IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 11, November 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.524

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311119|

Phishing Email Detection: Insights from a Systematic Review of NLP-based Machine Learning Methods

Shobha Agasibagil[1], B S Sahana[2], Lakshmi H R[3]

Assistant Professor, Department of Computer Science and Engineering, CIT, Gubbi, Tumkur, Karnataka, India[1]

U.G. Student, Department of Computer Science and Engineering, CIT, Gubbi, Tumkur, Karnataka, India[2]

U.G. Student, Department of Computer Science and Engineering, CIT, Gubbi, Tumkur, Karnataka, India[3]

ABSTRACT: A Systematic Literature Review on Phishing Email Detection Using Natural Language Processing Techniques" provides a detailed review of phishing detection methods, focusing on Natural Language Processing (NLP) techniques. It highlights the increasing threat of phishing, especially through emails, and assesses existing research trends, specifically on NLP-based phishing detection. By analyzing 100 studies from 2006 to 2022, it identifies key areas, such as feature extraction, machine learning algorithms, text features, and popular datasets like the Nazario phishing corpus. The review finds Support Vector Machines (SVMs) frequently used, with common NLP techniques including TF-IDF and word embeddings. It also points out gaps, especially for Arabic language phishing emails, suggesting future research needs.

KEYWORDS: phishing email detection, natural language processing (NLP), machine learning algorithms, feature extraction, text analysis, TF-IDF, word embeddings, support vector machines (SVMs), and cybersecurity in email communication.

I.INTRODUCTION

Phishing poses a major threat to internet security, exploiting users' lack of awareness to steal sensitive information, often through deceptive emails. Phishing attacks have risen significantly, with email being the preferred method due to detection challenges. The COVID-19 pandemic saw a sharp increase in phishing attacks, with attackers leveraging pandemic-related information to target victims. Despite over a decade of research, phishing remains prevalent, partly due to its complexity and the reliance on user unawareness.

This paper systematically reviews studies on phishing email detection using Natural Language Processing (NLP) and Machine Learning (ML). It explores key areas such as feature extraction, ML algorithms, NLP techniques, datasets, optimization methods, and evaluation criteria. The study identifies support vector machines (SVMs), TF-IDF, and word embeddings as commonly used methods and highlights datasets like the Nazario phishing corpus. It also addresses gaps in multilingual phishing detection, emphasizing the need for further research on diverse email components and mixed-language models.

The paper is structured to review relevant studies, classify phishing detection methods, analyze datasets, techniques, features, and evaluation metrics, and conclude with findings and recommendations for future work.

II. RELATED WORK

The main limitation of existing phishing surveys is their lack of focus on phishing email detection using Natural Language Processing (NLP) techniques. While numerous surveys have explored phishing detection in areas like URLs, websites, and general phishing, they do not comprehensively address the use of NLP for phishing emails. Additionally, none of the surveys have systematically reviewed NLP-based phishing detection techniques across multiple languages, particularly Arabic. This gap highlights the need for a focused study on phishing email detection using NLP, which this paper aims to address.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311119|

III.METHODOLOGY

The systematic review followed guidelines from [188], incorporating four phases: defining inclusion/exclusion criteria, data sources, quality assessment, and data analysis. Steps included identifying research, assessing quality, selecting studies, synthesizing data, extracting information, and verifying findings. Techniques from [189] ensured better organization through a structured review protocol, covering planning, execution, and analysis. The process emphasized precise search strategies, defining terms, and ensuring relevance to the study.

1. **Formulating Research Questions**: Specific research questions were established to guide the review process. These questions targeted primary aspects like feature extraction, commonly used ML algorithms, datasets, and performance metrics.
2. **Defining the Scope and Objectives**: The study began by identifying the need to systematically explore the use of NLP for phishing email detection. Objectives were clarified to assess prevalent NLP and ML techniques, key research areas, and performance evaluation methods in phishing email detection.
3. **Database Selection and Search Strategy**: To ensure comprehensive coverage, a range of databases such as IEEE, ACM, ScienceDirect, and Google Scholar were selected. The search terms included combinations like "phishing detection," "natural language processing," and "machine learning," yielding an initial collection of 1,125 articles.
4. **Screening and Selection Process**: Articles were screened in two stages: first, by eliminating duplicates, then by applying inclusion and exclusion criteria. Only studies that focused on phishing detection using NLP techniques and provided sufficient methodological details were included, reducing the pool to 100 articles.
5. **Quality Assessment**: A quality evaluation was conducted on the remaining studies using a checklist scoring system, where each study was rated from 0 to 7 based on criteria like clarity, methodological rigor, and relevance. This ensured that only high-quality studies contributed to the analysis.
6. **Data Extraction and Organization**: For each selected article, essential information was extracted, including publication year, NLP and ML techniques used, datasets, and evaluation metrics. Studies were categorized by research area, allowing for easier trend identification and analysis.
7. **Data Analysis and Synthesis**: Using a comparative approach, the studies were systematically analyzed to identify trends in popular NLP techniques, ML algorithms, and text features used in phishing detection. The synthesis focused on identifying research gaps, such as the limited studies on non-English phishing emails, and areas needing future exploration.
8. **Structured Review Framework**: The study adopted the PRISMA framework to document the article selection and review process, ensuring transparency and replicability in identifying, screening, and analyzing the studies.

This structured methodology allowed for a comprehensive overview of phishing email detection using NLP, highlighting trends, gaps, and opportunities for future research in this domain.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311119|

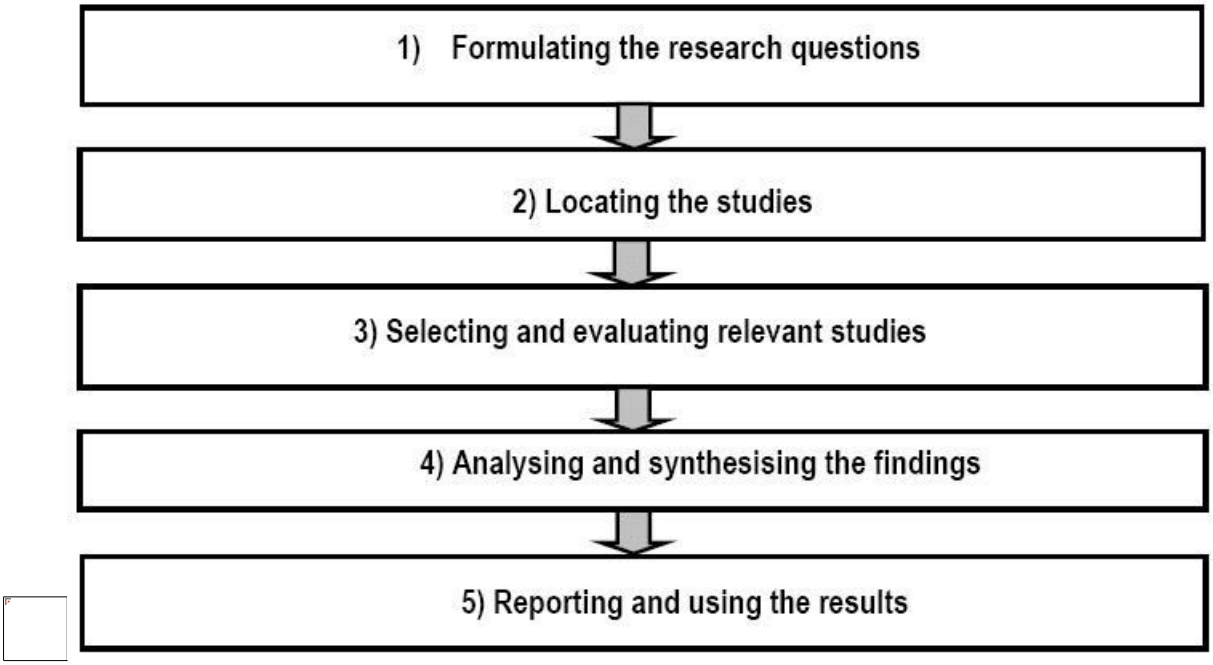

FIGURE 1 : SLR_METHODOLOGY:PROTOCOL REVIEW STAGES

IV.EXPERIMENTAL RESULTS

The systematic review addressed 10 research questions (RQs) related to phishing email detection using NLP techniques:

1. Key Research Areas (RQ1): The primary focus in phishing detection studies is on feature extraction/selection, observed in 86 studies. Phishing email detection itself is a significant area, with 68 studies, followed by classification techniques with 33 studies. However, only 14 studies focused on algorithm optimization, indicating limited research in this area.

2. ML Algorithms (RQ2): Among ML algorithms, Support Vector Machines (SVM) were the most commonly used (50 studies), followed by Naïve Bayes (33 studies). Decision Trees and Random Forest were employed in 29 studies each. Deep learning techniques such as RNNs, NNs, and CNNs were less frequently applied, appearing in 13, 8, and 5 studies, respectively.

3. Optimization Techniques (RQ3): The Adam optimizer was the most popular optimization technique, used in over 26% of the reviewed studies. Sequential Minimal Optimization (SMO) was the second most used, appearing in 21% of the studies, showcasing its role in text dataset processing.

4. Text Features (RQ4): The commonly used text features include Bag-of-Words (BoW) and Information Gain (IG), each employed in 10 studies, and Word2Vec in nine studies. Other features such as PCA, POS tagging, LDA, and LSA were less frequently utilized, ranging from 3 to 6 studies.

5. NLP Techniques (RQ5): Basic NLP tasks, including stopword removal, stemming, tokenization, and punctuation handling, were used in 59 studies. TF-IDF was applied in 36 studies, while word embeddings like Word2Vec were used in 12 studies.

6. Datasets (RQ6): The Nazario phishing corpus was the most commonly used dataset, appearing in 42 studies, followed by the SpamAssassin Public Corpus (40 studies) and the Enron dataset (23 studies). Smaller datasets like PhishTank and IWSPA-AP were also mentioned.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311119|

7. Evaluation Metrics (RQ7): Accuracy was the most commonly used metric, applied in 83 experiments. Other popular metrics included precision, recall, F1-score, and confusion matrix components like sensitivity and specificity, each used in 35–38 experiments.

8. Tools (RQ8): Python was the most widely used tool, appearing in 37 studies, followed by Weka (17), Scikit-learn (12), Keras (9), and Java (8). TensorFlow was used in six studies for deep-learning implementations.

9. Email Components (RQ9): The email body was the primary focus in 38% of the studies (93 studies), followed by the header (75 studies) and URLs (66 studies).

10. Trends Over Time (RQ10): Phishing email detection studies have significantly increased over the years, with 57% of studies published between 2018 and 2022. The highest number of publications was in 2020 (22 studies), reflecting the growing importance of this research area.

V.CONCLUSION

Presently, one of the most interesting topics in the domain of cybersecurity is assumed to be phishing email detection. In this research, journal, conference, and workshop papers were carefully analysed, published between 2006 and2022, with different techniques to investigate the trend of phishing email detection. A systematic literature review was employed to select 100 publications. All types of phishing email detection, for example, 'the domain name is mispelt,''the email is poorly written,' 'suspicious attachments or links' and 'phishing warning messages,' have been covered in our research. The background information regarding the phishing ecosystem is first presented along with the valuable phishing statistics. Next, the taxonomy of phishing detection schemes are then stated, and the phishing email detection datasets and phishing email detection features are discussed along with the detection algorithms and evaluation metrics. Lastly, recommendations are presented which can help with the phishing detection schemes' effective development, so that the compare-and-contrast schemes can be easily carried out.

REFERENCES

[1]. (2020) Anti-Phishing Working Group Phishing ActivityTrends Report 3rd Quarter 2020. Available: https://docs.apwg.org/reports/apwg_trends_report_q3_2020.pdf
[2]. Phishing Activity Trends Report 3rd Quarter 2021 Anti-Phishing Working Group. Available:https://docs.apwg.org/reports/apwg_trends_report_q3_2021.pdf
[3]. Brohi, and N. Zaman, "Ten deadly cyber security threats amid COVID-19 pandemic" TechRxiv, Tech. Rep., 2020, doi: 10.36227/techrxiv. 12278792 v1.N. A. Khan, S. N.
[4]. B. Pranggono and A. Arabo, "COVID-19 pandemic cybersecurity issues," Internet Technol. Lett., vol. 4, no. 2, Mar 2021, doi: 10.1002/12 247
[5]. D. Irani, S. Webb, J Giffin, and C. Pu, "Evolutionary study of phishing." in Proc eCrime Res. Summit. Oct. 2008, pp. 1-10, dor 10.1109/ECRIME 2008.4696967.
[6]. B. B. Gupta, A. Tewari, A. K. Jain, and D. P. Agrawal, ''Fighting against phishing attacks: State of the art and future challenges,'' Neural Comput. Appl., vol. 28, no. 12, pp. 3629–3654, Dec. 2017, doi: 10.1007/s00521-016-2275-y.
[7]. B. B. Gupta, N. A. G. Arachchilage, and K. E. Psannis, ''Defending against phishing attacks: Taxonomy of methods, current issues and future directions,'' Telecommun. Syst., vol. 67, no. 2, pp. 247–267, Feb. 2018, doi: 10.1007/s11235-017-0334-z.
[8]. D. Sahoo, C. Liu, and S. C. H. Hoi, ''Malicious URL detection using machine learning: A survey,'' 2017, arXiv:1701.07179.
[9]. E. S. Aung, C. T. Zan, and H. Yamana, ''A Survey of URL-based phishing detection,'' in Proc. DEIM Forum, 2019, pp. 2–3.
[10]. M. Khonji, Y. Iraqi, and A. Jones, ''Phishing detection: A literature survey,'' IEEE Commun. Surveys Tuts., vol. 15, no. 4, pp. 2091–2121, 4th Quart, 2013, doi: 10.1109/SURV.2013.032213.00009.
[11]. Qabajeh, F. Thabtah, and F. Chiclana, ''A recent review of conventionalvs. Automated cybersecurity anti-phishing techniques,'' Comput. Sci.Rev., vol. 29, pp. 44–55, Aug. 2018, doi: 10.1016/j.cosrev.2018.05.003.
[12]. L. Tang and Q. H. Mahmoud, ''A survey of machine learning-based solutions for phishing website detection,'' Mach. Learn. Knowl. Extraction,vol. 3, no. 3, pp. 672–694, Aug. 2021, doi: 10.3390/make3030034.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details