Phishing Email Detection based on Named Entity Recognition

Vít Listík[1], Šimon Let[2], Jan Šedivý[3] and Václav Hlaváč[3]
[1]*Faculty of Electrical Engineering, Department of Cybernetics, Czech Technical University in Prague, Technická 2, Prague 6, Czech Republic*
[2]*Faculty of Information Technology, Department of Theoretical Computer Science, Czech Technical University in Prague, Thákurova 9, Prague 6, Czech Republic*
[3]*Czech Institute of Informatics, Robotics and Cybernetics, Jugoslávských Partyzánů 1580/3, Prague 6, Czech Republic*

Keywords: Named Entity Recognition (NER), Phishing, Email.

Abstract: This work evaluates two phishing detection algorithms, which are both based on named entity recognition (NER), on live traffic of Email.cz. The first algorithm was proposed in (Ramanathan and Wechsler, 2013). It is using NER and latent Dirichlet allocation (LDA) as feature extractors for random forest classifier. This algorithm achieved 100% F-measure on the publicly available testing dataset. We are using this algorithm as the baseline for our newly proposed solution. The newly proposed solution is using companies detected by the NER and it is comparing URLs present in the email content to the company URL profile (based on history). The company URL profile contains domains which are frequently mentioned in legitimate traffic from that domain. The advantage of the proposed solution is that it does not need phishing dataset, which is hard to get, especially for languages other than English. Our solution outperforms the baseline solution. Both solutions are able to detect previously undetected phishing attacks. Combination of the solutions achieves 100 % F-measure on the portion of live traffic.

1 INTRODUCTION

Phishing is a fraudulent attempt to steal personal information. It is commonly used in email. This method is often using impersonation techniques. The attacker wants the victim to click on the malicious URL and fill in personal information. The attacker achieves that because the victim thinks that the email was sent by the legitimate and known organization. Phishing is causing huge financial and reputation losses globally.

1.1 Email.cz Traffic Statistics

Email.cz is the biggest freemail provider in the Czech Republic. The in-house anti-spam solution is analyzing approximately 50 million messages a day. The current anti-spam system it not focused on phishing detection. But the negative effects of those attacks affect not only the users but also the company. It became a bigger problem when Email.cz started offering *email on custom domain* because it is a service commonly used by companies.

Based on the language detection[1] 75% of incoming email messages at Email.cz are in Czech language and 15% in the English language. The rest is mostly undetected.

[1]https://github.com/CLD2Owners/cld2

1.2 Task

Our task is to create a model for detecting phishing emails in Email.cz traffic. As shown in Sec. 1.1 huge portion of the analyzed messages are in the Czech language. Therefore the proposed solution should be extensible for other languages than English and should not use phishing examples because they are almost impossible to obtain (described in Sec. 4.1).

2 STATE OF THE ART

Phishing detection was approached many times. The task solution may be based on metadata or content. We are focusing on the content based methods. Those methods may use email text, images contained in the email or the attachments. We are using natural language processing methods (NLP) for phishing detection. More specifically named entity recognition (NER) which was proven efficient for this task (Ramanathan and Wechsler, 2013).

Listík, V., Let, Š., Šedivý, J. and Hlaváč, V.
Phishing Email Detection based on Named Entity Recognition.
DOI: 10.5220/0007314202520256
In *Proceedings of the 5th International Conference on Information Systems Security and Privacy (ICISSP 2019)*, pages 252-256
ISBN: 978-989-758-359-9

Other possible approaches are based on metadata, the best example is information from SMTP traffic (i.e. sender reputation). Those approaches are successful for tasks like spam detection because the messages are sent in bulks big enough to build the reputation. Which does not apply to phishing which is sent in smaller batches and the SMTP servers are often stolen. That makes reputation approaches ineffective for phishing detection. Next argument for not exploring this path is that those mechanisms are already used in most anti-spam solutions, therefore, those attacks would be not delivered anyway.

Most commonly used approach for phishing detection is using blacklists, which is also true for spam detection. This method is simple to implement. It is using URLs present in the email body which are searched in the blacklist. The benefit of this method is very high precision because the lists are hand-curated, which is also its biggest downside, because the addition of new attacks is slow, causing a low recall.

2.1 Named Entity Recognition

Named entity recognition (NER) is a natural language processing technique which is mapping sequence of words to sequence of tags. Tags are called named entities i.e. personal names, places, companies, dates, and numbers.

We are using NER implementation called Nametag (Straková et al., 2014) because it achieves the best results for the Czech language. It also has good performance for English. We need to target Czech because most of Email.cz traffic is Czech (Sec. 1.1) and the Czech language is more complicated for machine processing then English.

Nametag is a two-pass algorithm based on maximum entropy Markov model. It is using morphological analysis and other features like word clustering, gazetteers and orthographic features (capitalization, punctuation) (Straková et al., 2014).

3 METHODS

We propose two methods for phishing detection. First one is described in the (Ramanathan and Wechsler, 2013). Authors achieved a great result of 100 % F-measure. This method is a random forest classifier based on features extracted by NER and latent Dirichlet allocation (LDA). We are re-implementing this method and validating its performance in live traffic. This method is dependent on a representative dataset which is hard to get for this task.

Our second proposed method is overcoming this precondition. We are detecting entity impersonation instead. This method is based on NER which is detecting companies and then comparing the detected company link profile to the email link profile. The email link profile is extracted from the email body. The company link profile is built from historical data and consists of domains which are referred in the legitimate emails sent by the company.

3.1 Random Forest Classifier

We trained model based on (Ramanathan and Wechsler, 2013). This model consists of three parts. The first part is named entity recognition. The second part is latent Dirichlet allocation (LDA). Third and the last part is random forest classifier.

First part is responsible for detecting names of people, places, and organizations. LDA is responsible for estimating topic probabilities. Those two models work well together because NER is considering the input word position and LDA is based only on word frequencies (position independent) (Ramanathan and Wechsler, 2013).

Stanford NER was used in the original paper (Jenny Rose Finkel and Manning, 2005). We are using Nametag because it has more granular tags and it was tested on the *email language* (described in 5.1) (Straková et al., 2014). We are using tags starting with ”g” (geographical names), ”i” (companies and institutions) and ”p” (personal names).

LDA model is using the term-document matrix as an input. This matrix is created with the Scikit count vectorizer (Pedregosa et al., 2011). We are using 1000 features after stop words removal. The vectorizer is trained on 90% of the dataset. We are using 90% of the corpus because the corpus is small and we do want most of the words to be present in the vectorizer. We are using 200 categories as supposed in the original paper. Achieved perplexity is 421 which is comparable to the original paper.

First two models are used to create an input feature vector for the final classifier. The feature vector consists of 200 topic probabilities predicted by the LDA model, and 40 NER based features consisting of the tags and the entities.

The final model consists of 200 weak learners of max depth 5. We trained it using the dataset described in Sec. 4.1 the same way as the original paper did. This model architecture was able to achieve 100% F-measure in the original paper. We achieved 94% F-measure (more details in 5.2), which is a sufficient result for the production test (described in Sec. 5.3).

Our implementation in Python is using scikit-learn

(Pedregosa et al., 2011) and is available on Github[2].

3.2 Impersonation Detection

Phishing is based on entity impersonation. The approach proposed in Sec. 3.1 may learn to detect the entity impersonation, because it is using NER features but it needs a lot of phishing samples to do that. As stated in Sec. 4.1 there is no dataset for phishing in the Czech language (and other languages) and the available datasets in English are old and not very representative. We still want to detect those attacks. Therefore we propose an impersonation detection method. This method is using the knowledge that phishing is impersonating some trusted entity to make a user click on the malicious link.

Our suggested method consists of three steps.

1. Detect company entities from the email body with use of NER (Sec. 3.3).
2. Map detected company entities (names) to the domains (Sec. 3.4).
3. Compare domains extracted from links to detected domain *link profiles* and report unexpected domains present in the email (Sec. 3.5).

Those steps are described in greater detail in the following sections.

3.3 Entity Detector

The first step of this system is to detect company entities from the email body. The email headers may be edited by the attacker, but the email body almost always contain the company entity to persuade the user that the entity sent the email. We are using NER for this task, specifically Nametag with the custom model described in Sec. 2.1. The output consisting of company entities (tag IF) detected by the Nametag model is passed to the Target Detector module (Sec. 3.4)

3.4 Target Detector

Second step of the system is called target detector. We are mapping company names detected with the entity detector (Sec. 3.3) to the corresponding domains (entity URL). Companies may be referred by many names, our goal is to normalize all company names to their canonical URL which may be used as an input to the next part of the algorithm (Sec. 3.5).

We built the mapping from the national registry of companies and added 50 selected international companies by hand. We are using entity name expansion (adding suffixes like ”co.”) because the official name of the company contains them, but it may be omitted in the email communication.

[2]https://github.com/tivvit/phishing-ner-lda

3.5 Domain Link Profile

Domain (company) link profile creation is an essential part of the phishing detection. We did choose 20 candidate domains which are commonly attacked. This list was created based on Phishtank reports and internal database of historic reports (OpenDNS,). The advantage of this approach is that new domains may be added (by hand or automatically) and it significantly lowers false positive detections.

For each of those domains, following steps are executed to build the link profile.

1. Take emails signed with DKIM (Kucherawy et al., 2011) matching to the domain of the analyzed company from the historic traffic (1 week in our case). This ensures that only emails legitimately sent from the company are used for the profile creation.
2. Extract all domains linked from those emails. Linked means present in the href attribute in the HTML body of the email.
3. Filter out the long-tail.

The detection part of this system is using the built domain link profile. It gets a list of links (URLs) from the email body and the detected domain for that email. If the DKIM signature matches the detected domain, the email is considered as valid (Kucherawy et al., 2011). If there is no signature or it does not match the module continues with the analysis. It extracts domains from the email body links and checks if they are present in the detected company profile, if not the email is considered as phishing.

4 DATA

This section describes datasets used for training and evaluation of the presented models.

4.1 Public Phishing Datasets

Most of the email phishing detectors are tested with the publicly available datasets.

The standard dataset of positive phishing samples is (Nazario,). This dataset consists of 4450 emails sent from 2004 to 2007. This dataset is no longer

available online, we published it again at Academictorrents[3] platform.

It is complicated to publish negative phishing samples because of the email private nature. Commonly used is the (spa,) dataset. It is a collection of ham and spam messages collected from 2002 to 2005. It consists of 6047 messages.

Most of the messages in those datasets are in the English language. To our knowledge, there are no publicly available phishing datasets for the Czech language.

4.2 Email.cz Named Entity Recognition Dataset

We have created a dataset of annotated entities from email conversations. This dataset was created for the internal use of Email.cz. It is based on 2 million email texts dumped from live traffic. Those emails were preselected and no personal messages were used in this dataset. Those messages were annotated by 4 people. We used entity tags based on Cnec 2.0 (Ševčíková et al., 2007), but used only 39 of the 45 original tags. We created this dataset because of poor results (described in Sec. 5.1) of the original model on email data. This dataset (annotated part) consists of 54724 sentences and 125711 tags.

5 EXPERIMENTAL RESULTS

In this section, we present results of NER on Email.cz dataset and also phishing detection models on publicly available datasets and Email.cz live traffic.

5.1 NER Models

We tested Nametag model (Straková et al., 2014) (trained on Cnec 2.0 (Ševčíková et al., 2007)) and our Nametag based model (trained on internal dataset (described in Sec. 4.2)) on testing data separated from the Email NER 2018-04 dataset (described in Sec. 4.2). Results for those models are shown in Tab. 1. Reported Nametag results for Cnec 2.0 is 77.35% F-score (for 45 tags) (Straková et al., 2014).

Table 1: Results for NER models on Email.cz dataset.

Model	Precision	Recall	F-score
Cnec 2.0	8.81	14.82	11.05
Email.cz	77.31	80.58	78.91

[3]http://academictorrents.com/details/\a77cda9a9d89a60dbdfbe581adf6e2df9197995a

Based on the results shown in Tab. 1 we used Email.cz NER model as the base for the other experiments. We assume that the low performance of the Cnec 2.0 based model is caused by the different language nature used in an email (informal speech) and Cnec 2.0 (based on news articles).

Because the domain link profile model (Sec. 3.2) relies on NER correctly detecting companies (IF tag) we have also done evaluation only for that tag in Tab. 2.

Table 2: Results for NER models on Email.cz dataset (only companies - IF tag).

Model	Precision	Recall	F-score
Cnec 2.0	26.57	21.11	23.53
Email.cz	75.16	75.67	75.41

5.2 Evaluation for Publicly Available Phishing Datasets

Results for our implementation of the phishing random forests classifier may be seen in the Tab. 3. The test was performed on randomly chosen 20% emails (testing sample not used for the training) from (Nazario,) dataset described in Sec 4.1.

Table 3: Results for phishing random forest classifier on publicly available dataset.

Metric	Ham	Phishing	avg / total
Precision	0.93	0.99	0.94
Recall	1.00	0.75	0.94
F-score	0.96	0.85	0.94
Support	1400	447	1847

Domain link profile based detection is not evaluated on publicly available datasets because it needs to build domain link profile, which would mean overfitting the dataset. Present domain link profile (built form Email.cz production data) also cannot be used because many of the targets in the dataset does not exist anymore or their profile changed significantly over the years.

5.3 Phishing Detection Results for Email.cz Traffic

Both solutions were used for analysis of the small portion of live traffic at Email.cz. Systems were evaluated for two days. 132000 messages were analyzed during that period.

Those emails were annotated by hand. Some of the malicious URLs were detected by Phishtank (OpenDNS,). Most of the analyzed emails are not

Table 4: Phishing detection results for Email.cz traffic.

Model	Precision	Recall	F-score	Support
Random forest	0.0002	1.00	0.0004	132000
Random forest filtered	0.33	1.00	0.50	77
Link profile	0.88	1.00	0.93	77
Combined	1.00	1.00	1.00	77

unique, therefore the annotation was simpler. 7 phishing attacks were found in this corpus although only 4 of the attacks were unique. We can see that the portion of the attacks is 0.0053%.

Results may be seen in 4. We can see that proposed solutions are able to detect all attacks. Random forest classifier (Sec. 3.1) itself achieves very low precision. When we filter only domains which are commonly attacked by phishing attacks (described in Sec. 3.5) the model achieves much better results.

In comparison model based on domain link profile (described in Sec. 3.2) is even more precise. Best result was achieved with the combination of both models.

6 CONCLUSIONS AND FUTURE WORK

We evaluated two phishing detection systems. First one is using random forest classifier with features extracted by NER and LDA and was proposed in (Ramanathan and Wechsler, 2013). The second solution is new in this work. This solution is based on NER detecting organizations. Then it is comparing company URL profile to the URL profile present in the email.

Both solutions were evaluated in live traffic of a freemail provider Email.cz. The first solution achieved 50% F-measure with 100% recall. Our proposed solution achieved 100% recall with 93% F-measure. Combination of the methods achieved 100% F-measure. None of the detected phishing attacks was detected by system currently used at Email.cz.

We optimized the solution to work also for the Czech language but over the testing period, there were only attacks in English. We suggest running this test for a longer period, which will generate more significant result. The detected attacks should be reported to Phishtank in the future (OpenDNS,). This system should label and store phishing attacks and create a multi-language public dataset which is currently missing.

REFERENCES

Spamassassin corpus. http://spamassassin.apache.org/publiccorpus/. [Online; accessed 2015-02-01].

Jenny Rose Finkel, T. G. and Manning, C. (2005). Stanford ner - incorporating non-local information into information extraction systems by gibbs sampling. urlhttp://nlp.stanford.edu/ner. [Online; accessed 2018-09-30].

Kucherawy, M., Crocker, D., and Hansen, T. (2011). DomainKeys Identified Mail (DKIM) Signatures. RFC 6376.

Nazario, J. Phishing corpus. http://monkey.org/~jose/wiki/doku.php?id=phishingcorpus. [Online; accessed 2015-02-01].

OpenDNS. PhishTank. urlhttp://www.phishtank.com. [Online; accessed 2018-09-30].

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ramanathan, V. and Wechsler, H. (2013). Phishing detection and impersonated entity discovery using conditional random field and latent dirichlet allocation. *Computers & Security*, 34:123–139.

Ševčíková, M., Žabokrtskỳ, Z., and Krůza, O. (2007). Named entities in czech: annotating data and developing ne tagger. In *International Conference on Text, Speech and Dialogue*, pages 188–195. Springer.

Straková, J., Straka, M., and Hajič, J. (2014). Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland. Association for Computational Linguistics.