electronics

MDPI

*Article*

# Multi-Language Spam/Phishing Classification by Email Body Text: Toward Automated Security Incident Investigation

**Justinas Rastenis [1,*], Simona Ramanauskaitė [2], Ivan Suzdalev [3], Kornelija Tunaitytė [3], Justinas Janulevičius [1] and Antanas Čenys [1]**

[1] Department of Information Systems, Vilnius Gediminas Technical University, Sauletekio al. 11, LT-10223 Vilnius, Lithuania; justinas.janulevicius@vilniustech.lt (J.J.); antanas.cenys@vilniustech.lt (A.Č.)
[2] Department of Information Technology, Vilnius Gediminas Technical University, Sauletekio al. 11, LT-10223 Vilnius, Lithuania; simona.ramanauskaite@vilniustech.lt
[3] Department of Aeronautical Engineering, Vilnius Gediminas Technical University, Sauletekio al. 11, LT-10223 Vilnius, Lithuania; ivan.suzdalev@vilniustech.lt (I.S.); kornelija.tunaityte@stud.vgtu.lt (K.T.)
* Correspondence: justinas.rastenis@vilniustech.lt; Tel.: +370-525-12-333

**Abstract:** Spamming and phishing are two types of emailing that are annoying and unwanted, differing by the potential threat and impact to the user. Automated classification of these categories can increase the users' awareness as well as to be used for incident investigation prioritization or automated fact gathering. However, currently there are no scientific papers focusing on email classification concerning these two categories of spam and phishing emails. Therefore this paper presents a solution, based on email message body text automated classification into spam and phishing emails. We apply the proposed solution for email classification, written in three languages: English, Russian, and Lithuanian. As most public email datasets almost exclusively collect English emails, we investigate the suitability of automated dataset translation to adapt it to email classification, written in other languages. Experiments on public dataset usage limitations for a specific organization are executed in this paper to evaluate the need of dataset updates for more accurate classification results.

**Keywords:** spam; phishing; classification; augmented dataset; multi-language emails

check for updates

**Citation:** Rastenis, J.; Ramanauskaitė, S.; Suzdalev, I.; Tunaitytė, K.; Janulevičius, J.; Čenys, A. Multi-Language Spam/Phishing Classification by Email Body Text: Toward Automated Security Incident Investigation. *Electronics* **2021**, *10*, 668. https://doi.org/10.3390/electronics10060668

Academic Editor: Krzysztof Szczypiorski

Received: 15 January 2021
Accepted: 7 March 2021
Published: 12 March 2021

## 1. Introduction

Despite new communication systems and solutions being constantly introduced to the market, email remains in leading positions for both business and personal use. This popularity attracts the attention of persons with malicious intentions—spam and phishing email attacks are one of the most popular cyber-security attacks: in the 3rd quarter of 2020 nearly 50% of email traffic was spam [1]; 98% of cyber-attacks rely on social engineering [2] which is mostly executed by sending phishing emails [3].

Email filtering systems have been improving continuously to follow malicious, unwanted content development to protect the end-users. However, existing solutions are focusing on spam and phishing email filtering out while further analysis and email labeling are not fully developed. Therefore, email-based attacks are either analyzed manually or not investigated at all.

The analysis of cyber-attacks is a must for detecting the attacker and preventing their further malicious activities. The digital information security forensics is a time- and resource-consuming process, therefore automation should be used as much as possible to reduce the investigation time as well as to increase its accuracy [4,5]. One of the first steps in the forensics is classification of obtained data and its prioritization. Taking into account the huge number of unwanted emails, the automated classification of malicious emails would work as initial prioritization of investigating incidents and would work as the initial phase for automated or semi-automated security incident investigation. The prioritization is important as the purpose of spam and phishing attacks are different—spam emails are oriented towards dissemination of advertising, while phishing attacks aim at victims' personal data collecting and its usage for other cyber-attacks. Therefore phishing emails should be investigated as fast as possible, with higher attention to them than spam emails. The automated classification between spam and phishing email would allow appropriate resource allocation.

This paper aims to automate the identification of phishing emails in spam/phishing mixed different language email flow. As a consequence, this would simplify email-based security attack investigation and would lead to a higher degree automation in the forensics process. To achieve this goal several research questions are raised: (i) are existing English language spam/phishing email datasets suitable for spam/phishing email classification in other languages? and (ii) do spam/phishing email text patterns change relating to a specific region and do they have to be updated to achieve a higher classification accuracy?

The further structure of the paper is organized as follows. Related work chapter summarizes existing research in the field of spam or phishing email automated classification as well as datasets, that are usually used to train spam or phishing email detection systems. Based on the existing solutions new research for spam and phishing email classification is presented along with the datasets. The paper does not propose a new classification method; however, it presents research for spam/phishing email following the steps comprising a common classification workflow (data preparation, text augmentation, text classification), applied for solving this specific problem. The performance of the proposed solution is evaluated and experiments on automated email dataset translation as well as the updates needed are investigated. The paper is summarized with conclusions and future work.

## 2. Related Work

Spam is undesired electronic information spread aiming to cause psychological and monetary harm to the victim [6]. While it can be spread within different channels, a spam email contains an advertisement or irrelevant text, sent by spammers having no relationship with the recipient [7]. While different definitions of spam exist it is mostly related to undesired commercial email, and therefore the end user is unsatisfied by receiving undesired content.

Meanwhile, phishing emails seek to mimic legitimate emails and influence the user to execute some intended actions and reveal their personal information. Phishing attacks are classified as social engineering attacks, where the attacker tries to affect the victim from making rational choices and force the victim to make emotional choices instead [8]. Therefore, phishing attacks are potentially more harmful in comparison to spam mails.

To classify the email automatically, some basic steps are executed: email preprocessing and email classification (with its performance evaluation).

### *2.1. Email Preprocessing*

An email has some specific properties which can be used for its classification to spam, phishing, legitimate email (ham), or any other category. An email can be presented in different file formats, therefore the property extraction should be prepared. However, for email classification, some additional processing might be used to obtain some specific features. For example, Ayman El Aassal et al. [9] divide phishing email-related features into two main categories: email features and website features. Email features are related to the data and metadata of the email and can be categorized into header, body, and attachment data. Meanwhile, website features are related to data, which can be gathered from the email body and links in it. Website features are based on the link and the websites the link points to. While most solutions [10–12] rely on the data which can be directly gathered from the email (the link uniform resource locator (URL) presented as internet protocol (IP), not domain name address; the number of different domains in the links; etc.), some solutions [9] go even further and analyze the website itself (the content of the website; script code; etc.) or use some additional tools to validate the URL [13].

To reduce the classification complexity, the number of extracted features is limited and expressed as numerical or binary values [14]. Therefore, different feature selection techniques are used [15,16] to obtain the most important features only and to eliminate non-significant ones. For example, Jose R. Mendez et al. [17] extracts the topic of the email and for spam email identification uses topics rather than the full bag of words of the email text. Sami Smadi et al. [18] uses 22 features, which are calculated, estimated based on a number or existence of some specific patterns; however, term meaning in the email body is not analyzed at all. Meanwhile, Andronicus A. Akinyelu and Aderemi O. Adewumi [19] define 7 features, which are based on the existence or number of some inspected elements in the email and add 2 features based on the existence of specific terms, words in the email body (one to define the direction to click some link; another related to action, which should be done after clicking the link). The proportion of email body content and other features depends on the author. For example, Saeed Abu-Nimeh et al. [20] and Devottam Gaurav et al. [21] use only email body features and by using text-mining their solution gathers the most frequent terms in the email body. To extract the most frequent terms, all hypertext markup language (HTML) code and unwanted terms (stop words), symbols are removed from the email body. Then the terms are processed to get the standard form (stemming). For later analysis, the frequencies or proportion of the specific terms are used as features.

Text analysis is very popular in the latest methods for spam and phishing classification and might include some additional text preprocessing to obtain more accurate classification results. For example, Ayman El Aassal et al. [22] takes into account the data from different datasets that might be associated with the email category, therefore they eliminated as much content as possible (organizations' or universities' names, recipients' names, domain names, signatures, etc.), which could associate it to the dataset. Another solution in email classification is the hierarchical classification [23,24] where, for example, first of all the email body is classified into some semantic categories and based on it the second layer identifies the email category itself.

### 2.2. Email Classification Solutions

Email classification can be implemented as a rule-based [25] system, however, it requires continuous support and updating. Therefore, hybrid [26] or machine learning [27] solutions take over where automated rather than manual rule, decision making logic updates are made. The machine learning solutions allow supervised learning when the model for email classification is designed based on the provided dataset.

In the field of spam, phishing, and ham email classification, the main classification methods are support vector machine (SVM), random forest (RF), decision tree (DT), naïve Bayes (NB), linear regression (LR), k-nearest neighbors (kNN) and other more specific solutions. The summary of classification method usage is presented in Table 1.

As seen, all email classification solutions are focused on the classification of legitimate, ham emails and unwanted, malicious (spam, phishing, or both) emails. The results of presented email classification solutions are high (F-score is 87 or more and even reaches 99.95), however no separation between spam and phishing is analyzed in scientific papers.

The lack of spam and phishing email separation is noticed in email datasets as well. While the Enron dataset is dedicated to legitimate ham emails, the University of California, Irvine (UCI) Machine Learning Repository has a dataset for spam emails, the Nazario dataset stores phishing emails, the SpamAssassin dataset has both spam and ham emails. Those two categories are separated in the SpamAssassin dataset, however, phishing emails are included inside of the spam emails. In most cases, some additional, personal email datasets are used to add variety and an ability to test the proposed solution with real situations, specific to some organization.

**Table 1.** Summary of recent papers on machine learning email classification solutions.

| Paper | Classification Categories | Classification Method | Dataset | MAX F-Score |
|---|---|---|---|---|
| El Aassal et al. [9] | phishing, ham | SVM, RF, DT, NB, LR, kNN, other | Enron [28],<br>SpamAssassin [29],<br>Nazario [30] | 99.95 |
| Li et al. [31] | phishing, ham | DT, NB, kNN | SpamAssassin, Nazario | 97.30 |
| Verma et al. [32,33] | phishing, ham | SVM, RF, DT, NB, LR, kNN | SpamAssassin, Nazario | 99.00 |
| Sonowal et al. [6] | phishing, ham | RF, other | Nazario | 97.78 |
| Gangavarapu et al. [34] | spam + phishing, ham | SVM, RF, NB, other | SpamAssassin, Nazario | 99.40 |
| Gaurav et al. [21] | spam, ham | RF, DT, NB | Enron, UCI Machine Learning Repository [35] | 87.00 |
| Ablel-Rheem et al. [36] | spam, ham | DT, NB, other | UCI Machine Learning Repository | 94.40 |
| Saidani et al. [24] | spam, ham | SVM, RF, DT, NB, kNN, other | Enron | 98.90 |
| Jáñez-Martino et al. [37] | spam, ham | SVM, NB, LR | SpamAssassin | 95.40 |
| Zamir et al. [23] | spam, ham | SVM, RF, DT, other | SpamAssassin | 97.20 |

Support vector machine (SVM), random forest (RF), decision tree (DT), naïve Bayes (NB), linear regression (LR), k-nearest neighbors (kNN).

## 3. Research on Text-Based Spam/Phishing Email Classification Solution

While methods for malicious email detection from legitimate emails exist and achieves high accuracy, there are no solutions to classify spam and phishing emails within the malicious email flow. Therefore, in this paper we propose a solution, dedicated to classifying unwanted emails to spam and phishing email categories. The proposed email classification solution incorporates existing classification solutions and is adapted to classify emails of different languages. In Lithuania, the largest portion of emails is written in Lithuanian, English and Russian, therefore the solution will be oriented to these three languages in this paper.

### *3.1. Email Dataset Preparation*

Both spam and phishing emails are undesired for the recipient and sent using very similar techniques. Therefore, the biggest difference between spam and phishing emails is their content. Therefore for spam and phishing email classification, we use email message body only.

We use supervised learning solutions and, therefore, a dataset of labeled spam and phishing emails is needed. The dataset was constructed by integrating three different datasets: (i) the Nazario dataset for list of phishing emails, (ii) the SpamAssassin dataset for a list of spam emails and (iii) an individual spam and phishing email dataset from Vilnius Gediminas Technical University (VilniusTech).

The Nazario dataset was used as it is to represent phishing email examples. Meanwhile, the SpamAssassin dataset includes spam and ham emails. We used the spam emails only; however, after inspecting them some phishing emails were found within the spam emails. Therefore, the dataset was relabeled to indicate spam and phishing emails.

VilniusTech dataset was collected and labeled by VilniusTech information technology specialists and includes emails from the period of 2018–2020.

All datasets were read by getting an email message body only (programming code to extract emails message body were written for each dataset). The emails additionally were preprocessed. Cleanup of email message body text was executed where all HTML, CSS (cascading style sheets), JavaScript code, special symbols were eliminated, leaving unformatted text only. As some emails contained personal information, it was eliminated too. This was done to avoid email message association to a specific dataset—the Nazario dataset has very common reference jose@monkey.org, in the VilniusTech dataset Vilnius Gediminas Technical University is mentioned etc. Therefore, all personal information (recipient's name, email address, organizations name) was replaced with keywords (NAME, EMAIL, ORGANIZATION), and dates (year) were removed from the text. This was done semi-automatically—part of the personal information was removed by using regex expressions and then all emails were revised manually.

Formatting and personal information removal revealed duplication of emails. Multiple instances of the same email templates were noticed and, therefore, unique messages were selected for the dataset while all duplicated versions were removed.

The individual VilniusTech dataset included emails written in different languages. The most popular languages (English, Lithuanian and Russian) were left while very rare cases of different languages (Latvian, German, Spanish, France, etc.) were eliminated from the dataset. Meanwhile, emails from the Nazario and SpamAssassin datasets were in English only. Therefore this dataset was translated (by using automated Google Translate service, integrated via application programming interface (API) into Python code, developed for preparation of the dataset) into Russian and Lithuanian languages. The keywords representing the recipient's personal information were not translated and left as keywords.

During the email filtering of unpopular languages and automated translation, each record in the dataset was assigned a new property—language. This property will not be used for email classification (in this paper), however will be used to form different test cases for the research.

Records from different datasets were combined into one dataset. The number of phishing emails in the combined dataset was much lower in comparison to spam emails (see Table 2). Therefore, random emails were selected from each category to obtain the same number of spam and phishing emails (see Table 2). This reduced the dataset from 3601 record to 1400, where 700 spam and 700 phishing emails are labeled.

**Table 2.** Summary of prepared spam and phishing dataset.

| Initial Dataset | Language | Before Balancing | | | After Balancing | | |
|---|---|---|---|---|---|---|---|
| | | Spam Emails | Phishing Emails | Total | Spam Emails | Phishing Emails | Total |
| SpamAssassin + Nazario | English | 692 | 182 | 874 | 150 | 150 | 300 |
| | Lithuanian (translated) | 692 | 182 | 874 | 150 | 150 | 300 |
| | Russian (translated) | 692 | 182 | 874 | 150 | 150 | 300 |
| VilniusTech | English | 559 | 205 | 864 | 200 | 200 | 400 |
| | Lithuanian | 40 | 38 | 78 | 35 | 35 | 70 |
| | Russian | 18 | 19 | 37 | 15 | 15 | 30 |
| Total | | 2693 | 808 | 3601 | 700 | 700 | 1400 |

For text-based classification all message texts were tokenized as separate terms (TF-IDF—term frequency-inverse document frequency) and pruning was applied. We removed very common (over 95% occurrence) and very infrequent terms (below 3% occurrence). The limit of attributes is not applied and reaches about 31,000 attributes (attribute presents relative, rather than the absolute occurrence of the term). The number of attributes was relatively large, however it presented words from three different languages. Taking into account the complexity and variety of word forms in Lithuanian language, the number of attributes was adequate but can be optimized in future.

### 3.2. Research Methodology and Results

As the dataset includes 700 spam and 700 phishing emails we do not use deep neural networks and concentrate on the usage of the most used classification methods. The research is divided into three main phases (see Figure 1): Figure 1a method selection Figure 1b multi-language-experiment Figure 1c concept-drift-experiment.

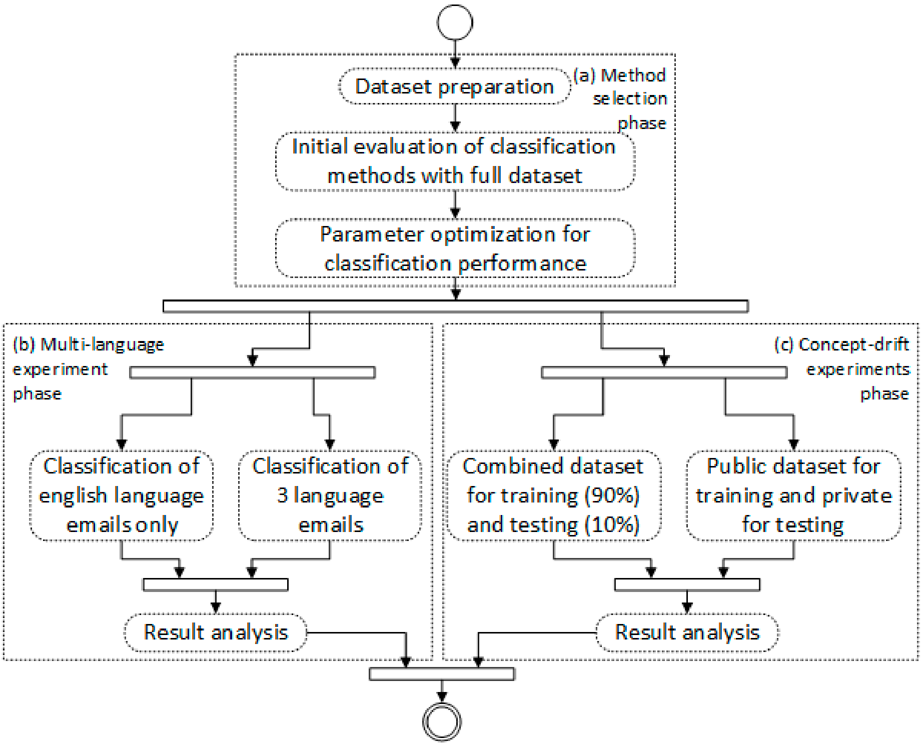
Dataset preparation
(a) Method selection phase
Initial evaluation of classification methods with full dataset
Parameter optimization for classification performance
(b) Multi-language experiment phase
(c) Concept-drift experiments phase
Classification of english language emails only
Classification of 3 language emails
Combined dataset for training (90%) and testing (10%)
Public dataset for training and private for testing
Result analysis
Result analysis

**Figure 1.** Workflow diagram of the research. (**a**) Method selection phase, (**b**) Multi-language experiment phase, (**c**) Concept-drift experiments phase.

In the first stage naïve Bayes, generalized linear model, fast large margin, decision tree, random forest, gradient boosted trees and support vector Machines methods were selected for the automatic identification of spam/phishing emails. Default settings and the full (balance of 1400 records) dataset was used in this step. The purpose of this step was to obtain the tendencies of classification performance and to select the methods we will be working on further.

For experiment execution, a RapidMiner tool was used to assure equal conditions for all methods (its standard implementation with possible settings). It was running on a 64-bit Windows 10 operating system on HP ProBook × 360 440 G1 Notebook PC with Intel core i3 processor and 8GM of RAM.

The results revealed (see Table 3), that 4 out of 7 analyzed solutions are not suitable to solve this problem as the accuracy does not exceed 60%. While ROC (receiver operating characteristic) curves (see Figure 2) and AUC (area under curve) values show naïve Bayes and decision tree methods are close to random solutions and the results obtained give no value in this situation.

**Table 3.** Classification methods performance in the initial experiment to classify spam and phishing emails.

| Methods | Accuracy, % | Precision, % | Recall, % | F Score, % | AUC, % | Training Time (1000 Rows), s | Scoring Time (1000 Rows), s |
|---|---|---|---|---|---|---|---|
| Naïve Bayes | 59.8 | 93.0 | 21.5 | 34.7 | 67.7 | 0.269 | 14 |
| Generalized Linear Model | 82.8 | 79.6 | 88.7 | 83.9 | 88.9 | 0.831 | 10 |
| Fast Large Margin | 83.2 | 79.1 | 90.7 | 84.4 | 92.5 | 0.157 | 15 |
| Decision Tree | 54.0 | 100.0 | 6.1 | 11.5 | 52.9 | 0.419 | 9 |
| Random Forest | 57.2 | 100.0 | 12.7 | 22.4 | 86.4 | 5.000 | 28 |
| Gradient Boost Trees | 57.0 | 93.0 | 13.7 | 23.5 | 98.2 | 15.000 | 9 |
| Support Vector Machine | 84.0 | 78.0 | 95.2 | 85.6 | 91.8 | 2.000 | 19 |

Area under curve (AUC).

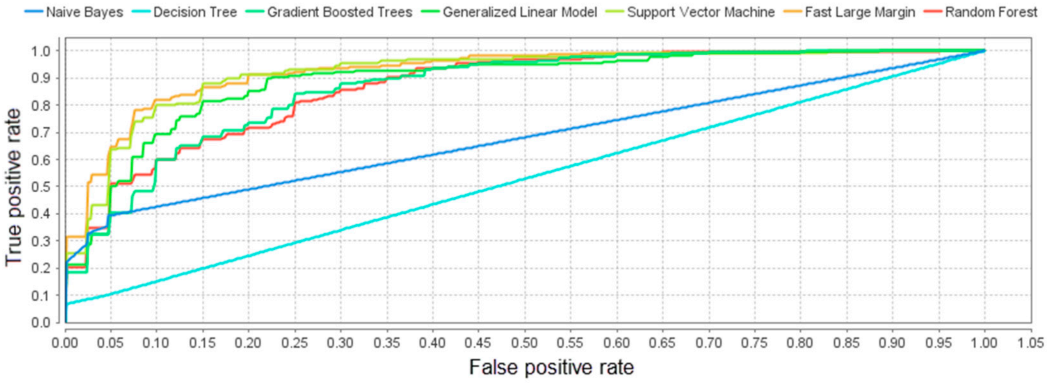

**Figure 2.** ROC (receiver operating characteristic) curves of different classification methods, used for initial email message classification to spam and phishing.

The support vector machine has the highest accuracy (84.0% ± 1.6%), however is one of the slowest solutions (for 1000 rows it takes 2s for training and 19s for scoring).

The next step of suitable classification method selection phase, a search for the most suitable parameters to increase the spam and phishing email classification performance, was executed with the generalized linear model, fast large margin and support vector machine. Different methods were used to analyze optimal parameters values—grid search, genetic algorithms [38], manual experiments. The best parameters were selected manually from the results obtained.

In this step the best accuracy was achieved with the fast large margin method (which was second in the initial experiment), using L2 SVM Dual solver, cost parameter $C = 1$, tolerance of the termination criteria $\varepsilon = 0.01$, identical class weights, and usage of bias. The cross-validation was executed with automatic sampling type and 10 fold as in the initial experiment. With these parameters, the accuracy increased to 90.07% ± 3.17%, and the confusion matrix of this classificatory is presented in Table 4.

**Table 4.** Confusion matrix and class prediction as well as class recall values of adjusted parameters for the fast large margin method.

| | **True Spam** | **True Phishing** | **Class Prediction** |
|---|---|---|---|
| Predicted Spam | 662 | 101 | 86.76% |
| Predicted Phishing | 38 | 599 | 94.03% |
| Class recall | 94.57% | 85.57% | |

The obtained configuration is used in parallel (independently) further in multi-language experiments (see Figure 1b,c).

In a multi-language experiment we investigated if the automated dataset translation was suitable for dataset augmentation and application for different language emails. This experiment was oriented to emails of three different languages, where part of the dataset was translated by Google Translate. If we applied the same model to the English language only, the accuracy was 89.2% ± 2.14%. This was the same result as in experiments with three languages and showed that the automated Google translation from English to Lithuanian and Russian languages was a suitable dataset augmentation method to adapt the dataset for spam/phishing email classification for different language emails.

The results similarity can be explained by two facts: (a) in most cases spam and phishing email templates are translated from the English language to other languages and in some cases, it is done with automated translation tools as well, therefore the augmented data in the dataset is similar to the data which would be sent in practice; (b) we use TF-IDF text vectorization where accuracies of separate terms are analyzed, not n-grams and, therefore the influence of translation quality is not as important.

A concept-drift experiment was concentrated on evaluating the need for dataset update. In this experiment, one dataset was used for training and another for testing. We took records from SpamAssassin and Narazio as the training set and VilniusTech email records as the testing set. In this situation, the accuracy decreased by more than 10%—if emails of only the English language were included, the accuracy was 74.94%, while if the augmented/translated SpamAssassin and Nazario datasets were used and tested with all records from VilniusTech dataset, the accuracy was 77.00%.

This shows that there are differences between the datasets which might be influenced by time, region or organization profile (the VilniusTech dataset is constructed from emails, obtained from university email boxes). The accuracy increase by using the augmented dataset can be explained by the increased number of records in the training dataset—there are 300 English language emails in the SpamAssassin and Nazario-based dataset while adding translations of two additional languages increases this to 900 emails.

## 4. Conclusions and Future Work

Analysis of the existing spam and phishing email classification solutions has revealed that there are multiple papers on this topic; however, all of them are focused on legitimate and malicious (spam and/or phishing) email separation from one email flow. There are no papers on automated spam and phishing email classification solutions. Spam and phishing emails sometimes are difficult to separate and the SpamAssassin dataset includes phishing emails as spam records. However, classification of spam and phishing emails would be beneficial as could be used to inform the user about the danger level of unwanted email as well as to assign priorities to the unwanted emails to investigate the cases.

Existing publicly available spam and phishing email datasets are English language only. This complicates its usage for email classification, which are written in different languages. The proposed solution with automated translation for dataset augmentation, adaptation for other languages prove the classification results do not decrease because of the automated translation—for English-only text, the accuracy was 90.07% ± 3.17% while for multi-language texts (English, Russian and Lithuanian) it was 89.2% ± 2.14%.

By training the spam and phishing classification model with the SpamAssassin and Nazario datasets and testing the model with the VilniusTech collected set of spam/phishing emails, the classification accuracy decreased more than 10% in comparison to a mixed dataset, used both for training and testing. This proves that the dataset should be updated, supplemented with data from the organization to obtain more accurate classification results.

For further directions, a deeper spam/phishing email classification performance analysis could be executed to increase the performance by adapting feature optimization (including header and formatting related features, feature number minimization or application of multi-level classification approaches), and deep-learning solution suitability for this task evaluation.

From the automated security incident investigation perspective, the emails could be classified based not only on spam/phishing classification but on potential thread recognition possibility, prevalence in the organization, and other features as well.

**Author Contributions:** Conceptualization, J.R. and S.R.; methodology, S.R.; software, J.R. and S.R.; validation, J.R., S.R. and K.T.; formal analysis, S.R.; investigation, J.R. and I.S.; resources, S.R.; data curation, J.R.; writing—original draft preparation, J.R and S.R.; writing—review and editing, J.R and S.R.; visualization, J.R. and S.R.; supervision, J.J.; project administration, A.Č. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** Dataset used in this experiment is available. It contains original SpamAssassin and Nazario records (dataset labeled "1"), its translation to Russian and Lithuanian languages (dataset labeled "2") and individual dataset, collected and labeled by VilniusTech information technology specialists during the period 2018–2020 (dataset labeled "3").

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Spam and Phishing in Q3 2020. Available online: https://securelist.com/spam-and-phishing-in-q3-2020/99325/ (accessed on 15 November 2020).
2. 2020 Cyber Security Statistics. Available online: https://purplesec.us/resources/cyber-security-statistics/ (accessed on 15 November 2020).
3. Social Engineering & Email Phishing–The 21st Century's #1 Attack? Available online: https://www.wizlynxgroup.com/news/2020/08/27/social-engineering-email-phishing-21st-century-n1-cyber-attack/ (accessed on 15 November 2020).
4. Carmona-Cejudo, J.M.; Baena-García, M.; del Campo-Avila, J.; Morales-Bueno, R. Feature extraction for multi-label learning in the domain of email classification. In Proceedings of the 2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), Paris, France, 11–15 April 2011; pp. 30–36.
5. Goel, S.; Williams, K.; Dincelli, E. Got phished? Internet security and human vulnerability. *J. Assoc. Inf. Syst.* **2017**, *18*, 22–44. [CrossRef]
6. Aassal, A.E.; Moraes, L.; Baki, S.; Das, A.; Verma, R. Anti-phishing pilot at ACM IWSPA 2018: Evaluating performance with new metrics for unbalanced datasets. In Proceedings of the IWSPA-AP Anti Phishing Shared Task Pilot 4th ACM IWSPA, Tempe, Arizona, 21 March 2018; pp. 2–10.
7. El Aassal, A.; Baki, S.; Das, A.; Verma, R.M. An In-Depth Benchmarking and Evaluation of Phishing Detection Research for Security Needs. *IEEE Access* **2020**, *8*, 22170–22192. [CrossRef]
8. Abu-Nimeh, S.; Nappa, D.; Wang, X.; Nair, S. A comparison of machine learning techniques for phishing detection. In Proceedings of the Anti-phishing Working Groups 2nd Annual Ecrime Researchers Summit, Pittsburgh, PA, USA, 4–5 October 2007; pp. 60–69.
9. L'Huillier, G.; Weber, R.; Figueroa, N. Online phishing classification using adversarial data mining and signaling games. In Proceedings of the ACM SIGKDD Workshop on CyberSecurity and Intelligence Informatics, Paris, France, 28 June–1 July 2009; pp. 33–42.
10. Peng, T.; Harris, I.; Sawa, Y. Detecting phishing attacks using natural language processing and machine learning. In Proceedings of the 2018 IEEE 12th international conference on semantic computing (icsc), Laguna Hills, CA, USA, 31 January–2 February 2018; IEEE: New York, NY, USA, 2018; pp. 300–301.
11. Weinberger, K.; Dasgupta, A.; Langford, J.; Smola, A.; Attenberg, J. Feature hashing for large scale multitask learning. In Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, QC, Canada, 14-18 June 2009; pp. 11–1120.
12. Zareapoor, M.; Seeja, K.R. Feature extraction or feature selection for text classification: A case study on phishing email detection. *Int. J. Inf. Eng. Electron. Bus.* **2015**, *7*, 60. [CrossRef]
13. Smadi, S.; Aslam, N.; Zhang, L. Detection of online phishing email using dynamic evolving neural network based on reinforcement learning. *Decis. Support Syst.* **2018**, *107*, 88–102. [CrossRef]
14. Toolan, F.; Carthy, J. Feature selection for spam and phishing detection. In Proceedings of the 2010 eCrime Researchers Summit, Dallas, TX, USA, 18–20 October 2010; IEEE: New York, NY, USA, 2010; pp. 1–12.
15. Verma, R.M.; Zeng, V.; Faridi, H. Data Quality for Security Challenges: Case Studies of Phishing, Malware and Intrusion Detection Datasets. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, London, UK, 11–15 November 2019; pp. 2605–2607.
16. Smadi, S.; Aslam, N.; Zhang, L.; Alasem, R.; Hossain, M.A. Detection of phishing emails using data mining algorithms. In Proceedings of the 2015 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), Kathmandu, Nepal, 15–17 December 2015; IEEE: New York, NY, USA, 2015; pp. 1–8.
17. Akinyelu, A.A.; Adewumi, A.O. Classification of phishing email using random forest machine learning technique. *J. Appl. Math.* **2014**, *2014*. [CrossRef]
18. Gangavarapu, T.; Jaidhar, C.D.; Chanduka, B. Applicability of machine learning in spam and phishing email filtering: Review and approaches. *Artif. Intell. Rev.* **2020**, *53*, 5019–5081. [CrossRef]
19. Li, X.; Zhang, D.; Wu, B. Detection method of phishing email based on persuasion principle. In Proceedings of the 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chongqing, China, 12–14 June 2020; Volume 1, pp. 571–574.
20. Verma, P.; Goyal, A.; Gigras, Y. Email phishing: Text classification using natural language processing. *Comput. Sci. Inf. Technol.* **2020**, *1*, 1–12. [CrossRef]
21. Sonowal, G. Phishing Email Detection Based on Binary Search Feature Selection. *SN Comput. Sci.* **2020**, *1*. [CrossRef] [PubMed]
22. Ablel-Rheem, D.M.; Ibrahim, A.O.; Kasim, S.; Almazroi, A.A.; Ismail, M.A. Hybrid Feature Selection and Ensemble Learning Method for Spam Email Classification. *Int. J.* **2020**, *9*, 217–223. [CrossRef]
23. Zamir, A.; Khan, H.U.; Mehmood, W.; Iqbal, T.; Akram, A.U. A feature-centric spam email detection model using diverse supervised machine learning algorithms. *Electron. Libr.* **2020**, *38*, 633–657. [CrossRef]
24. Gaurav, D.; Tiwari, S.M.; Goyal, A.; Gandhi, N.; Abraham, A. Machine intelligence-based algorithms for spam filtering on document labeling. *Soft Comput.* **2020**, *24*, 9625–9638. [CrossRef]
25. Saidani, N.; Adi, K.; Allili, M.S. A Semantic-Based Classification Approach for an Enhanced Spam Detection. *Comput. Secur.* **2020**, *94*, 101716. [CrossRef]

26. Jáñez-Martino, F.; Fidalgo, E.; González-Martínez, S.; Velasco-Mata, J. Classification of Spam Emails through Hierarchical Clustering and Supervised Learning. *arXiv* **2020**, arXiv:2005.08773.
27. Dada, E.G.; Bassi, J.S.; Chiroma, H.; Adetunmbi, A.O.; Ajibuwa, O.E. Machine learning for email spam filtering: Review, approaches and open research problems. *Heliyon* **2019**, *5*, e01802. [CrossRef] [PubMed]
28. Pérez-Díaz, N.; Ruano-Ordas, D.; Fdez-Riverola, F.; Méndez, J.R. Wirebrush4SPAM: A novel framework for improving efficiency on spam filtering services. *Softw. Pract. Exp.* **2013**, *43*, 1299–1318. [CrossRef]
29. Wu, C.H. Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks. *Expert Syst. Appl.* **2009**, *36*, 4321–4330. [CrossRef]
30. Enron Email Dataset. Available online: https://www.cs.cmu.edu/~{}enron/ (accessed on 22 October 2020).
31. SpamAssassin Dataset. Available online: https://spamassassin.apache.org/ (accessed on 22 October 2020).
32. Nazario Dataset. Available online: https://www.monkey.org/~{}jose/phishing/ (accessed on 23 October 2020).
33. UCI Machine Learning Repository. Available online: https://archive.ics.uci.edu/ml/datasets.php (accessed on 28 October 2020).
34. Asquith, A.; Horsman, G. Let the robots do it!–Taking a look at Robotic Process Automation and its potential application in digital forensics. *Forensic Sci. Int. Rep.* **2019**, *1*, 100007. [CrossRef]
35. Hayes, D.; Kyobe, M. The Adoption of Automation in Cyber Forensics. In Proceedings of the 2020 Conference on Information Communications Technology and Society (ICTAS), Durban, South Africa, 11–12 March 2020; IEEE: New York, NY, USA, 2020; pp. 1–6.
36. Syarif, I.; Prugel-Bennett, A.; Wills, G. SVM parameter optimization using grid search and genetic algorithm to improve classification performance. *Telkomnika* **2016**, *14*, 1502. [CrossRef]
37. Vinitha, V.S.; Renuka, D.K. Feature Selection Techniques for Email Spam Classification: A Survey. In Proceedings of the International Conference on Artificial Intelligence, Smart Grid and Smart City Applications (AISGSC), Coimbatore, India, 3–5 January 2019; Springer: Cham, Switzerland, 2020; pp. 925–935.
38. Mendez, J.R.; Cotos-Yanez, T.R.; Ruano-Ordas, D. A new semantic-based feature selection method for spam filtering. *Appl. Soft Comput.* **2019**, *76*, 89–104. [CrossRef]