Enhancing Online Educational Resource Security with BiGRU Attention Models

Simone Re[1,*,†], Matteo Olivieri[1,*,†], Ricardo Anibal Matamoros Aragon[2,3,*,†], Alessandro Solinas[2,4,*,†] and Francesco Epifania[2]

[1] *Informattiva S.r.l., Milan, Italy*
[2] *Social Things S.r.l, Milan, Italy*
[3] *Department of Computer Science, University of Milano Bicocca, Milan, Italy*
[4] *Politecnico di Milano, Milan, Italy*

Abstract

In today's interconnected digital landscape, the Internet plays a pivotal role in our daily human activities. However, the intricacy of the online communication network exposes vulnerabilities that can be exploited by malicious actors, who adopt increasingly sophisticated strategies to compromise cybersecurity. This issue extends to the domain of e-learning, where the protection of user personal data and the interaction with external educational resources become critical aspects.

In this context, we introduce an e-learning platform developed by Informattiva, integrated with an advanced cybersecurity mechanism. This mechanism is designed to analyze educational resources from external repositories, such as Merlot.org, aiming to identify potential insecurities based on URLs. To achieve this, we implemented a model based on the Bidirectional Gated Recurrent Unit (BiGRU) with attention mechanisms, focusing on the identification of potentially malicious web addresses. Preliminary results indicate that, through bidirectional processing and attention mechanisms, our methodology has the potential to effectively differentiate suspicious URLs from secure ones.

Keywords

Anomaly Detection, Artificial Intelligence, E-learning, Attention Mechanism

1. Introduction

In the interconnected world of today's digital era, where everything is connected, the internet stands as the cornerstone of modern communication and information dissemination. Its pervasive presence in our daily lives has revolutionized the way we learn, work, and interact with the world. Yet, as the internet continues to weave itself into the fabric of society, it concurrently exposes us to a growing spectrum of digital threats and vulnerabilities. Cybercriminals, in their relentless pursuit of exploiting these opportunities, constantly devise new tactics to breach our

AIABI 2023: 3rd Italian Workshop on Artificial Intelligence and Applications for Business and Industries, November 9, 2023, Milano, Italy

[*] Corresponding author.

[†] These authors contributed equally.

simone.re@smricercaselezione.com (S. Re); matteo.olivieri@smricercaselezione.com (M. Olivieri); ricardo.matamoros@socialthingum.com (R. A. M. Aragon); alessandro.solinas@socialthingum.com (A. Solinas); francesco.epifania@socialthingum.com (F. Epifania)

0000-0002-1957-2530 (R. A. M. Aragon); 0000-0002-5428-3187 (A. Solinas)

digital security, endangering both individuals and organizations alike.

In response to this ever-present cyber threat, the research team at Informattiva Srl has embarked on a mission to safeguard one of the most vital sectors of our digital realm: Elearning [1]. As the demand for remote learning and virtual classrooms has surged in recent years, the availability of open-source educational resources has grown exponentially. These resources offer educators an invaluable toolbox for enhancing the quality and effectiveness of their courses. Amongst this extensive collection of educational resources, there is a concerning lack of security controls which leaves resources vulnerable to exploitation. Recognizing this critical gap in online education, our research team has dedicated considerable efforts to address this issue head-on. We have developed a platform designed to empower educators with the means to fortify the security of their educational materials. At its core, our platform leverages a sophisticated security firewall capable of discerning malicious intentions by scrutinizing the URLs associated with online resources [2].

Researchers are currently exploring the use of machine learning for detecting malicious URLs. One notable study by Vanhoenshoven et al. [3] utilized a multi layer perceptron (MLP) for this purpose. The study discovered that varying feature sets can affect the accuracy of the results when working with the same dataset.

In their study, Azeez et al. [4] employed a naive Bayesian algorithm to identify malicious URLs by analyzing the syntax, vocabulary, hosts, and other content of the URL present in the email. Laughter et al. [5] incorporated the HTTP request features in the detection feature set by considering the process of visiting the website. In recent years, the growth of deep learning has brought new developments to detecting malicious web pages using those techniques [6]. In particular, Recurrent Neural Networks (RNN) are considered the best-performing and therefore most suitable models to perform anomaly detection due to their ability to capture sequential dependencies and temporal patterns in data, making them exceptionally adept at identifying deviations from expected patterns in various applications.

In this paper, we shed light on an innovative approach centered around the utilization of a Dropout Attention Bidirectional Gated Recurrent Unit (DA-BiGRU) model [7]. Our primary focus is on identifying potentially malicious web addresses within the vast sea of online educational resources. By harnessing the power of bidirectional processing and the precision of attention mechanisms, our approach showcases the potential to differentiate between suspicious URLs and harmless ones, thereby strengthening the security of online educational content.

As we delve into the intricacies of our research, we will explore the theoretical foundations of the DA-BiGRU model and its application in the realm of URL analysis. Through a comprehensive examination of this model and its experimental results, we aim to contribute to the growing body of literature addressing cybersecurity in the context of online education. Our work not only underscores the importance of securing educational resources but also demonstrates the transformative potential of cutting-edge machine learning techniques in the fight against digital threats [8].

In the following sections, we delve deeper into the methodology, results, and implications of our research, offering insights and recommendations that can pave the way for a safer and more secure online learning environment.

2. Datasets Utilized for Anomaly Detection in URL Analysis

When it comes to detecting anomalies in URLs, the data that is selected is crucial for keeping online systems and networks secure. It is important to have a good understanding of the typical patterns and behaviors of URLs so that any potentially harmful or unusual web traffic can be identified and dealt with. Having accurate and thorough data enables the anomaly detection algorithms to distinguish between genuine website interactions and suspicious activity, which helps to improve cybersecurity efforts and guard against eventual threats. To this end, we selected two open-source datasets from Kaggle.com about malicious URLs.
First, we used the Malicious URLs dataset[9], which contains 651,191 URLs with 34% of anomalies. This dataset will be divided into train, validation, and test. Then as an additional test and as proof of the model's scalability, we utilized the Malicious_n_Non-Malicious URL dataset [10] which is composed of 411,247 URLs and 18% of anomalies. The algorithm under consideration was validated using the previously described datasets. Subsequently, it was applied to the dataset from Merlot.org. This latter dataset represents a fundamental resource for the e-learning platform developed by Informattiva, allowing users to enrich and customize their educational paths by integrating external educational resources.

3. Model in-depth

In this chapter, we will summarize the DA-BiGRU attention model architecture, diving into the details of some key aspects. The meaning of the symbols used in this section is summarized in Table 1

3.1. BiGRU architecture

Introduced by Cho, et al. [11] in 2014, GRU aims to solve the vanishing gradient problem that comes with a standard recurrent neural network. Its introduction was made as an improvement of the Long Short-Term Memory (LSTM) architecture. The key components of GRU, summarized in Figure 2, are:

- **Hidden State**: to capture information from previous steps GRU maintains a hidden state h_t, as in traditional RNNs
- **Update Gate**: this is a crucial component of GRU that controls how much of the previous hidden state should be retained.
  It is computed through a sigmoid, as:

$$z_t = \sigma(W_{xz} \cdot x_t + W_{hz} \cdot h_{t-1} + b_z) \tag{1}$$

- **Reset Gate**: the reset gate determines how much of the previous hidden state should be reset or forgotten when computing the new candidate state. It is computed similarly to the update gate:

$$r_t = \sigma(W_{xr} \cdot x_t + W_{hr} \cdot h_{t-1} + b_r) \tag{2}$$

Symbol	Description
σ	Activation function
$W_{xr}, W_{hr}, W_{xz}, W_{hz}$	Weight parameters
b_r, b_z	Bias parameters
h_{t-1}	Hidden state in the previous timestamp
x_t	Current input

Table 1
Symbols meaning

The candidate hidden state is computed as:

$$\tilde{h}_t = \tanh(W_{zh} \cdot x_t + W_{hh} \cdot (h_{t-1} \odot r_t) b_h) \tag{3}$$

and then it is combined with the update gate in the computation of the new hidden state h_t as follows:

$$h_t = h_{t-1} \odot z_t + (1 - z_t) \odot \tilde{h}_t \tag{4}$$

From the last equation, is evident how the update gate (z_t) impacts the new hidden state. When it is closer to 1, the model retains most of the information of the hidden state at the previous timestamp (h_{t-1}), while if it approaches 0 most of the informations are retained from the candidate hidden state ($\tilde{h}_t$).

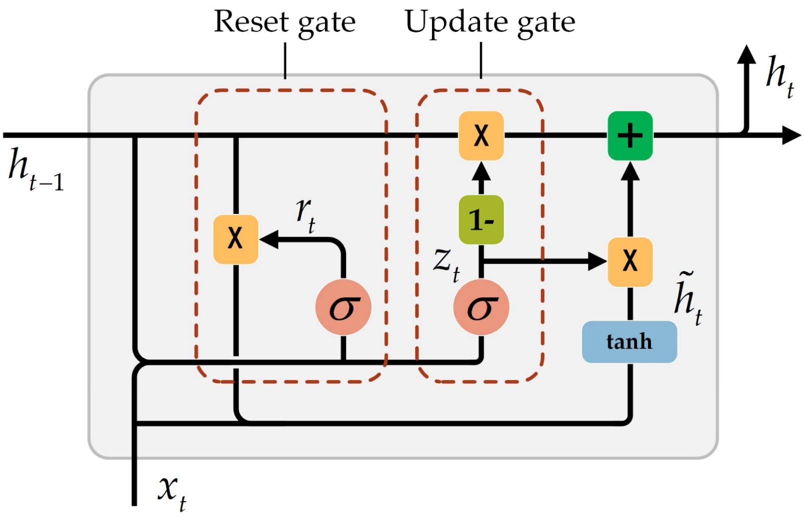

Figure 1: GRU architecture [12]

Bi-GRU combines two separate GRUs that process the input sequence in both the forward and backward directions simultaneously, enabling the model to capture contextual information from both past and future data points, making it particularly useful in tasks where understanding bidirectional context is crucial. By processing the URL sequence bidirectionally, Bi-GRU can identify patterns and relationships between different parts of the URL, such as domain names, subdomains, and query parameters. This bidirectional processing enables it to understand how different components of the URL relate to each other and extract valuable features for tasks like URL classification, parsing, or anomaly detection. Additionally, Bi-GRU's ability to model both past and future context ensures a comprehensive understanding of the URL.

3.2. Attention Mechanism for Enhanced URL Segment Analysis: Mathematical Formulation

URLs vary in structure across different locations, necessitating distinct specifications. An attention mechanism is introduced to comprehend the interdependence of words or symbols across diverse URL segments. This attention mechanism filters out irrelevant content and prioritizes crucial URL information, enhancing data utilization and ultimately elevating model accuracy [13, 14]. The mathematical formulation for this process is detailed below.

$$e_t = W^T \sigma(W_l \cdot x_t) \tag{5}$$

$$q_t = \frac{\exp(e_t)}{\sum_{t=1}^{T} e_t} \tag{6}$$

$$x_t^* = \sum_{t=1}^{T} q_t \cdot x_t \tag{7}$$

In Equation 5, the attention vector is computed using the input information at time t x_t, the learned weight matrices W_l, W^T and a hyperbolic tangent ($tanh$) as activation. Then the vector is normalized through a softmax function, as can be seen in Equation 7. Finally, the output x_t^* is obtained from element-wise multiplication of the input and attention vectors.

3.3. Dropout mechanism

Dropout is a regularization technique commonly employed in deep learning models to prevent overfitting. During training, it randomly deactivates a fraction of neurons or units in a neural network, effectively dropping them out, which encourages the network to become more robust and generalize better to unseen data. This stochastic dropout process helps prevent co-dependencies between neurons and promotes a more robust and reliable model.

3.4. Model structure

Within the context of deep learning architectures, the DA-BiGRU model emerges as a particularly advanced solution, characterized by a complex yet highly effective structure. The initial phase of the processing involves the preprocessing of input URLs. This critical phase employs the Word2Vec technique [15], a model renowned for its ability to transform text sequences into dense vector representations, commonly known as "embeddings". These embeddings allow the URL text to be represented in a format that can subsequently be processed by the model, ensuring a coherent and informative semantic representation.

Following this transformation phase, the input is introduced into a dropout layer. This layer, positioned before the BiGRU architecture, serves to prevent overfitting and enhance the model's

robustness. Within the BiGRU architecture, forward and backward propagation of the hidden state occurs, enabling the model to capture and process the temporal dependencies present in the data.
The output from the BiGRU structure is then fed into an attention layer. This layer plays a pivotal role in identifying and emphasizing the most relevant and pertinent features of the input, ensuring that the model focuses on the most informative aspects of the URLs.
Finally, the process concludes with a fully connected layer followed by a softmax function. This combination is responsible for the final classification, allowing the model to categorize the URLs based on the features learned during the training phase.

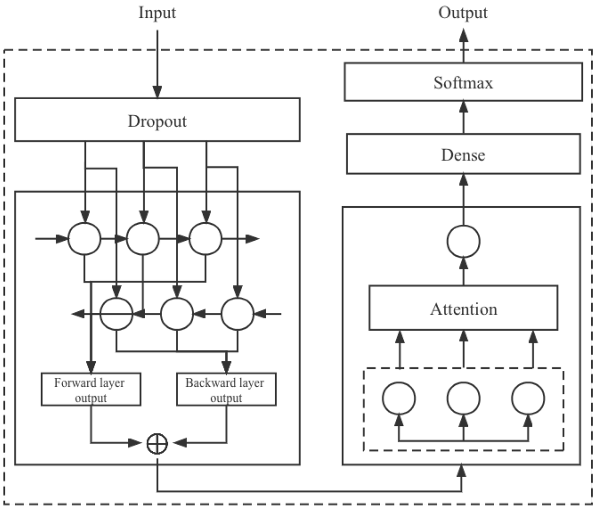

Figure 2: DA-BiGRU architecture [7]

4. Results

In the subsequent section, we delineate the outcomes procured during the model evaluation phase. Specifically, the model underwent training for 30 epochs, employing binary cross-entropy as the designated loss function, complemented by the Adam optimizer with a learning rate set at 10^{-3}.
Throughout the progression of each epoch, we meticulously observed pivotal performance indicators, encompassing loss, accuracy, precision, and recall, for both the training and validation datasets. A graphical representation of these metrics can be referenced in Figure 3. It's imperative to highlight that the model's preservation is predicated on the optimal validation loss, thereby rendering any overfitting tendencies in the concluding epochs inconsequential. The output of the model under consideration extends within the range between 0 and 1, where a higher value suggests a greater likelihood that a given sample is identified as an anomaly. To precisely determine which samples to categorize as anomalies, a specific threshold was defined. In this scenario, the precision metric, representing the ratio between true positive instances and the set of instances predicted as positive, assumes paramount importance. The

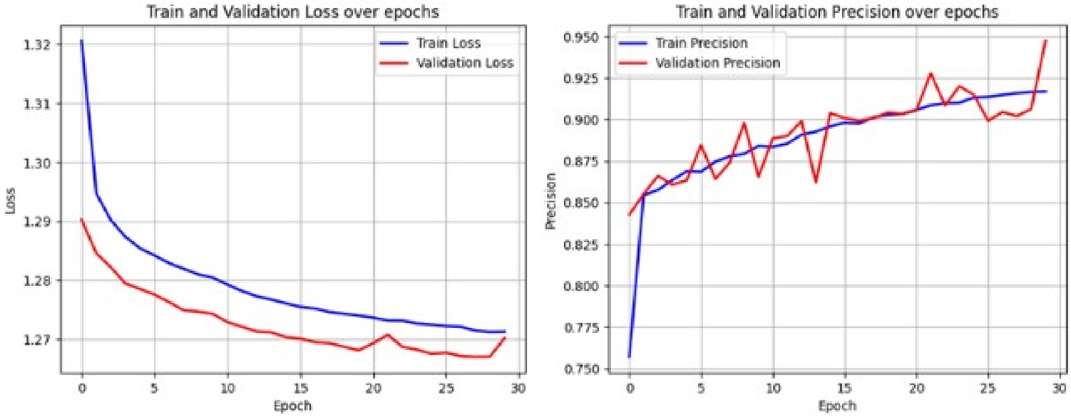

Figure 3: Model's loss and precision over epochs

primary objective was to favour precision over recall, with the intent to limit the number of false positives and prevent suboptimal resource allocation. After a weighted analysis, it was deduced that a threshold of 0.99 represents the ideal balance, effectively classifying samples as anomalies if their probability exceeds this value. The choice of this threshold aligns with the aim of ensuring a high level of reliability in anomaly detection while simultaneously reducing the danger of excluding valuable resources due to erroneous identifications.

In Table 2, we provide a detailed exposition of the metrics computed on the aforementioned distinct test samples. The composition of the second dataset was intentionally skewed, encompassing a mere 5% anomalies. This disproportionate dataset was meticulously curated to test the model's proficiency in anomaly detection under conditions that emulate real-world scenarios. Upon scrutinizing the outcomes, it is evident that for the inaugural dataset, our model manifests commendable efficacy, accurately categorizing 94% of websites. Notably, it evinces the adeptness to pinpoint 87% of malevolent websites. Moreover, the probability that websites adjudged as malicious by the model are indeed malicious stands at an impressive 94.5%. Transitioning to the evaluation on the second, highly imbalanced dataset, our model sustains elevated levels of accuracy and precision, with both metrics consistently surpassing the 90% mark. However, a marked diminution in recall is discernible. This attenuation in recall is attributable to our judicious selection of the threshold, a parameter that offers potential for optimization contingent on specific objectives. To elucidate, by hypothetically calibrating the threshold to 0.5, we attain a recall rate of 80%. This recalibration, nonetheless, incurs a decrement in precision, plummeting it to 88%. The determination of an optimal threshold necessitates a strategic balance between recall and precision, contingent upon the bespoke requirements and inherent limitations of the application in question.

Dataset	Accuracy	Precision	Recall
Test 1	0.9404	0.9456	0.8728
Test 2	0.9220	0.9275	0.6889

Table 2
Summary of the model's metrics

5. Conclusions

In conclusion, our investigations have illuminated significant insights regarding the enhancement of security measures applied to digital educational resources in today's interconnected online environment. While the Internet stands as a pivotal medium for education and communication, it subjects us to ever-evolving cyber threats, necessitating the adoption of proactive strategies to counter potential malicious activities. In response to this challenge, our research group has devised an advanced firewall system, specifically aimed at bolstering the security of educational content. Despite the widespread availability of open-source educational resources, the absence of adequate security controls has rendered such resources susceptible to exploitation. Our analysis has centered on the adoption of a Bidirectional Gated Recurrent Unit (BiGRU) attention model, expressly designed for the identification of potentially harmful web addresses. Leveraging the capabilities of bidirectional processing and attention mechanisms, the proposed methodology has showcased considerable potential in distinguishing between innocuous and potentially dangerous URLs. The results obtained underscore the essentiality of employing advanced machine learning methodologies in the realm of cybersecurity for educational resources. Such integration has facilitated significant advancements in strengthening the digital learning environment. Looking forward, the importance of continuous optimization of our model is evident, along with the need to modulate detection thresholds based on the specific security requirements of educational platforms and various digital contexts. As we continue refining our approach, we remain steadfast in our commitment to enhancing security measures in the digital age, with the aim of ensuring educators and learners can optimally utilize online resources in a context of trust and serenity.

References

[1] Bhatia, Meghna, and J. K. Maitra. ”E-learning platforms security issues and vulnerability analysis.” 2018 International Conference on Computational and Characterization Techniques in Engineering & Sciences (CCTES). IEEE, 2018.

[2] Tamjidyamcholo, Alireza, et al. ”Evaluation model for knowledge sharing in information security professional virtual community.” Computers & Security 43 (2014): 19-34.

[3] Malak Aljabri, Hanan S. Altamimi, Shahd A. Albelali, Maimunah Al-Harbi, Haya T. Alhuraib, Najd K. Alotaibi, Amal A. Alahmadi, Fahd Alhaidari, Rami Mustafa A. Mohammad, and Khaled Salah. *Detecting Malicious URLs Using Machine Learning Techniques: Review and Research Directions. IEEE Access*, Volume 10, 2022, Pages 121395-121417. DOI: 10.1109/ACCESS.2022.3222307.

[4] Nureni Ayofe Azeez, Balikis Bolanle Salaudeen, Sanjay Misra, Robertas Damaševièius,

and Rytis Maskeliûnas. *Identifying Phishing Attacks in Communication Networks Using URL Consistency Features. Int. J. Electron. Secur. Digit. Forensic*, Volume 12, Number 2, January 2020, Pages 200-213. ISSN: 1751-911X. DOI: 10.1504/ijesdf.2020.106318. URL: https://doi.org/10.1504/ijesdf.2020.106318.

[5] Ashley Laughter, Safwan Omari, Piotr Szczurek, and Jason Perry. *Detection of Malicious HTTP Requests Using Header and URL Features.* In: *Advances in Digital Forensics XVI*, Year 2021, Month January, Pages 449-468. ISBN: 978-3-030-63088-1. DOI: 10.1007/978-3-030-63089-8_29.

[6] Hou, Yung-Tsung, et al. ”Malicious web content detection by machine learning.” expert systems with applications 37.1 (2010): 55-60.

[7] Tiefeng Wu, Miao Wang, Yunfang Xi, and Zhichao Zhao. *Malicious URL Detection Model Based on Bidirectional Gated Recurrent Unit and Attention Mechanism. Applied Sciences*, Volume 12, Number 23, 2022, Article Number 12367. ISSN: 2076-3417. DOI: 10.3390/app122312367. URL: https://www.mdpi.com/2076-3417/12/23/12367.

[8] Musser, Micah, and Ashton Garriott. ”Machine learning and cybersecurity.” Center for Security and Emerging Technology: Washington, DC, USA (2021).

[9] Manu Siddhartha. (2016). *Malicious URLs dataset*, version 1. Retrieved in 2023 from https://www.kaggle.com/datasets/sid321axn/malicious-urls-dataset.

[10] antonyj. (2017). *Malicious_n_Non-Malicious URL*, version 1. Retrieved in 2023 from https://www.kaggle.com/datasets/antonyj453/urldataset.

[11] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. *Learning Phrase Representations using {RNN} Encoder{–}Decoder for Statistical Machine Translation.* In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing ({EMNLP})*, Doha, Qatar, October 2014. Publisher: Association for Computational Linguistics. DOI: 10.3115/v1/D14-1179. URL: https://aclanthology.org/D14-1179. Pages 1724-1734.

[12] Pengpeng Li, An Luo, Jiping Liu, Yong Wang, Jun Zhu, Yue Deng, and Junjie Zhang. *Bidirectional Gated Recurrent Unit Neural Network for Chinese Address Element Segmentation. ISPRS International Journal of Geo-Information*, Volume 9, Number 11, 2020, Article Number 635. ISSN: 2220-9964. DOI: 10.3390/ijgi9110635. URL: https://www.mdpi.com/2220-9964/9/11/635.

[13] Chorowski, Jan K., et al. ”Attention-based models for speech recognition.” Advances in neural information processing systems 28 (2015).

[14] Bahdanau, Dzmitry, et al. ”End-to-end attention-based large vocabulary speech recognition.” 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2016.

[15] Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. *Efficient Estimation of Word Representations in Vector Space. Proceedings of Workshop at ICLR*, Volume 2013, January 2013.