Spam Email Detection Using Machine Learning Techniques

Ioannis Moutafis[1], Antonios Andreatos[1] and Petros Stefaneas[1,2]
[1]Division of Computer Engineering and Information Science, Hellenic Air Force Academy, Dekeleia Air Force Base, Attica 13671, Greece
[2]Department of Mathematics, National Technical University of Athens, Politechneioupoli, Iroon Polytechneiou 9, Zografou 15772, Athens, Greece

moutafisioannis@yahoo.com
antonios.andreatos@hafa.haf.gr
petros@math.ntua.gr

Abstract: This paper focuses on the security of electronic mail, using machine learning algorithms. Spam email is unwanted messages, usually commercial, sent to a large number of recipients. In this work, an algorithm for the detection of spam messages with the aid of machine learning methods is proposed. The algorithm accepts as input text email messages grouped as benevolent ("ham") and malevolent (spam) and produces a text file in csv format. This file then is used to train a bunch of ten Machine Learning techniques to classify incoming emails into ham or spam. The following Machine Learning techniques have been tested: Support Vector Machines, k-Nearest Neighbour, Naïve Bayes, Neural Networks, Recurrent Neural Networks, Ada Boost, Random Forest, Gradient Boosting, Logistic Regression and Decision Trees. Testing was performed using two popular datasets, as well as a publicly available csv file. Our algorithm is written in Python and produces satisfactory results in terms of accuracy, compared to state-of-the-art implementations. In addition, the proposed system generates three output files: a csv file with the spam email IP addresses (of originating email servers), a map with their geolocation, as well as a csv file with statistics about the countries of origin. These files can be used to update existing organisational filters and blacklists used in other spam filters.

Keywords: Spam, Anti-spam filters, Machine Learning, Deep Learning, Blacklists, Geolocation, Classification

1. Introduction

Since the early days of the Internet, email has been one of the most popular applications; today email is the most important medium used in communications as it is widely used in almost every professional field like business, commerce, banking, education, as well as for interpersonal communications among Internet users.

The availability of bulk mailing software and lists of email addresses harvested from web sites, newsgroup archives, chain or bulk emails with visible the recipients, and service provider directories, allows messages to be sent automatically to millions of recipients at essentially no cost (Michelakis et al, 2004).

Spam is a major problem in today's world. Spam email is sent in bulk for advertising, scams, phishing or other purposes, by unknown senders. Spam messages usually contain commercial advertisements or malware that can harm users' systems and data (Ahmed et al, 2022). Phishing attacks, on the other hand, are a type of cybercrime that involves tricking individuals into sharing sensitive information or performing harmful actions through fraudulent means. The proliferation of spam messages over the past twenty years has been undermining constantly the usability of email. Apart from potential risks, spam messages are extremely annoying to most users, as they clutter their mailboxes, waste users' time and mailbox space, ISPs' bandwidth and CPU time. Therefore, the development of classifiers that can distinguish between spam and official emails is vital (Hemalatha et al, 2022).

According to "spamlaws.com", 86% of email addresses posted on websites are used by spammers to send unsolicited emails; 63% of all "unsubscribe" requests are not honoured. Spam accounted for 80.26% of all email traffic received in 2011 (SpamLaws, 2023). Figure 1 displays the global spam volume as percentage of total email traffic from 2011 to 2021.

Based on the detection method, three ways to mitigate spam attacks dominate: a) user awareness, b) blacklists, and c) spam detectors or anti-spam filters. A spam detector is a program used to detect unsolicited, unwanted, phishing and virus-infected messages automatically and prevent those messages from reaching users' inboxes.

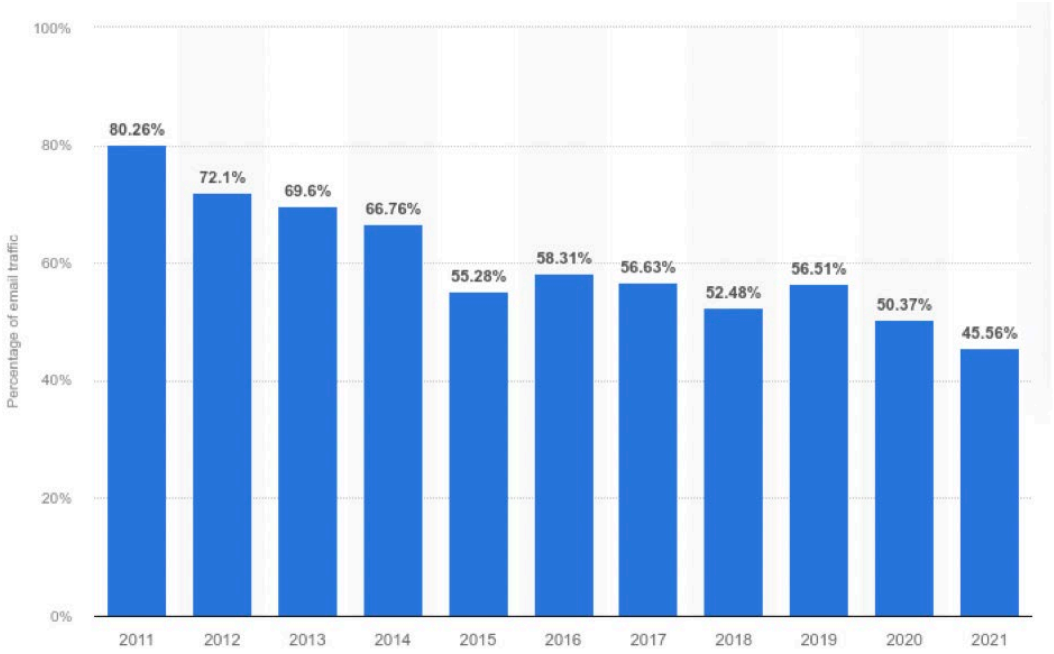

Figure 1: Global spam volume as percentage of total email traffic from 2011 to 2021

Source: https://www.statista.com/statistics/420400/spam-email-traffic-share-annual/

Anti-spam filters seem to be the most viable solution nowadays. Most commercially available filters of this type currently appear to rely on simple techniques such as whitelists of trusted senders, blacklists of known spammers, and hand-crafted rules that block messages containing specific words or phrases (Ahmed et al, 2022). On the other hand, the success of Machine Learning in text categorisation has led researchers to explore learning algorithms in anti-spam filtering (Ahmed et al, 2022). Learning-based anti-spam filters, initially based on Naïve Bayes (NB), have been in operation for a few years (Michelakis et al, 2004). Recently, various other machine learning methods have been tested by researchers and demonstrate superior performance, as we shall see below.

In this paper, an algorithm for the detection of spam and phishing messages with the help of machine learning is proposed. The algorithm accepts as input text email messages written in English, grouped as benevolent ("ham") and malevolent (spam), and produces a set of training data in csv format (comma separated values) with the following columns: id, sender, subject, content and label. The fifth column is a binary variable characterising the messages. Next, the algorithm uses ten supervised learning methods that produce input-output mapping functions. Once trained, the algorithm can automatically recognise spam emails with high accuracy. In addition, the algorithm produces as output a file with the spam email sender Ipext, addresses, a csv file with statistics about the countries of origin, as well as a map with the geolocation of originating spam email servers. These files can be used to update existing blacklists and websites such as "Abuse IP DB" (https://www.abuseipdb.com), a project dedicated to helping combat the spread of hackers, spammers, and abusive activity on the internet.

The remainder of this paper is organised as follows: Section 2 reviews the related work on available anti-spam filters, focusing on learning-based and novel alternative approaches. Section 3 presents the datasets used. Section 4 presents the proposed system. Section 5 summarises the results of our algorithm on the benchmark datasets. Section 6 concludes the paper and suggests directions of further research.

2. Related work

Since identification is a specific task of classification, machine learning is usually used to classify email into two classes, namely spam and benign messages, also called "ham". Several studies using machine learning-based spam classification have been published in the research bibliography (Ghosh & Senthilrajan, 2023; Karyawati et al, 2023). Previous research work on anti-spam filtering studied the performance of popular machine learning algorithms (Michelakis et al, 2004; Ahmed et al, 2022; Ghosh & Senthilrajan, 2023).

Machine learning is easy to implement and good for high-dimensional data. However, it requires large and unbiased amounts of data. In addition, parameter adjustment is required to select the best model (Karyawati et al, 2023).

Laorden (2012) developed a Word Sense Disambiguation preprocessing step before applying machine learning algorithms to detect spam data. Results indicated a 2 to 6% increase in the precision score when applied on Ling Spam and TREC datasets.

Several studies have used Support Vector Machines (SVM) for spam classification. Jáñez-Martino et al (2020) made a combined model of TF-IDF (Term Frequency - Inverse Document Frequency) and SVM showed a 95.39% F1 score and the fastest spam classification was achieved with the help of the TF-IDF and NB approach (Roiger, 2017; Tan et al, 2006).

Deep learning is a subfield of machine learning that uses neural networks to learn patterns in data, mimicking the human brain. Marie-Saint et al (2020) employed deep learning to solve the given firefly algorithm with SVM and Arabic text.

Nonlinear kernel functions are often used to convert the input data to a high-dimensional feature space where the input data becomes more separable than the original input space (Karyawati et al, 2023).

Pratiwi and Ulama (2016) compared k-Nearest Neighbour (kNN), linear kernel SVM and Radial Basis Function (RBF) kernel SVM methods. It was found that the kNN method with k=3 produced the best accuracy of 92.28% while the best accuracy in the SVM method (96.6%) was obtained using the SVM linear kernel.

Chrismanto and Lukito (2017) compared the Naïve Bayes method and the SVM method with the RBF kernel to identify Instagram comment spam. The results showed that the SVM method produced an accuracy of 78.49%, outperforming Naïve Bayes which achieved an accuracy of 77.25%.

Pratiwi and Ulama (2016) proposed a combination of kNN and SVM methods. They used a kNN-based sampling strategy to find close neighbors to improve the performance of the SVM method. The results of the study based on publicly available dataset (Dredze) showed that the accuracy increased to about 98%.

Reddy and Reddy (2021) proposed a new spam detection method that is effective in distinguishing spam from its content. The proposed classifier achieved an accuracy of 95.32%.

Spammers have introduced a new technique of embedding spam content into images attached to email messages. Harisinghaney et al (2014) implemented text and image-based spam emails with the help of the kNN, NB and reverse DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithms. Similarly, Zamil et al (2019) proposed a method for distinguishing the ham and spam images using a combination of SVM and kNN, achieving an average accuracy of 97.27% when K equals 20.

Hemalatha et al (2022) discussed the use of the Naïve Bayes classifier to detect spam. Their classifier used techniques like lemmatisation, word removal and lowercasing, to analyse the text of the email in order to determine whether it is spam or not. The RFD model has also been used in anti-spam filters. This model can identify spam patterns, non-spam patterns and common patterns to determine whether an email is spam or ham. The RFD model offers general patterns that users can use to decide whether an email is spam or not. The proposed system is currently for English language emails but could be expanded to other languages in the future. Additionally, the study discussed how images can be used in spam emails and how characteristics like file properties, histogram, and hough change can be used to detect them. The text emphasized the importance of detecting spam to avoid loss of important mail but also protect user systems from harm.

Benavides et al (2020) reviewed various deep learning techniques used for classifying phishing attacks. The study discussed the limitations of traditional approaches (such as rule-based and signature-based methods) to phishing detection. Next, the authors highlighted the potential of deep learning techniques for detecting phishing attacks. The article reviewed several studies that have employed deep learning techniques, including artificial neural networks, convolutional neural networks and recurrent neural networks (RNN) for detecting phishing attacks. The authors suggested that deep learning techniques can provide accurate classification of phishing attacks and outperform traditional approaches. The study also noted the limitations of current research, such as the lack of standardization in data sets and the need for more comprehensive evaluations of deep learning-based phishing detection solutions.

Karyawati et al (2023) investigated several kernel functions (i.e., linear, polynomial and RBF) to obtain the best SVM model for classifying spam emails. Some experiments were conducted to determine the effect of parameter

changes for each kernel function. The SVM performance was assessed using the Precision, Recall, and F-Measure metrics. They used a publicly available email dataset posted on Kaggle, which is based on the Enron corpus (https://www.kaggle.com/datasets/venky73/spam-mails-dataset). The total number of 1100 messages consists of 550 spam emails and 550 benign emails, of which 1000 emails were used as training data (i.e., to select the best model of each kernel), and 100 emails are used for testing (i.e., to evaluate the best model). The feature data used was word frequency. The data was in the form of text in csv format.

Karyawati et al. (2023) concluded that the evaluation of the best model of each kernel model (among five models: linear, polynomial of degree 1, 2, 3 and RBF kernel) using the testing data of identification Spam email showed that the linear kernel and the kernel of degree 1 polynomial produced the best Precision, Recall and F-Measure performances of 0.99, 0.99 and 98% respectively, compared to the RBF kernel which produced Precision, Recall and F1 score of 0.95, 0.95 and 0.94, respectively. The worst model was a degree 3 polynomial kernel which produced Precision, Recall and F1 score of 0.85, 0.79 and 0.78, respectively.

In a recent research, Ghosh and Senthilrajan (2023) implemented and compared thirteen state-of-the-art machine learning classifiers: Adaptive Booster, Artificial Neural Network, Bootstrap Aggregating, Decision Table, Decision Tree, J48, K-Nearest Neighbor, Linear Regression, Logistic Regression, Naïve Bayes, Random Forest, Sequential Minimal Optimization and SVM. They concluded that in terms of accuracy, the Random Forest classifier performs best and the performance of the Naïve Bayes classifier is substandard compared to the rest of the classifiers. The Random Forest (RF) classifier achieved the accuracy of 99.91% for the Spam Corpus and 99.93% for the Spambase datasets. The Naïve Bayes classifier achieved the accuracy of 87.63% for the Spam Corpus and 79.53% for the Spambase datasets.

Finally, Sahmoud and Mikki (2022) investigated the application of a powerful natural language processing (NLP) model that has been shown to be effective in a variety of NLP tasks, called BERT (Bidirectional Encoder Representations from Transformers) in detecting spam messages. The authors trained their model on many corpuses such as SpamAssassin corpus, SMS Spam Collection v.1, Enron corpus and the Ling-Spam corpus, and analysed the results. They reported very good results of using BERT in every corpus they tested (from 97.83% to 99.28%), outperforming other machine learning algorithms in many tasks. The authors attributed the high performance of their model to the fact that the algorithm can better understand the message context and correlate the words (Sahmoud & Mikki, 2022).

3. Training the model

Common datasets which have been used in the literature include: Enron Corpus, SpamAssassin Corpus, SMS Spam Detection Corpus, Ling-Spam Corpus (Sahmoud & Mikki, 2022). In our research, the following datasets were used to train the model:

1) A dataset from SpamAssasin containing 2551 ham and 501 spam emails available online from:

https://www.kaggle.com/datasets/veleon/ham-and-spam-dataset

2) The Enron1 dataset available online from: https://www2.aueb.gr/users/ion/data/enron-spam, containing 3228 ham and 3219 spam emails.

3) We also used a ready csv file available online from:

https://www.kaggle.com/code/mfaisalqureshi/email-spam-detection-98-accuracy/data

This csv file (spam.csv, 480.13 kB) prepared by Faisal Qureshi, contains 5157 unique instances of ham (87%) and spam (13%) messages.

4. The proposed system

The proposed system accepts two sets of emails as input (ham and spam), written in English. It uses preprocessing to prepare a csv file with selected fields of the messages, as well as their characterisation. Then it uses this csv file to feed a Machine Learning engine running ten different algorithms mentioned in the Abstract, which after training can be used to filter the user's incoming messages. Alternatively, our code also accepts ready csv files. As an output, the proposed algorithm generates a csv file containing the spam emails source IP addresses and the countries of origin, statistics in textual and graphical format (histogram), as well as a map with the geolocation of the spammers. The algorithm is written in Python. Figure 2 presents a block diagram of the proposed system.

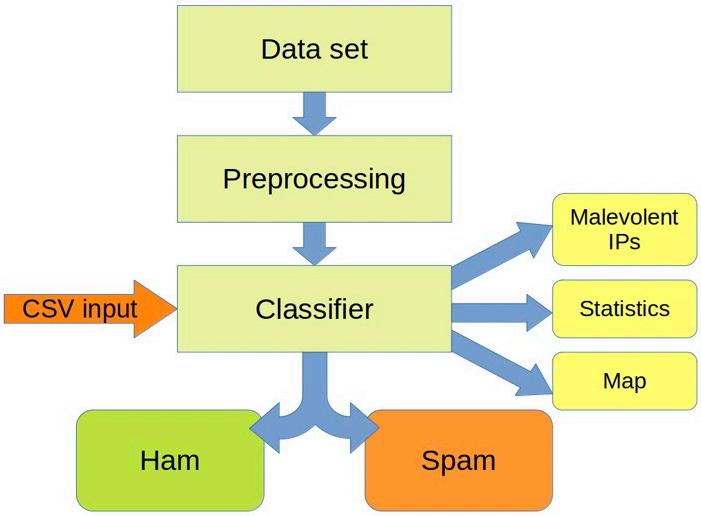

Figure 2: Block diagram of the proposed system

The code imports several Python libraries such as Json, Folium, Seaborn, Pandas and Nltk. It defines a function called "clean_text" which preprocesses email data by converting all the words to lowercase, removing URLs, punctuation marks, HTML tags, numbers and lemmatizing words. It also removes stopwords and filters out words that are less than four letters long. The function returns a clean text.

A function named "find_elements" extracts sender, subject, content and country information from emails and writes the extracted data into a csv file. Next, the "find_elements" function utilises regular expressions to extract IP addresses from emails and uses geolocation-db.com to obtain country names from IP addresses. Then, it calls the "clean_text" function to preprocess the email content. The function "create_stats_and_map" reads the csv file generated by the "find_elements" function and creates a map using the Folium library; it also uses a GeoJSON file that contains country boundaries and highlights each country according to the number of emails originating from it.

Figure 3 presents part of an output csv file. Figure 4 presents an example of the statistics, and Figure 5 presents a geolocation map with the places of the spammers of the Enron1 dataset.

	Sender	IP addr	Country
2	12a1mailbot1@web.de	203.122.2.197	India
3	ilug-admin@linux.ie	67.104.83.251	United States
4	sabrina@mx3.1premio.com	209.63.151.251	United States
5	wsup@playful.com	80.71.66.8	Denmark
6	social-admin@linux.ie	67.104.83.251	United States
7	Thecashsystem@firemail.de	24.214.168.210	United States
8	fort@bluemail.dk	217.127.249.196	Spain
9	OWNER-NOLIST-SGODAILY*JM**NETNOTE	209.216.124.212	Canada
10	Thecashsystem@firemail.de	24.214.168.210	United States
11	suz0123893616943@yahoo.com	65.113.29.188	United States
12	hurst@missouri.co.jp	202.101.163.34	China
13	simply-amateur-zzzz=example.com@free4por	205.252.89.51	United States
14	aileen@email2.qves.net	67.104.83.251	United States
15	OWNER-NOLIST-SGODAILY*JM**NETNOTE	168.143.181.42	United States
16	approvals@mindspring.com	200.173.221.83	Brazil
17	weseloh@bibsam.kb.se	200.24.85.9	Brazil
18	des34newsa@hotmail.com	63.186.16.221	United States
19	jjj@mymail.dk	61.97.34.233	South Korea
20	seko_mam@spinfinder.com	64.86.155.179	United States
21	safety33o@l11.newnamedns.com	64.25.38.81	United States
22	ilug-admin@linux.ie	209.63.151.251	United States
23	ilug-admin@linux.ie	64.86.155.148	United States
24	bell1hmed@yahoo.ca	64.86.155.148	United States
25	health104580m43@mail.com	135.5.185.20	United States
26	iq@insurancemail.net	65.217.159.103	United States
27	george300@Flashmail.com	61.129.68.17	China

Figure 3: Part of an output csv file

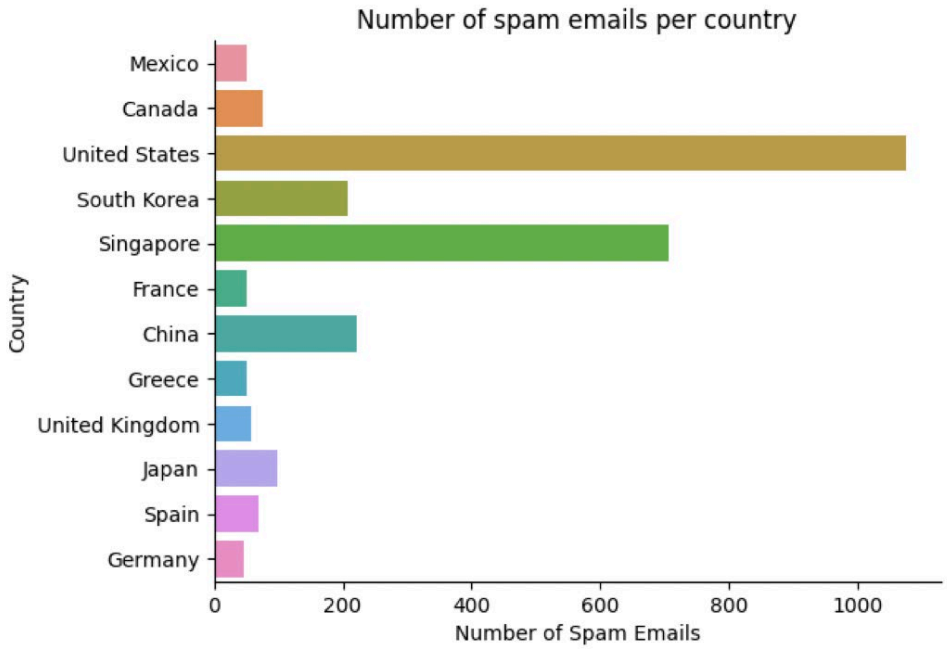

Figure 4: Statistics resulting from the spam emails (spam emails per country)

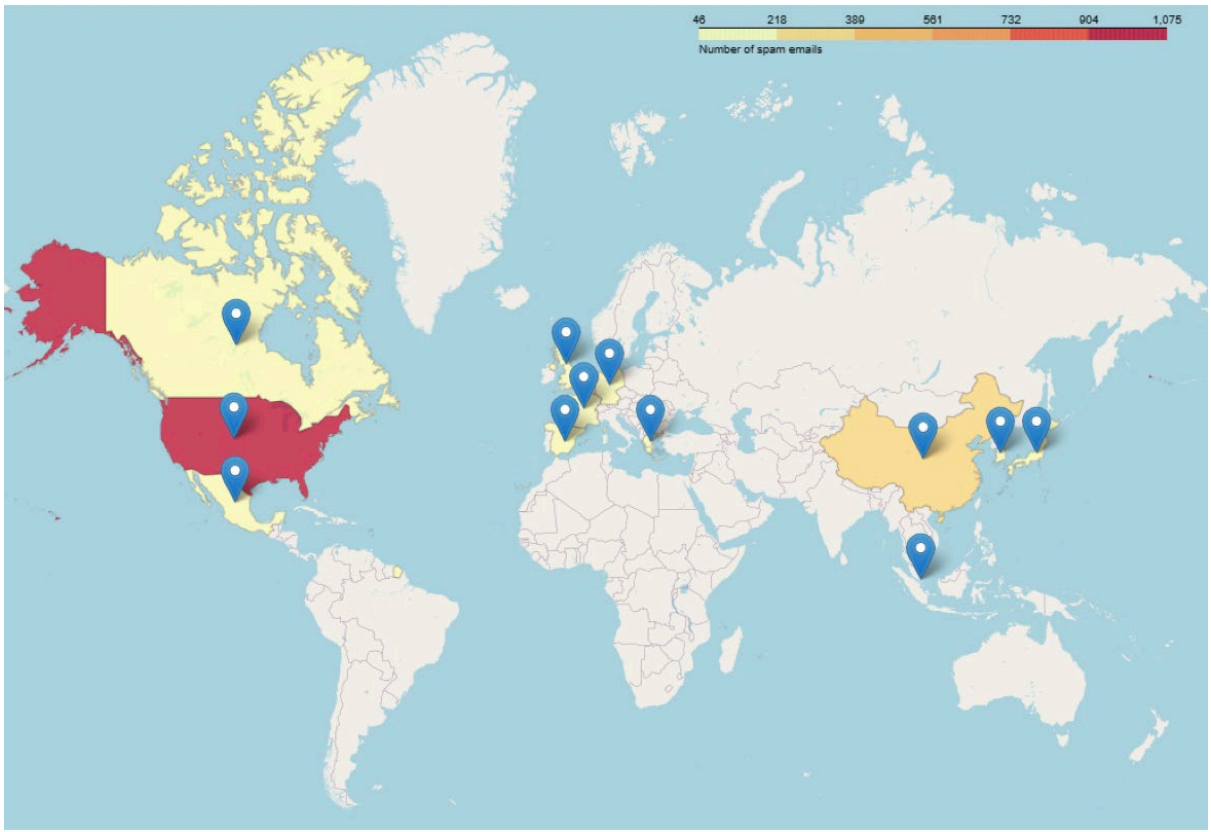

Figure 5: A geolocation map with the places of the spammers

5. Results

Table 1 shows our test results using several algorithms on the datasets under consideration, namely: Support Vector Machines (SVM), k-Nearest Neighbour (kNN, with k=1), Naïve Bayes (NB), Neural Networks (NN), Recurrent Neural Networks (RNN), Ada Boost (AB), Random Forest (RF), Gradient Boosting (RF), Logistic

Regression (LR) and Decision Trees (DT). The results are comparable and in some cases better than those reported in the recent bibliography (Ghosh & Senthilrajan, 2023; Ahmed et al, 2022).

Table 1: Accuracy of the proposed algorithm on various datasets under consideration

Method ► ▼ Dataset	SVM	kNN	NB	NN	RNN	AB	RF	GB	LR	DT
Spam Assasin	99.18%	95.9%	91.31%	99.51%	98.36%	98.2%	98.2%	95.9%	95.41%	95.73%
Enron1	99.38%	98.53%	98.3%	99.22%	99.3%	98.3%	97.2%	96.6%	99.22%	96.5%
csv by F. Qureshi	98.39%	95.51%	95.43%	98.23%	98.3%	98.38%	98.11%	95.9%	96.5%	97.85%

6. Conclusion

In this research work, an algorithm for the classification of raw emails consisting of spam and benign messages in English was presented. At a first stage, preprocessing produces a csv file with the basic email characteristics, which is next used to train ten popular machine learning classifiers to detect spam emails (Support Vector Machines, k-Nearest Neighbour, Naïve Bayes, Neural Networks, Recurrent Neural Networks, Ada Boost, Random Forest, Gradient Boosting, Logistic Regression and Decision Trees). Almost all classifiers demonstrate satisfactory performance, comparable to or superior than state-of-the-art implementations. Two publicly available datasets were used: the SpamAssasin dataset and the Enron1 dataset. Concerning the SpamAssasin dataset, the best performance (99.51 %) is achieved by NN. For the Enron1 dataset, the best performance (99.38 %) is achieved by SVM. Our algorithm also accepts ready csv files produced by other sources. We also tested our algorithm with a publicly available csv file; SVM and Ada Boost achieved the best performance (almost 98.4 %).

The proposed algorithm, implemented in Python, contains about 490 lines of code and can classify spam and ham emails. In addition, the proposed algorithm produces a csv file with the source IP addresses of the spam emails, statistics and graphs about the origin country of the spam emails, as well as a map with their geolocation. The IP addresses of the spam email senders can be used in spam email filters or to update spam email databases. In the future we intend to produce and test our own dataset, as well as work on spam filter personalisation.

References

Ahmed, N., Amin, R., Aldabbas, H., Koundal, D., Alouffi B. and Shah, T. (2022) "Machine Learning Techniques for Spam Detection in Email and IoT Platforms: Analysis and Research Challenges", *Hindawi Security and Communication Networks*, Vol. 2022, Article ID 1862888, https://doi.org/10.1155/2022/1862888

Benavides, E., Fuertes, W., Sanchez, S. and Sanchez, M. (2020) "Classification of Phishing Attack Solutions by Employing Deep Learning Techniques: A Systematic Literature Review". In Á. Rocha and R. P. Pereira (eds.), *Developments and Advances in Defense and Security, Smart Innovation, Systems and Technologies 152*, Springer Nature Singapore Pte Ltd. https://doi.org/10.1007/978-981-13-9155-2_5

Chrismanto, A. R. and Lukito, Y. (2017) "Identifikasi Komentar Spam Pada Instagram", Lontar Komputer : Jurnal Ilmiah Teknologi Informasi, Vol. 8, no. 3, p. 219, doi: 10.24843/lkjiti.2017.v08.i03.p08.

Ghosh, A. and Senthilrajan, A. (2023) "Comparison of machine learning techniques for spam detection", *Multimedia Tools and Applications*, 1-28. 10.1007/s11042-023-14689-3.

Harisinghaney, A., Dixit, A., Gupta, S. and Arora, A. (2014) "Text and image based spam email classification using kNN, Naïve Bayes and Reverse DBSCAN algorithm", 2014 International Conference on Reliability Optimization and Information Technology (ICROIT) pp 153-155.

Hemalatha, M., Katta, S., Santosh, R. S. and Priyanka, Ms. (2022) "E-Mail Spam Detection", *International Journal of Computer Science and Mobile Computing*, Vol. 11, Issue 1, Jan. 2022, pp. 36-44.

Jáñez-Martino, F., Fidalgo, E., González-Martínez, S. and Velasco-Mata, J. (2020) "Classification of Spam Emails through Hierarchical Clustering and Supervised Learning," *CoRR*, Vol. abs/2005.08773.

Karyawati, A. E., Wijaya, K. and Supriana, I W. S. (2023) "A Comparison of Different Kernel Functions of SVM Classification Method for Spam Detection", *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, 8, pp 91-97, doi: 10.33480/jitk.v8i2.2463.

Laorden, C., Santos, I., Sanz, B., Alvarez, G. and Bringas, P. G. (2012) "Word Sense Disambiguation for Spam Filtering", *Electron. Commer. Rec. Appl.*, Vol. 11, pp 290–298.

Marie-Sainte, S. L. and Alalyani, N. (2020) "Firefly Algorithm based Feature Selection for Arabic Text Classification", *J. King Saud Univ. Comput. Inf. Sci.*, Vol. 32, pp 320-328.

Michelakis, E., Androutsopoulos, I., Paliouras, G., Sakkis, G., and Stamatopoulos, P. (2004) "Filtron: A Learning-Based Anti Spam Filter", *International Conference on Email and Anti-Spam*.

Pratiwi, S. N. D. and Ulama, B. S. S. (2016) "Klasifikasi Email Spam dengan Menggunakan Metode Support Vector Machine dan k-Nearest Neighbor", *Jurnal Sains dan Seni ITS*, Vol. 5, No. 2, pp 344–349.

Reddy, G. A. and Reddy, B. I. (2021) "Classification of Spam Text using SVM", *Journal of University of Shanghai for Science and Technology*, Vol. 23, No. 08, pp 616–624, doi: 10.51201/jusst/21/08437.

Roiger, R. J. (2017) *Data Mining: A Tutorial-Based Primer*, 2nd ed. Boca Raton, CRC Press, Taylor & Francis Group.

Sahmoud, T. and Mikki, M. A. (2022) "Spam Detection Using BERT", preprint 10.48550/arXiv.2206.02443.

SpamLaws.com (2023) [online], https://www.spamlaws.com [Accessed 6 March 2023].

Tan, P.-N., Steinbach, M. and Kumar, V. (2006) *Introduction to Data Mining*. Boston: Pearson Addison-Wesley.

Zamil, Y. K., Ali, S. A. and Naser, M. A. (2019) "Spam image email filtering using K-NN and SVM", *International Journal of Electrical and Computer Engineering* (IJECE), Vol. 9, no. 1, pp 245-254, doi: 10.11591/ijece.v9i1.