Baghdad Science Journal

Volume 23 | Issue 5 Article 14

5-26-2026

A Novel FFUEAT Technique Enhancing the Performance of Multiple URL Classification and Cybersecurity using RNN and Transformer-based Models

Zafar Ali
Faculty of Artificial Intelligence, University Teknologi Malaysia, Kuala Lumpur, Malaysia, zmahar13@gmail.com

Siti Sophiayati Yuhaniz
Faculty of Artificial Intelligence, University Teknologi Malaysia, Kuala Lumpur, Malaysia, sophia@utm.my

Noureen Noureen
Faculty of Artificial Intelligence, University Teknologi Malaysia, Kuala Lumpur, Malaysia, noureen@graduate.utm.my

Ghulam Mujtaba
Department of Computer Science, Sukkur IBA University, Sukkur 65200, Pakistan, mujtaba@iba-suk.edu.pk

Husham M. Ahmed
College of Engineering University of Technology Bahrain Kingdom of Bahrain, Hushamahmed13@outlook.com

How to Cite this Article

Ali, Zafar; Yuhaniz, Siti Sophiayati; Noureen, Noureen; Mujtaba, Ghulam; and Ahmed, Husham M. (2026) "A Novel FFUEAT Technique Enhancing the Performance of Multiple URL Classification and Cybersecurity using RNN and Transformer-based Models," *Baghdad Science Journal*: Vol. 23: Iss. 5, Article 14.
DOI: https://doi.org/10.21123/2411-7986.5298

BAGHDAD SCIENCE JOURNAL 2026;23(5):1739–1759

RESEARCH ARTICLE

A Novel FFUEAT Technique Enhancing the Performance of Multiple URL Classification and Cybersecurity using RNN and Transformer-based Models

Zafar Ali [1,*], **Siti Sophiayati Yuhaniz** [1], **Noureen** [1], **Ghulam Mujtaba** [2], **Husham M. Ahmed** [3]

[1] Faculty of Artificial Intelligence, University Teknologi Malaysia, Kuala Lumpur, Malaysia
[2] Department of Computer Science, Sukkur IBA University, Sukkur 65200, Pakistan
[3] College of Engineering University of Technology Bahrain Kingdom of Bahrain

ABSTRACT

Thousands of new websites are published every day, which pose significant challenges for web classification and cybersecurity. URL classification datasets, including general and cybersecurity-specific ones, face challenges such as class imbalance, noise, and ambiguous data, which can significantly affect model performance. This study proposes a novel Fine-Tuned FastText Unsupervised Embedding Augmentation Technique (FFUEAT). Datasets (DMOZ and Phishing) were used to evaluate the performance of the proposed technique, achieving F1-scores of 0.8639 with DistilBERT and 0.8891 with BERT. The performance of sparse minority classes, achieving accuracy increases of 20.88% for 'home', 58.04% for 'news', and 11.18% for the 'kids' categories of the publicly available DMOZ dataset using the FFUEAT technique. When applied to the cybersecurity dataset (Phishing), leveraging a BiLSTM model, the proposed technique achieved F1-scores of 0.98 for legitimate sites and 0.99 for phishing URLs. The impact of class imbalance, noise and ambiguous data in URL classification datasets is reduced by applying the FFUEAT technique. This method offers a promising way to improve web classification and cybersecurity threat detection, contributing to better online content management and safety.

Keywords: Class imbalance, Cybersecurity, FastText, RNN and transformers, URL classification

Introduction

The exponential growth of digital content, with millions of web pages uploaded daily, poses challenges such as web page optimization, filtering, phishing, malware detection, and search engine optimization, underscoring the need for robust and scalable classification frameworks. This study focuses on multiple URL classification using the general DMOZ dataset and binary classification using the cybersecurity phishing dataset. The multiple URL classification task is more challenging than binary classification using the phishing dataset.

Traditional machine learning techniques for web page classification involve extracting features from web content and categorizing them into predefined classes or identifying phishing content. However, these techniques could not reach an ultimate solution due to the continuous rapid growth of the web pages. Yuan and Liu,[1] and Guo[2] developed a web crawler capable of efficiently collecting and analyzing unstructured and semi-structured data from

Received 13 July 2025; revised 5 December 2025; accepted 7 January 2026.
Available online 26 May 2026

* Corresponding author.
E-mail addresses: zmahar13@gmail.com (Z. Ali), sophia@utm.my (S. S. Yuhaniz), noureen@graduate.utm.my (Noureen), mujtaba@iba-suk.edu.pk (G. Mujtaba), Hushamahmed13@outlook.com (H. M. Ahmed).

https://doi.org/10.21123/2411-7986.5298

the web; however, the process was computationally intensive. Extracting features from web content is computationally intensive and faces scalability issues due to the rapid growth of internet resources.[3] Consequently, research shifted toward more efficient methods, focusing on feature extraction directly from URLs.[4,5] Web page classification has been widely studied, with advancements in feature engineering and machine learning classifiers. Li, D[4] identified an important point that URL domain names are often short, unstructured, and lack contextual relationships, making traditional word segmentation tools ineffective, while Zhang and Rao[5] proposed a method, called n-BiLSTM, that aims to improve text classification performance by leveraging both multi-scale feature representation (through n-grams) and contextual information (through BiLSTM). Cyprienna, Zo Lalaina Yannick[6] improve the efficiency of detecting malicious URLs using active learning, specifically the Query-based Committee (QBC) strategy. Despite these advancements, most studies focus on binary classification, leaving multi-class classification underexplored in terms of the multiple majority and minority classes. However, in the context of URL classification, binary classification is primarily employed in cybersecurity datasets that contain benign and malicious categories.

As discussed in the literature review section, prior studies have implemented approaches such as neural word embedding, Knowledge Graph with Tokenisation, Particle Swarm Optimisation (PSO), the Rabbit Optimisation Algorithm (ROA), and Hybrid Feature Engineering to improve the efficacy of URL classification. The results of prior approaches are satisfactory, with limited but balanced training and test sets. However, all prior techniques still struggle with three key factors of the dataset: minority classes, multifaceted feature classes, and noisy data. Therefore, these three factors severely affect the performance of deep learning models, so there is a need for a robust approach to address the persistent problem.

The adoption of deep learning techniques has further advanced URL classification by enabling the adaptation of several methods related to feature extraction. For example, Deng, Du, and Shen[7] used n-BiLSTM to improve text classification performance by leveraging both multi-scale feature representations. However, even with deep learning classifiers, multi-class URL classification struggles with sparse minority classes; the performance of the models suffers due to class imbalance, noise, and ambiguous data found in the URLs.

The DMOZ dataset, the largest and most imbalanced resource for URL classification, poses this challenge because 1.5 million URLs across 15 categories suffer from severe class imbalance, biasing models toward majority classes and compromising minority class performance.[8] Advanced techniques like resampling, synthetic data generation, and cost-sensitive learning are essential to address this imbalance.[9,10] Yet, research predominantly focuses on investigating several machine learning and deep learning models with different feature engineering techniques to address challenges found in URL classification datasets, such as class imbalance, noise, and ambiguous data that affect the performance of models. Most importantly, the class imbalance, the uneven distribution of the classes in multi-majority and multi-minority classes, is underexplored.[8] The cybersecurity datasets also face the challenge of class imbalance, as the benign category has more representation compared to the malicious category; however, the authors selected an equal sample from both categories. In this way, the features used are lost due to the reduction of the majority class in the dataset. Along with the class imbalance problem, URL classification is also affected by noise and misclassified data.[11–13] Therefore, Kexin, Liang[14] achieved a below-average efficacy of 0.50 F1 score.

The URL classification datasets contain hundreds of ambiguous and misclassified URLs, as shown in the URL classification dataset, with some misclassification examples illustrated in Table 1. These factors present significant challenges for deep neural networks in achieving optimal classification accuracy.[14,15] Previous research has examined various feature engineering techniques to address these issues.[16–18] Although phishing URL detection has achieved promising results, further improvements in model performance are still needed.[19,20] This study introduces a novel Fine-Tuned FastText Unsupervised Embedding Augmentation Technique (FFUEAT) to improve URL instances to a standardized dataset both contextually and semantically. The FFUEAT approach is intended to reduce class imbalance by tackling issues including noise, ambiguous data and sparse minority classes. This method improves the overall accuracy of URL classification, including general and cybersecurity datasets. The method represents a substantial advancement in dataset quality, facilitating more effective deep learning applications.

Related work

Machine learning and deep learning are the two primary categories of natural language processing (NLP) techniques used for URL classification. Because machine learning depends on manual feature engineering, researchers must find and extract crucial

Table 1. Ambiguous and misclassification data found in URLs.

URL	Reason to ambiguous	Category
http://lacasta.iespana.es	Spanish website; content unclear without translation.	'adult'
http://www.b4uhost.com	Hosting site; unclear if used for adult content.	'adult'
http://www.geocities.com/riverzildjian	The generic name and GeoCities domain indicate a personal page.	'arts'
http://www.textkit.com	This one is a borderline between the computer and writing category.	'arts'
http://www.solarattic	Could be confused for home and garden or even 'science	'business'
http://www.oilnergy.com	Sounds scientific because of the word energy	'business'
http://ubdesigner.com	It sounds more related to aviation.	'computers'
http://www.quickstart.com	Most like kids or games	'computers'
http://www.salute.co.uk	The generic name could be a military or event site.	'games'
http://www.mars.org	Academic document related to Mars - unclear how related to games	'games'
http://www.4everbeautiful.com	Refer to the beauty industry instead.	'health'
http://www.ckeepkidssafe.org	kid's safety, not necessarily a health	'health'

characteristics. Deep learning, on the other hand, can automatically identify and extract complex features from the data. Feature extraction from web-based content was the primary focus of early research.[21] Subsequent research started to directly extract features from URLs.[22,23] Since deep learning classifiers like convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are effective at learning, recent work has shifted from machine learning to deep learning approaches.[10,19]

The author, Kexin Liang[14] implemented URLNet using character embeddings, achieving 81.23% accuracy for five soft classes, and 74.41% accuracy for the five hard classes. Another study by Hung, Hung, and Diep[15] proposed a CNN-based model using URL features only for short-text in Vietnamese web content. It is observed that most publicly available URL classification datasets have an uneven distribution of classes, with a bias towards majority classes, such as WebKB, the Open Directory Project (DMOZ), and proxy datasets. Because the imbalanced dataset decreased the performance of the model, especially for underrepresented classes (minority classes), these datasets pose significant challenges for machine learning and deep learning models. However, current techniques often perform poorly on minority classes, and the presence of ambiguous and noisy data severely affects the performance of the models across both general and cybersecurity domains.[16–18] These factors significantly impede the accuracy and reliability of classification models. This research aims to bridge this gap by addressing these critical data factors and enhancing classification performance across all categories.

However, the most important data aspect in the multiple URL classification is the class imbalance. Thabtah, Hammoud[19] designed the Synthetic Minority Over-sampling Technique (SMOTE) to address a class imbalance in datasets. Wang and Awang[9] analyzed 35 datasets with varying imbalance levels (1.33%–10%), evaluating seven sampling techniques (e.g. RUS, ROS, SMOTE) and 11 classifiers such as SVM, Random Forest, Logistic Regression, Random Underdamping (RUS) and Random Oversampling (ROS) were found to perform best.[20] Traditionally, oversampling and under-sampling techniques have significant drawbacks, including overfitting to minority classes or losing information about majority classes.[21] In the multiple-class problem, one-versus-one (OVO) decomposition and spectral clustering were explored.[22]

Recent studies have also explored data augmentation techniques to address class imbalance. For example, Easy Data Augmentation (EDA)[23] and advanced synthetic data generation methods[24] were used to diversify training datasets. These approaches, while effective in image and text processing, have yet to be fully explored in URL classification. Similarly, word embeddings such as Word2Vec, GloVe, and FastText have been widely adopted in natural language processing (NLP) for their ability to capture semantic relationships.[25,26] The FastText neural word embedding, in particular, stands out for its use of character n-grams, enabling it to handle out-of-vocabulary (OOV) words effectively.[27,28] This makes it particularly effective for URL classification, which often contains short text, non-meaningful words, and special characters.[13,29] Arianto, Yuyun[30] found that both FastText and Word2Vec word embeddings demonstrate strong performance in retrieving relevant information; however, Word2Vec typically demonstrates slightly higher accuracy in Standard English language contexts compared to noisy corpus text. However, CNN with FastText embeddings[31] proves particularly effective for detecting short-, or noisy text.

Table 2, show research studies related to URL classification using machine learning, employing various methodologies, and outlines the results and limitations of these studies.

Table 2. Literature review URL classification using machine learning, and using URL classification (DMOZ and Phishing) datasets.

References	Model	Methodology	Results and Limitations
Alsenani, Ayon [32]	SVM, ANN, KNN	PSO is used to select the most relevant features	ANN achieved the highest accuracy of 97.81%, with binary class DMOZ
Sankaranarayanan, Sivachandran [33]	RF	22 Lexical Features extracted from URLs	Ensemble Classifier accuracy of 96.8%, DMOZ
Shahba, Heidary-Sharifabad and Mollahoseini Ardakani [34]	MLP, RF XGBoost	Rabbit Optimization Algorithm (ROA)	Precision: 97.52%–98.91%) Phishing Dataset
Pathak and Shrivas [35]	RF-PSO	Feature Selection Using PSO	F1-score: 98.69% Phishing Dataset
Giri and Banerjee [36]	Ensemble Model	Feature Extracted using the RF Regressor	F1-score: 97.07% Phishing Dataset
Sudar, Rohan and Vignesh [37]	Ensemble learning models	Feature Extraction and Selection Lasso Regressor	F1-score: 98.66% Phishing Dataset

Table 3. Literature review on deep learning networks and URL classification (DMOZ and Phishing) datasets.

References	Model	Methodology	Results Limitations
Rajalakshmi, Tiwari [13]	CNN with BRGU	n-grams Features	F1 Score: 82.04% DMOZ Dataset
Kexin, Liang [14]	Deep Neural Network	Char Embedding & URLNet	F1 Score: 81.23% 6 Soft Class DMOZ Dataset
Ghalechyan, Israyelyan [18]	BERT	Transfer learning techniques	Accuracy 97% Phishing Dataset
Kurt and Demirel [38]	CNN, LSTM, GRU	120K, 30K	15 Classes 0.7800 DMOZ Dataset 0.7600 0.7700
Buber and Diri [39]	5-Layer RNN with LSTM cells CNN LSTM GRU sequence modelling.	Meta-tags from web pages	85% Acc without transfer learning, 86% Acc with Transfer learning. Similar to Class DMOZ
Gupta and Bhatia [40]	BERT	Fuse knowledge Graph BERT	91% accuracy on the Reuters-RCV1 Limited 09 Classes ODP Variant
Nanda and Goel [41]	BiLSTM-GHA-CNN	Char-word Embedding	F1-Score 98%, Phishing Dataset
Kumar, Muthusamy and Jerald [42]	CNN & RNN	IWQPSO-FMRCNN.	F1-Score: 96.7% Phishing Dataset
Birthriya, Ahlawat and Jain [43]	CNN & GRU	Hybrid Feature Engineering	F1-Score: 99% Phishing Dataset

Table 2 compares various models and methodologies for URL classification across different studies. Alsenani, Ayon [32] applied SVM, ANN, and KNN with PSO, with ANN achieving 97.81% accuracy using the DMOZ dataset. Sankaranarayanan, Sivachandran [33] used RF with 22 lexical features, attaining 96.8% accuracy. Shahba, Heidary-Sharifabad, and Mollahoseini Ardakani [34] used MLP, RF, and XGBoost with ROA, achieving high precision (97.52% and 98.91%) on phishing data. Pathak and Shrivas [35] employed RF-PSO, achieving a 98.69% F1 score. A study [36] used an ensemble model with an RF regressor, achieving 97.07%, while another used Lasso regression, achieving a 98.66% F1 score on phishing data. [37] Despite these advancements, literature underscores persistent limitations, particularly in handling sparse minority classes, ambiguous data, and achieving consistent performance across diverse URL categories.

Table 3 presents key studies related to URL classification, which represent several deep learning classifiers and feature extraction methods, where Rajalakshmi, Tiwari [13] employed CNN, LSTM, achieving F1 scores of 82.04%, on the DMOZ dataset. Kurt and Demirel [38] using CNN, LSTM and GRU, obtained F1 score 0.78, 0.76 and 0.77. Advanced models like BERT, as explored by Gupta and Bhatia, [40] achieved an accuracy of 91% on the Reuters-RCV1 with limited classes, and Ghalechyan, Israyelyan [18] used phishing datasets and achieved 97% accuracy. Showcasing the efficacy of transfer learning and knowledge graph integration, hybrid models such as BiLSTM-GHA-CNN by Nanda and Goel [41] and CNN-GRU by, Birthriya, Ahlawat and Jain, [43] further pushed the boundaries, achieving exceptional F1 scores of 98% and 99.91% on phishing datasets. The word embedding has a significant role in providing contextual and semantic tokens to improve the performance of URL classification tasks. [44] However, challenges such as limited class diversity, noise and misclassified data constraints, and the need for robust methodology remain critical areas for future research.

According to the literature review, several URL classification datasets, including WebKB, DMOZ, Reuters-RCV1, Twitter, Newsgroup 20, Malicious,

Phishing, and Malware, have common issues such as class imbalance, noise, and ambiguous data. Recent improvements in machine learning, deep learning, and natural language processing have improved threat identification and automated response in cybersecurity.[45] Multi-modal neural networks improve defenses by combining diverse data sets for intrusion detection and malware analysis.[46] A hybrid ANN-BERT model achieved cutting-edge performance (98.0% accuracy) in detecting malicious URLs in social networks, offering important real-time protection against phishing, spam, and malware.[47] While numerous studies have worked on increasing the performance of cyber threads using various methods, tackling these data challenges remains critical in URL classification datasets.

The proposed FFUEAT technique utilizes data augmentation and semantic neighbors, leveraging Fine-Tuned FastText embedding's. This approach draws inspiration from synonym-based augmentation methods, as mentioned in Table 4, and incorporates elements of both augmentation and data noising strategies to enhance model robustness and performance.

In this study, a novel FFUEAT technique is empirically validated, and experimental results demonstrate its ability to improve dataset quality by reducing noisy and ambiguous data through augmenting the contextual and semantic features in the form of URL tokens. as a result, the performance of the sparse minority classes 'kids', 'adult', 'home', 'news', and cybersecurity class 'phishing' has been improved. Furthermore, the FFUEAT technique integrated with transformer-based models outperforms existing baseline studies, achieving superior classification accuracy and robustness.

Dataset for URL classification

This research classifies the URL across multiple categories using the publicly available standard DMOZ and phishing datasets. The DMOZ dataset is the largest and most imbalanced dataset available for the task of URL classification. It contains 1.5 million URLs distributed across fifteen (15) different categories—the details are given in Table 5. This table shows that the dataset is highly imbalanced and contains several uneven samples in training sets across multiple categories. Moreover, it shows that the categories 'arts,' 'society,' and 'business' contain vast training URLs. In contrast, categories such as 'kids,' 'adults,' 'home,' 'news' and phishing in the cybersecurity dataset contain fewer training samples, referred to as sparse minority classes, and are indicated by an asterisk symbol (*) in Tables 5 and 6. The phishing is noisier therefore, the data augmentation was applied to both 'legitimate' and 'phishing' categories referring to Tables 7 and 8.

To address this challenge, this study proposes a novel FFUEAT augmented technique that involves up-sampling sparse minority classes ('home', 'adult', 'news', 'kids', 'legitimate', and 'phishing'), reducing noise and ambiguous data; thereby enriching the standard DMOZ and phishing datasets. The data augmentation technique has been discussed in detail in the next section.

The proposed technique aims to increase the number of minority class samples while maintaining contextual and semantic relevance and enhancing the overall dataset features. Tables 7 and 8 show that the FFUEAT augmented technique achieved improvement in instances of sparse minority classes: 3.65% in 'home', 4.55% in 'adult', 1.40% in 'news', and 5.79% in 'kids', with the 'phishing' category also experiencing an increase.

The augmented DMOZ (ADMOZ) dataset represents a significant advancement over the standard DMOZ dataset, addressing the critical issue of sparse minority classes in URL classification. The augmented dataset comprises approximately 1.6 million records, representing a 9.2% increase from the standard DMOZ dataset. Also, the augmented Phishing

Table 4. Literature about data augmentation technique.

References	Data Augmentation Approach
Synonym Replacement, Insertion, Deletion and Swap[23,48]	EDA's four simple text augmentation operations— random insertion, swap, synonym replacement, and deletion—consistently improved text classification performance.
Similarity Insertion and Replacement[48]	The incorporation of tokens via augmentation techniques such as synonym replacement and Word2vec substitutions markedly improves text classification efficacy by broadening dataset diversity and enhancing model generalisation, particularly in low-resource or noisy data contexts.
Punctuation Insertion[49]	An Easier Data Augmentation) technique to help improve the performance on text classification tasks. AEDA inserts only the punctuation randomly into the original text.
Placeholder Replacement[50]	Data noising is a strong regularisation strategy for neural network language models that introduces controlled noise to input sequences, successfully replicating the smoothing behaviour of traditional n-gram models while capitalising on neural architectures.
Inversion and Passivization[51]	Dataset-reviewed

Table 5. DMOZ dataset class-wise representation.

Serial No	Category Name	Number URLs	Overall Representation
1	ARTS	252401	16.23
2	SOCIETY	242488	15.59
3	BUSINESS	239555	15.40
4	COMPUTERS	117014	7.52
5	SCIENCE	109758	7.06
6	RECREATION	106099	6.82
7	SPORTS	100971	6.49
8	SHOPPING	94980	6.11
9	HEALTH	59887	3.85
10	REFERENCE	57636	3.71
11	GAMES	56287	3.62
12	*KIDS	46002	2.96
13	*ADULT	35100	2.26
14	*HOME	28165	1.81
15	*NEWS	8947	0.58

Table 6. Phishing dataset class-wise representation.

Serial No	Category Name	Number URLs	Overall Representation
1	Legitimate	275932	77.33
2	*Phishing	80862	22.66
Total number of samples: 356794			

Table 7. ADMOZ dataset class-wise representation.

Serial No	Category Name	Number of URL Samples	Overall Representation
1	ARTS	252401	14.86
2	SOCIETY	242488	14.28
3	BUSINESS	239555	14.11
4	COMPUTERS	117014	6.89
5	SCIENCE	109758	6.46
6	RECREATION	106099	6.25
7	SPORTS	100971	5.95
8	SHOPPING	94980	5.59
9	HEALTH	59887	3.53
10	REFERENCE	57636	3.39
11	GAMES	56287	3.31
12	*KIDS	123051	5.79
13	*ADULT	93004	4.55
14	*HOME	72247	3.65
15	*NEWS	23760	1.40
Total Number of Samples: 1698369			

Table 8. A Phishing dataset class-wise representation.

Serial No	Category Name	Number URLs	Overall Representation
1	Legitimate	344881	69.34
2	*Phishing	152472	30.65
Total number of samples: 497353			

(APhishing) dataset is increased by 28.26% compared to the standard dataset. The augmented new data plays a crucial role in mitigating the impact of class imbalance, noise, and ambiguous data.

Training and testing set

The augmented dataset provides a notable improvement over standard datasets by solving the issue of missing or underrepresented minority classes and decreasing noise and ambiguity in URL classification. This novel technique enhances the coverage of the dataset while preserving the integrity of the standard datasets by increasing the total sample size by 9.2% in the DMOZ dataset and 28.26% in the phishing dataset. The training and test sets are derived from these augmented datasets. We randomly selected 6,000 equal testing samples from each category, regardless of whether they were majority or minority classes, in the DMOZ dataset, creating a robust testing set of 90K URLs; in contrast, the baseline study by Kexin, Liang,[14] used limited samples of the DMOZ dataset. For the phishing dataset, we utilized an 80/20 split for training and testing sets. This larger testing sample size in the current study provides more reliable evaluation metrics.

Proposed technique

In this study, a novel Fine-Tuned FastText Unsupervised Embedding Augmentation Technique (FFUEAT) is applied at two stages. In the first stage, it addresses class imbalance by oversampling the minority classes within the datasets, and in the second stage, it augments contextual and semantic features to all categories.

Pre-processing is an important part of the training and testing sets used for machine learning and model validation. Therefore, class label (15 classes 'business', 'sports', etc.) instances are in the form of characters, which may introduce bias during the training process. To ensure the validity, neutrality, and unbiased nature of the classification experiments; all class labels (15 classes 'business', 'sports', etc.) are assigned a label ID (1, 2, 3, etc.) as referred in Table 9. This approach prevents unintended semantic bias by eliminating linguistic meaning from labels, ensuring the model cannot exploit them.

FFUEAT technique

This section discusses the proposed novel Fine-Tuned FastText Unsupervised Embedding Augmentation Technique (FFUEAT), as demonstrated in Figs. 1 to 3. It introduces four robust algorithms, outlining

Table 9. Label encoding scheme for DMOZ classes.

#	Category Name	Experimental Label
1	Arts	0
2	Business	1
3	Society	2
4	Sports	3
5	Science	4
6	Computers	5
7	Shopping	6
8	Recreation	7
9	Reference	8
10	Health	9
11	Games	10
12	Home	11
13	Adult	12
14	News	13
15	Kids	14

Algorithm 1. Data preprocessing & extracting unsupervised tokens from both datasets.

Input : Standard DMOZ and Phishing Datasets
Output : Unsupervised Tokens of DMOZ and Phishing Datasets
Begin
read data from the Dataset
U ← total number of URLs
While u ≠ 0 do
x=Lower Case URL
x=Rmoval of Protocol, directory paths, dots, slashes, colons, and hyphens
x=Split URL Into Tokens
x=Extracting URL Tokens
end
End

Table 10. Illustration of unsupervised tokens.

Sample URLs	Unsupervised Token
http://silversabres.homestead.com	Silversabres homestead
http://ccn-irc.com	ccn irc
http://wtboxers.com	wt boxers
http://footprincess.co.uk	foot princess
http://fetishview.com	etish view
http://realdollpics.com	real doll pics
http://facialcum-shots.com	facial cum shots
http://xxxmovies.com	xxx movies
http://flat-sharer.com	Flat sharer
http://roommatesville.com	room mates
http://prideroommates.com	pride room
http://news.bbc.co.uk	news bbc
http://abcnews.go.com	abc news
http://starflats.co.uk	star flats
http://citizen.org	citizen
Cumulative Total - 7.98 Million Tokens	

the process of developing the FFUEAT augmented technique.

In Algorithm 1, all URLs were converted to lowercase. Subsequently, irrelevant data from the URLs was removed, including protocols (such as http, https, and www), directory paths, dots, slashes, colons, and hyphens. A 'token' here refers to a distinct substring or meaningful component derived from a URL after this pre-processing. The process shown in Algorithm 1 demonstrates the transformation of raw URLs into such meaningful tokens. Furthermore, it extracts approximately 6 million unsupervised tokens, including URL tokens, from the standard DMOZ dataset and 1.98 million unsupervised tokens from the standard phishing dataset.

Table 10 presents the URLs of the sample and their unsupervised tokens, where some contain a single token, while others consist of multiple tokens, contributing to a cumulative total of 7.98 million unsupervised tokens across both datasets.

The output of Algorithm 1 consists of 7.98 million unsupervised tokens derived from both datasets. These tokens were then fed as input to FastText word embedding for fine-tuning. This process enabled a closer semantic and coherent alignment with the unique characteristics of both datasets (as shown in Algorithm 2). The two most critical parameters during FastText fine-tuning are the dimension and the skip-gram. The dimension is set to 300, which ensures rich and informative word embeddings. Skip-gram architecture is chosen over continuous bag of words (CBOW) due to its superior ability to manage rare

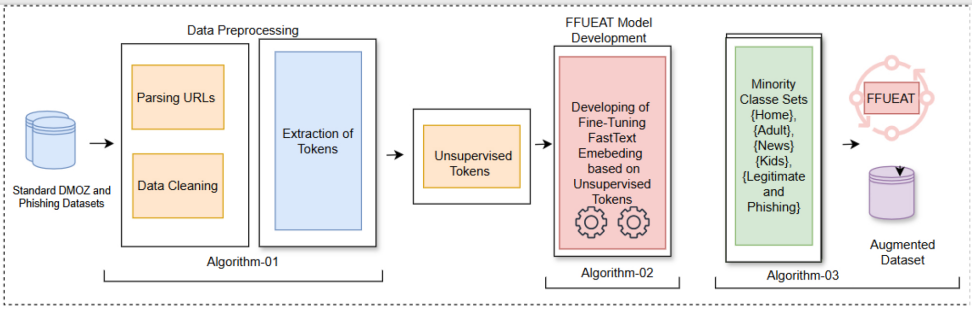


Fig. 1. Development and implementation of the FFUEAT augmentation technique.

Algorithm 2. Fine-Tuning Standard FastText Model using Tokens of DMOZ and Phishing Datasets.

Input : Unsupervised Tokens of DMOZ and Phishing Datasets

Load: Standard FastText Model

Output : Fine-Tuned FastText Model

Begin

Dimension = 300;

Model = Skipgram;

Epochs Iterates on Dataset = 10;

WordNgrams = 2;

Learning rate = 0.05;

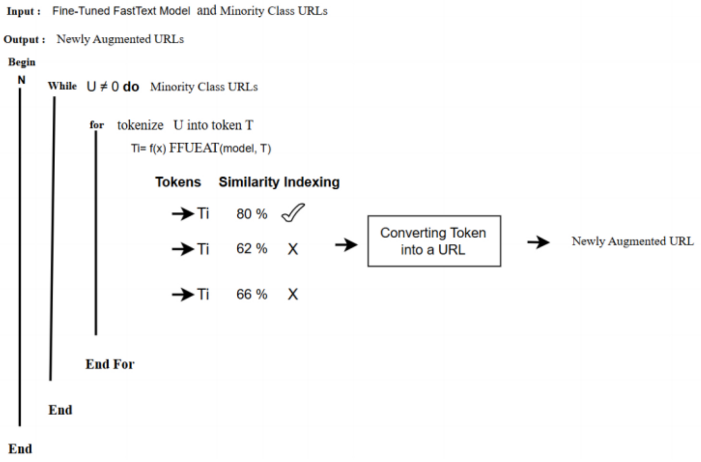
Loss=softmax;

Thread=4;

Bucket=2000000;

End

Algorithm 3. Up-sampling of the six categories ('home', 'adult', 'news', 'kids', and 'legitimate',and 'phishing').



or out-of-vocabulary (OOV) terms through sub-word representations.[28]

Algorithm 2, based on a pre-trained FastText word-embedding model, has the advantage of handling out-of-vocabulary terms and substrings to process complex data, further leveraging its pre-training on extensive corpora such as Wikipedia. To tailor FastText for the specific task of URL classification, the model was fine-tuned using 7.9 million unsupervised tokens extracted from the URL dataset.

Algorithm 3, receives two inputs: i) the FFUEAT model, ii) the minority classes — including cybersecurity 'home', 'adult', 'news', 'kids', 'legitimate', and 'phishing' — to increase the sample of underrepresented classed through an up-sampling strategy. Specifically, URLs from underrepresented classes were segmented into T tokens, each of which was fed into the FFUEAT model. This approach generates contextually similar tokens denoted by Ti token and original token represented by T into the Algorithm 3, effectively generating a new augmented URL for the minority classes. The process, illustrated in Fig. 2, ensures a more balanced representation of all categories, thereby improving the model's performance in classifying URLs across the spectrum of defined classes.

Importantly, not all of Algorithm 3's tokens are suitable for URL augmentation. A token must satisfy two requirements in order to be eligible: it must be contextually similar but not identical or the same,

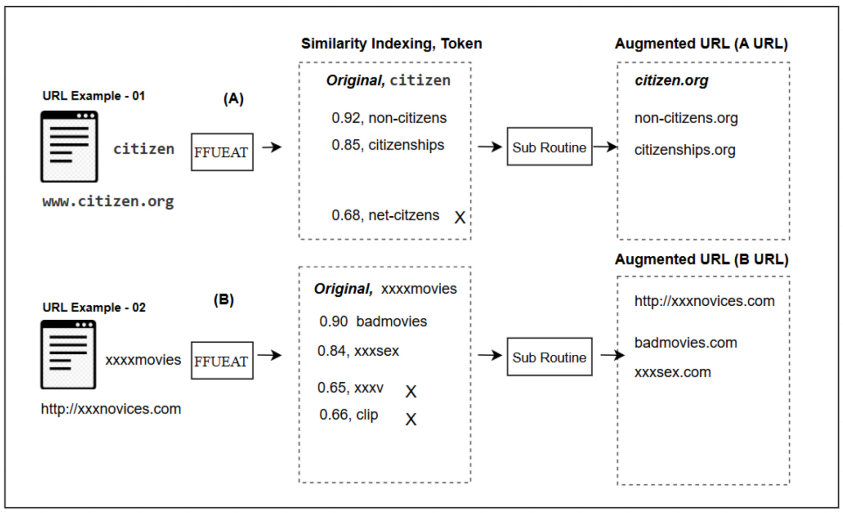


Fig. 2. URL Up-Sampling bases on tokenization and similarity indexing value.

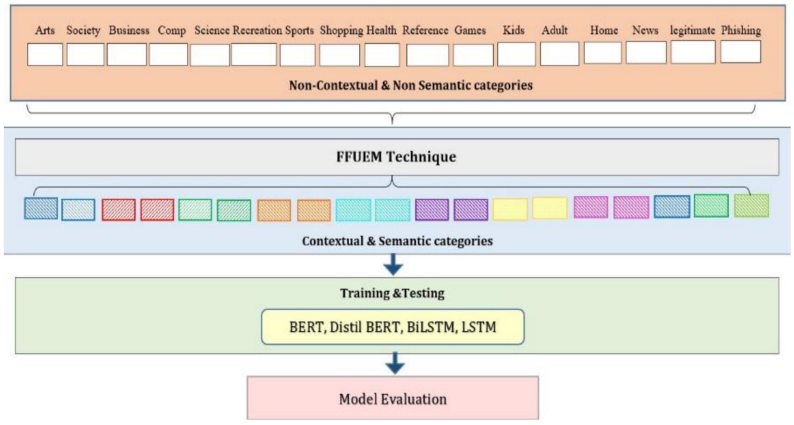


Fig. 3. Enhancing the quality of DMOZ and Phishing datasets by using FFUEAT Technique

even if the similarity score is high, and it must have at least a 70% similarity index to maintain its meaning close to an original token. These standards guarantee that enhanced tokens remain pertinent to the subject and do not yield redundant or pointless outcomes, preserving the high standard of the dataset.

Fig. 2 illustrates the generation of augmented URLs using the FFUEAT technique. In this example, two original URLs are provided as input, and four augmented URLs are produced as output. This demonstrates an important characteristic of the FFUEAT methodology: the number of augmented URLs generated from each source URL is neither fixed nor uniform. Instead, it varies based on both the structural composition and the semantic properties of the input URL. Depending on these factors, a single source URL may yield zero, one, or multiple augmented variants. The eligibility for augmentation is governed by two criteria described in the previous section, ensuring that only URLs meeting the required structural and semantic thresholds are selected for augmentation.

As shown in Fig. 1, the DMOZ dataset experiences a cumulative expansion of 9.2% and Phishing dataset of 28.26% in sample size, resulting in the augmented DMOZ and phishing datasets. This substantial augmentation increases the number of URLs for sparse minority classes (‘home’, ‘adult’, ‘news’, ‘kids’, ‘legitimate’, and ‘phishing’). Additionally, augmented URLs are semantically and contextually aligned with the standard datasets (DMOZ and Phishing).

In this study, we employ a novel Fine-Tuned Fast-Text Unsupervised Embedding Augmentation Technique (FFUEAT) at two stages. The first stage oversamples the minority classes. The second stage enhances data quality for URLs. Fig. 3 shows the workflow for improving data quality in URL categories. It incorporates contextual and semantic features across both datasets. Furthermore, this approach is inspired by the Random Insertion Synonym

Algorithm 4. Enhancing the data quality of the 17 Categories, including the DMOZ and Phishing datasets.

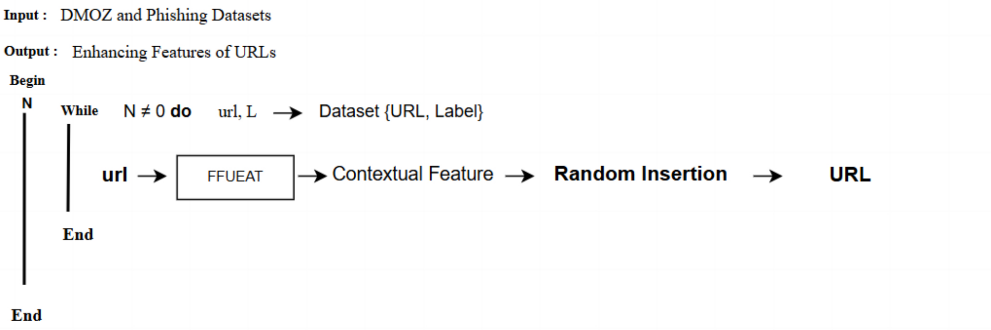


Table 11. Sample of the contextual and semantic features.

URL	FEATURES	SIMILARITY	Y	RANDOM INSERTION
ccn-irc.com	circ	0.69	0	ccn-circ.com
wtboxers.com	kickboxer	0.70	1	wtboxers-kickboxer.com
footprincess.co.uk	princes	0.80	1	footprincess-princes.co.uk
flat-sharer.com	flat-out	0.80	1	flat-out-flat-sharer.com
roommatesville.com	farmville	0.65	0	roommatesville..com
news.bbc.co.uk	newscast	0.79	1	news.bbc-newscast.co.uk
abcnews.go.com	cbbcnews	0.86	1	abcnews-cbbcnews.go.com
starflats.co.uk	starfox	0.85	1	starfox-starflats.co.uk

method proposed in prior studies.[49–51] To evaluate FFUEAT's performance, we use classifiers such as LSTM, BiLSTM, DistilBERT, and BERT. Algorithm 4 outlines the process for adding contextual and semantic features randomly to each URL using FFUEAT. Table 11 presents sample features from the standard DMOZ and phishing datasets. If the feature similarity index is above 69.99, the feature is added as a new contextual and semantic feature.

Despite its seemingly simple design, Algorithm 4 required approximately 100 hours to process 17 categories ('arts', 'business', 'society', 'sports', 'science', 'computers', 'shopping', 'recreation', 'reference', 'health', 'games', 'home', 'adult', 'news', 'kids', 'legitimate', 'phishing') of both datasets reflecting its rigorous computational demands. The feature extracted from the standard datasets using the FFUEAT augmented technique was a computationally intensive process.

The novel FFUEAT augmented technique provides the contextual and semantic URL classification datasets. To validate the performance of the FFUEAT augmented technique, we have conducted 12 experimental setups using the four deep neural networks: LSTM, BiLSTM, BERT, and DistilBERT.

BERT coupled with the FFUEAT technique

A novel FFUEAT technique has been validated and evaluated through 12 experimental setups, including BERT, which was executed using two distinct datasets: the publicly available standard DMOZ and the Phishing datasets. The BERT classifier outperforms other deep neural networks. As illustrated in Fig. 4, BERT processes URLs, which are subsequently tokenized into either enhanced or simple tokens, depending on the dataset. This process results in creating X-train and X-test sets, which serve as standard inputs for the model. The URL tokens are represented in Fig. 4 and are mathematically detailed in Table 12.

$$U = \{U, L\} \tag{1}$$

U represents the set of URLs, where each URL is denoted as U= $\{U_1, U_2, U_3 \ldots U_N\}$. This set contains a sample of 'n' URLs from the dataset. L represents the labels corresponding to each URL, defined as L= $\{L_1, L_2, .. L_n\}$. Each label L_i is associated with its respective URL U_i. In this formulation, each URL U_i is mapped to its corresponding label L_i, ensuring a clear relationship between the data points and their classifications.

Each URL comprised a Token denoted T_i as follow. In short, each URL can be expressed as below

$$U_i = \{U_{1:P} = T_1 \oplus T_2 \oplus T_3 \oplus \ldots T_p\} \tag{2}$$

The text data from the dataset is represented as X_{train}, Which corresponds to the training and testing

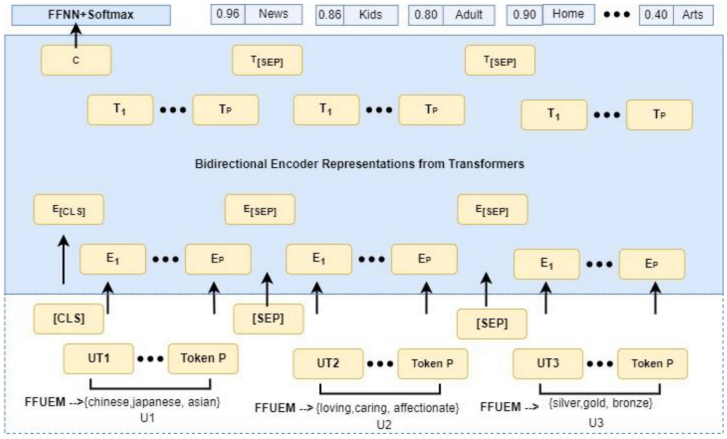


Fig. 4. BERT architecture for the URL classification.

sets, respectively.

$$X_{train} = \{U_1, U_2, \ldots, U_m\}, \quad X_{test} = \{U_1, U_2, \ldots, U_n\} \quad (3)$$

Where m is the number of training samples, and n−m is the number of testing samples.

Table 12. Mathematical model for BERT.

Steps	Equations
BERT tokenizer maps	Let the input text be (Xtrain, Xtest)
Tokenization	$x, m =$ Tokenization(x) (4)
Bert Model Output	$z = f\theta(x, m)$(5)
Loss Function	$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \log(softmax(z_i)_{yi})$ (6)
AdmW Update Rule	$\theta_{t+1} \leftarrow \theta_t - \eta(\nabla_\theta \mathcal{L} + \lambda\theta_t)$ (7)
Accuracy	Accuracy $= \frac{1}{N}\sum_{i=1}^{N} <(\arg\max z_i = y_i)$ (8)

The BERT sequence classification workflow includes tokenization (token IDs, attention masks), model architecture (Transformer encoder, classification layer), training dynamics (fine-tuning with Cross-Entropy Loss, AdamW), and evaluation metrics (accuracy, precision, recall, F1-score, ROC-AUC).

Experiments and results

All experimental procedures were executed within the Google Colab computational environment, with the specific configurations detailed in Table 13. To generate the augmented URLs for both the ADMOZ and Phishing datasets, a subset of these experiments was conducted on a local workstation equipped with a dedicated Graphics Processing Unit (GPU), Table 14.

In reference to Table 13, Ampere Architecture's compute capability of 8.0 is ideal for complex machine learning tasks, such as URL classification, especially with large datasets like the standard and augmented DMOZ, which contain 1.5 and 1.69 million records, respectively. With 6912 CUDA cores and 432 Tensor cores, the GPU accelerates parallel processing for URL feature evaluation and deep learning model operations.

Table 13. Google colab computational environment (A100 GPU).

Specification	Details
Architecture	NVIDIA Ampere
Compute Capability	8.0
CUDA Cores	6912
Tensor Cores	432 (3rd generation)
Memory	40GB
Memory Bandwidth	Up to 2039 GB/s
FP64 Performance	Up to 19.5 TFLOPS
FP32 Performance	Up to 19.5 TFLOPS
FP16 Performance	Up to 312 TFLOPS (with FP32 accumulate)
INT8 Performance	Up to 624 TOPS
Multi-Instance GPU	Supports up to 7 GPU instances
NVLink	3rd generation, up to 600 GB/s bidirectional

Table 14. Configuration of local machine.

Specification	Details
Processor	Intel Core i5-10300H, i5-11400H, i5-12500H
GPU	NVIDIA GeForce RTX 3050
RAM	64 GB RAM
Storage	512GB NVMe SSD

In Table 15, URL classification data pre-processing is an important step, for which libraries such as TensorFlow/Keras Tokenizer and NLTK are required for the process. The Tokenizer of TensorFlow/Keras transforms URLs into sequences and provides the possibility for padding sequences to a fixed length, allowing all inputs to be equal in size. This is important as we can guarantee input size (essential for batch processing). NLTK (Natural Language Toolkit) is useful to deal with noisy and non-informative ambiguous URLs as it is a popular package for tokenization, stemming and other text pre-processing

Table 15. Fundamental libraries used in Python.

Category	Details
Deep Learning	tensorflow, keras, transformers
Text Pre-processing	tensorflow.keras.preprocessing.text.Tokenizer, nltk.tokenize.word_tokenize
Machine Learning	sklearn.linear_model, sklearn.model_selection.train_test_split
Data Manipulation	pandas, numpy
Visualization	matplotlib.pyplot
Feature Engineering	gensim.models.FastText
Evaluation Metrics	sklearn.metrics.accuracy_score, sklearn.metrics.classification_report

Table 16. Hyper parameters used in the RNN models.

Hyper Parameter	BiLSTM Model	LSTM Model
Learning Rate	0.001 (Adam default)	0.001 (Adam default)
Batch Size	128	128
Number of Epochs	20	20
Optimizer	Adam	Adam
Loss Function	categorical_crossentropy	categorical_crossentropy
LSTM Units	maxlen	maxlen
Final Dense Units	15	15
Final Activation	Softmax	Softmax

tasks. An essential library for standard machine learning problems in URL classification is Scikit-learn, which provides different methods Decision Trees, Support Vector Machines (SVM), Logistic and so on.

As indicated in Table 15, text pre-processing is a crucial step in URL classification, and libraries like TensorFlow/ Keras Tokenizer and NLTK are necessary for this procedure. In order to ensure consistent input size, which is crucial for batch processing, the Tokenizer from TensorFlow/Keras converts URLs into sequences and additionally allows padding sequences. NLTK (Natural Language Toolkit) is particularly helpful for handling noisy and ambiguous URLs since it provides tools for tokenization, stemming, and other text pre-processing tasks. A fundamental library for conventional machine learning tasks in URL classification is Scikit-learn. It offers a variety of methods, including Decision Trees, Support Vector Machines (SVM), and Logistic Regression.

As mentioned in Table 16, both BiLSTM and LSTM architectures utilized identical hyperparameter configurations, including a learning rate of 0.001 via the Adam optimizer, a batch size of 128, and training over 20 epochs. The loss function was set to categorical crossentropy, with maxlen LSTM units, a final dense layer of 15 units, and softmax activation, ensuring consistent architectural and optimization conditions.

In reference to Table 17, the transformer-based models BERT and DistilBERT were configured with a lower learning rate of 2e-5, a reduced batch size of 64, and 16 training epochs, optimized with Adam and employing Categorical Crossentropy (with from_logits=True) as the loss function. Both transformer models were designed for 15 output classes

Table 17. Hyper parameters used in the transformers models.

Hyper Parameter	Bert Model	DistilBERT Model
Learning Rate	2e-5	2e-5
Batch Size	64	64
Number of Epochs	16	16
Optimizer	Adam	Adam
Loss Function	Categorical Crossentropy (from_logits = True)	Categorical Crossentropy (from_logits = True)
Metrics	accuracy	accuracy
Number of Classes	15	15

and employed accuracy as the evaluation metric, which reflects their different optimization needs and architectural efficiency in comparison to the recurrent models.

This section presents a detailed analysis of the experimental framework and results, highlighting the effectiveness of the FFUEAT augmented technique in enhancing the representation of experiments and their findings. This study compares the performance of augmented data characterized by contextual and semantic richness against the standard dataset, which lacks these qualities. A series of carefully designed experiments were conducted using four different datasets: the DMOZ and Phishing datasets, along with their augmented versions, ADMOZ and APhishing. These datasets were evaluated using advanced deep learning models, including LSTM, Bi-LSTM, BERT, and DistilBERT, to measure their ability to handle class imbalance, noise, and ambiguous data.

As shown in Tables 18 and 19, a comparison of key categories ‘arts’, ‘business’, ‘society’, ‘adult’, ‘news’, ‘kids’, ‘legitimate’, and ‘phishing’ demonstrates the effectiveness of the FFUEAT augmented technique in

Table 18. FFUEAT performance evaluation on standard and augmented DMOZ datasets using RNN.

URL Category	LSTM DMOZ	LSTM-ADMOZ	BiLSTM DMOZ	BiLSTM ADMOZ
Arts	0.8163	0.8294	0.8293	0.8725
Business	0.8436	0.8354	0.8023	0.8368
Society	0.8226	0.8292	0.8310	0.8413
Sports	0.9782	0.9762	0.9811	0.9752
Science	0.6930	0.7065	0.6893	0.6248
Computers	0.8573	0.8794	0.8755	0.8770
Shopping	0.9549	0.9645	0.9496	0.9741
Recreation	0.9256	0.9376	0.9238	0.9448
Reference	0.8123	0.8469	0.8069	0.8618
Health	0.9556	0.9552	0.9582	0.9628
Games	0.7177	0.8029	0.7181	0.8608
Home	0.9301	0.9445	0.9103	0.9570
Adult	0.5586	0.6603	0.5507	0.6661
*News	0.6526	0.6065	0.6588	0.6932
*Kids	0.4956	0.5998	0.5048	0.7219
F1-Score	0.8159	0.8273	0.8175	0.8511

enhancing the performance of classification models using the augmented datasets. For LSTM, significant improvements have been observed in 'adult' (an increase of 10.17), 'kids' (10.42), and 'arts' (01.31), while BiLSTM achieved even higher gains in 'kids' (21.71), 'adult' (11.54), 'games' (14.27), 'arts' (4.32), and 'business' (3.45). These results demonstrate strong performance in cybersecurity categories, such as 'legitimate' (4.32) and 'phishing' (3.45%). Overall, these findings highlight the FFUEAT augmented technique's ability to address class imbalance and improve the representation of sparse minority classes, such as 'adult' and 'kids', as well as cybersecurity classes, including 'legitimate' and 'phishing', which have also been enhanced.

The capabilities of APhishing to enhance model efficiency and deal with imbalanced classes and correctly classify Legitimate and Phishing URLs are shown in Table 19 below. The FFUEAT Augmented Technique provides additional contextual and semantic features that help in accurately classifying URLs. Tables 20 and 21 clearly illustrate how successfully DistilBERT and BERT algorithms are identify phishing and advanced phishing URLs, with BERT performing slightly better in categorizing valid URLs (0.9621).

The FFUEAT technique with the augmented DMOZ dataset consistently outperforms classification tasks when evaluated using both DistilBERT and BERT, as

Table 19. FFUEAT performance evalauation on standard and augmented phishing datasets using RNN.

URL Category	LSTM Phishing	LSTM APhishing	BiLSTM Phishing	BiLSTM APhishing
Legitimate	0.8692	0.9420	0.9443	0.9800
Phishing	0.8730	0.9532	0.9542	0.9900

Table 20. FFUEAT performance evaluation on standard and augmented DMOZ datasets using transformer models.

URL Category	DistilBERT	DistilBert ADMOZ	BERT DMOZ	BERT ADMOZ
Arts	0.8092	0.8667	0.8143	0.8610
Business	0.8613	0.8917	0.8742	0.8897
Society	0.8417	0.9002	0.8580	0.9047
Sports	0.9617	0.9826	0.9714	0.9756
Science	0.6722	0.6962	0.6791	0.6890
Computers	0.8910	0.9280	0.9119	0.9084
Shopping	0.9432	0.9595	0.9489	0.9585
Recreation	0.9246	0.9416	0.9308	0.9513
Reference	0.8522	0.9420	0.8620	0.9318
Health	0.9606	0.9611	0.9587	0.9634
Games	0.7327	0.7442	0.7363	0.8708
Home	0.9662	0.9663	0.9606	0.9688
Adult	0.5382	0.5499	0.5673	0.5502
News	0.7063	0.9627	0.7015	0.9604
Kids	0.4896	0.5218	0.5092	0.8012
F1-Score	0.8225	0.8639	0.8305	0.8891

Table 21. FFUEAT performance evaluation on standard and augmented phishing datasets using transformer models.

URL Category	DistilBERT Phishing	DistilBERT APhishing	BERT Phishing	BERT APhishing
Legitimate	0.8902	0.932	0.9302	0.9621
Phishing	0.9004	0.9298	0.924	0.9512

presented in Table 20, especially in critical categories such as 'news', 'kids', 'reference', and 'games'. For DistilBERT, the FFUEAT augmented technique shows significant improvements in 'news' (25.64%) and 'reference' (8.98%), while BERT achieves significant gains in 'kids' (29.2%) and 'news' (25.89%) compared to standard datasets. The superior performance of the FFUEAT augmented technique is attributed to its enhancement of data quality in terms of context and semantics. In contrast, the standard DMOZ dataset struggles in sensitive categories, such as 'adult' and 'kids', with only marginal improvements in areas like 'health' and 'home'.

As presented in Table 22, it can be seen that there were considerable performance enhancements obtained by all four models, LSTM, BiLSTM, DistilBERT, and BERT, upon applying them to the enhanced version of the ADMOZ dataset, wherein positive delta values imply improvements that occurred almost everywhere except in URL category 4 (adult), while BERT emerged as the best-performing model with an excellent ΔF1-score of $+5.86$ overall. It provided the best performance in some key categories such as "news" ($\Delta + 25.89$) and "kids" ($\Delta + 29.20$), showing its exceptional capability to adjust to the improved and diverse version of the ADMOZ dataset. Closely following BERT is DistilBERT, with a very good $\Delta + 4.14$ score, which is almost

Table 22. Delta performance of the deep learning models using (DMOZ → ADMOZ).

URL Category (DMOZ → ADMOZ)	Delta Performance LSTM	Delta Performance BiLSTM	Delta Performance DistilBERT	Delta Performance BERT
Arts	1.31	4.32	5.75	4.67
Business	–0.82	3.45	3.04	1.55
Society	0.66	1.03	5.85	4.67
Sports	–0.2	–0.59	2.09	0.42
Science	1.35	–6.45	2.4	0.99
Computers	2.21	0.15	3.7	–0.35
Shopping	0.96	2.45	1.63	0.96
Recreation	1.20	2.1	1.7	2.05
Reference	3.46	5.49	8.98	6.98
Health	–0.04	0.46	0.05	0.47
Games	8.52	14.27	1.15	13.45
Home	1.44	4.67	0.01	0.82
Adult	10.17	11.54	1.17	–1.71
*News	–4.61	3.44	25.64	25.89
*Kids	10.42	21.71	3.22	29.2
F1-Score	1.14	3.36	4.14	5.86

Table 23. Delta performance of the deep learning models using (Phishing → APhishing).

URL Category (DMOZ → ADMOZ)	Delta Performance LSTM	Delta Performance BiLSTM	Delta Performance DistilBERT	Delta Performance BERT
Legitimate	7.28	3.57	4.18	3.19
Phishing	8.02	3.58	2.94	2.72

equal to BERT's performance in some important categories, such as "news" (Δ + 25.64). These two models are efficient, highly capable tools for adapting to their domains more effectively. Furthermore, both LSTM (Δ + 1.14) and BiLSTM (Δ + 3.36) showed impressive performance improvements as well, especially in some categories, e.g., kids (Δ + 21.71) and reference (Δ + 5.49).

In reference to Table 23, it can be seen that in the category of cybersecurity, DistilBERT and BERT show remarkable efficiency in URL classification since their accuracy values exceed 0.89. In particular, BERT is effective at detecting legitimate URLs (0.932) and advanced phishing schemes (0.9621), while DistilBERT performs comparably and even surpasses BERT on basic phishing schemes (0.9004). It can be concluded that the transformer-based approach is valid for phishing detection regardless of whether the threats are simple or complex. This can be seen from the results table, where the FFUEAT-enhanced approach is efficient in classifying 'legitimate' and 'phishing'.

Measuring classification performance

A confusion matrix is a structured table that assesses a classifier's performance by linking actual class labels to predicted classifications, where rows indicate true categories and columns show predicted outcomes. It forms the foundation for calculating key evaluation metrics such as accuracy, precision, recall, and F1-score. The standard formulas are presented in Table 24.

Table 24. Confusion matrix description.

Description	Abbreviation
Instances correctly predicted True Positives.	TF
Instances wrongly predicted False Positives.	FP
Instances of class wrongly predicted as other classes.	FN
Instances correctly predicted as not a class.	TN

$$Accuracy = \frac{\sum_{I=1}^{N} TP}{Total\ Samples} \tag{9}$$

Overall classification accuracy across multiple categories

$$Precision_i = \frac{TP}{TP + FP} \tag{10}$$

The number of correctly predicted true instances for class X.

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

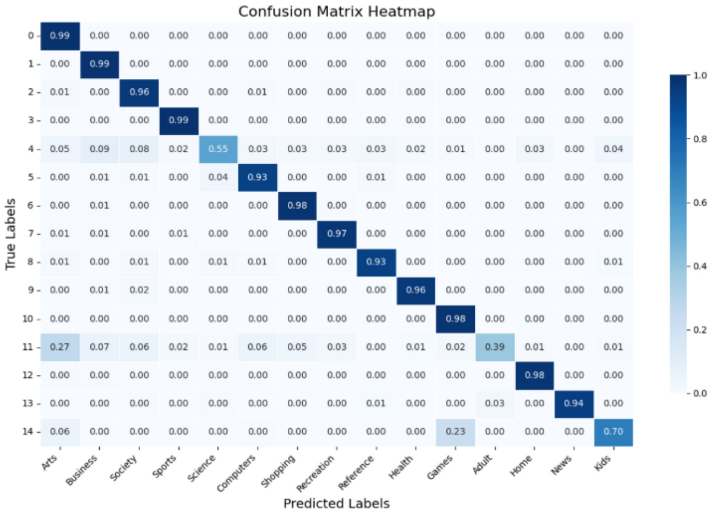


Fig. 5. Confusion matrix FFUEAT augmented technique using ADMOZ dataset with BERT setting

How many true instances of class i are correctly predicted

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (12)$$

Interesting behavioral trends are evident in the confusion matrix analysis of classifying the 'adult' and 'science' categories, offering valuable insights in Fig. 5. The accuracy rates are 39% for 'adults' and 55.3% for 'science'; they also highlight that the complex categories might have multi-dimensional domains. The misclassification of 'adult' content occurs in categories like 'arts' and 'society.' Similarly, using BERT, the 'science' category overlaps with 'arts,' 'business,' and 'society'. Nonetheless, using the FFUEAT technique, the BERT model obtained an F1 score of 55.02% for the 'adult' minority class, approximately 6.00% lower than the baseline model (shown in Table 25). However, with proposed FFUEAT technique, the RNN model outperformed by obtaining 5.61% improvement in F1-score compared to the baseline model (shown in Table 18). The possible reason behind the outperformance of the RNN model compared to the BERT model might be because the 'adult' class contains more noisy data compared to the rest of the classes. In addition, the baseline study[14] also implemented the RNN model using the DMOZ dataset.

In conclusion, the FFUEAT technique, along with classifiers, demonstrates strong overall performance, as evidenced by a weighted average F1-score of 0.8801 and an F1 score of 0.8891, reflecting its effectiveness in handling the FFUEAT technique (ADMOZ dataset). Our proposed FFUEAT technique demonstrated a significant improvement in sparse minority classes compared to baseline methodologies, as shown in Table 2. Specifically, the FFUEAT technique achieved significant improvements in the sparse minority classes, with accuracy gains of 20.88% in the 'home' category, 58.04% in 'news,' and 11.18% in 'kids' relative to the baseline techniques.

It is evident from Table 26 and Fig. 6 that the FFUEAT augmented technique (APhishing dataset)

Table 25. Classification report of FFUEAT augmented technique using ADMOZ dataset with BERT setting.

Category	Precision	Recall	F1-Score	Support
Arts	0.7605	0.9922	0.861	5735
Business	0.8107	0.9856	0.8897	4168
Society	0.8541	0.9617	0.9047	5401
Sports	0.9598	0.9919	0.9756	5658
Science	0.9132	0.5532	0.689	5192
Computers	0.8858	0.9322	0.9084	4320
Shopping	0.9341	0.9842	0.9585	5301
Recreation	0.9336	0.9698	0.9513	4969
Reference	0.9295	0.9342	0.9318	3878
Health	0.9629	0.9639	0.9634	4933
Games	0.7815	0.9833	0.8708	5317
*Home	0.9613	0.9765	0.9688	5701
*Adult	0.9609	0.3854	0.5502	3697
*News	0.978	0.9434	0.9604	1273
*Kids	0.9306	0.7034	0.8012	5735
Accuracy			0.8891	71278
Macro Average	0.9038	0.8841	0.879	71278
Weighted Average	0.8988	0.8891	0.8801	71278

Table 26. Classification report of the FFUEAT technique using Aphishing Dataset with BiLSTM.

	Precision	Recall	F1-Score	Support
Legitimate	0.99	0.98	0.98	30495
Phishing	0.99	0.99	0.99	68976
Accuracy			0.99	99471
Macro avg	0.99	0.99	0.99	99471
Weighted avg	0.99	0.99	0.99	99471

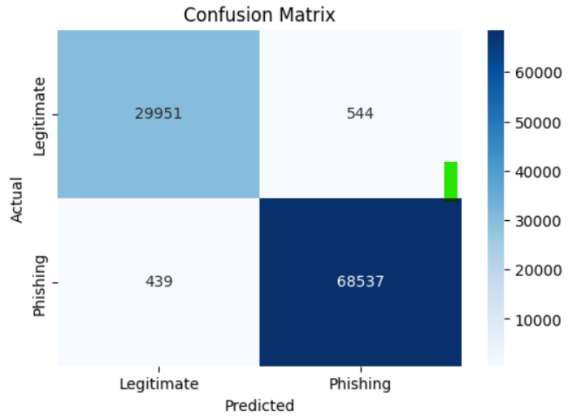


Fig. 6. Confusion Matrix FFUEAT augmented technique using Aphishing dataset with BiLSTM.

Table 27. Comparison of baseline techniques with the proposed FFUEAT augmented technique using DMOZ Dataset.

	Kexin, Liang[14]	Current Study FFUEAT Augmented Technique (ADMOZ)	
	CNN, RNN	Distil BERT	BERT
Arts	0.6400	0.8725	0.8610
Business	0.2600	0.8908	0.8897
Society	0.3700	0.8849	0.9047
Sports	0.7800	0.9791	0.9756
Science	0.6000	0.6709	0.6890
Computers	0.4900	0.9133	0.9084
Shopping	0.3000	0.9478	0.9585
Recreation	0.3600	0.948	0.9513
Reference	0.6800	0.933	0.9318
Health	0.4200	0.9548	0.9634
Games	0.3500	0.7645	0.8708
*Home	0.7600	0.9667	0.9688
*Adult	0.6100	0.5148	0.5502
*News	0.3800	0.9374	0.9604
*Kids	0.6900	0.5966	0.8012
F1 Score	0.5000	0.8632	0.8891

achieved the best performance with 98% F1 score for 'legitimate' and 99% for 'phishing'; it is well established that the FFUEAT augmented technique is effective not only for general categories but also for cybersecurity categories.

The transformer models, BERT and DistilBERT, using the FFUEAT augmented technique with the ADMOZ dataset, reveal significant advancements over Kexin, Liang,[14] particularly in sparse minority classes such as 'adult', 'news', and 'kids'. Kexin's study, which employed CCN and RNN, achieved lower performance in 'home' at 0.76, 'adult' at 0.61, 'news' at 0.38, and 'kids' at 0.69, and the overall F1 score is 0.50, as mentioned in Table 27. Still, the current study demonstrates substantial improvements, with BERT achieving 0.9688 in 'home', 0.9604 in 'news', and 0.8012 in 'kids', as well as DistilBERT achieving 0.9600 in 'home', 0.9374 in 'news', and 0.5900 in 'kids'. This highlights BERT's superior capability in handling challenging minority classes except the 'adult' category, especially those with lower representation, where the FFUEAT augmented technique significantly boosts performance. Additionally, BERT's overall F1 score of 0.8891 surpasses that of the baseline study.

Table 28 presents the findings from 20 comprehensive experiments based on URL classification datasets, utilizing machine learning and deep neural networks, including the current study's experiments that employ various advanced methods. In the context of the URL classification dataset, the performance of URL classification was thoroughly evaluated by integrating and hybridizing deep neural networks, including ANN, RNN, LSTM, BiLSTM, RNN with GRU, and Bidirectional GRU with CNN. However, when using a large dataset like the DMOZ, they struggled to achieve good performance. Gupta and Bhatia[40] selected the ten majority classes, whereas the current study includes fifteen classes, covering the more challenging categories. Kurt and Demirel[38] achieved 0.76, 0.77, and 0.78 with a limited sample size. Throughout all the studies, particularly those on the DMOZ dataset, researchers faced difficulties with performance because they did not address class imbalance towards minority classes, noise, and ambiguous data issues in URL classification. Therefore, the FFUEAT technique enhances the performance of minority classes and improves the overall quality of URL classification by reducing noise and ambiguous data while maintaining contextual and semantic relevance. The current study sets a new benchmark in URL classification by integrating FFUEAT with advanced deep learning models, achieving superior F1 scores up to 0.8891 and outperforming prior research in large-scale URL dataset classification.

The compute noise score function assigns a noise score to each URL based on multiple heuristics, including URL length, non-alphanumeric ratio, digit ratio, excessive punctuation, subdomain count, suspicious top-level domains (TLDs), and the presence of URL-encoded sequences.

After computing the noise scores for all URLs, each class sampled 5000, and a threshold (NOISE THRESHOLD = 0.5) is applied to classify URLs as either "noisy" or "clean". The scores are then aggregated per category (class) to compute the total number of URLs (n URLs), the mean noise score (mean noise), and the median noise score (median noise) as shown in Table 29.

Discussion

The FFUEAT augmented technique is a significant advancement in URL classification. It effectively man-

Table 28. Comparison of the current study with baseline studies using the URL classification datasets, including cybersecurity datasets.

Research Studies	Approach	Training, Testing	Class	F1 Score
Rajalakshmi, Tiwari,[13] 2020	RNN with GRU		Kids with the rest of all	0.7792
	RNN with Bidirectional GRU	46280 For Kids & Rest of URL. 80%, and 20%		0.7962
	Bidirectional GRU with CNN			0.8204
Kexin, Liang,[14] 2021	TextCNN	1.5 Million	15 Classes	0.5000
	TextCNN		6 Soft classes	0.8107
	TextCNN		6 Hard classes	0.7346
Kurt and Demirel.[38] 2022	CNN	120K, 30K	15 Classes	0.7800
	LSTM			0.7600
	GRU			0.7700
Gupta and Bhatia,[40] 2022	BERT fused	Reuters-RCV1	10 Classes	0.9100
	Knowledge Embeddings	WebKB		0.8700
	With Deep Inception	newsgroup	Limited Samples	0.9000
		Conference		0.7900
Kumar, Muthusamy and Jerald,[42] 2024	CNN & RNN	Phishing	Binary Classes	0.9670 Phishing Dataset
Ghalechyan, Israyelyan,[18] 2024	BERT	Phishing	Binary Classes	0.9700 Phishing Dataset
Current study	LSTM,	1.6+ Million, 90K	15 Classes	0.8172
	BiLSTM,			0.8511
	DistilBERT and			0.8639
	BERT with Augmented DMOZ Dataset			0.8891
	BiLSTM with Augmented with Phishing Dataset	0.49 Million		0.9900

Table 29. Noisy class rankings in the DMOZ dataset.

Rank	Classes	Samples	Median Noise	Mean Noise
1	adult	5000	0.169318	0.172716
2	science	5000	0.165942	0.170375
3	kids	5000	0.162103	0.169648
4	computer	5000	0.160870	0.164192
5	reference	5000	0.158374	0.165000
5	games	5000	0.158374	0.162726
7	business	5000	0.155072	0.157838
7	recreation	5000	0.155072	0.158760
7	society	5000	0.155072	0.159851
7	arts	5000	0.155072	0.158764
7	news	5000	0.155072	0.158159
7	sports	5000	0.155072	0.159127
13	health	5000	0.152049	0.156674
13	home	5000	0.152049	0.157041
15	shopping	5000	0.149275	0.151038

ages class imbalance, reduces ambiguous data, and incorporates noise into the URLs. It also integrates contextual and semantic augmented data into the 17 categories of both datasets, as shown in Fig. 3. The FFUEAT technique enhances the representation of sparse minority classes and handles noise in cybersecurity categories, such as 'legitimate' and 'phishing'. This technique is effective with both RNN and transformer models.

The classes have multifaceted problems and limited data representation, even though some of them ('adult', 'science') have multidimensional features. Meanwhile, phishing URLs have more noisy data; nevertheless, empirical evaluation shows that the proposed technique demonstrates enhanced performance on both complex datasets (DMOZ and Phishing) for URL classification. The RNN model is more effective on the 'adult' and 'science' datasets than the Transformer model; this suggests that RNNs are better at capturing the sequential nature of these datasets or converge more efficiently with fewer samples. However, Transformer models (DistilBERT and BERT) demonstrated enhanced performance in the 'news' (25.89%) and 'kids' (29.2%) classes. The RNN's architecture, based on sequential data, is more effective for noisy URL classes than Transformers. Therefore, the 'adult' category achieved a 0.6661 F1-score using the RNN model, compared to the Transformer model, which obtained 0.5502, as shown in Table 18.

Conclusion

The experimental results show that the FFUEAT technique significantly improves the performance of the URL classification dataset, particularly for sparse minority and cybersecurity classes, by reducing class imbalance, noise, and ambiguous data. Using deep learning classifiers such as LSTM, BiLSTM, BERT, and DistilBERT shows consistent performance on aug-

mented datasets. Interestingly, the RNN classifiers (LSTM and BiLSTM) seem to be effective on the noisy and multi-feature classes; therefore, the RNN classifier obtained an F1 score for 'phishing' (0.99), 'adult' (0.6661) and 'science' (0.7065). Whereas BERT from Transformer obtained 25.89% in 'news', and 29.2% in 'kids' using the augmented dataset, compared to the standard DMOZ and Phishing datasets.

Future work will propose a unified model trained on the general and cybersecurity-specific URL classification datasets, enabling the model to be used for URL classification tasks such as web content filtering, access control, quality of service, and addressing cybersecurity threats such as malicious detection, blocking URLs used in spam emails, social engineering attacks, or fraudulent websites.

Acknowledgment

The acknowledgement of the FAI (Faculty of Artificial Intelligence) at Universiti Teknologi Malaysia, Kuala Lumpur, for providing the necessary support for this research.

Authors' declaration

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are ours. Furthermore, any Figures and images that are not ours have been included with the necessary permission for re-publication, which is attached to the manuscript.
- No animal studies are present in the manuscript.
- No human studies are present in the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee at University Teknologi.

Authors' contributions statement

Zafar Ali was responsible for conceptualization, as well as designing and developing the technique. Siti Sophiayati Yuhaniz contributed to the research design and conducted a critical analysis of the results. Noureen prepared the original draft and performed experiments and evaluations. Ghulam Mujtaba Shaikh supervised the research work. Husham provided a thorough review of the manuscript.

Data availability

The DMOZ dataset is available on Kaggle at https://www.kaggle.com/datasets/shawon10/url-classification-dataset-dmoz. Additionally, the Phishing dataset is publicly available at https://data.mendeley.com/datasets/vfszbj9b36/1. The dataset of FFUEAT of the minority classes is availbe at https://www.kaggle.com/datasets/zafarmahar/ffueat-based-augmented-dataset-of-minority-classes. The BERT model and tokenizer files are very large; therefore, they will be shared upon request from researchers.

References

1. Yuan Z, Liu Y. Extraction of Feature Information and Noise Reduction Strategies for Web Pages. 2024 14th International Conference on Information Technology in Medicine and Education (ITME); 2024 13–15 Sept. 2024;41–44. https://doi.org/10.1109/ITME63426.2024.00018.
2. Guo H. Research on Web Data Mining Based on Topic Crawler. J. Web Eng. 2021;20(4):1193–206. https://doi.org/10.13052/jwe1540-9589.20411.
3. Dhar T, Mazumder S, Dhar S, Karak S, Chatterjee D. An Approach to Design and Implement Parallel Web Crawler. 2021 International Conference on Smart Generation Computing, Communication and Networking (SMARTGENCON). 2021;1–4. https://doi.org/10.1109/SMARTGENCON51891.2021.9645918.
4. Li Y, Du T, Zhu L, Qu S. An Efficient Minimal Text Segmentation Method for URL Domain Names. Sci. Program. 2021;volume 2021(1):9946729. https://doi.org/10.1155/2021/9946729.
5. Zhang Y, Rao Z. n-BiLSTM: BiLSTM with n-gram Features for Text Classification. 2020 IEEE 5th information technology and mechatronics engineering conference (ITOEC); 2020:IEEE. 1056–1059 https://doi.org/10.1109/ITOEC49072.2020.9141692.
6. Cyprienna RA, Yannick RZL, Randria I, Raft RN. URL Classification based on Active Learning Approach. 2021 3rd International Cyber Resilience Conference (CRC); 2021 29–31 Jan. 2021;1–6. https://doi.org/10.1109/CRC50527.2021.9392555.
7. Deng L, Du X, Shen Jz. Web page classification based on heterogeneous features and a combination of multiple classifiers. Front Inf Technol Electron. Eng. 2020;21(7):995–1004. https://doi.org/10.1631/FITEE.1900240.
8. Lango M, Stefanowski J. What makes multi-class imbalanced problems difficult? An experimental study. Expert Syst Appl. 2022;199:116962. https://doi.org/10.1016/j.eswa.2022.116962.
9. Wang J, Awang N. MKC-SMOTE: A novel synthetic oversampling method for multi-class imbalanced data classification. IEEE Access. 2024;12:196929–38. https://doi.org/10.1109/ACCESS.2024.3521120.
10. Elsadig M, Ibrahim AO, Basheer S, Alohali MA, Alshunaifi S, Alqahtani H, *et al.* Intelligent deep machine learning cyber phishing url detection based on bert features extraction. Electronics. 2022;11(22):3647. https://doi.org/10.3390/electronics11223647.

11. Zieni R, Massari L, Calzarossa MC. Phishing or not phishing? A survey on the detection of phishing websites. IEEE Access. 2023; 11: 18499-519. https://doi.org10.1109/ACCESS.2023.3247135.
12. Wang J, Awang N. A Novel Synthetic Minority Oversampling Technique for Multiclass Imbalance Problems. IEEE Access. 2025;13:6054–66. https://doi.org/10.1109/ACCESS.2025.3526673.
13. Rajalakshmi R, Tiwari H, Patel J, Kumar A, Karthik R. Design of Kids-specific URL Classifier using Recurrent Convolutional Neural Network. Procedia Comput Sci. 2020;167:2124–31. https://doi.org/10.1016/j.procs.2020.03.260.
14. Kexin X, Liang B, Rai A, Chan A. URL Classification with Deep Learning. 2021. National University of Singapore. April 27th 2021:1–5.
15. Hung PD, Hung ND, Diep VT. URL Classification Using Convolutional Neural Network for a New Large Dataset. International Conference on Cooperative Design, Visualization and Engineering; 2022:Springer. 103–114. https://doi.org/10.1007/978-3-031-16538-2_11.
16. Abroshan H, Devos J, Poels G, Laermans E. Phishing happens beyond technology: The effects of human behaviors and demographics on each step of a phishing process. IEEE Access. 2021;9:44928–49. https://doi.org/10.1109/ACCESS.2021.3066383.
17. Barik K, Misra S, Mohan R. Web-based phishing URL detection model using deep learning optimization techniques. Int J Data Sci Analyt. 2025:1–23. https://doi.org/10.1007/s41060-025-00728-9.
18. Ghalechyan H, Israyelyan E, Arakelyan A, Hovhannisyan G, Davtyan A. Phishing URL detection with neural networks: an empirical study. Sci Rep. 2024;14(1):25134. https://doi.org/10.1038/s41598-024-74725-6.
19. Thabtah F, Hammoud S, Kamalov F, Gonsalv es A. Data imbalance in classification: Experimental evaluation. Inf Sci. 2020;513:429–41. https://doi.org/10.1016/j.ins.2019.11.004.
20. Nemoto S, Kitada S, Iyatomi H. Majority or Minority: Data Imbalance Learning Method for Named Entity Recognition. IEEE Access. 2025;13:9902–9. https://doi.org/10.1109/ACCESS.2024.3522972.
21. Bujang SDA, Selamat A, Ibrahim R, Krejcar O, Herrera-Viedma E, Fujita H, *et al.* Multiclass Prediction Model for Student Grade Prediction Using Machine Learning. IEEE Access. 2021;9:95608–21. https://doi.org/10.1109/ACCESS.2021.3093563.
22. Aguilar-Ruiz JS, Michalak M. Classification performance assessment for imbalanced multiclass data. Sci Rep. 2024;14(1):10759. https://doi.org/10.1038/s41598-024-61365-z.
23. Wei J, Zou K. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. arXiv preprint arXiv:1901.11196 2019. https://doi.org/10.48550/arXiv.1901.11196.
24. Vo MT, Nguyen T, Vo HA, Le T. Noise-adaptive synthetic oversampling technique. Appl Intell. 2021;51(11):7827–36. https://doi.org/10.1007/s10489-021-02341-2.
25. Selva Birunda S, Kanniga Devi R. A review on word embedding techniques for text classification. Innov Data Commun Technol Appl. 2021:267–81. https://doi.org/10.1007/978-981-15-9651-3_23.
26. Kabakus AT. Towards the Importance of the Type of Deep Neural Network and Employment of Pre-trained Word Vectors for Toxicity Detection: An Experimental Study. J Web Eng. 2021;20(8):2243–68. https://doi.org/10.13052/jwe1540-9589.2082.
27. Nasution NA, Nababan EB, Mawengkang H. Comparing LSTM Algorithm with Word Embedding: FastText and Word2Vec in Bahasa Batak-English Translation. 2024 12th International Conference on Information and Communication Technology (ICoICT); 2024:IEEE.306–313 https://doi.org/10.1109/ICoICT61617.2024.10698481.
28. Rana A, Pant A, Rawat N, Rawat P, Vats S, Sharma V. Semantic Similarity Analysis using FastText. 2024 IEEE 3rd World Conference on Applied Intelligence and Computing (AIC); 2024: IEEE. 454–460. https://doi.org/10.1109/AIC61668.2024.10731025.
29. AL-Jumaili ASA, Tayyeh HK. Character-Level Embedding using FastText and LSTM for Biomedical Named Entity Recognition. 2024;25(6):5258–64. https://doi.org/10.12694/scpe.v25i6.3365.
30. Arianto WR, Yuyun, Abrar A, Umar N, Nasrullah, Nurfaedah, *et al.* Comparative Study of Word2Vec, FastText, and Glove Embeddings for Synonym Identification in Bugis Language. 2024 Beyond Technology Summit on Informatics International Conference (BTS-I2C); 2024 19–19 Dec. 2024. 555–560. https://doi.org/10.1109/BTS-I2C63534.2024.10942212.
31. Sadiq S, Aljrees T, Ullah S. Deepfake Detection on Social Media: Leveraging Deep Learning and FastText Embeddings for Identifying Machine-Generated Tweets. IEEE Access. 2023;11:95008–21. https://doi.org/10.1109/ACCESS.2023.3308515.
32. Alsenani TR, Ayon SI, Yousuf SM, Anik FBK, Chowdhury MES. Intelligent feature selection model based on particle swarm optimization to detect phishing websites. Multimed Tools Appl. 2023;82(29):44943–75. https://doi.org/10.1007/s11042-023-15399-6.
33. Sankaranarayanan S, Sivachandran AT, Mohd Khairuddin AS, Hasikin K, Wahab Sait AR. An ensemble classification method based on machine learning models for malicious Uniform Resource Locators (URL). Plos one. 2024;19(5):e0302196. https://doi.org/10.1371/journal.pone.0302196.
34. Shahba L, Heidary-Sharifabad A, Mollahoseini Ardakani M. Detection of fake web pages and phishing attacks with rabbit optimization algorithm. J Supercomputing. 2024;81(1):313. https://doi.org/10.1007/s11227-024-06658-w.
35. Pathak P, Shrivas AK. Development of Proposed Model Using Random Forest with Optimization Technique for Classification of Phishing Website. SN Comput Sci. 2024;5(8):1059. https://doi.org/10.1007/s42979-024-03388-x.
36. Giri S, Banerjee S. Ensemble Learning Approach for Phishing Website Detection Using an Optimal Greedy Stacking Model. J Inst Eng India Ser B. 2024. https://doi.org/10.1007/s40031-024-01134-8.
37. Sudar KM, Rohan M, Vignesh K. Detection of adversarial phishing attack using machine learning techniques. Sādhanā. 2024;49(3):232. https://doi.org/10.1007/s12046-024-02582-0.
38. Kurt MS, Demirel EY. Web page classification with deep learning methods. Uludağ Üniversitesi Mühendislik Fakültesi Dergisi. 2022;27(1):191–204. https://doi.org/10.17482/uumfd.891038.
39. Buber E, Diri B. Web Page Classification Using RNN. Procedia Comput Sci. 2019;154:62–72. https://doi.org/10.1016/j.procs.2019.06.011.
40. Gupta A, Bhatia R. Knowledge based deep inception model for web page classification. J Web Eng. 2021;20(7):2131–68. https://doi.org/10.13052/jwe1540-9589.2075.

41. Nanda M, Goel S. URL based phishing attack detection using BiLSTM-gated highway attention block convolutional neural network. Multimed Tools Appl. 2024;83(27):69345–75. https://doi.org/10.1007/s11042-023-17993-0.
42. Kumar SS, Muthusamy P, Jerald MPA. A Hybrid Framework for Improved Weighted Quantum Particle Swarm Optimization and Fast Mask Recurrent CNN to Enhance Phishing-URL Prediction Performance. Int J Comput Intell Syst. 2024;17(1):251. https://doi.org/10.1007/s44196-024-00663-w.
43. Birthriya SK, Ahlawat P, Jain AK. Enhanced Phishing Website Detection Using Dual-Layer CNN and GRU with Attention Mechanism and Lexical NLP Features. SN Comput Sci. 2024;5(7):929. https://doi.org/10.1007/s42979-024-03282-6.
44. Huspi SHH, Ali Z. Sentiment analysis on roman urdu students' feedback using enhanced word embedding technique. Baghdad Sci J. 2024;21(2):46. https://doi.org/10.21123/bsj.2024.9822.
45. Ali S, Wang J, Leung VCM. AI-driven fusion with cybersecurity: Exploring current trends, advanced techniques, future directions, and policy implications for evolving paradigms– A comprehensive review. Information Fusion. 2025;118:102922.https://doi.org/10.1016/j.inffus.2024.102922.
46. Ali S, Wang J, Leung VCM, Bashir F, Bhatti UA, Wadho SA, Humayun M. CLDM-MMNNs: Cross-layer defense mechanisms through multi-modal neural networks fusion for end-to-end cybersecurity—Issues, challenges, and future directions. Information Fusion. 2025;122: 103222. https://doi.org/10.1016/j.inffus.2025.103222.
47. Afzal S, Asim M, Beg MO, Baker T, Awad AI, Shamim N. Context-aware embeddings for robust multiclass fraudulent URL detection in online social platforms. Computers and Electrical Engineering. 2024;119: 109494. https://doi.org/10.1016/j.compeleceng.2024.109494.
48. Marivate V, Sefara T, editors. Improving short text classification through global augmentation methods. International Cross-Domain Conference for Machine Learning and Knowledge Extraction; 2020: Springer. https://doi.org/10.1007/978-3-030-57321-8_21.
49. Karimi A, Rossi L, Prati A. AEDA: An easier data augmentation technique for text classification. arXiv preprint arXiv: 210813230. 2021. https://doi.org/10.48550/arXiv.2108.13230.
50. Xie Z, Wang SI, Li J, Lévy D, Nie A, Jurafsky D, Ng AY. Data noising as smoothing in neural network language models. arXiv preprint arXiv:170302573. 2017. https://doi.org/10.48550/arXiv.1703.02573.
51. Min J, McCoy RT, Das D, Pitler E, Linzen T. Syntactic data augmentation increases robustness to inference heuristics. arXiv preprint arXiv:200411999. 2020. https://doi.org/10.48550/arXiv.2004.11999.

تقنية حديثة لتعزيز أداء تصنيف عناوين URL المتعددة والأمن السيبراني باستخدام النماذج القائمة على الشبكات العصبية المتكررة (RNN) والمحولات (Transformers)

ظفر علي[1]، ستي صوفياتي يوهانيز[1]، نورين[1]، غلام مجتبى[2]، هشام محمد أحمد[3]

[1]كلية الذكاء الاصطناعي، الجامعة التكنولوجية الماليزية، كوالالمبور، ماليزيا.
[2]قسم علوم الحاسب الآلي، جامعة سوكور IBA، سوكور 65200، باكستان.
[3]كلية الهندسة، الجامعة التكنولوجية، البحرين، مملكة البحرين.

الخلاصة

تُنشر آلاف المواقع الإلكترونية الجديدة يوميًا، مما يُشكّل تحديات كبيرة لتصنيف المواقع الإلكترونية والأمن السيبراني. وتواجه مجموعات بيانات تصنيف عناوين المواقع، سواءً العامة منها أو الخاصة بالأمن السيبراني، تحدياتٍ مثل عدم توازن الفئات، والتشويش، والبيانات الغامضة، والتي قد تؤثر بشكلٍ كبير على أداء النموذج. تقترح هذه الدراسة تقنيةً جديدةً لتحسين تضمين النصوص السريعة المُعدّلة بدقة (FFUEAT). استُخدمت مجموعتا بيانات (DMOZ وPhishing) لتقييم أداء التقنية المقترحة، حيث حققت نتائج F1 بلغت 0.8639 باستخدام DistilBERT و0.8891 باستخدام BERT. كما تم تحسين أداء الفئات الأقل تمثيلًا، حيث حققت زيادةً في الدقة بنسبة 20.88% لفئة "الرئيسية"، و58.04% لفئة "الأخبار"، و11.18% لفئة "الأطفال" في مجموعة بيانات DMOZ المتاحة للجمهور باستخدام تقنية FFUEAT. عند تطبيقها على مجموعة بيانات الأمن السيبراني (التصيد الاحتيالي)، باستخدام نموذج BiLSTM، حققت التقنية المقترحة نتائج F1 بلغت 0.98 للمواقع الشرعية و0.99 لعناوين URL الخاصة بالتصيد الاحتيالي. وقد قللت تقنية FFEUEAT من تأثير عدم توازن الفئات والتشويش والبيانات الغامضة في مجموعات بيانات تصنيف عناوين URL. توفر هذه الطريقة وسيلة واعدة لتحسين تصنيف مواقع الويب والكشف عن تهديدات الأمن السيبراني، مما يُسهم في إدارة أفضل للمحتوى الإلكتروني وتعزيز السلامة.

الكلمات المفتاحية: عدم توازن الفئات، الأمن السيبراني، FastText، الشبكات العصبية المتكررة والمحولات، تصنيف عناوين URL .