Data Acquisition and Corpus Creation for Phishing Detection

I. Dunđer*, S. Seljan* and M. Odak**
* Faculty of Humanities and Social Sciences, University of Zagreb, Department of Information and Communication Sciences, Zagreb, Croatia
** Faculty of Humanities and Social Sciences, University of Mostar, Department of Information Sciences, Mostar, Bosnia and Herzegovina
ivandunder@gmail.com, sanja.seljan@ffzg.hr, marko.odak@sum.ba

***Abstract* - Detecting phishing attacks is not straightforward, since there are many obstacles that derive from language complexity and technical aspects. Studying phishing attacks and other related issues heavily relies on computer datasets, i.e. digital corpora that reflect these linguistic and technical intricacies. Diverse studies using phishing datasets have been performed, but mainly for the English language. Research for other languages is scarce, and especially for not widely spoken languages. For the Croatian language there is an evident lack of corpora that are essential for diverse analyses and for constructing models that are capable of recognizing phishing attacks and protecting users. These datasets are necessary for natural language processing and building machine learning workflows, where results largely depend on corpora that must be specifically crafted for this purpose. Therefore, creating high-quality domain-specific corpora is of great importance in the domain of information security. Such corpora can be employed for teaching purposes in various courses in higher education, and could be analyzed in numerous ways in order to understand the underlying principles of phishing attack strategies. The aim of this paper is to demonstrate the entire process of data acquisition and corpus creation for the phishing detection domain. In addition, an analysis of the corpus is presented with regard to different aspects, such as descriptive attributes, terminology characteristics, metadata and language.**

Keywords - data acquisition; digital corpus creation; computational data analysis; natural language processing; phishing; information privacy; information security

I. INTRODUCTION

A phishing attack is a type of social engineering in which an attacker poses as a reliable entity in order to deceive a victim into disclosing sensitive information [1], including login passwords, personal data or financial information [2]. Phishing detection, on the other hand, is the effort to recognize and stop phishing attacks on people and business enterprises, and as such represents a crucial endeavor in the field of information and cyber security [3].

Nowadays, phishing attacks can be assessed in a variety of ways, as the idea is to gauge how well phishing attacks and phishing detection methods work [4]. However, detecting phishing attacks is not an easy task, since there are many difficulties that arise not only from linguistic complexity and diversity, but also from technical and technological aspects [5].

In the process of analyzing and evaluating the effectiveness of phishing attacks and phishing countermeasures, computer datasets that contain a large number of phishing messages play a significant role [6]. These datasets, also known as digital corpora, are purposefully crafted and prepared in order to account for these linguistic and technical nuances.

The importance of digital corpora in the fields of natural language processing (NLP), data science, artificial intelligence (AI) and especially machine learning (ML), cannot be overstated, as they enable computers (machines) to comprehend, interpret and manipulate data, but also to interact with human languages. All these fields offer various techniques and methods that are suitable for detecting phishing attacks [7].

Phishing datasets include electronic messages, such as e-mails that have been previously marked as malicious or suspicious [8]. They can be fed into security systems that use machine learning so that, based on prior observations, they can learn to differentiate between valid and illegitimate messages [9]. In that way, digital corpora help to determine the level of effectiveness of phishing detection technologies and approaches, which can then be utilized to implement any necessary adjustments in security systems.

II. MOTIVATION

Timely analysis and comprehension of incoming data, the discovery of trends and changes in data content and communication volume can support the development of security campaigns in business enterprises.

Processing enormous amounts of data from many sources, such as e-mails, social media platforms, internet portals, product reviews, web feeds or news articles, is one of the main advantages of computers and modern technology, such as natural language processing and machine learning [10].

Digital corpora are crucial for the study of phenomena in human languages, as they enable researchers access to a vast and varied collection of textual data, which can be employed for building machine learning models that are

applicable to several tasks [11], like text filtering, classification, clustering, summarization, sentiment analysis, language translation etc. These models would be significantly less accurate and practical without access to a wide range of digital corpora.

Implementing phishing detection requires extensive training and education on how to spot phishing attacks and how to react to them [12, 13]. This knowledge, however, derives from several disciplines, such as computer science, information science, computational linguistics, statistics etc. Therefore, such specially crafted corpora can also be used for teaching purposes in various courses in higher education [14], where they could be examined in numerous ways in order to comprehend the underlying fundamentals of phishing attack methods and strategies.

III. Related Work

Diverse studies have been carried out using phishing datasets, but primarily for the English language. A recent paper has presented two dataset variations that consist of almost 150,000 websites labeled as legitimate or phishing, which allow training of classification models, building phishing detection systems, and mining association rules [15].

Another research explained key dimensions of data quality relevant for security, illustrated them with several popular datasets for phishing, intrusion detection and malware, and presented operational methods for assuring data quality in datasets for security challenges [16].

One paper addressed the absence of benchmark datasets for phishing detection. The authors stated that this is due to the fact that phishing websites are short-time living and dead URLs cannot be used in content-based analysis. A dataset was constructed by following a set of proposed guidelines, and afterwards it was utilized for assessing the level of effectiveness in systems based on the random forest classifier algorithm [17].

Another paper presented datasets and tools that are helpful for researching phishing. The authors describe the problem of creating high-quality, diverse and representative phishing datasets. They also discuss the problems of datasets that already exist on the market. Then a benchmarking framework is proposed that automates the extraction of more than 200 features, that implements more than 30 classifiers and a dozen evaluation metrics for the detection of phishing [18].

Ref. [19] suggested a technique that analyzes ULRs, extracts several features, reduces the problem dimensionality, and that when integrated with the support vector machine classifier showed high efficiency in phishing detection. The proposed approach exhibited good results when benchmarking with a range of standard phishing datasets.

Research for other languages is performed less, particularly for those that are not commonly spoken. For the Croatian language there is an evident lack of recent research in the field of phishing detection, and especially in research of corpora that are essential for conducting analyses and for building models that are capable of identifying phishing messages, stopping attacks and protecting users.

One Croatian research has focused on examining the familiarity of users in Croatia with threats in form of social engineering and phishing attacks. Here a practical assessment of users' capabilities to identify phishing attacks was conducted. Also, the research investigated the potential features that are problematic for users when trying to detect phishing attacks [20].

Another Croatian research examined the various types of phishing attacks, such as e-mail phishing, instant messaging phishing, smishing, bulk phishing, spear phishing, whaling, vishing and pharming. The paper also proposed machine learning algorithms suitable for phishing detection, such as decision trees, random forest, support vector machine, k-nearest neighbors etc. [9].

One approach that could be used for discerning phishing messages was presented in a recent Croatian study. The paper focused on the suitability of online services for machine learning of models for predicting classification outcomes [21].

IV. Research, Results and Discussion

Since phishing messages are written in human languages, it is not surprising that creating high-quality and domain-specific textual corpora is of great importance for various analyses in security-related domains and protective strategies.

The aim of this section is to demonstrate the entire process of acquiring data and creating a digital corpus for the task of phishing detection, followed by content analysis using text mining techniques [22] for topic detection. In addition, the resulting corpus is presented and discussed with regard to descriptive attributes, terminology characteristics, metadata and language.

A. Dataset Acquisition and Corpus Creation

The initial dataset was acquired and crafted by extracting data that derived from several personal official e-mail accounts. It consists of 520 phishing e-mails written in Croatian and English dating from the year 2022. In addition to the 520 phishing e-mails, there were some phishing e-mails that were written in other languages, such as French, German and Spanish, but they were less represented and therefore discarded.

The dataset was divided into two subsets containing e-mails that were manually checked and verified as phishing:

- **Set A** which contains 260 e-mails originally written in Croatian (not necessarily grammatically correct), and
- **Set B** which contains 260 e-mails originally written in English and translated automatically into Croatian with Google Translate and without any post-editing.

Each subset contains some duplicates in order to reflect the frequency of certain topics of phishing messages. The initial dataset was pre-processed which

included operations, such as lowercasing, removing accents and numbers, and text filtering.

A stop word list of 290 words was created, consisting of exclamations, conjunctions, pronouns, auxiliary verbs, particles (so-called function words) etc. in order to obtain a corpus of semantically full words, and to filter out all the redundant words.

The resulting phishing corpus consists exclusively of e-mails with the corresponding information and text body, whereas attached files, such as executables, PDF files and textual documents were discarded. Descriptive attributes of the phishing corpus and metadata are presented in Table I.

TABLE I. DESCRIPTIVE ATTRIBUTES AND METADATA

Attributes	Set A	Set B
Sentence number	1,471	939
Token number	37,886	21,427
Word number	31,532	17,660
Mentioning e-mail address(es)	68	51
Containing web link(s)	48	95
Containing attached file(s)	12	6
Containing image(s)	2	5
To field: *personal*	45	85
To field: *undisclosed-recipients*	167	48
To field: *other or not available*	48	67

When comparing the two subsets, Table I presents information on the number of phishing e-mails sent directly to personal e-mail accounts (more than 40 in Set A and more than 80 in Set B), undisclosed recipients (almost 170 in Set A and almost 50 in Set B) and when the recipient was obfuscated (almost 50 in Set A and almost 70 in Set B).

It also shows the number of phishing e-mails that mentioned other e-mail addresses in the text itself (more present in Set A), that contained web links in the body, attached files or images. It is evident that Set A contained more attachments and is larger in terms of sentence, token and word number, whereas Set B contains more web links (95) and images (5). Overall, Set A and Set B consist of almost 60,000 tokens, divided into 2410 sentences, i.e. 1471 in Set A and 939 in Set B.

B. Frequency of tokens

Word clouds as a method for analyzing frequencies of tokens are very helpful, as they offer a visual depiction of the most commonly used terms in a given text or group of texts, which makes it simpler to comprehend and swiftly evaluate important themes and subjects in large datasets [23].

Fig. 1 presents a word cloud with the 50 most frequent tokens from Set A (after pre-processing). The word cloud shows semantic words related to financial fraud, personal data, credit cards, lottery wins, business offers, fundings and bank payments.

Figure 1. Word cloud for Set A

Fig. 2 presents a word cloud with the 50 most frequent tokens from Set B (after pre-processing). The word cloud reveals tokens and semantic words that are related to obscene and indelicate content, e.g. girls, craigslist, service, teenagers, hot, cute, dating, escort, secret, night etc.

Figure 2. Word cloud for Set B

C. Frequent Terminology

The identification of frequent terminology enables more fine-grained evaluation with the distinct aim of differentiating the distribution of words and concepts from the field of phishing attacks. Applying such a straight-forward method allows researchers to interpret the structure of phishing messages with their unique style, vocabulary and specific terminology. This can provide a deeper, objective, unbiased and consistent insight into the ways of crafting phishing messages.

However, such analyses are altogether based on word occurrences and this approach yields problems related to the qualitative spectrum of data analysis [24]. Therefore, it is important to warrant an investigation of concordances that can explain the diverse contexts of such words in a phishing message.

Two word lists containing the most frequent meaningful lemmas of nouns and verbs, reflecting the content of e-mails from Set A and Set B are given in Table II.

TABLE II. MOST FREQUENT TERMINOLOGY

Set A		Set B	
Nouns	**Verbs**	**Nouns**	**Verbs**
banka	kontaktirati	djevojka	željeti
fond	moći	e-pošta	tražiti
e-pošta	poslati	stranica	razgovarati
ime	moliti	račun	podijeliti
račun	dobiti	pratnja	poslati
adresa	željeti	usluga	kontakirati
sredstvo	primiti	pozdrav	moliti
broj	odgovoriti	transakcija	pogledati
podatak	pomoći	url	pozivati
pozdrav	dati	mreža	pomoći
kartica	potvrditi	trošak	koristiti
iznos	odlučiti	aplikacija	kliknuti
plaćanje	obavještavati	datoteka	dati
dolar	učiniti	poziv	javiti
zemlja	kliknuti	zajednica	upoznati
milijun	prijaviti	veza	čekati
novac	donirati	novac	znati
informacija	dijeliti	dolar	primiti
poruka	osvojiti	nsa	morati

The extracted word lists presented in Table II display the content of phishing e-mails. Word list from Set A contains mostly nouns and vocabulary related to financial fraud (*banka – bank*, *fond – fund*), valuable data (*račun – account*, *adresa – address*, *broj – number*), verbs reflecting communication of receiving and giving, sending, helping, replying, and mentions e-mails as a means of communication (*e-pošta – e-mail*, *podatak – data*, *sredstvo – medium*/*way*).

Word list from Set B contains mostly obscene vocabulary (*djevojka – girl*, *usluga – service*, *pratnja – escort*), mentions websites and e-mails as means of communication (*e-pošta – e-mail*, URL, *mreža – network*), transaction-related data (*račun – account, transakcija – transaction*), and greetings at the end of messages (*pozdrav – greetings*). Verbs reflect more a polite tone of communication, using words, such as *željeti – wish*, *razgovarati – talk*, *moliti – beg*, *pomoći – help*, *tražiti – ask*, *podijeliti – share*, *poslati – send*.

D. N-gram analysis

In order to detect how and what words are frequently combined in a language, collocations can be used. They are commonly applied in various natural language processing tasks, text mining and linguistic analyses.

Collocations can be examined through n-grams, which typically represent a sequence of *n* words in a text. They can help to capture the context and understand word semantics.

In this research, the focus is on bigrams and trigrams. Bigrams are sequences of two words that appear together in a text, and can be employed for identifying common word combinations and text patterns. A trigram is a sequence of three words and can be used to find more intricate structures and patterns in a text.

When it comes to Set A, the most frequent bigrams are related to financial fraud, personal data and personal communication, such as *poštovani korisniče – dear user*, *bankovni račun – bank account*, *banka africa – bank Africa*, *afrička unije – African Union*, *broj telefona – phone number*, *kartice adresu – card address*, *atm kartica – ATM card*, *pošaljite informacije – send information*, *godine spol – years gender*, *milijun dolara – million dollars*, *pošaljite podatke – send data*, *visa kartice – Visa cards*, *prijenos sredstava – money transfer*, *najbliža rodbina – close relatives*, *odmah odgovorite – answer immediately* etc. A few expressions in Set A are also written in English, e.g. *united bank – united bank*.

Trigrams reveal the same content as bigrams, but provide a broader context, such as *united bank africa – united bank Africa*, *atm kartice adresa – ATM card address*, *telefon godine spol – phone years gender*, *pošaljite podatke visa – send Visa data*, *iznos milijun tisuća – amount million thousands*, *poštovani korisniče računa/pošte – dear account/mail user*, *odgovorite united bank – answer united bank*, *međunarodni monetarni fond – International Monetary Fund* etc.

Most frequent bigrams in Set B are primarily related to obscene content, such as *zgodna pratilja – handsome companion*, *povremeni spojevi – occasional dates*, *craigslist zakona – craigslist law*, *tajna zajednica – secret community*, *nabaviti zgodne – get a pretty*, *usluga pratnje – escort service*, *noć povremena – occasional night* etc.

Trigrams in Set B are, e.g. *noć povremene spojeve – night occasional dates*, *odjeljku craigslist zakona – craigslist law section*, *rastuća tajna zajednica – growing secret community*, *nabavite zgodne djevojke – find hot girls*, *djevojka na poziv – call girl*, *tajna zajednica nsa – secret NSA community* etc.

E. Context analysis

Concordances are very useful when data analysis requires examining and comprehending context in which a certain word, n-gram or phrase is used in a text. This helps to spot patterns and trends in a text, such as how a term is used repeatedly or in distinct contexts, how word collocations are formed, or when and what words frequently follow each other in a text.

In this research a few keywords in context (KWIC) [24] were chosen. Table III shows concordances from Set A and Set B with regard to four keywords (KWIC) with responding contexts: odmah – immediately, *sada – now*, *razgovarati – talk*, *podijeliti – share*. The integer in parentheses indicates the number of occurrences of a particular word.

TABLE III. CONCORDANCE ANALYSIS

Set A	Set B
odmah (44) – immediately	**odmah (5) – immediately**
savjetujemo da ***odmah*** pošaljete svoje ime, kontakt adresu i broj mobilnog telefona.	ljubazno mi odgovorite ***odmah***. Hvala.
ako budem plaćen, ***odmah*** ću uništiti video	Odgovorite mi i nazovite me da ***odmah*** dobijem besplatnu ponudu
(MMF) Afričke regije ***odmah*** unutar sljedećih 168 sati.	Dopustite mi da počnem ***odmah*** od točke
Napomena: ***odmah*** po zaključku transakcije imate pravo na 45%	kliknite ovdje za nadogradnju i ***odmah*** ponovno potvrdite svoj račun
kontaktirajte me ***odmah*** za daljnju komunikaciju	***Odmah*** potvrdite svoj interes
sada (60) – now	**sada (41) – now**
Sada kontaktirajte Službu za korisnike UBA banke	Chatajte ***sada*** (samo unesite svoju e-poštu)
Sada vaše novo plaćanje, broj odobrenja	Možemo li ***sada*** uspostaviti video poziv
Sada želim da hitno kontaktirate	Provjerite ***sada*** gdje možete
Sada se obratite generalnom direktoru	**Sada** sam spreman za poziv
Sada sam u skrovištu samo da zaštitim svoj život i	***Sada*** dostupno WhatsApp Facebook Instagram
razgovarati (10) – talk	**razgovarati (44) – talk**
Mogu li ***razgovarati*** s vama putem ove e-pošte?	pregledajte našu stranicu i **razgovarajte** -> https:
kako bismo ***razgovarali*** o mojoj hitnoj potrebi za nasljednikom	privatne fotografije ***Razgovarajte*** sada (samo unesite svoju e-poštu)
investicijskog projekta o kojem smo ***razgovarali***	želite li **razgovarati** ?? Ako se slažete, pošaljite mi svoje ime, dob
podijeliti (10) – share	**podijeliti (44) – share**
kako bi mi pomogli ***podijeliti*** ovih 5,5 milijuna dolara	***podijelila*** je datoteku s vama
spreman sam ***podijeliti*** novac s vama.	***Podijelit*** ću slike i više detalja o sebi čim mi se javite
novac vraćen od prevaranata ***podijeli*** među 100 sretnih ljudi	povjeriti vama i ***podijeliti*** s vama ovaj povjerljivi posao.
ovaj će se posao ***podijeliti*** u ovom omjeru	bih želio ***podijeliti*** s tobom ako samo možeš odgovoriti na moju poštu

Many phishing e-mails ask for an urgent response, and therefore oftentimes contain a call for an emergency action. Such messages include the words *odmah – immediately* and *sada – now*. This is especially noticeable in Set A (originally written in Croatian), where these words appear 44 and 60 times, respectively.

However, there are some differences in the use of such words between the two subsets. Urgency is more present in Set A, which is more focused on financial fraud. On the other hand, most phishing messages from Set B (phishing e-mails originally written in English and translated into Croatian) are in the style of begging for help or asking for a conversation, hence the of use of verbs, such as *razgovarati – talk* and *podijeliti – share* (both used more than 40 times). This was accompanied by a call to click on links that were integrated into the text body of e-mails (95 times out of 260 e-mails).

V. CONCLUSION

This paper presents the process of data acquisition and corpus creation for the phishing detection domain. Collected phishing e-mails were divided into two subsets of data: e-mails originally written in Croatian (Set A), and e-mails translated from English into Croatian (Set B). Both subsets were evaluated in a number of ways, e.g. by analyzing n-grams, collocations, and frequencies of e-mails and web links that were integrated into phishing e-mails.

In this dataset, phishing e-mails written in Croatian were predominantly sent to undisclosed recipients, whereas those originally written in English were mostly sent to personal e-mails. The authors assume that these e-mail addresses were probably extracted from public research papers, presentations or speeches.

Phishing e-mails originally written in the Croatian language refer more to other e-mail addresses as a malicious means of connecting with victims, whereas those originally written in English contain more web links, enticing potential victims to click on them and instructing victims to disclose private or sensitive information.

Set A reveals content that is related to finance, personal data, bank details, business offers, funding and payments, while Set B contains more mature and obscene content.

Bigrams and trigrams show collocations belonging to the same domains. Phishing e-mails that are related to financial transactions and credit cards are often marked as urgent, and ask for an immediate response (frequent use of terms *odmah – immediately* and *sada – now*). Phishing messages that are, however, related to indecent content are more in a kind and begging tone, not focusing on urgency. Here the intention is to lure victims into sharing data, either by clicking on web links that are integrated into the text body, or by inviting them to friendly conversations.

For future research, the authors of this paper plan to increase the volume of the phishing dataset, and then to categorize messages into different topic classes, to examine the sentiment that is present in phishing messages, and to apply machine learning algorithms for predicting various classes of phishing e-mails for the Croatian language. In addition, the authors intend to collect phishing e-mails in other languages, and then follow similar analyses and processes outlined in this paper in order to explore the common characteristics that occur across various languages, and in order to obtain

valuable insights into how to improve phishing models for Croatian.

REFERENCES

[1] Australian Cyber Security Centre, "ACSC Annual Cyber Threat Report: July 2019 to June 2020", https://www.cyber.gov.au/sites/default/files/2020-09/ACSC-Annual-Cyber-Threat-Report-2019-20.pdf

[2] M. Dadkhah, S. Shamshirband, and A. Wahab, "A hybrid approach for phishing web site detection", The Electronic Library, vol. 34, no. 6, pp. 927–944, 2016.

[3] M. Rosenthal, "Must-Know Phishing Statistics: Updated 2022", Tessian, https://www.tessian.com/blog/phishing-statistics-2020/

[4] Z. Alkhalil, C. Hewage, L. F. Nawaf, and I. Khan, "Phishing Attacks: A Recent Comprehensive Study and a New Anatomy", Frontiers in Computer Science, vol. 3, article 563060, 2021.

[5] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges", Multimedia Tools Applications, vol. 82, pp. 3713–3744, DOI: https://doi.org/10.1007/s11042-022-13428-4, 2023.

[6] A. ALmomani, T.-C. Wan, A. Manasrah, A. Altaher, M. Baklizi, and S. Ramadass, "An enhanced online phishing e-mail detection framework based on "Evolving connectionist system"", International Journal of Innovative Computing, Information and Control (IJICIC), vol. 9, no. 3, pp. 1065–1086, 2013.

[7] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from URLs", Expert Systems with Applications, vol. 117, pp. 345–357, 2019.

[8] W. A. Awad, and S. M. ELseuofi, "Machine Learning methods for E-mail Classification", International Journal of Computer Applications, vol. 16, no. 1, pp. 39–45, 2011.

[9] A. Kovač, I. Dunđer, and S. Seljan, "An overview of machine learning algorithms for detecting phishing attacks on electronic messaging services", 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO), Croatia, 2022, pp. 954–961, DOI: 10.23919/MIPRO55190.2022.9803517.

[10] R. H. Hariri, E. M. Fredericks, and K. M. Bowers, "Uncertainty in big data analytics: survey, opportunities, and challenges", Journal of Big Data, vol. 6, no. 44, DOI: https://doi.org/10.1186/s40537-019-0206-3, 2019.

[11] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text Classification Algorithms: A Survey", Information, vol. 10, no. 4, DOI: https://doi.org/10.3390/info10040150, 2019.

[12] M. Adil, R. Khan, and M. A. Nawaz Ul Ghani, "Preventive Techniques of Phishing Attacks in Networks", 3rd International Conference on Advancements in Computational Sciences (ICACS), Pakistan, 2020, pp. 1–8, DOI: 10.1109/ICACS47775.2020.9055943.

[13] A. Sumner, and X. Yuan, "Mitigating Phishing Attacks: An Overview", ACM Southeast Conference (ACM SE '19). Association for Computing Machinery, USA, 2019, pp. 72–77, DOI: https://doi.org/10.1145/3299815.3314437.

[14] R. Jaworski, S. Seljan, and I. Dunđer, "Towards educating and motivating the crowd – a crowdsourcing platform for harvesting the fruits of NLP students' labour", 8th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, vol. 1, 2017, pp. 332–336.

[15] G. Vrbančič, I. Jr. Fister, and V. Podgorelec, "Datasets for phishing websites detection", Data in Brief, vol. 33, ISSN 2352-3409, DOI: 10.1016/j.dib.2020.106438, 2020.

[16] R. M. Verma, V. Zeng, and H. Faridi, "Data Quality for Security Challenges: Case Studies of Phishing, Malware and Intrusion Detection Datasets", ACM SIGSAC Conference on Computer and Communications Security (CCS '19). Association for Computing Machinery, USA, 2019, pp. 2605–2607, DOI: https://doi.org/10.1145/3319535.3363267.

[17] A. Hannousse, and S. Yahiouche, "Towards benchmark datasets for machine learning based website phishing detection: An experimental study", Engineering Applications of Artificial Intelligence, vol. 104, ISSN 0952-1976, DOI: https://doi.org/10.1016/j.engappai.2021.104347, 2021.

[18] V. Zeng, S. Baki, A. El Aassal, R. Verma, L. F. Teixeira De Moraes, and A. Das. "Diverse Datasets and a Customizable Benchmarking Framework for Phishing", Sixth International Workshop on Security and Privacy Analytics (IWSPA '20). Association for Computing Machinery, USA, pp. 35–41, DOI: https://doi.org/10.1145/3375708.3380313.

[19] J. Rashid, T. Mahmood, M. W. Nisar, and T. Nazir, "Phishing Detection Using Machine Learning Technique", First International Conference of Smart Systems and Emerging Technologies (SMARTTECH), Saudi Arabia, 2020, pp. 43–46, DOI: 10.1109/SMART-TECH49988.2020.00026.

[20] J. Andrić, D. Oreški, and T. Kišasondi, "Analysis of phishing attacks against students", 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Croatia, 2016, pp. 1423–1429, DOI: 10.1109/MIPRO.2016.7522363.

[21] I. Zatezalo, and I. Dunđer, "Online service for accessible machine learning of prediction models", Zbornik radova Međimurskog veleučilišta u Čakovcu, vol. 12, no. 2, p. 10, 2021.

[22] M. Pejić Bach, Ž. Krstić, S. Seljan, and L. Turulja, "Text Mining for Big Data Analysis in Financial Sector: A Literature Review", Sustainability, vol. 11, no. 5, pp. 1-27, DOI: 10.3390/su11051277, 2019.

[23] M. A. Hearst, E. Pedersen, L. Patil, E. Lee, P. Laskowski, and S. Franconeri, "An Evaluation of Semantically Grouped Word Cloud Designs", in IEEE Transactions on Visualization and Computer Graphics, vol. 26, no. 9, pp. 2748–2761, 2020, DOI: 10.1109/TVCG.2019.2904683, 2020.

[24] M. Pavlovski, and I. Dunđer, "Is Big Brother watching you? A Computational Analysis of Frequencies of Dystopian Terminology in George Orwell's 1984", 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Croatia, 2018, pp. 638–643, DOI: 10.23919/MIPRO.2018.8400120.