Procesamiento del Lenguaje Natural, Revista nº 72, marzo de 2024, pp. 155-158 recibido 15-12-2023 revisado 20-01-2024 aceptado 27-01-2024

Analysis and classification of spam email using Artificial Intelligence to identify cyberthreats

Análisis y clasificación de correo electrónico no deseado mediante Inteligencia Artificial para la identificación de ciberamenazas

Francisco Jáñez Martino
Departamento de Ingeniería Eléctrica y de Sistemas y Automática, Universidad de León
Campus de Vegazana, s/n, 24007 León, España
francisco.janez@unileon.es

Abstract: Summary of the Ph.D. thesis written by Francisco Jáñez Martino and supervised by Prof. Dra. Rocío Alaiz Rodríguez and Dr. Víctor González Castro at Universidad de León. The defense of the thesis was in León (Spain) in 21st of December 2023 by a committee formed by Dr. Arturo Montejo Ráez (Universidad de Jaén, Spain), Dr. Petr Motlicek (Idiap Research Institute, Switzerland), and Dra. Laura Fernández Robles (Universidad de León, Spain). An international mention was garnered following a six-month tenure at the Universitá di Bologna under the supervision of Dr. Alberto Barrón Cedeño. This Ph.D. thesis was awarded an outstanding Cum Laude grade.
Keywords: Spam email classification, Machine Learning, Attention models, Natural Language Processing, Persuasion detection, Risk classification, Cybersecurity

Resumen: Tesis doctoral realizada por Francisco Jáñez Martino y supervisada por la Prof. Dra. Rocío Alaiz Rodríguez y el Dr. Víctor González Castro en la Universidad de León. La defensa de la tesis se realizó en León (España) el 21 de diciembre de 2023 ante un tribunal compuesto por el Dr. Arturo Montejo Ráez (Universidad de Jaén, España), el Dr. Petr Motlicek (Idiap Research Institute, Suiza), y la Dra. Laura Fernández Robles (Universidad de León, España). Se obtuvo la mención internacional tras una estancia de 6 meses en la Università di Bologna bajo la supervisión del Dr. Alberto Barrón Cedeño. La tesis obtuvo una calificación de sobresaliente Cum Laude.
Palabras clave: Clasificación de correos spam, Aprendizaje Automático, Modelos de atención, Procesamiento del Lenguaje Natural, Detección de la persuasión, Predicción del riesgo, Ciberseguridad

1 Introduction

Spam email has been a problem since the creation of this popular communication medium. Traditionally, these unwanted and unsolicited emails contained advertisements, strange chains or just annoying messages. Due to the rise of Internet and electronic devices, cybercriminals leverage the accessibility of free payment, anonymity and massive use of email services to spread malware, phishing or spoofing attacks among other scams. This turns spam into a big data problem as well as a current cybersecurity challenge.

The main solution for detecting spam email are the anti-spam filters, which showed high performance in the literature. These filters are currently based on Natural Language Processing (NLP) and Machine Learning (ML) models (Dada et al., 2019). However, users still report attacks rooted in spam emails. Hence, understanding, analysing and classifying how spammers design these emails has become a mandatory stage, not only to enhance filtering but also to improve the extraction of information.

In this Thesis, we introduced novel models, methodologies, approaches, and datasets for the analysis and identification of emerging cybersecurity threats in spam

ISSN 1135-5948 DOI 10.26342/2024-72-12

emails. Motivated by our collaboration with the Spanish National Institute of Cybersecurity (INCIBE), our dedication lies in creating applications and conducting research to enhance the early detection of risky and malicious emails. Our approach heavily relies on the application of NLP, as well as Machine and Deep Learning techniques, mainly centred around supervised learning methods.

Several contributions outlined in this dissertation are intended to be integrated into tools being developed by Law Enforcement Agencies (LEAs) and INCIBE. These tools aim to provide more comprehensive and timely alerts to organizations and citizens regarding potential risks posed by spam email. This thesis proposed models aimed at ensuring the security, integrity, and privacy of users in the face of cyberattacks originating from spam emails.

Our main objectives were: a) classifying spam emails according to their cybersecurity topic, b) spotting both the presence of persuasion and the specific techniques employed and c) extracting potentially useful information from both their headers and body to spot risky emails. Additionally, many of the data mining and NLP techniques can be utilized for similar issues, such as smishing, fraudulent content on websites, or social media.

2 Thesis Overview

This thesis consists of seven chapters, which are described as follows:

Chapter 1 We outlined the objectives and motivation behind the thesis. Our motivation moving away from the traditional spam filtering to provide support for cybersecurity organizations to comprehend the properties of spam emails and present models to expedite and enhance their analysis.

Chapter 2 We reviewed the state-of-the-art anti-spam filters, and found that they showed high performance on outdated datasets during their evaluation. However, their assessment did not consider two challenging problems in the spam domain: dataset shift and spammer strategies to deceive these filters. Our review encompassed the investigation of dataset shift in ML models considering adversarial environments and works related to detecting specific spammer strategies. In depth, we reviewed the study of spammer tricks like obfuscated words, poisoning text, hidden text, image-based spam and other emerging trends. Moreover, we carried out an empirical experimentation to provide supporting evidences. Finally, we explored the existing cybersecurity challenges associated with spam email. The review and experimentation of this Chapter has been published in Artificial Intelligence Review journal (Jáñez-Martino et al., 2022).

Chapter 3 In this Chapter, we addressed the development of a text classifier capable of identifying the cybersecurity topic of a spam email. We used a hierarchical clustering and manual inspection to define eleven cybersecurity classes for the first time in the literature.

We conducted an evaluation (per language) of the combinations of two traditional approaches, Term Frequency - Inverse Document Frequency (TF-IDF) and Bag of Words (BoW) and two word embeddings models, word2vect and BERT (Devlin et al., 2018), as text representation techniques along with four popular ML classifiers: Support Vector Machine (SVM), Naïve Bayes, Logistic Regression and Random Forest.

We also provided the confusion matrices, an evaluation of the models performance per class and a data augmentation analysis using both reducing the majority classes and increasing the minority class. The work of this Chapter has been published in Applied Soft Computing journal (Jáñez-Martino et al., 2023). A preliminary study was published on Arxiv (Jáñez-Martino et al., 2020).

Chapter 4 In the fourth Chapter, we sought to identify persuasive elements in spam email. Upon reviewing the state of the art, we found out theoretical and psychological works associated with different kinds of spam emails like phishing emails. Due to this fact, we started from relating the persuasive principles presented in (Ferreira, Coventry, and Lenzini, 2015) to the growing attention in developing automatic models to spot persuasion and those techniques involved in news articles (Da San Martino et al., 2019). Theses works set the basis of our hypothesis, which is to analyze the role of persuasion in manipulating users to perform an specific action such as clicking in an external link or opening an attachment.

We designed NLP models at three levels of granularity: full email, sentences and span text (one or more words but always shorter than a sentence). We detailed our approach to use the datasets and models de-

rived from persuasion in news articles for full email and span text classification. For sentence classification, we described the creation of a manually annotated dataset following binary and multilabel annotations and adjusted pre-trained models.

This chapter has been covered by a paper presenting the whole study and submitted to a journal.

Chapter 5 In this Chapter, we aimed to extract further information from spam email to improve the warnings launched by cybersecurity agencies to report organization and citizens about spam campaigns and frauds involving harmful and risky emails (Gallo et al., 2021). We followed two approaches: a) binary classification (high and low risk) and regression (scaling the email in a level of risk range from 1 to 10).

We analyzed the spam email through a NLP feature extraction according to reported key points pointed out by cybersecurity experts. We also used the previous cybersecurity topics as features and conducted an extend investigation to determine the quality of email address of spammer senders. We explained every feature and analyzed the relevance of each one and group for classification.

The extended work on address classification was presented at the Document Engineering 2021 conference (Jáñez Martino et al., 2021). The paper presenting the whole system has been submitted to a journal.

Chapter 6 We highlighted our eight main findings and future work, emphasizing the expansion of some research lines. In addition, we expressed our interest in applying the methodologies developed in this thesis to other domains, such as social media or instant messaging.

Chapter 7 In compliance with university requirements, we translated the conclusions and future work presented in Chapter 6 into Spanish.

3 Contributions

We enumerated the principal contributions of this thesis as follows:

We outlined the spam filtering, spammer strategies and dataset shift problem. We empirically demonstrated how these factors negatively impact the evaluation of anti-spam filters. We compared the performance of filters when being trained on one dataset and evaluated on other dataset (using five of the most used spam emails datasets, both back and forth in time). This review underscored the importance of comprehending spam properties to enhance both the filters and their assessment. Additionally, it also highlighted the need to study the detection of spammer strategies such as hidden text, word obfuscation, or text embedded in images, as well as other emerging tricks like mixing languages.

We semi-automatically labeled a novel dataset in the spam emails domain by using hierarchical clustering and visual inspection through a collection of INCIBE spam emails. Our dataset called Spam Email Classification (SPEMC) holds almost 15k spam emails per language (both English and Spanish) divided into eleven cybersecurity topics. These are academic media, extortion hacking, fake reward, health, identity fraud, money making, pharmacy, service, sexual content dating, work offer and other.

We presented a text classification pipeline based on traditional text representation techniques and word embeddings along with four popular ML algorithms to detect the eleven cybersecurity topics. This pipeline includes an email processing stage to extract all textual content from the subject, body and images. We considered the appearance of spammer strategies, such as image-based message or hidden text, and we applied Optical Character Recognition (OCR) techniques to extract only the visible text.

We developed automatic systems to detect persuasion and its techniques at different levels of granularity: full email, sentences and text spans. We replicated ML models based on NLP features as well as fine tuning pretrained attention models. For sentences classification, we manually annotated sentences of spam emails based on binary, persuasive or not, and multilabel perspective, containing eight persuasion techniques labels plus the negative one.

We introduced a novel set of 56 features based on NLP to discriminate those spam emails with more potential risk for individuals and organizations. We divided the features into five groups: headers, text, attachments, URLs and protocols. We developed models following two approaches: classification and regression.

We manually annotated two spam email datasets collected in different sources, one private from INCIBE resources and one pu-

blic from Spam Archive of Bruce Guenter[1], based on their potential risk. We labeled them for (i) a regression problem using a scale of risk (1-10) and (ii) a classification problem distinguishing two classes, low and high risk. Low risk spam refers to messages that closely resemble traditional ones containing advertisements and annoying content, but without the presence of malware or scams that could end exposing leaked data of users. While high risk level include cybersecurity attacks such as spreading ransomware, phishing, spoofing or extortion.

We evaluated three classifier and three estimators using our novel set of features as input. We conducted an analysis of feature importance for the classification approach by systematically removing or retaining one set of features. We also evaluated the relevance of removing each individual feature one by one to establish a cutoff number of features. Due to the high relevance of the address classification according to cybersecurity experts, we carried out an extended investigation. The objective was to classify the address of a spam sender into low and high quality. To do this, for the first time in the literature, we presented a set of 18 features extracting information from the username, domain and top-level domain (TLD) of each address and fed up four ML classifiers for evaluation using a manually labeled dataset call Email Address Quality - 6k. A high quality is given when the address contains popular brands or email services, truthful TLDs and imitate common user's address without random number, characters or letters and short username or domains.

Acknowledgements

This work was supported by the framework agreement between the Universidad de León and INCIBE (Spanish National Cybersecurity Institute) under Addendum 01 and the Predoctoral Grant of Junta de Castilla y León.

References

Da San Martino, G., S. Yu, A. Barrón-Cedeño, R. Petrov, and P. Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646.

Dada, E. G., J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa. 2019. Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon*, 5(6):e01802.

Devlin, J., M. Chang, K. Lee, and K. Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805:1–16.

Ferreira, A., L. Coventry, and G. Lenzini. 2015. Principles of persuasion in social engineering and their use in phishing. In T. Tryfonas and I. Askoxylakis, editors, *Human Aspects of Information Security, Privacy, and Trust*, pages 36–47, Cham. Springer International Publishing.

Gallo, L., A. Maiello, A. Botta, and G. Ventre. 2021. 2 years in the anti-phishing group of a large company. *Computers & Security*, 105:102259.

Jáñez Martino, F., R. Alaiz-Rodríguez, V. González-Castro, and E. Fidalgo. 2021. Trustworthiness of spam email addresses using machine learning. In *Proceedings of the 21st ACM Symposium on Document Engineering*, DocEng '21, page 4, New York, NY, USA. Association for Computing Machinery.

Jáñez-Martino, F., R. Alaiz-Rodríguez, V. González-Castro, E. Fidalgo, and E. Alegre. 2022. A review of spam email detection: analysis of spammer strategies and the dataset shift problem. *Artificial Intelligence Review*, 56:1145–1173.

Jáñez-Martino, F., R. Alaiz-Rodríguez, V. González-Castro, E. Fidalgo, and E. Alegre. 2023. Classifying spam emails using agglomerative hierarchical clustering and a topic-based approach. *Applied Soft Computing*, 139:110226.

Jáñez-Martino, F., E. Fidalgo, S. González-Martínez, and J. Velasco-Mata. 2020. Classification of spam emails through hierarchical clustering and supervised learning.

[1]http://untroubled.org/spam/ retrieved December 2023