BABEŞ-BOLYAI UNIVERSITY, CLUJ-NAPOCA, ROMÂNIA
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

Detection and Analysis of Malicious Web Links

–PhD Thesis Summary–

Author
Claudia-Ioana Coste
PhD Supervisor: Prof. PhD. Anca-Mirela Andreica

2026

Contents

Chapter 1

Introduction

The last couple of decades were marked by the rise of Internet technology. Nowadays, life without the Internet cannot be imagined. According to Statista [26], there are 5.52 billion Internet users worldwide, approximately 67.31% of the world's population (8.2 billion [31]). Regular users surf the online domains to complete their everyday tasks, for example, shopping, socializing, public administration services, education, entertainment, and so on. However, humans cannot keep up with the pace at which the Internet is expanding. Thus, many users have poor online literacy, and they are vulnerable targets for attacks, which are becoming more and more sophisticated.

Most online threats are based on phishing, aiming to deceive consumers and trick them into clicking links, downloading malware programs without their knowledge, stealing personal information, and propagating false content to influence their behavior. Malicious web links identify and point to websites stored on the World Wide Web (WWW), performing any type of bad-intended behavior. For example, malicious links can host cloned websites used to mislead users into filling in their credentials. Moreover, some links run malicious code, downloading programs without consent, following the consumers' actions online, performing large-scale operations in the browser, which consume many computational resources, etc. Additionally, malicious websites can contain low-quality content (i.e., clickbait) or untrue stories contributing to the disinformation phenomenon.

Moreover, these mischievous sites are propagated into the online environment by clickbait, sensational and dramatic short messages or titles specifically created to mislead consumers and make them access the link and share it. Usually, clickbait is mostly used with low-quality journalist articles to achieve more reads, clicks, and more financial benefits from publicity. Besides the monetary gains, clickbait can be used to spread fake news to propagate false information and maliciously affect the users' behavior. Fake news is news containing false information and created with the vicious intent to deceive readers [27].

1.1 Motivation

Since consumers are estimated to spend approximately 4 hours a day on their mobile devices using different social media applications and the Internet, according to Insider Intelligence [25], they can be vulnerable targets for different malware attacks. Unfortunately, according to European data [22], media literacy is rather low in the European Union (EU) with an average of 60%. This percentage represents the individuals with basic and above-basic digital skills. This ranking [22] places Romania and Bulgaria in the last positions, with almost 25% of individuals having confidence in their digital abilities. The online environment is evolving at a faster pace than humans are able to develop new online skills. Thus, web attacks are becoming more and more sophisticated, and the task of spotting them has become far more challenging in recent years. According to Forbes [25], in 2023, there were 1.11 billion Uniform Resource Locators (URLs) available on the Internet, from which 202 million were active, providing resources, services, and interactions. Additionally, the AVG website is reporting that almost 1 million unique phishing websites were created in just the first quarter of 2024 [19]. The main propagating vector for web-malware is stated to be malicious links integrated in deceiving messages distributed through social media and other platforms [19]. Other propagating vectors for web threats are infected web pages. For example, in the case of WordPress websites, there are estimates that indicate that around 700,000 WordPress pages contained at least one malicious file [19]. WordPress is one of the most popular content management systems, and almost 40% of all websites are created using it [19].

Malicious web links can be associated with a large variety of online threats, such as malicious redirection, code execution to use the browser's resources, drive-by-downloads, cloned and phishing websites, etc. Additionally, they can share fake news and low-quality articles with attractive titles.

Clickbait is defined by the Oxford dictionary as material available online, which was created with the aim to draw attention and persuade consumers to access the web link which points to a specific website [23]. Clickbait can denote journalistic articles below the quality standard, which promise something from the title but fail to reach the readers' expectations. They aim to generate clicks and views such that more financial gains are obtained from publicity. There are multiple techniques used to attract attention and gain clicks, and many of them are being used in journalistic media outlets, on television, or in the written press. The content of these articles does not live up to the expectations set by the headline and may present false information or pseudoscience.

Clickbait can be further used to share false news. According to [27], fake news is news presenting factually wrong information that can be proven to be so. Moreover, they are fabricated with the

malicious intent to deceive readers. Fake news proved to influence the 2016 United States elections and the United Kingdom's referendum concerning Brexit. For example, in the United Kingdom during the time the European Union referendum was organized, the campaign against the EU had more followers and was more often shared than the one for the EU, according to [29]. Moreover, nowadays, because of the lack of proper regulations, fake news is influencing the voting behavior of consumers, as it happened in the United States in 2016. In [17], it is mentioned that the most distributed article in 2016 was a blog entry entitled "Why I am Voting For Donald Trump", written by a well-known conservative woman. The strategy seems to be repeated during the 2024 elections in the United States and other countries (e.g., Romania). In Romania, the results of the first round of elections were canceled due to proven Russian influence by spreading deceiving information on TikTok [16]. Often, fake news is popularized through social media, making use of trolls, bots, malicious links, clickbait, and others. Altogether, news presenting false information can be used to interfere with the consumption, health, or lifestyle behavior of the readers.

1.2 Objectives

The present thesis considers malicious web links detection domains from multiple perspectives. It elaborates and aims to bring contributions to multiple stages of the pipeline for malicious web links classification. First, the feature extraction process is investigated, and there is an experimental insight into feature importance analysis, information extraction from web links, and transformation of the Uniform Resource Locator (URL) to images. Secondly, machine learning algorithms are applied, and novel ideas are proposed together with relevant comparisons. Thirdly, malicious web links are analyzed considering a broader spectrum of deceiving web attacks, such as clickbait and fake news. Nevertheless, it is important to mention the survey conducted for the literature articles in this domain. Papers were analyzed and categorized according to the features used and the classification algorithms. Moreover, comparisons between multiple approaches were carried out.

The objectives could be elaborated from the perspective of malicious web links and the clickbait and fake news point of view. Thus, our objectives regarding malicious web links could be formulated as follows:

O1) Review the literature published for malicious web links detection [7];

For this objective, the analysis should follow the detection pipeline, by including the dataset, data collection, and annotation processes, features used in modelling, and the classification algorithm. The characteristics extracted and the classification methods should be classified and split into

relevant categories. A brief description of each depicted approach would be suitable, as well.

O2) Experiment with multiple Machine Learning (ML) algorithms [10], [4], [3], [11], [13], and [12];

The tests should consider a diverse range of artificial intelligence methods, such as: K-Nearest Neighbor (KNN), Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), Naive Bayes (NB), Adaptive Boosting (ADA), Gradient Boosted Tree (XGB), Support Vector Machine (SVM), and other machine learning methods, different ensemble models (e.g., Majority-Voting ensembles [4], Weighted Majority-Voting ensembles [13], [12], and Particle Swarm Optimization (PSO) ensemble [3]), deep learning architectures (e.g., Convolutional Neural Networks (CNNs), Long Short Term Memory (LSTM), ResNet101, ResNet50V2, VGG16) [11], and different Large Language Models (LLMs) [12].

Furthermore, a sound methodological idea should be proposed to deliver the experiments [10].

O3) Use multiple datasets [3], [11];

Experiments should be passed through multiple datasets, multiple types of datasets.

O4) Experiment with different feature categories;

Multiple approaches for the detection of malicious web links can be delivered by using different categories of features. For example, features that are automatically extracted through machine learning [11] or formulae [3], or manually engineered [13] and [12]. Extracted characteristics can undergo transformations and preprocessing to enhance the classification model.

The objectives designed for clickbait and fake news detection are the following:

O5) Investigate the viability of automatically collecting non-clickbait samples using Squid (i.e., a proxy web server) logs [1];

O6) Experiment with multiple machine learning methods (e.g., RF in [2] and [9], LR, NB, DT, SVM) for clickbait and fake news detection [6], [5], [8];

O7) Extract hand-crafted features to be language-independent [6], [5], [9], [8];

O8) Rank the most important properties in the clickbait or fake news classification [6], [5], [9], [8];

O9) Investigate the association between clickbait, fake news, and malicious links as a distribution vector.

1.3 Original Contributions

The original contributions of the thesis could be split into four main domains, considering the problem approached. The categories are malicious web links detection, clickbait detection, fake news detection, and finally, the relation between links, clickbait, and false journalistic articles.

First, for the malicious web links domain, one could outline the following contributions:

- Survey of the most relevant approaches on malicious web links detection field of research [7];
  - Classification and description of the solutions from the literature;
  - Categorization and features presentation;
  - Comparative analysis and discussions regarding paper reproductivity, advantages and disadvantages of the approaches, etc.
- Design a new methodology for calibration and comparisons [10] on the dataset proposed in [30];
  - Refined a methodology for malicious web links detection that may be used to experiment with ML algorithms;
  - Previous methodology used for training and comparing several machine learning algorithms, such as RF, DT, and KNN. The best one was RF with 87.6% precision;
  - Host-based features proved to be more relevant compared to lexical features;
- Developing a new Majority-Voting ensemble-based classification algorithm [4] on the dataset provided by [30];
  - Compared and calibrated 10 machine learning algorithms (e.g., LR, NB, Ada Boost, XGB, Linear Discriminant Analysis (LDA), Multi-layer Perceptron (MLP), and SVM);
  - The best single model was XGB with 95.07% precision. Majority-Voting ensembles were built in any combination possible (120 different variants), and the best one (ensemble Adaptive Boosting - Gradient Boosted Tree - Support Vector Machine with Radial Basis Function (ADA-XGB-SVM-rbf)) achieved 96.31% precision. It outperformed the individual models and other compared solutions;
- Proposal of a novel Particle Swarm Optimization ensemble used for maliciousness detection [3];
  - Compared 12 single machine learning models, such as: LR, SVM, ADA, RF, DT, KNN, Perceptron, Nearest Centroid (NC), Passive Aggressive Classifier (PAC), Stochastic Gradient Descent (SGD), KMeans, and different variants for NB;

  - The PSO ensemble was built;
  - Tested against two datasets: D1 [21] and D2 [28];
  - For D1 [21], PSO-model reached 97.05% accuracy and for D2 [28], 91.12% accuracy;
  - Results outperformed other literature solutions;

- An approach for using LLMs for malicious web links detection in [13] and [12];
  - Enriched classification model with hand-crafted features for D2 [28] with a label generated by OpenAI's model (GPT-4 and fine-tuned GPT-2);
  - Experiments with multiple LLMs (Generative Pre-trained Transformer (GPT) models: GPT-3.5-turbo, GPT-4, o1-mini, GPT-4o-mini, and a fine-tuned GPT-2);
  - Compared 13 machine learning algorithms and Weighted Majority Voting ensembles to observe if the GPT's feature leads to a better performance;
  - The experiments proved a slight improvement in performance for most algorithms;

- Elaboration of an innovative approach by transforming URLs to images and applying deep learning methods for malicious web links detection [11];
  - Experimented on two datasets D1 [21] and D2 [28];
  - Two types of images: RGB and grayscale;
  - Multiple configurations for the deep learning methods applied (e.g., CNNs, LSTM, VGG16, ResNet50V2, ResNet101) and fine-tuning;
  - 96.82% accuracy for D2 [28] surpassing its counterparts;
  - No significant differences for the Red, Green, and Blue (RGB) or the greyscale classification;
  - LSTM may be influenced if the information encoded in the images is filled in by row or by column.

Secondly, for clickbait detection, several makings were reached:

- Collected non-clickbait samples based on heuristics from users' online activity patterns generated from Squid logs [1];
- Proposal of a language-independent strategy for clickbait detection using hand-crafted features [2, 6, 5];

- Several experiments conducted on a large English dataset training multiple machine learning algorithms (e.g., RF [6], LR, NB, SVM), taking into consideration different feature categories [5];
- Reached a performance of 73.93% accuracy for LR and SVM, considering lexical features, and a performance of 71.81% for RF, taking into account all feature categories [5];
- Developed a feature importance analysis, which managed to confirm other literature results for clickbait detection [2, 6, 5].

Taking into account the fake news detection problem, some important research contributions were achieved in [8].

- Proposal of a feature-engineered classification model;
- Tests on two datasets in different languages (i.e., English and German);
- Compared multiple machine learning algorithms (e.g., RF, LR, NB and SVM with linear and Radial Basis Function (RBF) kernel);
- Achieved at most 75.38% accuracy for the RF model;
- Feature importance analysis.

Finally, the relation between malicious links, clickbait, and fake news was explored, and the resulting analysis could be summarized as:

- Clickbait is a journalistic technique to help the spread of false news and other types of low-quality media articles;
- Clickbait and fake news presented in a European and global context [2];
- Links can host fake news or questionable articles, and because of this, one could consider the links to be malicious.

1.4 Publication List

The evaluation standards used are those valid as of October 1st, 2018. The list of conferences, as well as their categorization, is based on the international CORE classification. The journal list is the one used by UEFISCDI for awarding articles published in international scientific journals.

1. Kiraly, A., **Coste, C-I.**, Șotropa, D-F. and Bufnea, D.-V. (2026). Towards Efficient Phishing Email Detection with LLM-Based Models. In Proceedings of the Intelligent Decision Technologies Conference (KES-IDT-26). KES Smart Innovation Systems and Technologies. Accepted for publication. In press. - Proceedings indexed in Scopus and Web of Science, conference Rank C, points: 2/(4-2) = 1;

2. **Coste, C-I.**, Chira, C. and Bufnea, D.-V. (2026). Malicious Web Links Detection: A Survey. SUBMITTED. - points: 0;

3. **Coste, C-I.** and Kaisser, T. (2026). Malicious Web Links Detection with Large Language Models. Lecture Notes in Business Information Processing. Accepted for publication. In press. - journal indexed in Scopus, Not found in any quartile from the UEFISCDI lists of indexed journals, points: 1;

4. Moldovan, D.-A. and **Coste, C-I.** (2025). Music Emotion Recognition with Deep Learning Techniques. In: Chira, C., Matei, O., Pop, F., Pop-Sitar, P. (eds) Innovative Perspectives on Computational Intelligence and Data Science. InnoComp 2025. Communications in Computer and Information Science, vol 2793. Springer, Cham. 10.1007/978-3-032-12478-4_11 - journal indexed in Scopus, Not found in any quartile from the UEFISCDI lists of indexed journals, points: 1;

5. **Coste, C-I.** (2024). Malicious Web Links Detection Based on Image Processing and Deep Learning Models. In Proceedings of the 2024 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), ISBN 979-8-3315-0494-6/24, pages 445-449. - conference Rank B (short paper), points: 2/3*4 = 2.66;

6. Kaisser, T. and **Coste, C-I.** (2024). Using Chat GPT for Malicious Web Links Detection. In Proceedings of the 20th International Conference on Web Information Systems and Technologies, ISBN 978-989-758-718-4, ISSN 2184-3252, pages 425-432. - Awarded with Best Student Paper - conference rank C (short paper), points 2/3*2 = 1.33;

7. **Coste, C-I.** (2024). Using Ensemble Models for Malicious Web Links Detection. In Proceedings of the 16th International Conference on Agents and Artificial Intelligence - Volume 3, ISBN 978-989-758-680-4, ISSN 2184-433X, pages 657-664. - conference Rank B (short paper), points: 2/3*4 = 2.66;

8. **Coste, C-I.**, Andreica, A-M., and Chira, C. (2023). Malicious Web Links Detection Using Ensemble Models. In Proceedings of the 19th International Conference on Web Information

Systems and Technologies, ISBN 978-989-758-672-9, ISSN 2184-3252, pages 268-275. - conference rank C (short paper and poster presentation), points: 2/3*2 = 1.33;

9. **Coste, C-I.** (2023). Malicious Web Links Detection - A Comparative Analysis of Machine Learning Algorithms. Studia Universitatis Babeș-Bolyai Informatica, ISSN 2065-9601. [S.l.], v. 68, n. 1, pages 21-36. - BDI journal, points: 1;

10. **Coste, C-I.**, and Bufnea, D.. (2021). Advances in Clickbait and Fake News Detection Using New Language-independent Strategies. Journal of Communications Software and Systems 17.3, pages 270-280. - journal indexed in Scopus, Found in Q4 (quartile 4) from the UEFISCDI lists of indexed journals, points: 2;

11. **Coste, C-I.**, Bufnea, D., and V. Niculescu. (2020). A new language independent strategy for clickbait detection. In Proceedings of the International Conference on Software, Telecommunications and Computer Networks (SoftCOM), Hvar, Croatia, pages 1-6, 10.23919/SoftCOM50211.2020.9238342; - conference rank D, points: 1;

12. **Coste, C-I.** (2019). Online bad habits: fake news and clickbait. In Proceedings of the Student Interdisciplinary Conference The European Union and Global Order, Cluj University Press, pp. 179-193, April 5th, 2019, Cluj-Napoca, Romania; - journal, points: 0;

13. **Coste, C-I.** (2018). Controlling the click bait. In Proceedings of the International Student Conference StudMath-IT, ISSN 2602-1579, pages 11-18, Arad, 15th - 16th November 2018 - student conference, points: 0.

Total points: 13 points

1.5 Summary

The current summary of the thesis is split into six chapters.

The first one is an introduction to the approached research niche together with the motivation for choosing it. Moreover, the objectives of this thesis, the most important original contributions, the summary, and the publication list are included.

The second chapter thoughtfully goes through the state-of-the-art reporting on the most important solutions for malicious web links detection. It discusses the main ideas, splitting them into categories (i.e., dynamic, static).

The third chapter, entitled "Systematic Comparison and Analysis of Malicious Web Links Detection Techniques", discusses and gives a critical analysis of the literature by stating some general observations about the common pipeline for maliciousness detection.

The chapter including the main contributions is the sixth, presenting the most relevant solutions proposed in this thesis. Information about the datasets depicted, about the preprocessing stage, about different methods applied for feature extraction, together with different types of classification algorithms, is described in detail.

The next chapter presents the investigation regarding clickbait and fake news, including data collection using heuristics and Squid logs, algorithms for classification, and feature importance analysis.

Finally, the last chapter highlights the conclusions, whether the proposed objectives were attained, and the main contributions proposed by the current thesis.

Chapter 2

State-of-the-Art for Malicious Web Links Detection

This chapter proposes to tackle multiple types of solutions regarding malicious web links detection. There are discussions regarding honeypots, sandbox approaches, antiviruses, blockedlist-based solutions, complex networks, and machine learning.

By surveying most of the articles in the literature concerning malicious web link detection, it was observed that all approaches could be split into two major categories: dynamic and static, as in [18]. In the dynamic category, we include approaches that let the link run its code and analyze the behavior, Hypertext Markup Language (HTML), Cascading Style Sheets (CSS), or JavaScript code; these solutions classify the link by accessing it as well. The dynamic approach consists of three approaches: honeypots, sandbox-based solutions, and machine learning applying behavioral features. Honeypot approaches propose to attract and mislead potential malicious links to be further analyzed in a safe sandbox environment. The studies using just sandboxes for the detection have the aim to first execute the malicious link in a safe environment and observe its behavior. Based on the behavior, it will be marked as suspicious or not. Static approaches do not involve running the link at first; they annotate the links, taking into consideration other static properties. From the static solutions, we can recall machine learning approaches, antivirus engines, and complex network analysis. The resemblance between the two approaches, static or dynamic in the case of ML solutions, is that both extract features that will be later given as input to intelligent models [7].

A suggestive visual representation for the classification of the malicious web links detection solutions is displayed in Figure 2.1. As an observation, there are many papers proposing approaches that could be included in more than just one category. For instance, there are antivirus solutions that combine

the signature approach with relevant heuristics and even with some intelligent methods. Moreover, there are algorithms using complex network theory combined with machine learning algorithms [7].

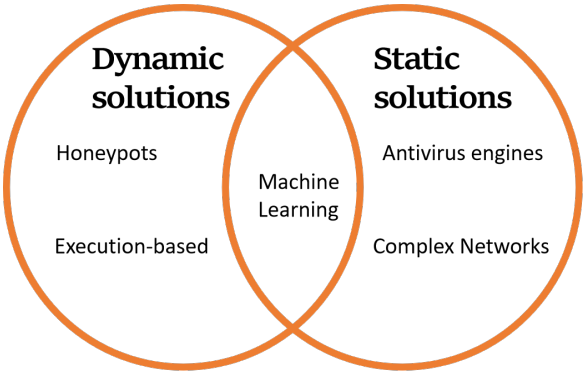

Figure 2.1: Classification of the surveyed solutions as presented in [7].

Chapter 3

Systematic Comparison and Analysis of Malicious Web Links Detection Techniques

The literature review chapter proposes to critically analyze and categorize the solutions found in the research field of malicious web links detection. All approaches will be dissected according to the stages described by our paper [7]. These stages include the datasets, feature extraction and selection, classification algorithms, paper reproducibility and relevant comparisons.

Taking into consideration our research from [7], most approaches use already available datasets, as well as aggregating some of them with the aim of obtaining a bigger dataset. The single-dataset detection systems used primarly [20] as the source for data. Additionally, half of the investigated research articles used a private dataset, which was not publicly available. However, in most cases, they specify their sources [7].

Solutions that choose to collect their records and features associated with a URL use a web crawler to retrieve information from the Internet [7]. Most benign records are retrieved from site ranking websites such as Alexa's Top [15]. In addition, other benign samples are collected from search engines as in [14]. Malicious samples are taken from public blocked lists.

The annotation process is done using sandboxes, and Application Programming Interface (API)s provided by antivirus engines or ranking services.

Considering the information used in solving the classification and constructing the intelligent model, usually, approaches take into consideration the URL. Thus, many studies focus more on content-independent features. Still, some approaches consider the content of a web link relevant, mostly

JavaScript (JS) and HTML scripts.

The feature extraction process is the most relevant and the knowledge-providing step in the classification process. The URL, together with the content of the link, must be transformed into numbers, such that it will be given as input to the detection algorithm. Feature extraction steps can pull off consistent and valuable information for the later classification process. This process can be done either automatically by using ML tools (e.g., autoencoders, embeddings, convolutional neural networks, etc.) or by manual engineering, where, based on heuristics and expert domain information, certain features are retrieved. Additionally, relevant information can be depicted from other methods: complex networks, graphs, mathematical formulas, binary encoding, etc.

Regarding paper reproducibility, just around five of all the articles mentioned (more than 50) gave enough implementation details and made publicly available the code they used for experiments. The dataset was provided for around 37% of the surveyed papers [7].

Additionally, comparisons were made between different types of solutions considering the effects of the attacks, explainability, 0-day attacks, evolution in time, obfuscation techniques, computational resources, and storage needed.

Chapter 4

Proposed Approaches for Malicious Web Links Detection

The present chapter will tackle our approaches proposed regarding malicious web links detection. First, the datasets used are described. For all the approaches proposed in this thesis, three datasets were used. The initial dataset [30] already has host-based and lexical features extracted. The second and third datasets are larger, though they contain just the URLs. The first dataset (D1) can be found at [21] and the second one (D2) at [28].

Next, the preprocessing methods are detailed. For the Urcuqui et al. dataset [30], the pre-processing included four stages according to our paper [10]: cleaning; data transformation; categorical features, and data normalization. Additionally, the preprocessing in [11] consisted of transforming URLs into images taking into account the American Standard Code for Information Interchange (ASCII) code which was translated to a greyscale or colored hue. For the experiments presented in another paper of ours [3], the dataset D2 [28] was preprocessed with Term Frequency and Inverse Document Frequency (TF-IDF) vectorizer, from the Python Sklearn library [24].

In the following step, the empirical methodologies are presented. The proposed algorithms are described in detail in what follows. Multiple ML methods are introduced, such as K-Nearest Neighbor, Decision Tree, Random Forest, Adaptive Boosting, Logistic Regression, different variants for Naive Bayes and Support Vector Machine, Linear and Quadratic Discriminant Analysis, Gradient Boosted Tree, Multilayer Perceptron, etc. Furthermore, a strategy for experiments is developed and applied to three ML algorithms as detailed in [10]. The strategy takes into account the fine-tuning process of multiple ML algorithms.

Additionally, three proposed ensemble types are exposed. The classification result is based on

distinct combinations of diverse individual algorithms. A Majority-Voting ensemble is proposed in [4], where the final class is chosen based on the the majority of the output of all classifiers. Particle Swarm Optimization ensembles are discussed in [3], where the weights of each classifier forming the enesemble are generated using the PSO algorithm. A Weighted-Majority Voting ensemble is developed by using a Large Language Model (LLM)-extracted characteristic in [13] and in [12]. Each annotation done by each classifier has a coefficient associated with it, a trust rate. The final class is computed as the weighted average of all labels that were output by the classifiers. In [13], we employed two OpenAI models such as GPT 3.5-turbo and GPT 4. In [12], two GPT mini models were added for tests: GPT-4o-mini (i.e., GPT-4o-mini-2024-07-18) and o1-mini (i.e., o1-mini-2024-09-12).

Moreover, deep learning solutions are also taken into account to classify images where the information contained in the URL is encoded [11]. Multiple models were proposed, such as a Convolutional Neural Network (CNN) with three convolutional layers, and other well-known architectures such as ResNet50V2, ResNet101, and VGG16. Finally, models were enriched with a LSTM layer with the intention of improving performance and better capturing the information encoded in the image.

Then, in the rest of the chapter, the experiments and their results are related. Finally, comparisons and critical discussions are recounted, positioning our proposed approaches in the literature.

Chapter 5

Malicious Web Links in Relation to Clickbait and Fake News

In what follows, the thesis's perspective on the relation between malicious web links, clickbait, and fake news is presented. First, the Clickbait Detection domain is described, along with the research and experiments proposed by this current thesis. Our proposal on the data collection and annotation process is shown in [1] and [9], where we derived an heuristic to automatically label non-clickbait samples using navigation patterns extracted from the logs generated from the Squid proxy. Furthermore, our ideas in clickbait detection considering language-independent and manually engineered characteristics are detailed in [2], [9], [6], [5], and [8]. Our methodology included detecting the language first and then extract the features from the clickbait message or the headline of the news. Then, the Fake News Detection domain is presented in conjunction with experiments, where ML approaches are tested, and a feature importance analysis is conducted [8]. Additionally, several ML algorithms (such as LR, RF, NB, DT, SVM) were applied for both clickbait and fake news detection.

In the last section of the current chapter, the relation between clickbait and fake news is explained considering the global and European context [2], and next, their association with malicious web links is discussed.

Chapter 6

Conclusions and Future Work

The present chapter will describe the main conclusions from this thesis and indicate some ideas for future research worth investigating.

6.1 Conclusions

Malicious web links detection is a complex and challenging domain, and the present thesis is considered to advance the research in this field. The investigation started by surveying the most relevant papers written in the current domain. All solutions and features used in the detection systems were categorized and described. Several comparisons are made between types of solutions and between solutions, when appropriate. Furthermore, other observations are generated, such as paper reproducibility.

Concerning the proposed objectives for malicious web links detection, these were mostly attained. A systematic review of the literature was delivered with multiple analyses on the used datasets, features, and a large range of approaches (objective O1). Multiple categories of artificial intelligence algorithms were tested for objective O2 such as Random Forest, Decision Tree, variants for Naive Bayes, different kernels for Support Vector Machine, Logistic Regression, Adaptive Boosting, Gradient Boosting Tree etc., multiple types of ensemble models (e.g., Majority-Voting, Weighted Majority-Voting, Particle Swarm Optimization ensembles) and deep learning strategies (e.g., ResNet50V2, ResNet101, VGG16, Convolutional architectures). Additionally, the experiments took into account multiple datasets (objective O3), such as a dataset with lexical and host-based features already extracted and two larger datasets (i.e., D1 and D2) containing just the string for the web link and the label. Furthermore, the features chosen as input were diverse (objective O4), some lexical and host-based ones already extracted from the dataset. Other experiments incorporated hand-crafted lexical and host-based characteristics gathered from the web links. Another set of tests used an embedding formula (i.e., Term

Frequency - Inverse Document Frequency), and lastly, another work considered encoding the URL data into a greyscale or RGB image.

By doing experiments on the dataset with features already extracted from web links samples, K-Nearest Neighbor improved the state-of-the-art solution, and RF achieved the best precision with 87.6%. Moreover, there was a proposed methodology for training and comparing algorithms. It was applied to K-Nearest Neighbor, Random Forest, and Decision Tree. Additionally, host-based features proved to be more relevant compared to the lexical attributes. On the same dataset, a comparison between other machine learning algorithms (e.g., Logistic Regression, Naive Bayes, Linear Discriminant Analysis, Gradient Boosted Tree, Multilayer Perceptron, and Support Vector Machine) was added. Then, heterogeneous ensembles were built from 3 classifiers. All models were calibrated, and all combinations were tried. Ensembles improve the performance of the single models, reaching at best 96.31% precision with the ADA-XGB-SVM-RBF ensemble. This solution improves the state-of-the-art by taking into account the Multilayer Perceptron, Logistic Regression, and Gaussian Naive Bayes models.

The empirical studies on the URL dataset, D1, managed to outperform other previous solutions. The proposed approach considers heterogeneous ensembles formed by multiple machine learning classifiers with weights generated by the Particle Swarm Optimization algorithm. The solution reached 97.05% accuracy.

For the second and largest URL dataset, D2, experiments with hand-crafted attributes achieve 94-95% accuracy on the whole dataset. Moreover, by adding OpenAI's output for classification as a feature, slight improvements were observed for most machine learning algorithms run and the Weighted Majority-Voting ensembles. In addition, regarding the D2 dataset, a novel solution was proposed by transforming URLs into RGB or grayscale images. Then, a variety of deep learning architectures were selected for experiments, such as Convolutional Neural Networks, Long Short Term Memory, VGG16, ResNet50V2, and ResNet101 architectures. The best model proved to be ResNet50V2 with the Long Short Term Memory layer, reaching a performance of 96.82% accuracy. It had a higher score compared with other literature results.

Moreover, about malicious web links, clickbait, and fake news detection was also investigated using intelligent algorithms and features engineered by hand. Considering the objectives proposed, collecting non-clickbait samples through the Squid Hypertext Transfer Protocol (HTTP) proxy server and the proposed heuristics was not suitable for the real-life scenario, so few samples were collected in this way (objective O5). Thus, for the next experiment, the Clickbait Challenge dataset was used. Another objective (O6) was the usage of multiple machine learning algorithms (e.g., Random Forest, Logistic

Regression, Naive Bayes, and Support Vector Machine), which was completed. Features were manually extracted in such a way as to be language-independent (objective O7). The obtained results, accuracy between 71% and 74%, precision within 72% and 75%, recall in the [68-71]% interval, and F_1 score between 70% and 73%, computed for Logistic Regression, Support Vector Machine, and Random Forest, are pertinent and comparable with other solutions to encourage more examinations in this direction. The analysis done on the most significant attributes is confirmed by similar results obtained in the scientific literature (objective O8).

Taking into account fake news detection, the solution proposed involves training a feature-engineered model on a merged English-German database. The characteristics used for modeling were considered to be independent of the language of the dataset (objective O7). By comparing multiple machine learning algorithms (e.g., Random Forest, Logistic Regression, Naive Bayes, and Support Vector Machine) and aiming for objective O6, an accuracy of 75.38% was obtained for the Random Forest Classifier for the all-features model. Additionally, the most relevant features were ranked and discussed in comparison with other results from related work (objective O8).

Taking into account the relation between malicious web links, fake news, and clickbait messages or headlines, this association is not very common; it is a niche scenario, really difficult to capture in real life. However, some articles and opinions were found for this research niche (objective O9).

6.2 Future Work

As a future direction for the literature survey on malicious web links detection (objective O1), one could continue to research and add an exhaustive analysis, including statistics, the impact of the survey articles on industry solutions, and the scalability and robustness of the proposed approaches.

For future work regarding the malicious web links detection domain and objective O2, there could be more research directed towards explainable artificial intelligence models, helping both computer science specialists and users to be better informed. Moreover, explainability can contribute to raising awareness of the deceiving tactics employed by attackers. Additionally, more ensemble types should be elaborated, containing machine learning and especially deep learning classifiers. Ensembles combined with nature-inspired algorithms could be a fruitful direction of investigation. Another topical research niche is represented by the usage of Large Language Models for different tasks, such as link classification, together with clickbait and fake news detection. Large Language Models could be employed for additional services to help and improve the Machine Learning classification.

Large-scale environments can enhance the model's accuracy and robustness (objective O3). Also,

sizable datasets can test the detection approach for scalability to prove that in real life, the solution is powerful and accurate enough. In addition, the link classification system should be tested for adversarial attacks.

Besides, the idea to transform URLs to images has potential and could be further explored in order to capture the web link logic into images better. For example, the image could have different regions where specific information from a URL will be encoded there (e.g, protocol, IP address if used, port, domain name, etc.). Other types of features could be employed, especially automatically extracted features using autoencoders, Large Language Models, and others. Additionally, collecting URLs can sometimes be a relevant direction for investigation since many mediums where links are spreading are end-to-end encrypted, for example, the case for WhatsApp. Thus, in these environments, the web link interception can be a novel idea for further analysis. This could help advance our results obtained for objective O4.

Considering collecting non-clickbait samples (objective O5), a future direction could be developing more complex heuristics or even using unsupervised methods to analyze the network traffic, tracking down what websites are clickbait or not. For improving objective O6, one could experiment with deep learning architectures, and even Large Language Models, to improve and do the classification for clickbait and fake news. Additionally, for objective O7, the system should include other types of features, such as the images embedded in the news article. On top of that, features could be automatically extracted and fed into the classification algorithm. The automatically-crafted characteristics have the advantage of being language-independent. Further advancing with objective O8 could involve testing the algorithms on supplementary data, collected from a vast variety of topics, to have a more significant outlook about the most relevant clues when performing clickbait and fake news detection.

Having in mind the association between clickbait, fake news, and malicious links, interdisciplinary efforts should be explored to further investigate this field, to advance with objective O9. An integrated system collecting and classifying fake news, clickbait, and links could be implemented over the web. Another interesting research idea, not very well explored, is considering clickbait and malicious links as a regression problem, or as ordinal classification.

Appendix A

List of abbreviations

ADA Adaptive Boosting

ADA-XGB-SVM-rbf Adaptive Boosting - Gradient Boosted Tree - Support Vector Machine with Radial Basis Function

API Application Programming Interface

ASCII American Standard Code for Information Interchange

CNN Convolutional Neural Network

CNNs Convolutional Neural Networks

CSS Cascading Style Sheets

DT Decision Tree

EU European Union

GPT Generative Pre-trained Transformer

HTTP Hypertext Transfer Protocol

HTML Hypertext Markup Language

JS JavaScript

KNN K-Nearest Neighbor

LDA Linear Discriminant Analysis

LLM Large Language Model

LLMs Large Language Models

LR Logistic Regression

LSTM Long Short Term Memory

ML Machine Learning

MLP Multi-layer Perceptron

NB Naive Bayes

NC Nearest Centroid

PAC Passive Aggressive Classifier

PSO Particle Swarm Optimization

RBF Radial Basis Function

RGB Red, Green, and Blue

RF Random Forest

SGD Stochastic Gradient Descent

SVM Support Vector Machine

TF-IDF Term Frequency and Inverse Document Frequency

URL Uniform Resource Locator

URLs Uniform Resource Locators

WWW World Wide Web

XGB Gradient Boosted Tree

Bibliography

All our publications

[1] **Coste, Claudia Ioana**. "Controlling the click bait". In: *Proceedings of the International Student Conference StudMath-IT*. 2018, pp. 11–17.

[2] **Coste, Claudia Ioana**. "Online bad habits: fake news and clickbait". In: *Proceedings of the Student Interdisciplinary Conference The European Union and Global Order*. Cluj Universitary Press, 2019, pp. 179–193.

[3] **Coste, Claudia Ioana**. "Using Ensemble Models for Malicious Web Links Detection." In: *ICAART (3)*. 2024, pp. 657–664.

[4] **Coste, Claudia Ioana** and Andreica, Anca-Mirela and Chira, Camelia. "Malicious Web Links Detection Using Ensemble Models." In: *WEBIST*. 2023, pp. 268–275.

[5] **Coste, Claudia Ioana** and Bufnea, Darius. "Advances in clickbait and fake news detection using new language-independent strategies". In: *Journal of Communications Software and Systems* 17.3 (2021), pp. 270–280.

[6] **Coste, Claudia Ioana** and Bufnea, Darius and Niculescu, Virginia. "A new language independent strategy for clickbait detection". In: *2020 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*. IEEE. 2020, pp. 1–6.

[7] **Coste, Claudia Ioana** and Chira, Camelia and Bufnea, Darius-Vasile. "Malicious Web Links Detection: A Survey". In: *SUBMITTED* -.- (2026).

[8] **Coste, Claudia-Ioana**. "Clickbait & Fake News - a cross language approach". Masters Thesis. Babeș-Bolyai University, Cluj-Napoca, Romania, 2021.

[9] **Coste, Claudia-Ioana**. "Contribuții privind identificarea legăturilor web de tip „clickbait" (eng., Contributions Regarding Clickbait Web Links)". Bachelor's Thesis. Babeș-Bolyai University, Cluj-Napoca, Romania, 2019.

[10] **Coste, Claudia-Ioana**. "Malicious Web Links Detection - A Comparative Analysis of Machine Learning Algorithms". In: *Studia Universitatis Babeș-Bolyai Informatica* 68.1 (2023), pp. 21–36. ISSN: 2065-9601. DOI: 10.24193/subbi.2023.1.02. URL: https://www.cs.ubbcluj.ro/~studia-i/journal/journal/article/view/86.

[11] **Coste, Claudia-Ioana**. "Malicious Web Links Detection Based on Image Processing and Deep Learning Models". In: *2024 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. IEEE. 2024, pp. 445–449. ISBN: 979-8-3315-0494-6/24. DOI: 10.1109/WI-IAT62293.2024.00071.

[12] **Coste, Claudia-Ioana.** and Kaisser, Thomas. "Malicious Web Links Detection with Large Language Models." In: *Lecture Notes in Business Information Processing* Accepted for publication.In press. (2026).

[13] Kaisser, Thomas. and **Coste, Claudia-Ioana.** "Using ChatGPT for Malicious Web Links Detection." In: *WEBIST*. 2024, pp. 425–432. ISBN: 978-989-758-718-4.

The rest of the publications

[14] Husam Adas, Sachin Shetty, and Waled Tayib. "Scalable detection of web malware on smartphones". In: *2015 International Conference on Information and Communication Technology Research (ICTRC)*. IEEE. Abu Dhabi, UAE: IEEE, 2015, pp. 198–201.

[15] Alexa Internet, Inc. *The top 500 sites on the web*. Alexa-TopSites (https://www.alexa.com/topsites) (last accessed February 2, 2023). 2023.

[16] Veronica Anghel. *Why Romania just canceled its presidential election*. News article (https://www.journalofdemocracy.org/online-exclusive/why-romania-just-canceled-its-presidential-election) (last accessed Feb 13, 2025). Dec. 2024.

[17] Dominic DiFranzo and Kristine Gloria-Garcia. "Filter bubbles and fake news". In: *XRDS: crossroads, the ACM magazine for students* 23.3 (2017), pp. 32–35.

[18] Birhanu Eshete, Adolfo Villafiorita, and Komminist Weldemariam. "Malicious website detection: Effectiveness and efficiency issues". In: *2011 First SysSec Workshop*. IEEE. Amsterdam, The Netherlands: IEEE, 2011, pp. 123–126.

[19] Jessica Valasek Estenssoro. Cybersecurity Threats and Trends & Malware Statistics 2024 (https://www.avg.com/en/signal/malware-statistics) (last accessed Nov 24, 2024). Nov. 2024.

[20] Kaggle Inc. *Kaggle*. Kaggle (https://www.kaggle.com/) (last accessed March 20, 2025). 2022.

[21] GTKlondike. *Machine-Learning-for-Security-Analysts*. Dataset website (https://github.com/NetsecExplained/Machine-Learning-for-Security-Analysts) (last accessed Jun 22, 2024). June 2019.

[22] Davide Mancino and Frederica Fragapane. Digital Literacy in the EU (https://data.europa.eu/en/publications/datastories/digital-literacy-eu-overview) (last accessed Nov 24, 2024). Dec. 2023.

[23] Oxford University Press. *Clickbait noun - definition, pictures, pronunciation and usage notes*. Clickbait definition (https://www.oxfordlearnersdictionaries.com/definition/english/clickbait?q=clickbait) (last accessed Dec 20, 2024). Dec. 2024.

[24] F. et al. Pedregosa. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[25] Lexie Pelchen and Samantha Allen. *Internet usage statistics in 2024*. Forbes - Internet Statistics (https://www.forbes.com/home-improvement/internet/internet-statistics/) (last accessed Nov 24, 2024). Nov. 2024.

[26] Ani Petrosyan. *Internet and social media users in the world 2024*. Statista article (https://www.statista.com/statistics/617136/digital-population-worldwide/) (last accessed Nov 12, 2024). Nov. 2024.

[27] Kai Shu et al. "Fake News Detection on Social Media: A Data Mining Perspective". In: *SIGKDD Explor. Newsl.* 19.1 (Sept. 2017), pp. 22–36. ISSN: 1931-0145. DOI: 10.1145/3137597.3137600.

[28] Manu Siddhartha. *Malicious URLs dataset*. Kaggle - Malicious URLs dataset (https://www.kaggle.com/datasets/sid321axn/malicious-urls-dataset) (last accessed May 2, 2024). July 2021.

[29] Damian Tambini. "Fake news: public policy responses". In: *London School of Economics and Political Science* (2017).

[30] Christian Urcuqui et al. "Machine Learning Classifiers to Detect Malicious Websites." In: *SSN*. Spain: SSN, 2017, pp. 14–17.

[31] *World Population Clock: 8.2 Billion People (LIVE, 2024) - Worldometer*. Worldmeter website (https://www.worldometers.info/world-population/) (last accessed Nov 12, 2024). Nov. 2024.