Journal of ELECTRICAL ENGINEERING, VOL 73(2022), NO4, 292–298

sciendo

PAPERS

Spam filter based on geographical location of the sender

Tomáš Caha, Martin Kovařík[1]

Spam annoys users and poses a security threat. This article proposes a spam filter based on geographical location of the sender determined by IP geolocation. This filter was implemented as a plugin to the SpamAssassin anti-spam software. The plugin allows to define a penalty score for specific countries sending spam. The proposed filter was tested on a dataset of 1500 e-mails consisting of 1200 spam and 300 legitimate e-mails. The Matthews correlation coefficient of the filter has a value of 0.222. This indicates that the proposed spam filter contributes to the correct spam filtering.

K e y w o r d s: spam filter, IP geolocation, Matthews corrrelation coefficient

1 Introduction

Electronic mail (e-mail) is one of the means of communication of modern times. Users who use e-mail encounter unsolicited messages (spam) daily. These unsolicited messages may contain offers for various products or services, which is annoying for the user. However, through spam, spammers also very often try to obtain sensitive information (payment card details, internet banking login credentials) or financial resources (fraudulent invoices, payment requests, threats) from the recipient.

One of the protections against unsolicited messages are anti-spam filters, which decide whether a message is legitimate (ham) or unsolicited (spam). The topic of spam detection is quite extensive and there are many types of filters. Currently, research is focused on improving current techniques or on proposing completely new methods, which are described below. The aim of these filters is to improve the quality of use of electronic communication via e-mail and to increase user security.

This article presents a method to classify e-mails based on the geographical location of the sender. Each e-mail is transferred through one or more mail servers on its way from the sender to the recipient. The simple mail transfer protocol (SMTP) protocol used by these servers requires each server to include trace fields in the e-mail header. Each server includes its digital footprint containing its Internet Protocol (IP) address in the message. Therefore, it is possible to determine the sender with a public IP address. The trustworthiness of the trace fields and the possibility of spoofing must be considered. Public IP addresses can be geographically located. Geographical location (abbreviated as geolocation) is the process of determining the physical location of a network device based on network parameters (*eg* IP address), [1].

Geolocation of the IP address has a wide range of applications. Advertising platforms can better target advertise ments for products or services in the user's area. Websites can be configured to automatically preselect language or currency depending on where the visitors are coming from. Another location-based content can be weather forecast or local news. Some content distributed through audio or video streaming platforms may be served to licensed territories only and therefore providers must ensure that no digital rights are violated. In these cases, the content may be blocked based on the user's location, [2]. Geolocation is also involved in a cybersecurity and detection systems. Sometimes a two-factor authentication may be enforced when a user attempts to perform some action from an unusual location, [3].

We propose an actual spam filter, including the introduction of the developed plugin to the SpamAssassin anti-spam solution. The plugin is available for download on GitHub, [4]. The results of applying the proposed method are presented and tested on a custom dataset of legitimate (ham) and spam e-mails.

2 Related work

Spammers are constantly adapting to new anti-spam solutions to get around them. Researchers try to improve existing spam filters or find new solutions. Current developments in spam detection techniques focus on artificial intelligence (AI) elements using machine learning (ML) or deep learning (DL) and natural language processing (NLP).

In [5], authors came up with filtering e-mails based on deep learning techniques to make their classifier to discriminate between three classes ham, spam and phishing e-mails. They trained their artificial neural network

[1] Department of Telecommunications, Faculty of Electrical Engineering and Communication, Brno University of Technology, Technická 12, 616 00 Brno, Czech Republic, tomas.caha1@vut.cz,190001@vut.cz

https://doi.org/10.2478/jee-2022-0038, Print (till 2015) ISSN 1335-3632, On-line ISSN 1339-309X

(ANN) model with 4 layers (1 input, 2 hidden, and 1 output) with building time of 8.77 seconds for one dataset with 0.999 validation accuracy.

In [6] an enhancement of naive Bayes classifier with employing longest common subsequence (LCS) was proposed. In some cases, spammers spell some words intentionally to deceive spam filters. Words such as discount or disc0unt instead of discount) can be readable by a human but not a machine. They implemented LCS logic into the classifier to identify the correct form of a word. They increased specificity by 7.6% and sensitivity by 7.54% for the LCS spam filter over the simple Bayesian spam filter.

Designed an online subject-based weighted naive Bayesian (WNB) classifier is in [7]. This type of filter is faster than filters scanning the whole e-mail body. They accomplished an accuracy of 95.7% on Enron-spam datasets.

Geolocation can be performed in active or passive ways. Active methods are based on the measurement of network transmission parameters between reference points and the network device for which we are trying to determine its physical location. The reference points are network devices with a known location. The most measured parameter is latency, the time it takes for a single data segment to be transferred from the source to the destination and back called round-trip time (RTT). Latency is caused by transmission links (speed, utilization, distance), intermediate nodes, and end devices (performance, load). Measurements are usually made from several reference points, because then the approximate geographical position can be estimated from the measured values,[8].

In [9], an active geolocation using RIPE IPmap single-radius engine is described and its accuracy against two commercial geolocation databases was evaluated. Their results showed that this method had city-level accuracy higher for 80.3% of estimated geolocations than both geolocation databases. RIPE IPmap single-radius engine consists of several steps like finding a set of RIPE Atlas probes close to the given IP address, performing ping measurements, converting and filtering the results, calculating the distance between a probe with minimum latency to the given IP address, ranking cities within the calculated distance from the probe and picking up the highest ranked city.

Passive methods are based on searching databases that store information about network devices. These databases can be public or commercial. Geolocation databases usually contain blocks of IP addresses mapped to a geographical location. Searching for this information is then faster and easier compared to active methods, as nothing needs to be measured or computed. Non-commercial databases are usually free of charge but provide a limited amount of information with less accuracy and less up-to-date information. Commercial databases, on the other hand, are fee-based, guarantee availability, and provide more and more accurate information.

In [10], the consistency and coverage of four geolocation databases (namely IP2Location DB11.LITE, MaxMind GeoIP2, MaxMind GeoLite2 and Digital Element NetAcuity) using a dataset of 1.64 M router interface addresses is examined. According to their analysis IP2Location DB11.LITE and Digital Element NetAcuity have almost perfect country and city-level coverage. Freely accessible IP2Location DB11.LITE and MaxMind GeoLite2 databases are comparable with 77.5% to 78.6% country-level geolocation accuracy.

In [3], author investigated a retrospective IP address geolocation. Approach referred to as late location means locating an IP address in the past by a current geolocation database. The median location error increased by a minor value of up to 3 years to the past for IPv4 addresses and up to 1 year to the past for IPv6 addresses. The median location lifetime duration was about 46 days for IPv4 addresses and 24 days for IPv6 addresses.

3 Proposal of a spam filter

We tested an idea of spam filtering based on the geographical location of the sender. First we dissect the email parts that are important for our filter. The following is a description of the proposed filter.

3.1 Header fields of an e-mail

E-mail is an electronic message delivered to the recipient from the sender. It is essentially a text file whose format is defined in RFC 5322, [11]. This document specifies that the message is divided into two parts, namely the header and the body. The body of the message contains the actual text of the message including possible attachments.

The header usually contains information about the sender, the recipient and other additional information. Header fields consist of a header name separated by a colon and followed by a value. An example of the header of an unsolicited e-mail is shown below.

- Listing 1: An example of the header of an unsolicited e-mail

```
Received: from allbygrace.com (dzgualoil.xyz
[5.181.80.128])
   by emai lsmtpd19.ko.seznam.cz (Seznam SMTPD
1.3.136) with ESMTP;
   Thu, 28 Apr 2022 13:18: 07+0200 (CEST)
To: example@seznam.cz
Subject: Pick and PackService/Fulfillment
Message-ID: <4ea213e9236576317a90fefe300ed7b0
@stanleysistrunk.com>
Return-Path: nathanov@rossendental.com
Date: Thu, 28 Apr 2022 13:06:53+0200
From: ''Jenna Franco''<nathanux@rossendental.com>
Reply-To: info@727770.com
```

The most important header fields are briefly described below. More detailed information can be found in RFC 5322 section 3.6.

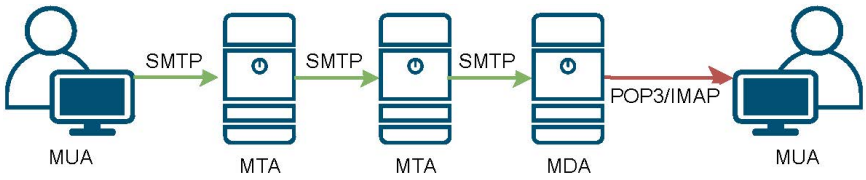

Fig. 1. Simplified representation of the transmission of an e-mail from the sender to the recipient

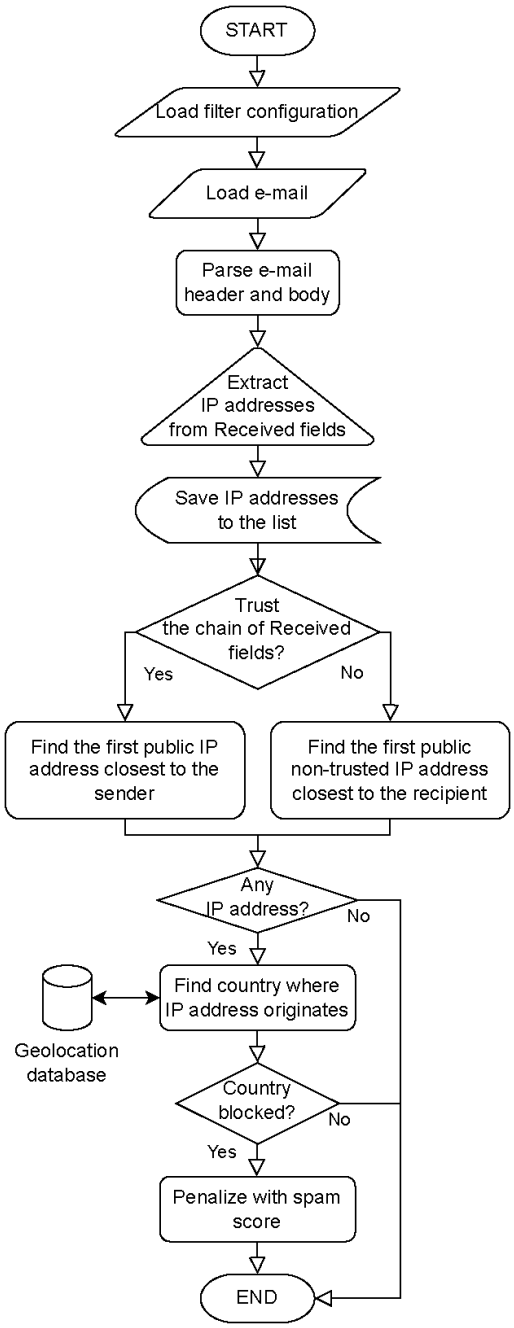

Fig. 2. Diagram of proposed filter

- To: Contains the e-mail address(es) of one or more primary recipients of the message.
- From: specifies the address of the sender of the message.
- Subject: contains a short text string identifying the subject of the message.
- Date: contains the date and time the message was composed and ready to be sent.
- Received: is one of the trace fields that carry information about the path of the message through the SMTP servers from the sender to the recipient.

The format of the content of this field is defined in RFC 5321, [12]. Each SMTP server that receives a given message for processing (forwarding to another server, delivery, other processing) must insert its own *Received*: field at the beginning of the message, in which it inserts its identification and other information. Servers shall not modify *Received*: fields previously inserted by other servers in any way. The *Received*: field typically contains the following information:

- From – server the message was received from, its name and IP address,
- By – which server received the message and inserts this *Received*: field and in parentheses is the software the server uses (for example Postfix),
- With – used protocol (*eg* SMTP, ESMTP),
- For – contains the recipients address,
- Id – identifier,
- Timestamp – date and time usually given in local server time.

Not all these parts need to be in the *Received*: field, or they may take a different form.

3.2 Message transfer

Figure 1 shows a simplified representation of the transmission of an e-mail from the sender to the recipient. The user creates a message in a mail program (Microsoft Outlook, Mozilla Thunderbird) generally referred to as mail user agent (MUA). Once sent, the message is routed through one or more mail servers to the destination server. These servers, generally referred to as mail transfer agent (MTA), are responsible for receiving the message and determining where it should be forwarded. The mail submission agent (MSA), which is usually a part of the MTA, takes over the message from the MUA. When a message is received by an MTA server, the *Received* field is inserted into the message header. From the destination MTA server, the message is forwarded to the mail delivery agent (MDA) server, which ensures delivery of the

message to the user's mailbox, where users then download incoming messages using their MUA mail client. The figure also outlines the commonly used SMTP, POP3 and IMAP protocols.

3.3 Origin of e-mail

When trying to determine the geographical location of the origin of the e-mail, the trace fields in the header of the message can be used. The *Received* fields can be used, because as mentioned above, every mail server (SMTP server) must alter the incoming message by prepending such a field. Listing 2 shows an example of chain of the *Received* fields. This field closest to the recipient is the highest (first), the other one closest to the sender is the lowest (last).

RFC 5321 explicitly states the following: "An Internet mail program MUST NOT change or delete a Received: line that was previously added to the message header section. SMTP servers MUST prepend Received lines to messages; they MUST NOT change the order of existing lines or insert Received lines in any other location", [12]

- Listing 2: Example of a chain of the *Received* fields

```
Received: from mailserver.callsoft.be (mailserver.
callsoft.be [81.95.119.145])
by email-smtpd30.ng.seznam.cz (Seznam SMTPD 1.3.136)
with ESMTP;
Tue,19 Apr 2022 00:52:05 +0200 (CEST)
Received: from localhost (localhost.localdomain
[127.0.0.1])
by mailserver.callsoft.be (Postfix) with ESMTP id
7C70E8E214C;
Tue,19 Apr 2022 00:52:04 +0200 (CEST)
Received: from mailserver.callsoft.be ([127.0.0.1])
by localhost (mailserver.callsoft.be [127.0.0.1])
(amavisd-new, port 10032) with ESMTP id 5T1egK4Fc8K9;
Tue, 19 Apr 2022 00:52:04 +0200 (CEST)
Received: from localhost (localhost.localdomain
[127.0.0.1])
by mailserver.callsoft.be (Postfix) with ESMTP
id EE7888E214D;
Tue, 19 Apr 2022 00:52:03 +0200 (CEST)
Received: from mailserver.callsoft.be ([127.0.0.1])
by localhost (mailserver.callsoft.be [127.0.0.1])
(amavisd-new, port 10026) with ESMTP id Fh1jkGyR8m5q;
Tue, 19 Apr 2022 00:52:03 +0200 (CEST)
Received: from celqjl (unknown [77.40.3.33])
by mailserver.callsoft.be (Postfix) with ESMTPSA
id 204368E214C; Tue, 19 Apr 2022 00:52:02 +0200 (CEST)
```

If the previously cited conditions are met, the geographical location of the sender would be equal to the result of geolocation of IP address 77.40.3.33. Zhuang et al. in 2008 pointed out that a malicious relay mail server can easily tamper the chain of relaying IP addresses. Therefore, the first IP address in the chain (the closest to the sender) may not be trustworthy. They proposed the following method. First, the trusted sender's IP address is the closest to the recipient. Then the chain of *Received* fields is walked through and tested if reported sender's IP address is on the trust list. First unrecognized IP address in the chain is then taken as the e-mail source, [13].

When considering the previously mentioned method, the geographical location of the sender would be equal to the result of geolocation of IP address 81.95.119.145.

3.4 Method steps

The whole filter flow is shown in Fig. 2. Due to the above-mentioned problem with trustworthiness of a chain of the *Received* fields, we believe that it should be possible to let the mail server administrator, decide whether the filter should trust the chain or not. Therefore, our proposal covers both cases.

If the filter is configured not to trust the chain of the *Received* fields, then the origin of the e-mail (the sender's IP address) is the first public non-trusted IP address closest to the recipient because we assume that all private IP addresses closest to the recipient belong to the mail system under recipient's control and therefore, they are safe. In case of trusting the chain of the *Received* fields, then the origin of the e-mail (the sender's IP address) is the first public IP address closest to the sender (private IP addresses must be omitted as they can not be geolocated). If no public IP address in the chain of the *Received* fields was found at all, we assume that the e-mail was transferred internally in the mail system under recipient's control and therefore, this e-mail is a ham.

Finally, the geolocation database is searched for the IP address of the origin of the e-mail. Country where the searched IP address originates is compared to the list of blocked countries. If matched, the e-mail is penalized with given spam score.

3.5 Custom plugin

The proposed filter was also implemented in a plugin named *Geolock*, which was developed for the SpamAssassin anti-spam solution in version 3.4.6. SpamAssassin is a popular and frequently used open-source software, written entirely in Perl and can be extended with custom plugins written in the same language.

The plugin consists of two files *Geolock.pm* and *Geolock.cf*. The latter file, shown in Listing 3, contains configuration parameters such as the penalty score, whether the administrator does not trust the chain of the *Received* fields (1 – do not trust, 0 – trust), and a list of countries from which to evaluate e-mails as a spam. The *Geolock.pm* file contains the actual Perl implementation of the proposed filter, including the connection to the locally stored IP2Location DB5.LITE geolocation database. This geolocation database is free, available for download and provides a good accuracy in country-level IP address location estimation, [10].

To enable the developed plugin, it is necessary to modify the configuration of SpamAssassin, specifically to add one line into to init.pre file according to Listing 4.

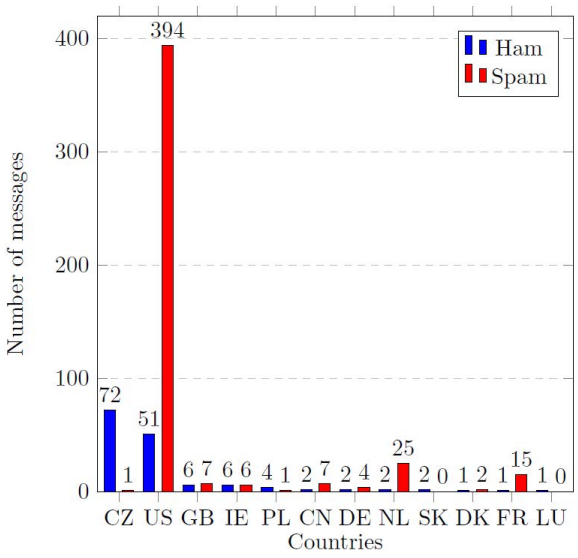

Fig. 3. Distribution of countries sending ham in combination with spam in first part of the dataset

Table 1. Distribution of countries sending spam in first part of the dataset

Country	Spam
US	394
FI	58
NL	25
FR	15
RU	13
NG	12
IN	8
CN, GB	7
IE	6
CA, ID	5
DE, IT	4
VE, VN	3
BG, BR, CH, DK, KR, MY, SG, TG	2
AU, BF, BJ, CZ, EC, HK, JP, KH, LT, LV, PL, PR, SI, TW, ZA	1

- Listing 3: Example of configuration file Geolock.cf

```
header BLOCKED_COUNTRY eval:get_country()
score BLOCKED_COUNTRY 5.0
describe BLOCKED_COUNTRY The country of origin
is blocked
rule 0 IN,CN,AU
add_header all Country_MYTAG_
```

- Listing 4: Loading plugin in init.pre

```
loadplugin Geolock Geolock.pm
```

This site or product includes IP2Location LITE data available from http://www.ip2location.com.

If everything is set up correctly, *X-Spam-Country* field will be added to the header of the message with additional information about where the message originates from, what IP address was geolocated and whether it was penalized.

4 Validation and results

The functionality of the proposed filter was validated using the developed *Geolock* module. A total of 1500 e-mails from the mailboxes of authors and colleagues from the period 2021/2022 were used as a dataset. The dataset consisted of 1200 spam and 300 ham e-mails. The dataset was randomly divided into 2 equal-sized parts, thus 2 groups of 600 spam and 150 ham. The geolocation database used was the IP2Location DB5.LITE from October 2021. The filter was set in trust mode in the chain of received fields. In the results in this chapter, the countries are represented in the form of a two-letter country code according to ISO 3166-1 alpha-2.

To begin with, the first part of the dataset was analyzed. Table 1 shows the distribution of countries sending spam. Most of the spam originated from the United States of America (394), Finland (58) and the Netherlands (25). Many countries were represented by a small number of spam (up to 5 messages). Figure 3 is a graph showing the distribution of countries with legitimate mail in combination with the observed amount of spam according to Tab. 1.

- Listing 5: Spamming countries

```
AU, BF, BG, BJ, BR, CA, CH, EC, FI, HK, ID,

IN, IT, JP, KH, KR, LT, LV, MY, NG, PR, RU,

SG, SI, TG, TW, VE, VN, ZA
```

Based on this observation, all countries from which spam was received while no ham was received were listed as spamming countries. The list of spamming countries can be seen in Listing 5.

The second part of the dataset was tested with the settings of the spamming countries according to Listing 5. The results are written in the confusion matrix in Tab. 2, together with the assessment parameters: specificity, sensitivity, precision, and accuracy.

The specificity (1) says that truly legitimate messages are labeled as legitimate with a good value of the probability. The sensitivity (2) says that true spam is marked as a spam with the given probability, which implies that a large amount of spam is not intercepted. The precision (3) indicates that the probability of correctly marking a message is relatively high. And the accuracy (4) indicates what proportion of messages are classified correctly. As a result, approximately 37.3% of the messages in the second half of the dataset were marked correctly.

Table 2. Confusion matrix of results in second part of the dataset, and assesment coefficients (1) to (5) used to clasify the decision process

		Predicted	
	Total	Spam	Ham
	$T = S + H = 750$	$S = 132$	$H = 618$
Actual	$S = S_{\mathrm{T}} + H_{\mathrm{T}} = 600$	$S_{\mathrm{T}} = 131$	$H_{\mathrm{T}} = 469$
	$H = S_{\mathrm{F}} + H_{\mathrm{F}} = 150$	$S_{\mathrm{F}} = 1$	$H_{\mathrm{F}} = 149$

Specificity: $S_{\mathrm{H}} = H_{\mathrm{F}}/H = 0.933$ (1)

Sensitivity: $S_{\mathrm{S}} = S_{\mathrm{T}}/S = 0.218$ (2)

Precision: $$P = \frac{S_{\mathrm{T}}}{S_{\mathrm{T}} + S_{\mathrm{F}}} = 0.992 \quad (3)$$

Accuracy: $$A = \frac{S_{\mathrm{T}} + H_{\mathrm{F}}}{S + H} = 0.373 \quad (4)$$

$$M_{\mathrm{CC}} = \frac{S_T H_F - S_F H_T}{\sqrt{(SH(S_F + H_T)}} = 0.223 \quad (5)$$

Matthews correlation coefficient M_{CC}, (5) is used to measure the quality of a binary classifier. It is also very useful when classifying groups of data of different sizes. It ranges from -1, being the worst decision, to 1, as an excellent prediction, [14].

Since the M_{CC} value is positive, the proposed filter slightly contributes to the protection against spam. This result is highly dependent on the setting of the spamming countries. The creation of the list of spamming countries should be based on the traffic analysis of a specific mail server or environment and may change over time. Global statistics cannot be used. In our case, the list of spamming countries was built based on the assumption that it is feasible to block countries from which spam is coming, but there is no legitimate communication with them.

Table 3. Comparison of the proposed filter with other techniques

Reference	S_{H}	S_{S}	P	A	M_{CC}
Proposed filter	0.993	0.218	0.992	0.373	0.222
[5], (300 epochs) all features SpamBase	N/A	0.997	0.997	0.995	0.852
[6], LCS filter July data	0.932	0.918	N/A	N/A	N/A
[7], WNB_N Enron-S4 feature set C	0.877	0.984	0.960	0.957	N/A

Table 3 shows evaluation data of binary classifiers which were mentioned above. These values can be calculated only if any filter was applied. It can be seen that the specificity (S_{H}) of the proposed filter is good (0.993) when compared to other techniques (0.932, 0.877). It is thanks to the low false positives (actual ham marked as a spam). On the other hand, the sensitivity (S_{S}) of the proposed filter is low (0.218). Other techniques reached high values of (S_{S}) (0.997, 0.918 and 0.984). The precision (P) is very competitive; 0.992 compared to 0.997 and 0.960. Due to the low sensitivity of the proposed filter, the accuracy (A) and M_{CC} values are low when comparing to the other techniques. Usually, a combination of several techniques is used to filter spam. The high specificity and high precision of this filter suggest that the proposed filter can be a good complement to filters with high sensitivity.

5 Future work

The presented results show that the proposed spam filter contributes to correct spam filtering. Future directions of research will be focused on improving the classification accuracy. Currently, we are analyzing a large dataset of spam and based on the results we would like to select additional e-mail characteristics to be used as the input to the classifier of this filter. We will mainly focus on features such as the number of Received fields in the header, the transfer time from the sender to the recipient, the duration from the message creation to its sending, and the language used in combination with the geographical location of the sender of the message.

6 Conclusion

The article deals with the description of the proposal of a spam filter that classifies messages based on the country they originate from. The problem of determining the sender of an e-mail from the Received fields in the trace fields of the e-mail header was mentioned, considering the possibility of spoofing. The proposed spam filter was implemented as a plugin to the anti-spam software SpamAssassin called Geolock publicly available from GitHub, [4]. The plugin was connected to the IP2Location DB5.LITE geolocation database.

The proposed spam filter was tested on a custom dataset of 1500 e-mails (1200 spam, 300 ham). The dataset was divided into 2 equal sized groups. The first group was analyzed in the perspective of which countries the e-mails were sent from. Based on this investigation, the countries from which messages should be classified as spam were identified. They were the countries where only spam and no ham came from. The spam filter was set to these spamming countries and applied to the second part of the dataset. The results show that the filter had a very good specificity value of 0.993 and precision value of 0.992. However, the sensitivity was quite low with a value of 0.218 and the accuracy representing the proportion of correctly classified messages was 37.3%. The Matthews correlation coefficient indicating the quality of the binary classifier was 0.222, which says that the proposed filter slightly contributed to the protection against spam.

The above-mentioned results are highly dependent on the determination of spamming countries. The list of spamming countries should not be based on general global statistics. An analysis of the traffic of a specific mail server or environment should be made. It must be considered that the list of spamming countries may change overtime. Now, it can be assumed that a suitable list of spamming countries can be constructed in such a way that a spamming country is the one from which only spam and no ham comes. It is not advisable to use this filter as the only filter. Usually, multiple different filters are used, and this one can complement the existing ones.

Future research on this filter will focus on improving the sensitivity and accuracy of the classification. This may be aided using other e-mail characteristics such as the transfer time of the message from the sender to the recipient or the language used in combination with the geographical location of the sender of the message.

References

[1] I. Poese, S. Uhlig, M. A. Kaafar, B. Donnet, and B. Gueye, "Ip geolocation databases", *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 2, pp. 5356, 2011-04-15. https://dl.acm.org/doi/10.1145/1971162., 1971171.

[2] J. Taylor, J. Devlin, and K. Curran, "Bringing location to IP addresses with IP geolocation,", *Journal of Emerging Technologies in Web Intelligence*, vol. 4, 08 2012.

[3] D. Komosny, "Retrospective ip address geolocation for geography-aware internet services", *Sensors*, vol. 21, no. 15, https://www.mdpi.com/1424-8220/21/15/4975, 2021.

[4] M. Kovařík, "GitHub MartinKovarik/Geolock: Plugin for SpamAssassin for blocking e-mails based on the geolocation of the sender using a IP2Location database", https://github.com/MartinKovarik/Geolock, 2022.

[5] S. Magdy, Y. Abouelseoud, and M. Mikhail, "Efficient spam and phishing emails filtering based on deep learning", *Computer Networks*, vol. 206, pp. 108826, https://www.scie ncedirect.com/science/article/pii/S1389128622000469, 2022.

[6] K. Roy, S. Keshari, and S. Giri, "Enhanced Bayesian spam filter technique employing lcs", *International Conference on Computer, Electrical Communication Engineering (ICCECE)*, pp. 16, 2016.

[7] C.-N. Lee, Y.-R. Chen, and W.-G. Tzeng, "An online subject-based spam filter using natural language features", *IEEE Conference on Dependable and Secure Computing*, pp. 479487, 2017.

[8] I. Youn, B. L. Mark, and D. Richards, "Statistical geolocation of internet hosts", *Proceedings of 18th International Conference on Computer Communications and Networks*, IEEE, pp. 16. http://ieeexplore.ieee.org/document/5235373/, 2009.

[9] B. Du, M. Candela, B. Huffaker, A. C. Snoeren, and K. Claffy, "Ripe ipmap active geolocation: Mechanism and performance evaluation", *SIGCOMM Comput. Commun. Rev*, vol. 50, no. 2, pp. 310, May, https://doi.org/10.1145/3402413.3402415, 2020.

[10] M. Gharaibeh, A. Shah, B. Huffaker, H. Zhang, R. Ensafi, and C. Papadopoulos, "A look at router geolocation in public and commercial databases", *Proceedings of the Internet Measurement Conference*, ser. IMC 17. New York, NY, USA: Association for Computing Machinery, pp. 463469. https://doi.org/10.1145/3131365.3131380, 2017.

[11] P. W. Resnick, "Internet message format", Internet Requests for Comments, RFC Editor, RFC 5322, October, http://www.rfc-editor.org /rfc/rfc5322.txt, 2008.

[12] J. Klensin, "Simple mail transfer protocol", Internet Requests for Comments, RFC Editor, RFC 5321, October, http://www.rfc-editor.org/rfc/rfc5321.txt.http://www.rfc-editor.org/rfc/rfc5321.txt, 2008.

[13] L. Zhuang, J. Dunagan, D. R. Simon, H. J. Wang, and J. D. Tygar, "Characterizing botnets from email spam records", *Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats*, ser. LEET08. USA: USENIX Association, https://www.microsoft.com/en-us/research/wp-content/uploads/2016/12/zhuang.pdf, 2008.

[14] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation", *BMC Genomics*, vol. 21, no. 1, https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-6413-7, 2020.

Received 30 June 2022

Tomáš Caha was born in Brno, Czech Republic, in 1993. He received his MSc degree in Communications and Informatics from the Faculty of Electrical engineering and Communication at Brno University of Technology, Czech Republic, in 2019. His PhD research is focused on methods of detecting suspicious communication content. He leads practices of a course dealing with IP networks and object-oriented programming. He also specializes in developing business information systems.

Martin Kovařík was born in Opočno, Czech Republic, in 1996. He received his master degree in Information Security from the Faculty of Electrical Engineering and Communication at Brno University of Technology, Czech Republic, in 2022. He works in a cybersecurity team as an incident handler and a cybersecurity specialist.