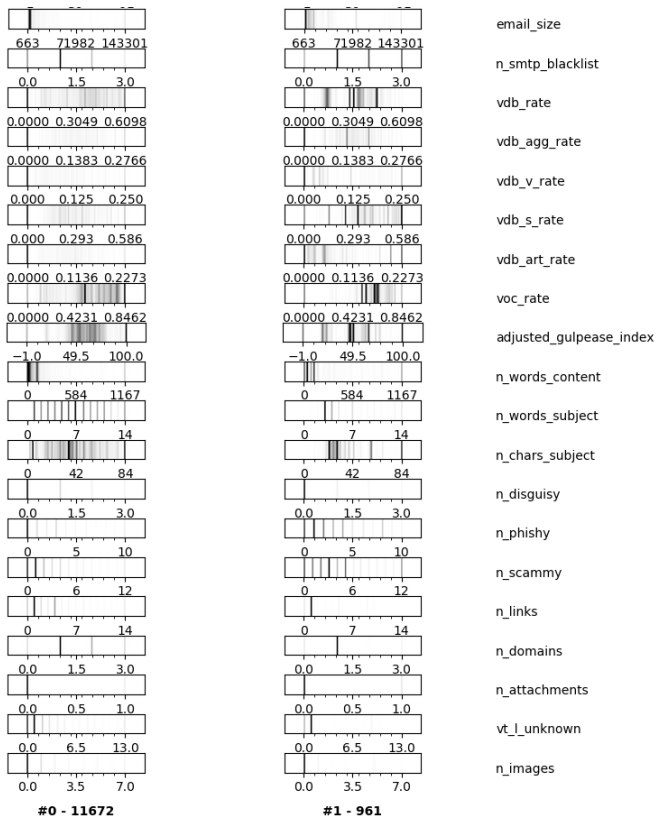


Figure 2: Heatmap by feature of the distribution of samples

Previous studies [10] show that the words listed in table 2 are those most used to capture the attention of the scammed target and it has been manually verified that this is also true in our dataset. Finally, we used the information from TIP and VirusTotal: the former

Phishing words				
account	security	user	verify	service
valid	required	credentials	attention	request
suspended	company	bank	deposit	post
Scamming words				
$	€	£	customer	prize
donate	buy	pay	congratulation	death
please	response	dollar	looking	urgent
warning	win	offer	risk	money
transaction	sex	nude		

Table 2: words considered misleading (for scam and phishing purposes)

is an internal threat intelligence platform managed by the company's security department which collects IOCs [1] ; the latter is an online malware and url analysis service [2].

Fig. 2 shows how samples are distributed along the feature ranges: to make the image more readable only the most significant features are represented, either because they were relevant to the feature selection process performed later, or because they show an interesting characteristic of how spam emails are built. For example, note that the recipient is often unique which means that in order to dispatch a single spam mail to different recipients, attackers prefer to send the same mail multiple times to a single recipient. The heatmap divides the samples by classes so that it shows collectively how a spam email that can create a security incident is made: the malicious content is most often sent via a link and not with an attachment. Generally, the link is unique and does not point to a notoriously malicious site, but it is often a link to shared repositories that still leads to the download of a malware. The second important content shown by the figure 2 concerns that potentially critical emails have as their main characteristic a well written content and subject. It is actually logical to think that the most readable, correct and deceptive emails have created security incidents. Generally, images do not have the same effect.

4 SELECTING SUPERVISED MACHINE LEARNING MODELS

We trained different Machine Learning models to perform the binary classification explained above, with the aim of choosing the best and conducting more in-depth experiments on them. We used Scikit-learn libraries to calibrate the following ML-based algorithms and to perform the evaluations: Linear SVM, RBF SVM, Decision Tree, Random Forest, Naive Bayes, MLP Neural Net. These ML models have been preliminarily selected, on the basis of the experiments shown by other works concerning the spam detection. We address the interested reader to the survey [1] and to the Scikit-learn documentation [14] for additional information on the algorithms and their configuration parameters.

We tested the classification capabilities of these six supervised approachs by computing the True and False Positive Rates (TPR/FPR), using as input the full set of features. The Figure 3 depicts the Receiver Operating Characteristic (ROC) curves obtained with each model. All presented results correspond to 10-fold cross validation. One of the metrics used to evaluate the performance of these approaches is the "Area under Curve (AUC)", which states that the two best approaches are Random Forest (98%) and RBF SVM (96%). Random Forest has been configured with 140 trees in the forest and 6 variables in the random subset at each node, following the optimization process proposed by *Lee et al.* [8]; RBF SVM has been configured with the gamma coefficient to 0.7 and the penalty parameter C to 5. Since the dataset is unbalanced, the only AUC cannot properly evaluate performance [13]. For this reason it has been used only for a preliminary selection of the best models, then all the following results are shown in terms of Precision and Recall. For the best two approaches, the Precision and Recall metrics are shown below when the class weights and features used change.

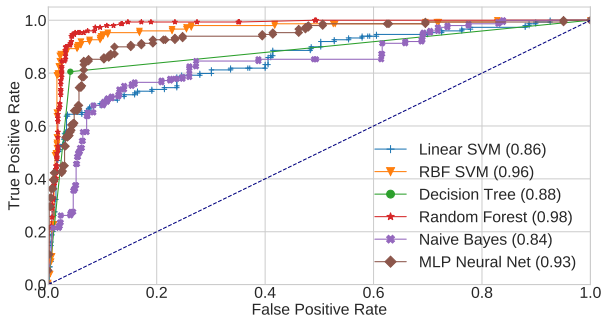


Figure 3: ROC curves of different ML-models (AUC values)

4.1 Classification Performance

This section shows the results obtained with the two best approaches previously selected. The metrics measured are Precision and Recall, as the dataset being quite unbalanced accuracy is not a good measure of quality of the classification. All metrics were obtained using a 10-fold cross validation procedure. The figure 4 shows the performance of Precision, Recall and F-measure of RBF SVM and Random Forest, varying the weights assigned to the two classes.

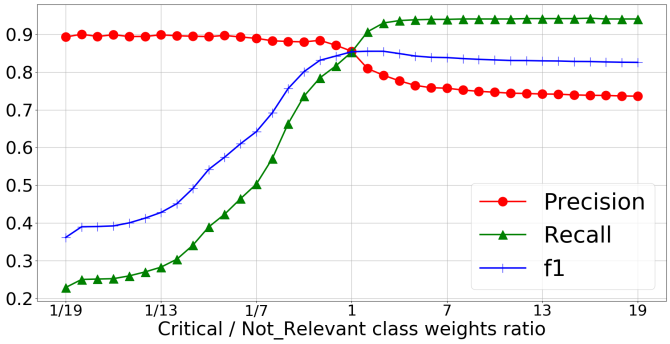
According to the figure 3, Random Forest has slightly better performance in general (F-measure up to 89%), but those of SVM have a more regular and predictable trend. In addition, with SVM it is possible to obtain very high Recall values at the expense of those of Precision. The choice therefore of the best of the two and the best configuration, depends on which metric you want to maximize: in a context where you want to minimize the risk is better SVM configured to obtain high values of recall (up to 97%), but if you want to minimize the time used to analyze alarms of this type, it is preferable to use a configuration that does not generate too many false alarms and then maximize the Precision (up to 92%).

4.2 Feature Selection

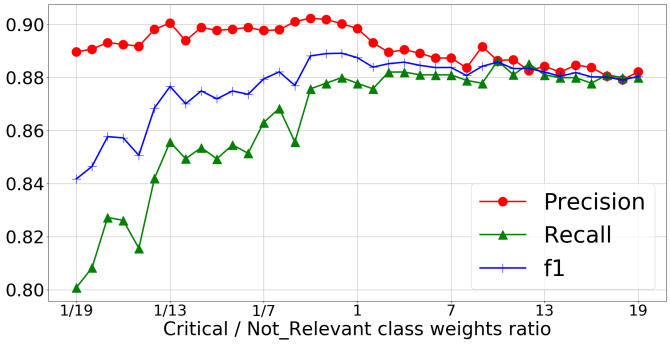
Using a large number of features does not always correspond to an improvement in performance, due to redundant information, noise in the data and overfitting. Using fewer features also reduces the complexity of the processing to be performed, therefore costs and execution times decrease aswell. For this reason, we have studied how the classification performances vary as the number of features decreases, choosing them through standard feature selection techniques.

First of all, we have calculated for each feature individually the linear correlation with the positive class. The figure 6a shows the classification performance as the number of features used increases: note that only the best 21 features (listed in the table 4) are needed to obtain performance very similar to that obtained using the entire

[1] Indicator of compromise: in computer forensics is an artifact (e.g. antiviral signatures, malicious domains or IP Addresses etc.) observed on a network or in an operating system that, with high confidence, indicates a computer intrusion [12]. In this context IoCs are antiviral signatures, malicious IP Addresses, MD5 hashes that uniquely identify a malicious file, URLs and/or domain names from which an attack has been carried or to which a malware connects once activated.
[2] VirusTotal https://www.virustotal.com/gui/home/upload
Replaceable with an equivalent sandbox and antivirus system



(a) RBF SVM



(b) Random Forest

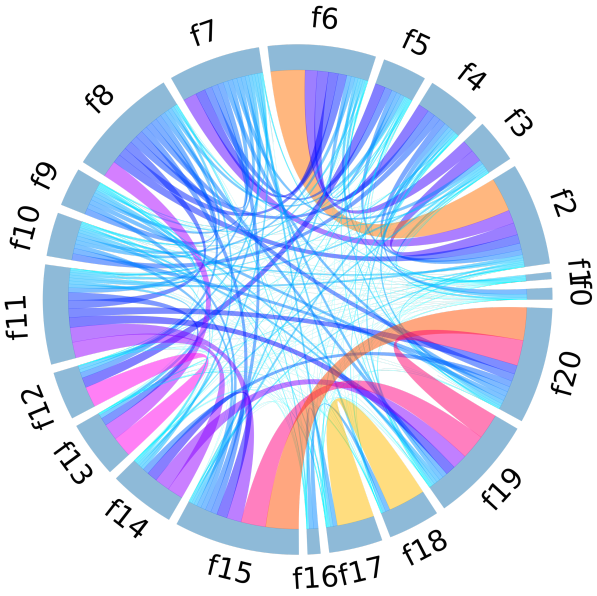
Figure 4: Performance with different class weights

set of features. Thinking about the possibility of further reducing the number of features without suffering an excessive degradation of performance, we measured the existing correlation between features: as you can see in the figure 5a, the information of some of these features are closely related. Therefore, using a Recursive Feature Elimination procedure, each iteration one these 21 features that had a high correlation with at least one other is eliminated. In this way, we have obtained a subset of features that are loosely correlated with each other (as shown in figure 5b) and highly correlated with the positive class. Using this set of features, the classification suffers a minimum degradation of performance, summarized in the table 3.

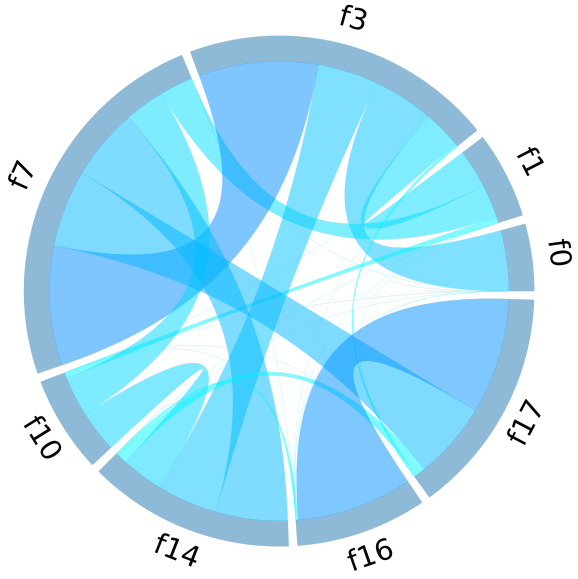
Model	Features	AUC	F1
Random Forest	Full set	0.981	0.892
	Best 21 Features	0.973	0.874
	8 Features best sub-set	0.961	0.852
RBF SVM	Full set	0.966	0.866
	Best 21 Features	0.934	0.850
	8 Features best sub-set	0.919	0.812

Table 3: Performance with different feature sets

Another important advantage in using a reduced number of features is the reduction of sample classification times, mainly consisting of the time taken to extract feature values from the raw data. Although the classifier takes a few milliseconds to perform the classification, the extraction of features is in the order of seconds, due to the interaction with third-party services. As the figure 6b shows, using different feature sets significantly reduces time. The full feature set can normally be used for spam reports; while to classify emails even without being reported, which are in much



(a) Best 21 Features



(b) 8 Features best sub-set

Figure 5: Correlations between features

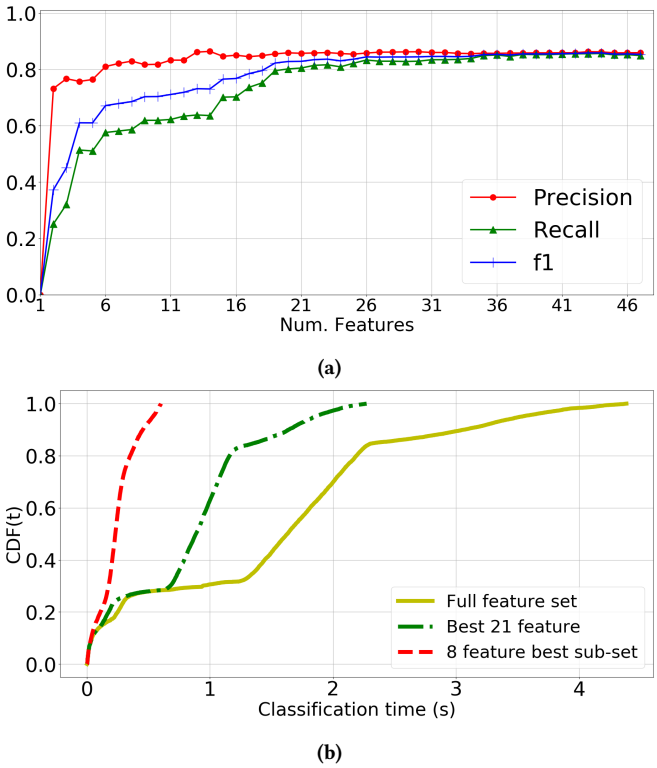


Figure 6: (a) Performance using the best X features; (b) Performance comparison with different feature set in terms of classification time (mainly consisting of feature extraction time)

higher number, it is better to use a reduced feature set that to get the information much faster.

5 CONCLUDING REMARKS

Anti-spam filters do not solve the problem of cyber fraud by spam emails, which are still widely used to spread malware and steal confidential data. Even in the small portion of spam emails that pass the spam filter check, there can be real threats. In case of large companies this portion is wide enough to consider unreasonable a manual analysis. To ensure a high level of security, a collaborative approach is necessary. Through the continuous monitoring of systems by human analysts, spam emails that have created a security incident have been labeled as critical. Using this labeled dataset, we have shown that some types of machine learning algorithms can well classify them as critical highlighting the threats. We have shown the main features that make a spam email effective. We used both legacy and novel features and sorted them by relevance and correlation to the target. Using the entire feature set maximizes classification performance of supervised approaches, up to 92% precision and 97% recall; however, by applying appropriate feature selection techniques we have identified reduced feature sets that greatly reduce execution times and degrade performance little: as often happens only a few dimensions are important to capture the problem.

#	Name	Best sub-set	Description
f0:	n_smtp_blacklist	✓	the number of smtp servers traversed in the blacklists
f7:	gulpease_index	✓	readability index (Flesh formula for english text)
f8:	n_chars_subject		number of characters in the subject
f14:	vt_l_clean	✓	number of links not considered malicious by all engines Virus Total
f3:	vdb_agg_rate	✓	the rate of adjectives within the content
f12:	vt_l_rate		rate of links considered malicious for at least one engine of Virus Total
f11:	n_domains		number of link domains
f17:	n_tip	✓	number of entities in TIP
f10:	n_phishy	✓	number of misleading words, related to phishing, in the content and subject
f6:	voc_rate		the rate of words of the content in the vocabulary (correctly written)
f13:	vt_l_maximum		maximum number of Virus Total engines that consider a link malicious
f5:	vdb_art_rate		the rate of articles within the content
f9:	is_non_ASCII_subj		if the subject contains non-ASCII characters
f18:	n_tip_a		number of attachments in TIP
f16:	vt_a_clean	✓	number of attachments not considered malicious by all Virus Total engines
f1:	n_recipients	✓	the number of recipients
f2:	vdb_rate		the rate of words of the content within the basic vocabulary (most used words)
f19:	n_images		number of images
f20:	n_images_link		number of images as links
f4:	vdb_s_rate		the rate of nouns within the content
f15:	vt_l_unknown		number of unknown links to Virus Total

Table 4: Best 21 Features (sorted)

The contributions shown in this work can definitely lead to a greater awareness of the risks faced by companies and, above all, can lead to the automation of the detection of real threats in spam emails, both in a reporting system context and in a context of Managed Security Services.

REFERENCES

[1] Enrico Blanzieri and Anton Bryl. 2008. A survey of learning-based techniques of email spam filtering. *Artificial Intelligence Review* 29, 1 (01 Mar 2008), 63–92. https://doi.org/10.1007/s10462-009-9109-6

[2] Asaf Cidon, Lior Gavish, Itay Bleier, Nadia Korshun, Marco Schweighauser, and Alexey Tsitkin. 2019. High Precision Detection of Business Email Compromise. In *28th USENIX Security Symposium (USENIX Security 19)*. USENIX Association, Santa Clara, CA, 1291–1307.

[3] Michael Crawford, Taghi M. Khoshgoftaar, Joseph D. Prusa, Aaron N. Richter, and Hamzah Al Najada. 2015. Survey of review spam detection using machine learning techniques. *Journal of Big Data* 2, 1 (05 Oct 2015), 23. https://doi.org/10.1186/s40537-015-0029-9

[4] Yuli Dai, Shunsuke Tada, Tao Ban, Junji Nakazato, Jumpei Shimamura, and Seiichi Ozawa. 2014. Detecting Malicious Spam Mails: An Online Machine Learning Approach. In *Neural Information Processing*. Springer International Publishing, Cham, 365–372.

[5] Haruna Chiroma Shafi'i Muhammad Abdulhamid Adebayo Olusola Adetunmbi Opeyemi Emmanuel Ajibuwa Emmanuel Gbenga Dada, Joseph Stephen Bassi. 2019. Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon* 5, 6 (2019), e01802. https://doi.org/10.1016/j.heliyon.2019.e01802

[6] Special Agent Vicki D. Anderson FBI Cleveland. 2016. FBI Warns of Rise in Schemes Targeting Businesses and Online Fraud of Financial Officers and Individuals. https://www.fbi.gov/contact-us/field-offices/cleveland/news/press-releases/fbi-warns-of-rise-in-schemes-targeting-businesses-and-online-fraud-of-financial-officers-and-individuals.

[7] Wilfried Gansterer and David Pölz. 2009. E-Mail Classification for Phishing Defense. *Proceedings of the 31st European conference on information retrieval (5478)*, 449–460. https://doi.org/10.1007/978-3-642-00958-7_40

[8] S. M. Lee, D. S. Kim, J. H. Kim, and J. S. Park. 2010. Spam Detection Using Feature Selection and Parameters Optimization. In *2010 International Conference on Complex, Intelligent and Software Intensive Systems*. 883–888. https://doi.org/10.1109/CISIS.2010.116

[9] Basavaraju Mallikarjunappa and Dr R. Prabhakar. 2010. A Novel Method of Spam Mail Detection using Text Based Clustering Approach. *International Journal of Computer Applications* 5 (08 2010). https://doi.org/10.5120/906-1283

[10] Dr. Sarwat Nizamani, Nasrullah Memon, Mathies Glasdam, and Dong Duong Nguyen. 2014. Detection of fraudulent emails by employing advanced feature abundance. *Egyptian Informatics Journal* 15 (08 2014). https://doi.org/10.1016/j.eij.2014.07.002

[11] R. Shirey. 2007. RFC4949: Internet Security Glossary, Version 2.

[12] RSA. 2012. *Understanding Indicators of Compromise (IOC) Part I.* https://blogs.rsa.com/understanding-indicators-of-compromise-ioc-part-i/

[13] Takaya Saito and Marc Rehmsmeier. 2015. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE* 10, 3 (03 2015), 1–21. https://doi.org/10.1371/journal.pone.0118432

[14] scikit-learn developers. [n.d.]. Documentation of scikit-learn. https://scikit-learn.org/stable/documentation.html.

[15] Symantec. 2019. 2019 Internet Security Threat Report. https://www.symantec.com/security-center/threat-report.

[16] Amber van der Heijden and Luca Allodi. 2019. Cognitive Triaging of Phishing Attacks. In *28th USENIX Security Symposium (USENIX Security 19)*. USENIX Association, Santa Clara, CA, 1309–1326.