Classifying ransomware victims' nationalities based on leak page entries

LUC DOP, University of Twente, The Netherlands

Ransomware is a type of malware that prevents a user from accessing their files by encrypting them. This is done to extort the victim. Some malware strains go beyond this and post the victim's personal information and files online to add extra pressure to pay. Pages dedicated to the posting of such information are called leak pages. These leak pages can provide a lot of information about the victim, such as their nationality. In this research, a refined data set of features related to a victim's nationality is created from a set of leak page entries. This data set is then used to train a classification model to classify a victim by country. Afterward, the results of this classification model are analyzed and it is shown that the model has a mean accuracy of 91%

Additional Key Words and Phrases: Ransomware, Leak page, Classification model

1 INTRODUCTION

Over the past decades, the internet has become an increasingly important part of our lives. With this has come an increasing amount of people and organizations being threatened by malware. One of the more commonly occurring forms of malware is ransomware. Ransomware has existed since 1989 when the 'AIDS' trojan was shared on floppy disks [14, 20]. This early ransomware was unsuccessful due to difficulties in spreading the malware and paying the ransom. Since then ransomware has developed into a more sophisticated and widespread threat, which has raised concern among law enforcement agencies. In 2022 alone the Federal Bureau of Investigation (FBI) received 2,385 complaints of ransomware with adjusted losses of 34.3 million dollars [13]. This is corroborated by Europol, which states that ransomware reports have increased in recent years and have become more focused on targeting public institutions and large corporations [11]. However, since these cases are self-reported the actual number of victims and losses may be much higher.

Studies have been done to classify ransomware victims and give info about what countries are more likely to be targeted. These find that it is mostly European and North-American countries that are targeted, with the top 3 being the US, the UK, and Canada [3, 5, 13]. It is noteworthy that these studies rely on self-reported data by ransomware victims. This is problematic, as cybercrime is notoriously underreported. Reasons for this include the victim not knowing when or how they have been targeted, the sentiment among victims that an investigation will not lead to satisfying results, and a lack of feedback during the cybercrime reporting process from law enforcement [6].

To alleviate this ransomware's lack of reporting, researchers and law enforcement must turn to alternative ways of getting information about ransomware victims. One such alternative method is the analysis of ransomware leak pages. These are dedicated sites where cybercriminals post information about their target to extort them further. From these leak pages, information can be learned about which countries are more likely to be targeted by ransomware groups. Knowing what countries are prone to ransomware attacks is crucial for criminal justice professionals to correctly deploy resources and assess the effectiveness of different programs in those countries. An issue with analysis is that any data gathered from leak pages must be enriched by data analysts to obtain actual valuable knowledge. It would be beneficial if this enrichment process could be automated.

1.1 Research Questions

This research aims to explore the possibility of identifying a ransomware victim's country of origin based on leak page entries. This leads to the following research question:

What information can be gained about a victim's nationality based on the URL posted in the leak page entry?

This question will be answered with the following sub-questions:

- **RQ1:** What information can be gained about a victim's nationality using internet measurements?
- **RQ2:** What information can be gained about a victim's nationality when scanning the victim's webpage?
- **RQ3:** What information can be gained about a victim's nationality when using their name as a search term in the chamber of commerce?

In this research, content analysis is done to study a data set of ransomware leak page entries, obtained from eCrime [7]. This site monitors ransomware groups and extracts information about their victims, specifically companies and organizations targeted by ransomware groups. They then enrich this data by providing among other things the victim's country. From these entries, indicators can be found that can link the victim to a certain country based on the name of the victim and a URL. Because eCrime provides the victim's country these indicators can be immediately validated. We then use these indicators as features in a classification model, where the target class is the victim's country of origin. This research paper starts with an exploration of related works. It will then explain how the country indicators are extracted from the leak page entries as well as how the classification model functions. This is followed by an analysis of how the model performs based on different metrics. These results will then be discussed and compared to previous literature. The paper closes with the conclusion of the research as well as possible future research avenues.

TScIT 39, July 7, 2023, Enschede, The Netherlands

2 RELATED WORK

The Domain Names System (DNS) refers to both a namespace from which unique names can be assigned to users, and a protocol for matching these names to specific Internet protocol (IP) addresses in response to queries from Internet users' computers. The DNS is a hierarchical system with Top Level Domains (TLDs) being the top level of the hierarchy. These TLDs have a subset that is linked to countries called country code Top Level Domains (ccTLDs). To date, several studies have investigated the relationship ccTLDs and their respective countries. Many countries retain or are interested in being the sole authority of their ccTLD [15]. Some ccTLDs have made themselves into 'quasi-generics' by exploiting similarities between their country code and other meanings. However, most countries strive to keep their ccTLDs firmly linked to them, as a means of exerting political power and expressing their sovereignty [8].

As the DNS matches domain names to IP addresses, one or more IP addresses can be obtained from a domain name using forward DNS [10]. The resulting IP address(es) can then be associated with geographical locations using IP geolocation techniques [18]. These can be divided into active and passive techniques. Active techniques are more accurate than passive ones but are significantly more time-consuming. Passive geolocation or database-driven geolocation takes less time to produce results and is provided either for free or as a commercial tool. However, the databases that are used in this technique often have questionable accuracy [16, 23].

Various studies have assessed the efficacy of extracting information from websites using web scraping. This is a set of techniques used to automatically get some information from a website instead of manually copying it [26]. This is typically done by composing an HTTP request to the target website and analyzing the response. The information gained can then be used to classify a multitude of concepts using machine learning algorithms [24, 25, 28].

While there are papers that give insight into what countries and organizations get targeted more by ransomware, most of this information comes from reports made by the victims themselves [5, 12]. There are some papers describing how ransomware groups put extra pressure on their victims beyond extorting them by encrypting their files. They can post the encrypted files on dedicated leak pages, perform DDoS attacks, and even reach out to the target's customers and stakeholders to extort their victims even further [19, 27].

3 DATASET PREPERATION

The classification model created during this research was trained using data gathered from a set of leak page entries from eCrime [7]. Most entries in the eCrime data set provide the company name of the victim, the country of origin of the victim, and a URL leading to a webpage related to the victim. From this, a variety of features can be gathered that indicate the country of origin of the victim. These include the following:

- A country code top-level domain included in the URL.
- The IP address of the URL's webpage.
- The language in which the web page is written.
- Country code top-level domains included in e-mail addresses found on the web page.
- Country calling codes included in phone numbers found on the webpage.
- Mentions of countries and/or cities in the webpage's text.
- Mentions of countries and/or cities in the first document found on LexisNexis when searching by the victim's name.

Only the entries containing all three of these things were refined, to avoid too much missing data. Entries are also skipped if the URL's webpage is inaccessible. Being inaccessible means that the URL was incorrect, connecting to the webpage provided by the URL times out, or the webpage provided by the URL tries to redirect too often.

3.1 Collecting Data

3.1.1 Country code top-level domain. URLs contain a top-level domain (TLD). In this research, we specifically look at a subset of TLDs called country code top-level domains. A country code top-level domain (ccTLD) is an Internet top-level domain generally used or reserved for a country, sovereign state, or dependent territory identified with a country code. Most of these ccTLDs use a two-letter code using the Latin alphabet, though in recent years internationalized ccTLDs have been added [1]. The conditions for the use of ccTLDs can be complex and vary per country, with some being exclusive to citizens and others being available to everyone. Those ccTLDs available to everyone are referred to as Generic ccTLDs. For this research, an algorithm was written that checks the URL of every leak page entry for a Latin character ccTLD. The algorithm then looks it up in a dictionary linking each ccTLD to its respective country. Any TLDs encountered which do not belong to a country are labeled as 'generic'.

3.1.2 IP address. An IP address is a unique address that identifies a device on the internet or a local network. These IP addresses can be translated to domain names and vice-versa by Domain Name System (DNS) servers [29]. Using this the IP address of the URL mentioned in a leak page entry can be acquired. Because IP addresses are tied to a general location, they can be used to locate the victim's country of origin. In this research, the IP address is retrieved by issuing the 'nslookup' command in the Windows command-line together with a URL mentioned in the leak page. Once an IP address has been obtained a GET request is sent to country.is, which is a free open-source API with access to a geolocation database. The API then responds with the ISO 3166-1 alpha-2 code of the country related to the IP address.

3.1.3 Language. The spoken language of a person or organization can be an indicator of what country they are from. For this reason, we try to detect the language in which the victim's webpage is written. Most webpages contain an attribute called 'lang' which tells web browsers the language of elements on the page. In this research, code has been written to find this attribute and get its value. If this element does not exist, we instead try to detect the language of the text. Over 55 languages can be detected.

3.1.4 E-mail. Like URLs, e-mail addresses can contain country code top-level domains. Using the same dictionary mentioned in 3.1.1, these domains can be tied to countries. HTML allows webpages to

create links that redirect users to e-mail addresses called 'mailto' links. By finding these mailto links e-mail addresses can be obtained from the victim's webpage. Any country code top-level domain present in the e-mail address will be extracted and the corresponding country added to a list. Any TLDs encountered which do not belong to a country are labeled as 'generic'. The mode of all the countries added to the list is then taken.

3.1.5 Phone number. Country calling codes are telephone number prefixes for reaching telephone subscribers in foreign countries or areas via international telecommunication networks. By finding phone numbers and looking for these country codes a victim can be linked to a certain country. HTML allows webpages to create clickable phone links. By finding these links phone numbers on a victim's webpage can be obtained. Because not all phone numbers on a website have clickable links, regular expressions are used as an added method for finding phone numbers. A regular expression is a sequence of characters that specifies a pattern in text. By creating one which specifies the pattern of a phone number more of them can be retrieved from the victim's webpage. Any phone numbers found are then validated. If a fitting country code is found, the corresponding country's ISO 3166-1 alpha-2 code is returned.

3.1.6 Location. A victim's webpage contains information about who they are and what they do. This often includes the cities, states, and countries they are located in. By using natural language processing it is possible to find these mentions of locations. In the case of cities, we try to get what country a city is located in. The mode is then taken of all countries found. The code tries to validate any cities and countries passed to it, in case the natural language processing marked an unrelated word as a location.

3.1.7 LexisNexis. Organizations and companies must register themselves at a chamber of commerce or other government body to be able to operate. This means providing information about themselves, including where they are located. Documents containing these registration details are often stored in online databases. A company that provides access to such a database is LexisNexis [2]. By looking up the name of the target mentioned in the leak page it is to discover their country of origin. In this research, the victim's name is queried on LexisNexis, and the top resulting document is accessed. Then using the same method as mentioned in 3.1.6 the page is scanned for locations and the mode of all countries found is returned.

3.2 Dataset pre-processing

In this research, the Python library scikit-learn [21] is used for pre-processing and model building. First, all data samples are removed from countries that occur less than 4 times. This is done to allow for an even distribution of all countries in both the test and train subsets. If this is not done, we run into the problem of countries appearing in the training subset but not the test subset and vice-versa. For the model, 75 percent of the data is used for training and 25 percent for testing.

Once the data has been split it gets encoded. The dataset created for this research works exclusively with categorical variables. All the variables in the dataset are nominal, which means they are categorical but there is no order in the categories. For this reason and because scikit-learn supports it, one-hot encoding has been used in this research.

Finally, missing values must be dealt with. For many leak page entries, one or multiple features mentioned in 3.1 could be missing. These missing values will cause the machine-learning algorithm to fail and must therefore be removed. There are a variety of different methods to deal with missing values [22]. For this research mean/mode imputation (MMI) is used. Here the missing data for a given attribute is replaced by the mode of all known values of that attribute.

3.3 Dataset

We considered 6000 leak page entries from the Ecrime dataset. Of these 6000 entries, 3568 are found containing a functioning URL, the name of the targeted victim, and their country of origin. This set contains 113 unique countries/geographical regions. With the removal of all entries whose country appears less than four times, we end up with a total of 3484 refined data entries to use for the classification model. This final dataset features entries from 59 countries in total. Table 1 shows the top 5 most commonly occurring countries. A table of all countries and their distribution can be found in the appendix. What stands out in the table is that the dataset contains far more victims from the United States, with more than 7 times the amount of entries than the second largest entry of the United Kingdom.

Table 1. Top 5 most common countries in dataset

Country	Occurences
United States	1554
United Kingdom	212
Germany	193
Canada	174
France	153

4 THE CLASSIFICATION MODEL

For this research, the random forest machine learning algorithm is used. A Random Forest classifier is an ensemble learning method that combines multiple decision trees to make predictions. It is accurate, robust to overfitting, and allows for the estimation of the importance of individual features for overall predictive performance.

To improve accuracy further hyperparameter tuning is performed. This involves selecting the optimal values for the various hyperparameters that control the behavior and performance of the model. Hyperparameters are parameters that are not learned from the data but are set by the user before training the model. For this research hyperparameter tuning is done using random search. Here random combinations of hyperparameters are selected from a range of values whereafter the model's performance is observed. The set of hyperparameters that yields the best performance is then selected,

which is determined based on the model's accuracy. The following hyperparameters are considered in the random search:

- The number of trees to use in the random forest.
- The number of features to consider at each split.
- The maximum depth of the trees.
- The minimum number of samples required to split at a node.
- The minimum number of samples required at each leaf node.
- To use bootstrapping when sampling data. This involves creating multiple subsets of the original training data by randomly sampling with replacement.

5 RESULTS AND EVALUATION

To assess the performance of the model stratified 4-fold cross-validation is used. Here the dataset is divided into k subsets or 'folds' which are iteratively used for training and validation. Making the cross-validation stratified means the folds are made by preserving the percentage of each country in the folds. This gives a more reliable performance estimation because its performance is based on multiple iterations, which reduces the impact of data variability. In addition to this it allows for effective use of data: In k-fold cross-validation, all data points are utilized for both training and validation, ensuring that each sample is used for validation exactly once. This maximizes the use of available data and minimizes the risk of overfitting or underfitting. From each fold, a number of metrics are calculated. The accuracy, precision, recall, and f1 scores are calculated as follows:

$$Accuracy(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} 1(y_i = \hat{y}_i) \tag{1}$$

$$Precision(y, \hat{y}) = \frac{|y \cap \hat{y}|}{\hat{y}} \tag{2}$$

$$Recall(y, \hat{y}) = \frac{|y \cap \hat{y}|}{y} \tag{3}$$

$$F1(y, \hat{y}) = 2 * \frac{Precision(y, \hat{y}) * Recall(y, \hat{y})}{Precision(y, \hat{y}) + Recall(y, \hat{y})} \tag{4}$$

Here n is the number of samples, $\hat{y}$ are the predicted values, and y are the corresponding true values. 1(x) is the indicator function. Cohen's Kappa is calculated to measure the performance of machine learning classification models based on assessing the perfect agreement and agreement by chance between the two raters (a real-world observer and the classification model)[9]. This is given with the following equation:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{5}$$

Here p_o is the empirical probability of agreement on the label assigned to any sample (the observed agreement ratio), and p_e is the expected agreement when both annotators assign labels randomly. p_e is estimated using a per-annotator empirical prior over the class labels[4]. To measure the model's ability to distinguish correctly between countries the Area Under the Curve (AUC) is used as a metric. The AUC score is calculated by getting the average AUC score of all pairwise combinations of countries. This is weighted by the prevalence of the countries in the dataset. The equation proposed by [17] is used to calculate this:

$$\frac{1}{c(c-1)} \sum_{j=1}^{c} \sum_{k>j}^{c} p(j \cup k)(AUC(j|k) + AUC(k|j)) \tag{6}$$

The mean scores of the folds can be seen in Figure 1.

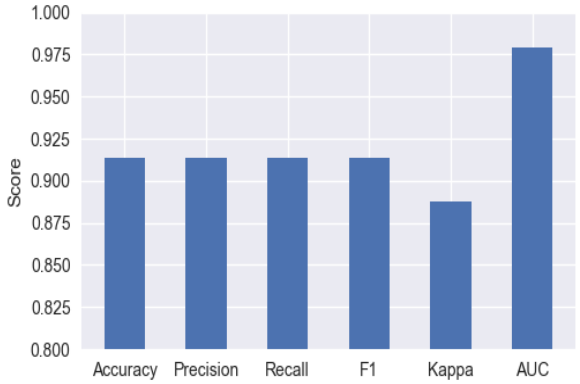

Fig. 1. Evaluation metrics of the 4 fold cross validation

To improve the model's interpretability we model which features contribute the most to the accuracy of the model based on permutation importance. This is done by shuffling the features ten times and observing how this influences the model's accuracy. The importance of the individual features is given in the mean decrease in the accuracy of the model when they are removed. The results of this can be seen in Figure 2.

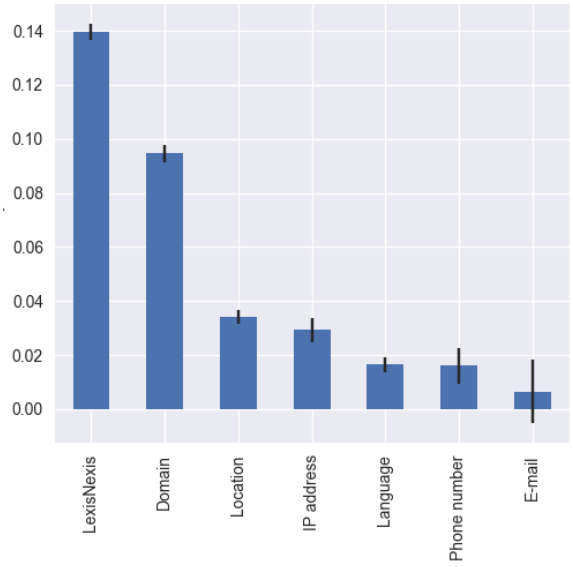

Fig. 2. Mean feature importance in the model

The second most important feature in the model is the Domain feature. When a Top level domain is found that is not tied to a country this feature is labeled as 'generic'. By splitting the training data into subsets with generic top-level domains and non-generic ones and evaluating its scores more insight can be gained into the importance of this feature. To evaluate performance the same metrics are used as in Figure 1 except for AUC. This is because the subsets of generic and non-generic data do not have the same distribution of countries. The blue bar shows the scores of the total subset, the green bar those of the generic subset, and the red those of the non-generic subset.

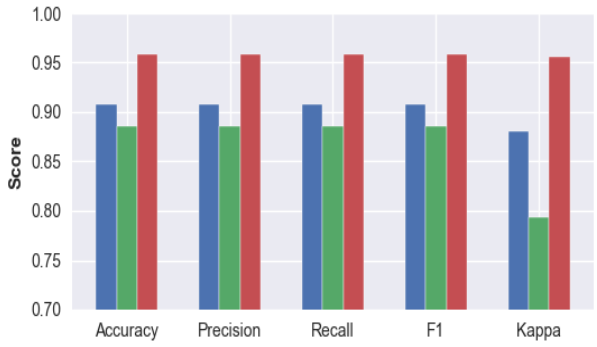

Fig. 3. Evaluation of Generic vs non-generic domain

To analyze the model's data sufficiency and complexity a graph is made to visualize the learning curve against the training set size. This helps determine if the model would benefit from more data or if it is already reaching its performance limit. Performance is measured based on the accuracy of the predictions.

Finally, a manual analysis is done on a set of wrong predictions and what could be the cause for the error. For this, we specifically look at missing/wrong features and try to find out why these were incorrect. For this, 75 wrong predictions are looked at and any problems encountered are gathered as well as their number of occurrences. The percentage of features incorrect in these 75 predictions can be seen in Table 2. The top five most occurring problems can be found in Table 3 with the full list being in the appendix.

Table 2. Percentage wrong of each feature

Feature	Percentage incorrect
Domain name	88.0%
IP address	78.7%
Phone number	74.7%
E-mail	94.7%
Location	62.7%
LexisNexis	58.7%

Table 3. Top 5 problems in wrong predictions

Problem	Amount
LexisNexis found the wrong adress. Either a different company with the same name or incorrect information in the Ecrime dataset.	26
The LexisNexis data was outdated, and the current code actually gets the correct country.	6
A company's local branch was targeted, but the site to the main branch was given. This caused LexisNexis to return the wrong country	6
The algorithm used to access LexisNexis cannot handle '&' in company names, and cuts off the request.	4
There is a phone number on the main page or contact page, but it is a regional number and thus does not get recognized.	4

6 DISCUSSION

Figure 1 shows the results obtained from the 4-fold cross-validation. What stands out in this figure is that the mean accuracy, precision, recall, and f1 score are all 0.91. While the accuracy is similar to the models discussed in Section 2[25, 28], the precision, recall, and f1 scores all being the same is another matter. This is because we calculate these metrics globally by counting the total true positives, false negatives, and false positives. It is also possible to calculate these metrics for each country individually and take the mean result of each metric. This would significantly change the results of these metrics due to the imbalanced nature of the dataset, as shown in section 3.3. Countries that are less prevalent in the dataset and on which the model performs worse would significantly lower the scores of the metrics. Figure 1 also shows a mean Cohen's Kappa score of 0.88. This shows that the model has a good inter-annotator agreement. Finally, Figure 1 shows an AUC score of 0.98. This indicates that the model has a strong ability to discriminate between positive and negative instances, suggesting good predictive performance and reliable rankings.

Figure 2 shows the mean importance of each feature used in the model. It is apparent that the LexisNexis feature is by far the most important feature in the dataset with a mean accuracy decrease of 14.0% ± 0.3. As mentioned in the literature review the country code Top Level Domain is a good indicator of what country a victim is from, being the second most important feature in the model. The examined literature expressed skepticism towards passive IP address geolocation. This is somewhat corroborated by Figure 2, which shows the IP address feature to be less than half as important as the domain feature. The data for the 4 least important features were all obtained using web scraping. This is in contrast to examined studies that feed web-scraped data to classification models and get a high accuracy metric. Interestingly, the e-mail feature was observed to have a negative mean accuracy decrease when shuffled if we look at the standard deviation. This means this feature could have a negative impact on the overall accuracy of the model.

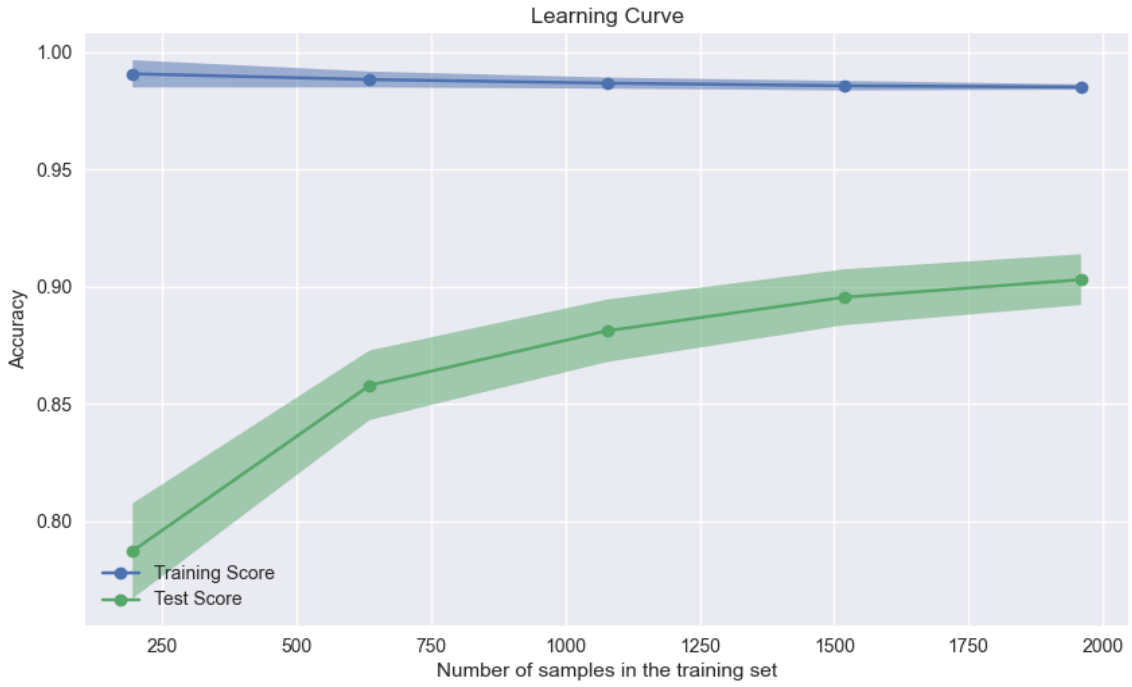

Fig. 4. The learning curve of the model against training set size

Figure 3 shows that the non-generic data improves on all metrics over the total dataset with a score on all metrics of 96.4%. The generic data on the other hand scores lower on all metrics, with accuracy, precision, recall, and f1 score of 88.1%. Furthermore, Cohen's Kappa score significantly decreases to 79.3%. A possible explanation for this might be an imbalance in the countries featured in the generic data. The subset of generic data samples consists of a sizable part of victims from the United States.

Figure 4 shows that the test score starts to plateau as the training size reaches its regular size, though it does not decrease yet. This indicates that performance could yet be increased with more data.

Table 3 shows that the most occurring problem is LexisNexis returning the wrong address for a victim. This may be due to the importance of this feature, which means that once the LexisNexis feature is incorrect the predicted country is more likely to be incorrect as well. Furthermore, the high percentage of incorrect e-mails as seen in Table 2 further underlines the e-mail feature as being prone to errors.

7 CONCLUSION

In this paper, we present an algorithm capable of extracting features related to a ransomware victim's origin from a set of leak page entries. In addition, we present a classification model capable of classifying the ransomware victim's country of origin with a mean of 91% accuracy based on these features. This section reflects on the research questions established in the introduction.

RQ1: *What information can be gained about a victim's nationality using internet measurements?* From the URL the country code Top Level Domain (ccTLD) and IP address can be extracted. The ccTLD is strongly linked to a country or geographical location for political reasons and the IP address provides a victim's country using IP address geolocation. Both are found to be useful when classifying ransomware victims by their country.

RQ2: *What information can be gained about a victim's nationality when scanning the victim's webpage?* From the victim's webpage this research extracts language, phone numbers, mentions of cities or countries, and e-mail addresses. All of these features are shown to be somewhat useful in classifying the victim's country of origin, with the exception of the e-mail addresses.

RQ3: *What information can be gained about a victim's nationality when using their name as a search term in the chamber of commerce?* This research uses LexisNexis[2] to search for the victim's company profile provided by the chamber of commerce. These profiles are very useful when classifying a victim's country of origin.

8 FUTURE WORK

This research paper focuses on classifying ransomware victims by their nationality, but there are other aspects on which targeted companies/organizations can be classified. For example, the documents provided by LexisNexis often also provide information about what area of the economy the victims are active in. This can be used in conjunction with web scraping techniques to create a model that classifies victims by their economic sector.

ACKNOWLEDGMENTS

The data used in this research is provided by eCrime [7].

REFERENCES

[1] 2020. IANA — Root Zone Database. https://www.iana.org/domains/root/db. Accesed: 13-06-2023.

[2] 2023. LexisNexis. https://www.lexisnexis.com/nl-nl. Accessed: 03-05-2023.

[3] Bander Ali Saleh Al-rimy, Mohd Aizaini Maarof, and Syed Zainudeen Mohd Shaid. 2018. Ransomware threat success factors, taxonomy, and countermeasures: A survey and research directions. *Computers & Security* 74 (2018), 144–166. https://doi.org/10.1016/j.cose.2018.01.001

[4] Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics* 34, 4 (2008), 555–596. https://doi.org/10.1162/coli.07-034-R2

[5] Amir Atapour-Abarghouei, Stephen Bonner, and Andrew Stephen McGough. 2019. Volenti non fit injuria: Ransomware and its Victims. In *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 4701–4707. https://doi.org/10.1109/BigData47090.2019.9006298

[6] Morvareed Bidgoli and Jens Grossklags. 2016. End user cybercrime reporting: what we know and what we can do to improve it. In *2016 IEEE International Conference on Cybercrime and Computer Forensic (ICCCF)*. IEEE, 1–6. https://doi.org/10.1109/ICCCF.2016.7740424

[7] Cosin Camichel. 2022. Ecrime. https://ecrime.ch/. Accessed: 03-05-2023.

[8] George Christou and Seamus Simpson. 2009. New governance, the internet, and country code top-level domains in Europe. *Governance* 22, 4 (2009), 599–624. https://doi.org/10.1111/j.1468-0491.2009.01455.x

[9] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46. https://doi.org/10.1177/001316446002000010

[10] Ovidiu Dan, Vaibhav Parikh, and Brian D Davison. 2021. IP geolocation through reverse DNS. *ACM Transactions on Internet Technology (TOIT)* 22, 1 (2021), 1–29. https://doi.org/10.1145/3457611

[11] Europol. 2021. *Internet Organised Crime Threat Assessment (IOCTA 2021)*. European Union Agency for Law Enforcement Cooperation (Europol). https://www.europol.europa.eu/publications-events/main-reports/internet-organised-crime-threat-assessment-iocta-2021

[12] Cath Everett. 2016. Ransomware: to pay or not to pay? *Computer Fraud & Security* 2016, 4 (2016), 8–12. https://doi.org/10.1016/S1361-3723(16)30036-7

[13] FBI. 2023. *Internet crime report 2022*. Federal Bureau of Investigation (FBI). https://www.ic3.gov/Media/PDF/AnnualReport/2022_IC3Report.pdf

[14] Alexandre Gazet. 2010. Comparative analysis of various ransomware virii. *Journal in computer virology* 6 (2010), 77–90. https://doi.org/10.1007/s11416-008-0092-2

[15] Michael Geist. 2004. Governments and country-code top level domains: a global survey. *Feb, Surveydistributed tomembersof ITU* (2004). http://www.michaelgeist.ca/resc/geistgovernmentcctlds.pdf

[16] Bamba Gueye, Steve Uhlig, and Serge Fdida. 2007. Investigating the imprecision of IP block-based geolocation. In *Passive and Active Network Measurement: 8th Internatinoal Conference, PAM 2007, Louvain-la-neuve, Belgium, April 5-6, 2007. Proceedings 8*. Springer, 237–240. https://doi.org/10.1007/978-3-540-71617-4_26

[17] David J Hand and Robert J Till. 2001. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine learning* 45 (2001), 171–186. https://doi.org/10.1023/A:1010920819831

[18] Ethan Katz-Bassett, John P John, Arvind Krishnamurthy, David Wetherall, Thomas Anderson, and Yatin Chawathe. 2006. Towards IP geolocation using delay and topology measurements. In *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*. 71–84. https://doi.org/10.1145/1177080.1177090

[19] Tom Meurs, Marianne Junger, Erik Tews, and Abhishta Abhishta. 2022. Ransomware: How attacker's effort, victim characteristics and context influence ransom requested, payment and financial loss. In *Symposium on Electronic Crime Research, eCrime 2022*. https://doi.org/10.1109/eCrime57793.2022.10142138

[20] Philip O'Kane, Sakir Sezer, and Domhnall Carlin. 2018. Evolution of ransomware. *Iet Networks* 7, 5 (2018), 321–327. https://doi.org/10.1049/iet-net.2017.0207

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?ref=https:/

[22] Liu Peng and Lei Lei. 2005. A review of missing data treatment methods. *Intell. Inf. Manag. Syst. Technol* 1 (2005), 412–419. https://spu.fem.uniag.sk/cvicenia/ksov/prokeinova/MBA-Business%20Modelling/Lecture%201/Missing%20values/missing%20values.pdf

[23] Ingmar Poese, Steve Uhlig, Mohamed Ali Kaafar, Benoit Donnet, and Bamba Gueye. 2011. IP geolocation databases: Unreliable? *ACM SIGCOMM Computer Communication Review* 41, 2 (2011), 53–56. https://doi.org/10.1145/1971162.1971171

[24] Saurabh Sahu, Km Divya, Neeta Rastogi, Puneet Kumar Yadav, and Yusuf Perwej. 2022. Sentimental Analysis on Web Scraping Using Machine Learning Method. *Journal of Information and Computational Science* 12 (2022). https://doi.org/10.12733/JICS.2022/V12I08.535569.67004

[25] Cepy Slamet, Rian Andrian, Dian Sa'adillah Maylawati, Suhendar, Wahyudin Darmalaksana, and Muhammad Ali Ramdhani. 2018. Web scraping and Naïve Bayes classification for job search engine. In *IOP Conference Series: Materials Science and Engineering*, Vol. 288. IOP Publishing, 012038. https://doi.org/10.1088/1757-899X/288/1/012038

[26] Eloisa Vargiu and Mirko Urru. 2013. Exploiting web scraping in a collaborative filtering-based approach to web advertising. *Artif. Intell. Res.* 2, 1 (2013), 44–54. https://doi.org/10.5430/air.v2n1p44

[27] Arun Warikoo. 2023. Perspective Chapter: Ransomware. In *Malware*, Eduard Babulak (Ed.). IntechOpen, Rijeka, Chapter 5. https://doi.org/10.5772/intechopen.108433

[28] Ferry Wahyu Wibowo, Akhmad Dahlan, and Wihayati. 2021. Detection of Fake News and Hoaxes on Information from Web Scraping using Classifier Methods. In *2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*. 178–183. https://doi.org/10.1109/ISRITI54043.2021.9702824

[29] Hao Wu, Xianglei Dang, Lidong Wang, and Longtao He. 2016. Information fusion-based method for distributed domain name system cache poisoning attack detection and identification. *IET Information Security* 10, 1 (2016), 37–44. https://doi.org/10.1049/iet-ifs.2014.0386

A APENDIX

Table 4. Full list of problems found in predictions

Problem	**Amount**
LexisNexis found the wrong adress. Either a different company with the same name or incorrect information in the Ecrime dataset.	26
The LexisNexis data was outdated, and the current code actually gets the correct country	6
A company's local branch was targeted, but the URL to the main branch was given. This caused Lexisnexis to give the wrong country	6
The algorithm used to access LexisNexis cannot handle '&' in company names, and cuts off the request.	4
There is a phone number on the main page or contact page, but it is a regional number and thus does not get recognized.	4
Cannot reach the contact page because it is accessed by appending some variation of 'contact us' to the current URL.	4
The webpage contains no text or states it is inactive	3
Cannot reach the contact page because 'contact' is written in another language.	2
The target company is located in a city with a country name and was therefore labeled with the wrong location e.g. Peru, Indiana.	2
The contact page was renamed 'get in touch' and was therefore not found.	1
The algorithm marked the word 'Us' as the United States.	1
The main page of the victim's website is a login page with no contact information.	1
A valid phone number was not detected because it did not start with '+'.	1
The code for validating country names does not know the country Scotland.	1
The code for validating country names sees Hong Kong as its own country but LexisNexis data sees it as a part of China.	1

Table 5. Countries in data set p1

Country	Amount
United States	1554
United Kingdom	212
Germany	193
Canada	174
France	153
Italy	133
Spain	96
Australia	82
Brazil	61
Switzerland	57
India	49
Japan	49
Netherlands	42
Belgium	39
Austria	39
Thailand	37
China	35
Taiwan, Province of China	29
Mexico	25
Turkey	20
United Arab Emirates	20
South Africa	19
Sweden	19
Hong Kong	18
Argentina	18
Portugal	17
Indonesia	16
Singapore	16
Israel	16
Poland	14
Greece	14
Denmark	14
Malaysia	14
New Zealand	12
Chile	12
Colombia	12
Saudi Arabia	11
Ireland	11
Peru	10
Norway	9
Vietnam	8
Kuwait	8
Ecuador	8
Philippines	8
Finland	7

Table 6. Countries in data set p2

Country	Amount
Czech Republic	7
Costa Rica	6
Lebanon	6
Venezuela	6
Qatar	6
Hungary	6
South Korea	6
Egypt	5
Bulgaria	5
Romania	5
Luxembourg	4
Iran	4
Bosnia and Herzegovina	4
Dominican Republic	4
Pakistan	3
Cyprus	3
Nicaragua	3
Botswana	3
Tanzania	3
Puerto Rico	3
Nigeria	3
Morocco	3
Isle of Man	2
Bahamas	2
Uganda	2
Barbados	2
Czechia	2
Senegal	2
Guatemala	2
Serbia	2
Sri Lanka	2
Bahrain	2
Croatia	2
Jamaica	2
Panama	2
Estonia	2
Slovakia	1
Gibraltar	1
Iraq	1
Mali	1
Trinidad and Tobago	1
Gambia	1
Gabon	1
Bolivia	1
Montenegro	1

Table 7. Countries in data set p3

Country	**Amount**
Sint Maarten	1
Monaco	1
Paraguay	1
Kenya	1
Greenland	1
North Macedonia	1
Tunesia	1
Seychelles	1
Ivory Coast	1
Angola	1
Ethiopia	1
Zambia	1
Honduras	1
Ukraine	1
Lithuania	1
Brunei	1
Mongolia	1
Myanmar	1
Jordan	1
Slovenia	1
Democratic Republic of the Congo	1
Cayman Islands	1
Oman	1