INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 8, August 2026 | ISSN: 2456-4184 | IJNRD.ORG

A Systematic Review of Lexical Analysis Techniques for Phishing Website Detection: Datasets, Methodologies, and Open Challenges

[1]Lalit Upadhyay, [2]Rahul Savla, [3]Dr. Shivkumar Goel

[1]Research Student, [2]Research Student, [3]Head of Department
[1] Department of Computer Application Vivekanand Education Society's Institute Of Technology,
[1]Mumbai, India

Abstract : Phishing remains one of the most common and economic cybersecurity problems due to its exploitation of the human element via social engineering. Though conventional ways use reactive methods with the help of blacklists, the current atten-tion in academia is devoted to proactive approaches with the use of machine learning techniques. Among those techniques, there is the lexical approach when a Uniform Resource Locator (URL) is analyzed but not the page. The paper gives the comprehensive review of the available information on lexical-based phishing detection. Analyzing popular datasets and methodologies of machine learning algorithms, we reveal major gaps in this area. Our findings show that, despite the claimed high accuracy of the results by the authors, there is a lot of space to improve and investigate further. The main problem is lack of reliable benchmark datasets for this task 98.

I. INTRODUCTION

The development of digital services, cloud computing, and e-commerce has transformed modern global communication significantly. Correspondingly, the increasing dependency on the internet has led to increasing vulnerability to cyber-fraud. One of the most common types of the cyber-fraud in which criminals disguise themselves as trusted people and try to steal the confidential data (login/passwords, bank credentials) is called phishing. Nowadays, the phishing attacks include spear-phishing, smishing, and dynamic web hosting.

For a long period of time, the anti-phishing strategy based on the list of malicious websites (blacklist/whitelist) was implemented in Google Safe Browsing and other similar com-panies. These solutions work effectively against discovered threats but suffer from the one significant drawback. They do not protect the users from "zero-day" phishing domains because the lists are asynchronous and take a certain amount of time to be created.

To solve this problem, the content-based machine learning approaches were proposed. Even though they were highly pre- digital.

II. LITERATURE REVIEW

Evaluation of the current state of the academic research reveals that the current approach involves using complex eco-systems of open datasets, various feature extraction method-ologies, and constantly improving machine learning algo-rithms.

A. Existing Dataset Assessment

The effectiveness of machine learning algorithms relies greatly on the training datasets they are using. In case of malicious URLs, researchers use two sources - PhishTank and OpenPhish. Although plentiful, earlier studies revealed that PhishTank dataset can be noisy and contain offline or be-nign websites from time to time. Benign URLs are gathered through scraping of the Alexa Top 1 Million (Tranco list), or Common Crawl corpus

B. Lexical Feature Extraction Methodologies

The task of lexical feature extraction involves conversion of raw strings into numerical vectors. They can be classified into different domains in the literature as follows: Structural Features: Probably, the most obvious feature of malicious URLs is their structural anomalies. The length of the URL appears to be an extremely important characteristic, as the attacker uses long URLs to hide the path off-screen. Special Characters: Some special characters in the URL appear to play a great role. For example, the legitimate URL will never have @ sign, which is utilized by the attacker to disguise his target. Hyphens (-) are often used in typo-squatting attacks (chase-bank-login.com). Semantic Tokens: The researchers conduct light NLP analysis of the URL path and seek the following suspicious tokens: login, secure, verify, account, and update.

III. PROPOSED SYSTEM

A. Problem Statement

Even though good results can be reached via the usage of machine learning tech-niques, there are some problems that arise when dealing with contemporary phish-ing data that prevent the application of such techniques. The first problem is called dataset decay. It means that phishing websites are usually eliminated immediately after being identified. Thus, it is very difficult for researchers to conduct experi-ments on currently existing datasets because most of them contain broken links. Another challenge connected with working with phishing data is the problem of imbal-ance in the distribution of classes. In real life, most of the websites that are vis-

ited by people are safe. However, researchers create a balance between classes in laboratory settings, i.e., they include both malicious and safe URLs in equal propor-tion (50%). This leads to an impression that machine learning algorithms have extremaly high accuracy, although it works only under controlled conditions and gives too many false positives in real life. Finally, there is a problem of extremely fast evolu-tion of methods used by attackers to bypass detection. For instance, they use URL shortening, homograph attacks, and typo squatting. Moreover, recent methods of detection based on deep learning algorithms function as black boxes and do not ex-plain why a certain URL has been blocked.

B. Scope

The purpose of this literature review is to investigate the ap-plication of lexi-cal analysis techniques to phishing detection. The aim of this paper is to Analyze the quality and reliability of publicly available datasets. Research effective ways to con-struct lex-ical features. Compare machine learning and deep learning methods. Identify re-search gaps and suggest future research directions. The scope of this paper is focused on client-side phishing detection techniques using lexical analysis of URLs. It does not include such approaches as analysis of webpages' contents This email analysis is meant to maintain the focus on comparing URL-based.

IV. METHODOLOGY

Methodology of this research was developed in order to find out what changes took place in the area of applica-tion of lexical analysis techniques during the previous years, what kinds of datasets and algorithms are mostly ap-plied in the practice of detection and what performance they demonstrate, as well as some practical limitations of such methods. However, it is important to note that systematic review methodology needs to be structured and transparent to guar-antee reliability, relevance, and representativeness of reviewed literature. Experimental research would imply design-ing and implementation of machine learning algorithm, but this research focuses on critical analysis of litera-ture concerning lexical analysis techniques for phishing websites detection. Methodology of the research involves the following main stages: identification of relevant literature, selection of proper studies, quality assessment of stud-ies, data extraction and comparison of findings. Such methodology guarantees unbiasedness of the results and makes it possible to conduct a meaningful comparison of results obtained by different researchers.

Methodology of this research was developed in order to find out what changes took place in the area of applica-tion of lexical analysis techniques during the previous years, what kinds of datasets and algorithms are mostly ap-plied in the practice of detection and what performance they demonstrate, as well as some practical limitations of such methods.

A. Literature Search Strategy

The first stage of systematic review involved identifica-tion of academic literature related to phishing website de-tection by means of lexical analysis. Considering that phish-ing detection is an interdisciplinary field including se-curity, machine learning, artificial intelligence, data mining, and web technologies, the literature was collected from various reputable academic databases in order to make sure that broad coverage would be provided. Main sources considered for the review include: • IEEE Xplore Digital Library • ACM Digital Library • SpringerLink • ScienceDirect • Wiley Online Library • Google Scholar • MDPI Journals • Taylor Francis
• ResearchGate (for publicly available research papers) These databases were chosen due to the fact that they provide peer-reviewed journal articles, conference pro-ceedings, technical reports and review papers by recognized researchers and organizations. Structured keyword-based approach was used for the search by combining the following search terms: • "Phishing Website Detection" • "Lexical Analysis" • "URL Classification" • "Machine Learning for Phishing Detection"
• "URL Feature Engineering" • "Phishing URL Detection" • "Cybersecurity Machine Learning" • "URL-based Detection"
• "Lexical Feature Extraction" • "Deep Learning Phishing Detection" Search accuracy was enhanced by use of Boolean operators (AND, OR, NOT). Some examples of search expres-sions included: • ("Phishing Detection" AND "Lexical Anal-ysis") • ("URL Classification" AND "Machine Learning")
• ("Phishing Website Detection" AND "Random Forest")
• ("URL Feature Engineering" OR "Lexical Features") • ("Client-side Phishing Detection" AND "Deep Learn-ing"
This search strategy ensured the inclusion of both classical and up-to-date studies in the literature review.

B. Inclusion Criteria

After searching for research papers in different databases, certain inclusion criteria were determined for assessing the relevance of the found studies in relation to the aims of this literature review.
The selected studies were satisfying at least one or several of the following criteria:

- Publications in peer-reviewed journals and international reputable conferences.
- Publications within the period from 2018 to 2025.
- Oriented primarily to the detection of phishing websites.
- Lexical analysis or URL features extraction usage.
- Usage of machine learning or deep learning algorithm.
- Evaluation of the model's performance according to accuracy, precision, recall, F1-score, ROC-AUC, false positive rate.
- Methodology information (dataset/feature extraction).

Limitation to the peer-reviewed papers allowed us to ensure the credibility and reliability of analyzed papers

1) Exclusion Criteria : 1) : Various studies were excluded from the review to maintain the consistency and relevancy. This list of studies is the following:

- Papers related to the detection of phishing emails.
- Webpage content analysis without lexical features.

- Malware detection instead of phishing website detection.
- Duplicate papers found in various databases.
- Papers with abstracts without full methodology descrip-tion.
- Papers without experimental evaluation.
- Not English-language studies.
- Unreliable opinion papers or blogs.

These exclusion criteria help to focus more on the papers using lexical analysis approaches rather than general cyber-security issues.

C. Study Selection Process

The study selection process included several stages.
Firstly, a great amount of research articles was found using keyword searches. Then, duplicate publications were deleted from the search results.
The rest of the articles were screened according to their title and abstract for relevance. Studies that were clearly not related to our topic (spam filtering, malware analysis, email security) were eliminated from further consideration.
The selected articles were analyzed in detail, including their full methodology, experiments performed, feature extraction process, used datasets, machine learning algorithms, and evaluation metrics.
Only those papers that were directly related to lexical phishing detection were included in the systematic review.

1) Data Extraction Process : After selecting the final set of research papers, important information was extracted sys-tematically to facilitate comparison between different studies. The following parameters were recorded for each publication:

- Year of publication
- Authors
- Objective of the research
- Dataset used
- Amount of phishing URL
- Amount of legitimate URLs
- Approach to feature engineering
- Machine learning technique
- Deep learning network (if applies)
- Evaluation metrics
- Accuracy achieved
- Precision
- Recall
- F1-Score
- ROC-AUC
- Limitations revealed by authors
- Suggestions for the future research

Collecting these attributes in a systematic way allowed making a meaningful comparison of multiple scientific papers despite various research methods applied.

2) Quality Assessment : 1) : Not all the scientific papers published are equal in terms of contribution to science. In order to synthesize the research results, quality assessment has been done prior to that step. Every scientific paper has been evaluated according to several quality indicators, including:

Quality of Dataset

Preferably, those scientific papers applying large-scale pub-licly available datasets with proper documentation about data collection and data preprocessing have been preferred.

Experimental Reproducibility

Preferably, the scientific papers with detailed description of data preprocessing, feature extraction procedure, selection of hyperparameters and evaluation method have been chosen.

Evaluation Metrics

Preferably, the scientific papers presenting several measures of model performance have been chosen rather than those focusing only on classification accuracy.

Novelty

Special attention has been paid to the scientific papers offering new approaches to feature engineering, new machine learning models, explainable AI or phishing detection systems capable of adaptation.

Practical Relevance

Preferably, the scientific papers solving the deployment-related problems, such as concept drift, adversarial attacks, computa-tional efficiency and scalability have been chosen.

D. Identification of Research Gaps

Apart from comparing existing research, one more goal of the review is to identify research gaps in the field of lexical phishing detection. Multiple gaps were identified, including but not limited to:

- Aging of datasets

- Dataset imbalance
- Frequency of false positives
- Issues with generalization
- Concept drift
- Adversarial attacks on URLs
- URL shortening services Homographic attacks based on Unicode
- Problem of model explainability
- Privacy preserving
- Issues with real-time deployment .

The lack of benchmarks Identification of these gaps allows finding promising directions of future research.

V. ANALYSIS AND FINDINGS

Apart from comparing existing research, one more goal of the review is to identify research gaps in the field of lexical phishing detection. Multiple gaps were identified, including but not limited to:

- Aging of datasets
- Dataset imbalance
- Frequency of false positives
- Issues with generalization
- Concept drift
- Adversarial attacks on URLs
- URL shortening services
- Homographic attacks based on Unicode
- Problem of model explainability
- Privacy preserving
- Issues with real-time deployment .

The lack of benchmarks Identification of these gaps allows finding promising directions of future research.

A. Analysis of Public Datasets

The quality of the dataset has a direct impact on the performance of the model – that is the main conclusion one can draw from the surveyed articles. Researchers typically use publicly available sources to get phishing URLs and domain names of legitimate websites. Popular phishing URLs datasets include:

- PhishTank
- OpenPhish
- URLHaus
- APWG datasets

Domain names of legitimate websites are often taken from:

- Tranco Top Sites
- Cisco Umbrella Top Domains
- Common Crawl
- Alexa Top Websites (for old articles)

These resources have been established benchmarks due to their public availability and regular updates. However, at the same time, some researchers noted certain limitations of these datasets.

1) Dataset decay: Phishing websites usually exist online for only a short period of time before being taken down. As a result, datasets compiled a few months or even years ago may already con-tain inactive websites. This effect is known as dataset decay and makes researches less reproducible since subsequent researchers are unable to confirm the validity of original web pages used by their predecessors in experiments.

2) Class imbalance: Another important issue is class im-balancedness. Practically, most of the internet traffic con-sists of legitimate websites while many experimental datasets have an equal number of phishing and non-phishing URLs. Even though class balancing is favorable for machine learning, it does not represent the real world environ-ment in which the system will be deployed. Consequently, classifiers work well in lab environment but have increased rate of false positives in practice.

3) Duplicate samples: Another issue pointed out by sev-eral researchers is duplicate samples between different reposito-ries. The same phishing website could be in both PhishTank, OpenPhish and URLHaus at the same time and lead to unexpected overlap of train and test data in case there is no preprocessing performed. This issue proves that there should be created new continuously updated and preprocessed bench-marks for further investigation of lexica-phishing detection systems.

4) Analysis of Lexical Features : It is clear from the literature that feature extraction is a key component in the successful development of lexica-phishing detection sys-tems. Various papers have suggested various features of the URL that can be used to differentiate phishing URLs from the legitimate ones.

5) Structural Features: Structual features remain the most popular ones. Some common structural characteristics are:

- URL length
- Hostname length
- Subdomain count

- Directory depth
- Dot count
- Path length
- Query length
- Digit count
- Slash count
- Domain length

Most of the phishing URLs are significantly longer than legitimate URLs according to all studies since the attack-ers use long URLs to conceal malicious information. Also, phishing websites tend to have several subdomains in or-der to disguise themselves as trusted institution and hide their true registered domain. For example, https://secure-login.bank.verify.user-account.example.com looks like legiti-mate website even though it belongs to completely different domain.

Character-level features: Some researchers studied distri-bution of characters in URLs. Com Character-level Features Instead of structural characteristics only, most of recent stud-ies utilize NLP techniques in order to reveal hidden suspicious words in the URLs. Some examples of frequently appearing phishing words are:

- login
- verify
- secure
- account
- update
- payment
- signinbanking
- authentication
- password

The presence of many suspicious words increases the like-lihood that a URL is malicious. There is no doubt that many studies succeeded in detecting malicious URLs by using both structural and semantic features.

Statistical Features: In addition to the manual way of extracting features, some researchers have introduced some methods for statistically representing the URLs such as:

- Shannon Entropy
- Character frequency vector
- N-gram frequencies
- TF-IDF representations
- Character embeddings

These techniques help the machine learning algorithms to capture complex patterns that cannot be extracted by manual feature engineering methods.

1) Machine Learning Algorithms : One of the objectives of this review paper is to compare machine learning algorithms used for detecting phishing through lexicons.

2) Logistic Regression: Logistic Regression is a benchmark model since it is easy to implement and interpret. Advantages of this model include:

- Fast training
- Very low computational cost
- Simple implementation

However, this linear model is not able to model relationships between lexical features.

3) Decision Trees: These trees classify URLs according to the decision hierarchies.

Advantages:

- Simple visualization
- Interpretability
- Fast inference

However, disadvantages of this algorithm include:

- Overfitting
- Generalization
- Sensitivity to noisy datasets

4) Support Vector Machines: They are famous for their accurate classification especially for medium sized dataset. Advantages of this model include:

- Very accurate
- Efficient for dealing with high dimensional features
- Well-established theory

However, the computational complexity of training becomes very high for millions of URLs.

5) Naive Bayes: This algorithm is suitable for classifying URLs because of the efficiency of its computation. Advantages of this model include:

- Fast inference
- Real-time classification
- Outstanding performance on small datasets Limitations of this model include:
- Independence assumption
- Lower accuracy than ensemble approach

6) Random Forest Classifier: This algorithm showed out-standing success when comparing with the other traditional machine learning algorithms for classification of URLs. This is due to the following reasons:

- High robustness
- Lower overfitting
- Feature importance calculation
- Scalability
- Good performance on noisy datasets

It was found that the accuracy of classification in many studies varied from 97% to 99%. This made it one of the benchmark algorithms.

7) Boosting Algorithms: Recently developed modern algo-rithms like:

- XGBoost
- LightGBM
- CatBoost become popular.

Such algorithms improve sequentially weak classifiers and generally outperform conventional classifiers. The benefits of such algorithms include:

- Predictive performance
- Efficient missing values handling
- Computational speed
- Regularization
- Scalability

According to several studies, XGBoost yields results similar or superior to Random Forests, while consuming less compu-tational resources in prediction.

Therefore, the proposed system provides robust perfor-mance with accurate assessment and effective feedback, and hence it can be deployed in practice in academic or placement preparation context.

1) Deep Learning approaches : Deep learning represents the most recent research direction in the field of lexical phishing detection. Unlike classical machine learning ap-proaches, which heavily rely on manually crafted features, neu-ral networks automatically learn representations directly from the raw data (URLs in our case).

2) Convolutional Neural Networks (CNN): CNN archi-tectures analyze URLs as character sequences. Advan-tages:

- Feature extraction
- Superior classification performance
- Effective local pattern recognition Disadvantages:
- Need for larger datasets
- Increased training time

VI. LIMITATIONS AND FUTURE SCOPE

Limitations of Existing Lexical Phishing Detection Sys-tems: Although lexical analysis proved to be one of the most efficient methods of detecting phishing websites, based on the analysis of current literature, currently available systems suffer from several limitations that prevent their widespread deployment. Most of machine learning algorithms perform effectively in experimental setting, but in dynamic setting, their efficiency drops down due to several reasons. Such limitations are caused by changes in attacks, issues with datasets, com-putational limitations and inherent limitations of URL-based detection.

The first limitation is data dependence. Machine learning algorithms depend on historical data sets of phish-ing and non-phishing URLs for training. Since most phish-ing websites operate for only a short period of time, publically available datasets soon become outdated. With changing attack tech-niques and URLs structure models trained on outdated data cannot detect new attacks. This issue is known as dataset aging.

The second limitation is the issue of class imbalance. Most stud-ies that use machine learning approach for detecting phishing employ balanced datasets which have equal num-ber of phish-ing and legitimate URLs. In reality, though, the proportion of legit URLs is much higher than that of phishing ones. Hence, models that are trained on artificial balance dataset produce high error rates when applied to real-world tasks. False positives not only affect users' convenience, but also destroy their trust to automated phishing detection systems.

Lexical detection also faces another kind of limitation, namely, the fact that attackers try to avoid it using various manipulations with URL structure. Bitly, TinyURL and Re-brandly services that create shortened URLs eliminate most of lexical features that were learned by models in the course of training. As a result, models that rely on lex-ical information misclassify such shortened URLs.

Another advanced type of attack is IDN homograph attack. In it cybercriminals use Unicode characters from different languages that are close to standard Latin alphabets. There-fore, by replacing Latin characters with those from Cyrillic or Greek al-phabet one can obtain identical in looks, but different in character encodings domain names that will go unnoticed by lexical feature extrac-tion methods because the latter only work with ASCII characters.

In a similar vein, another challenge is posed by typosquat-ting attack. Here attackers register domain names which contain misspellings of popular websites. For example, they may omit one letter in the spelling of domain name. Although such domain names look completely normal to users, they may not be detected by classic lexical classifiers due to insufficient feature engineering.

Adversarial machine learning has become increasingly pop-ular among cybercriminals recently. Recent studies demon-strated that changing certain parts of URL can affect models' predictions in a way that they classify URL as legitimate. Hence, adversarial examples demonstrate vulnerabilities of machine learning approaches and call for develop-ment of new robust classifiers.

Another problem that should be mentioned is that of inter-pretability of machine learning models. Nowadays many of the most advanced models, particularly, those based on ensemble or deep learning, are black box models. While providing very good

classification performance they do not offer any explanation of how they came to their decision. This fact is crucial for practical application of machine learning tech-niques to phishing detection because cybersecurity requires justifications of decisions.

However, computational efficiency is a serious issue as well. Although the processing of a string via lexical analysis does not involve considerable computational resources consump-tion, the deep learning algorithms such as LSTM network and Transformer are resource-consuming. The fact results in impossibility of implementing them on low-capability devices such as smartphones, embedded systems or limited-computing-resource browsers.

Furthermore, the lack of standardization of benchmarking practices makes further research in this area even more chal-lenging. Since each researcher applies his or her own dataset, pre-processing technique, approach to feature extraction and evaluation measure, it is difficult to compare performances of the models by means of direct comparison of the obtained values. In this case, there is the possibility that a model with 99% accuracy for one dataset will perform significantly worse on the other dataset collected in a different way.

However, taking into account the current situation in this field, it can be stated that there are many promising research directions in relation to lexical phishing detection techniques. Firstly, the development of XAI (Explainable Artificial In-
telligence) should be mentioned. At the moment, it is pos-sible to apply machine learning explanation techniques that include such modern techniques as SHAP (SHapley Additive ex-Planations) and LIME (Local Interpretable Model-Agnostic Explanations) to ensure models' abilities to provide explana-tions of their decision-making process. Instead of applying a 'malicious' label, the system will be able to provide an explanation of why it made such a conclusion – due to the long URL, presence of unusual keywords, high entropy, multiple nested sub-domains, etc.

Secondly, the development of hybrid phishing detection systems should be mentioned. As opposed to lexical analysis only, the developed system will contain more complex features that can be obtained on the basis of additional contextual information, namely DNS record, domain age, SSL certificate metadata, WHOIS and reputation information. In such a way, lexical analysis can become the first stage of filtering while the following steps can be applied when some doubts exist.

Finally, the development of online learning and incremen-tal machine learning becomes a matter of interest among researchers. Contrary to the classical machine learning paradigm that requires a full re-training of a model after obtaining new data, online learning algorithm will update itself after receiving new phishing URLs. Thus, it will be able to cope with concept drift problem and changing phishing strategies.

Also, the rising popularity of distributed computing environ-ments adds to the popularity of federated learning. Federated learning implies collaboration in the training of the machine learning models without exchange of any personal data, in-cluding browsing information. In simple words, each device calculates its model updates which are then aggregated in one device.

Another area is development of graph neural networks (GNN). While in classical approach the relations among the various components of URLs like domains, hosting, IP-addresses, hyperlinks, and DNS records were not considered at all, using the graph approach will give an opportunity to include the relational information into the model. Graphs can help in detecting coordinated phishing campaigns.

Development in natural language processing can improve the task of lexical features extraction. Using big language models and context-aware embeddings can make it possible to discover complicated relationships in URL tokens auto-matically without the need for any manual work of feature engineering. The combination of classical structural features with context-aware representa-tions can increase effectiveness in detection of phishing attack patterns.

Creation of benchmark datasets is another important area for future research. It would be necessary to develop datasets that would be open access, up-to-date and well-balanced and would reflect the current internet traffic situation. Standardiza-tion of the benchmark will help to compare the performance of various machine learning models and will help to reproduce the results in independent research groups.

Despite the current state of affairs, there are still many directions for further research into lexical phishing detection approaches.

Firstly, the development of XAI (Ex-plainable Artificial Intelligence) is required. With the recent advancements of ma-chine learning, it becomes possible to employ such explainabil-ity frameworks as SHAP (SHapley Additive ex-Planations) and LIME (Local Interpretable Model-Agnostic Explanations) and develop self-explainable models. Instead of simply labeling the website as malicious, the model will provide an explanation for such behavior of the website – it contains long URL, keywords, high entropy, several subdomains, and so on.

The second direction is development of hybrid phishing detection systems. While lexical analysis doesn't account for any contextual features, hybrid systems will use much more sophisticated features such as DNS records, age of the domain, SSL certificate metadata, WHOIS and reputation. Lexical analysis will become just a starting point in the filtration process, and other steps will only be taken if there is some doubt about the website.

Meanwhile, the scientific community is turning its focus onto online learning and incremental machine learning. Contrary to classical machine learning, which requires full re-training each time the new sample appears, online learning will update the model each time a new phishing URL appears. Consequently, such approach will work much better with the problem of concept drift adaptation.

The increase in the popularity of distributed computing also attracts more attention towards federated learning. Federated learning implies cooperation in the training of machine learn-ing algorithms without transferring any data. In other words, model updates will be calculated locally on each device and then aggregated in one place.

Another important development is graph neural networks (GNN). Contrary to the traditional approach that considers different domains, hosting services, IP-addresses, hyperlinks and DNS records unrelated to each other, GNN allows to use all this information in the model. Thus, GNN might become useful in detecting coordinated phishing attacks.

Future advancements in the field of NLP could help improv-ing lexical feature extraction process. Big language models and context embeddings allow extracting complex relations between URL tokens automatically. Thus, in combination with traditional structural features, it can help finding new patterns of phishing attacks.

The development of more accurate benchmark datasets is also one of the key future research directions. The datasets used in all future works should be accessible, constantly updated and balanced, as well as should represent the current internet traffic situation. Standardized benchmarks will allow comparing different machine learning models and reproduc-tion of their results.

VII. Conclusion

Phishing remains one of the major types of cyberattacks to date. While classical blacklists are vital for the cyberse-curity sector, these methods fail to cope with novel zero-day phishing domains emerging every day. Lexical analysis has been proven to be a productive and secure method for the client-side detection of phishing attempts based only on URL structure and semantics. Nevertheless, implementation of such technology into practice is far from being easy. We have to stop working with artificial datasets and black box approach. With explainable AI, online adaptability and benchmarking, we will be able to develop novel phishing attack detection tools that will help users to withstand advanced methods used by cyber criminals today.

References

[1] Abutaha, M., et al. ”URL Phishing Detection using Machine Learning Techniques based on URLs Lexical Analysis.” International Conference on Information and Communication Systems (ICICS), 2021.

[2] Kustiawan, Y. A., & Ghauth, K. I. ”Feature Engineering for Phishing Website Detection Using Machine Learning: A Systematic Review.” IEEE Access, vol. 13, 2025.

[3] Vung, C. G., & Win, Y. Y. ”URL Classification Based on Lexical Features by Machine Learning.” International Conference on Computer Applications (ICCA), 2023.

[4] Jain, A. K., & Gupta, B. B. ”Towards detection of phishing websites on client-side using machine learning based approach.” Telecommunication Systems, vol. 68, no. 4, 2018.

[5] Korkmaz, M., et al. ”Detection of Phishing Websites by Using Machine Learning-Based URL Analysis.” International Conference on Comput-ing, Communication and Networking Technologies (ICCCNT), 2020.