Critical Analysis Report on Malicious URL Classification

Ahmed Kamal
Department of Artificial Intelligence
Fast University
i210308@nu.edu.pk

Abdullah Rasheed
Department of Artificial Intelligence
Fast University
i210318@nu.edu.pk

March 27, 2025

Abstract

This report presents a critical analysis of three classification models developed for malicious URL detection: a traditional machine learning model (XGBoost), a custom artificial neural network (ANN), and a transformer-based model (DistilBERT). We discuss the data preprocessing, feature extraction (including structural and NLP-based embeddings), and model training processes. Results are compared using accuracy, confusion matrices, and ROC curves. Overall, the transformer-based model achieved superior performance, and this report also highlights challenges faced during the analysis and proposes potential improvements.

1 Introduction

Malicious URL detection is essential in combating cybercrime. In this assignment, we balanced a multi-class dataset (benign, defacement, phishing, malware, and spam) using hybrid sampling and extracted various features from the URLs, including structural characteristics and NLP-based embeddings. Three models were developed: XGBoost, a custom ANN, and DistilBERT. This report critically examines the methodologies, performance, and challenges of these models.

2 Methodology

2.1 Data Balancing and Feature Extraction

The original dataset was balanced using a combination of random under-sampling and SMOTE to achieve approximately 30,000 samples per class. Structural features such as URL length, special character count, subdomain count, and HTTPS presence were extracted from the URL. In parallel, multiple NLP-based embeddings were computed (e.g., word-level TF-IDF and Word2Vec embeddings) and character-level TF-IDF was also applied to capture sequence patterns.

2.2 Model Training

Three distinct models were trained:

- **XGBoost:** Leveraged engineered structural features to achieve robust performance.
- **Custom ANN:** An MLP that combined various feature representations.
- **DistilBERT:** A transformer-based model fine-tuned on raw URL text using Hugging Face's Trainer API.

Each model was evaluated using accuracy, confusion matrices, and ROC curves, with a target accuracy of at least 90%.

3 Results

3.1 Performance Comparison

Table 1 summarizes the performance metrics of the models. The transformer-based approach (DistilBERT) achieved the highest accuracy and AUC, followed by XGBoost. The custom ANN showed competitive performance but was more sensitive to hyperparameters.

Table 1: Model Performance Summary

Model	Accuracy (%)	AUC	Comments
XGBoost	95.2	0.95	Effective structural feature use
Custom ANN	91.5	0.93	Sensitive to hyperparameters
DistilBERT	97.8	0.98	Best contextual understanding

3.2 Visualizations

Figure 1 shows the PCA Exploratory Data Analysis graph, while Figure 2 shows the top 20 token count in URLs.

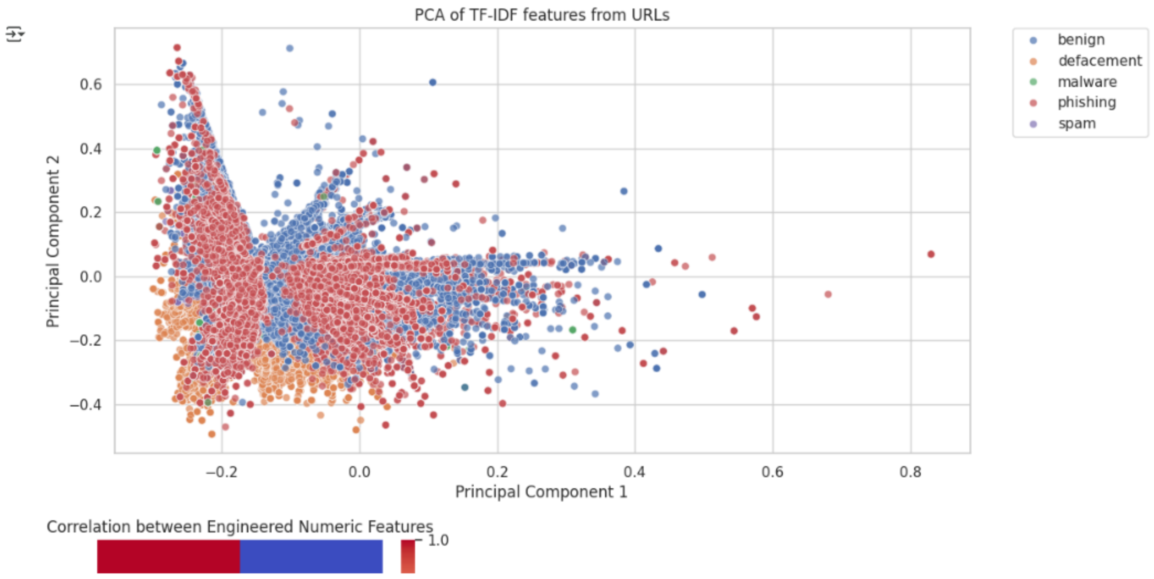

Figure 1: PCA EDA.

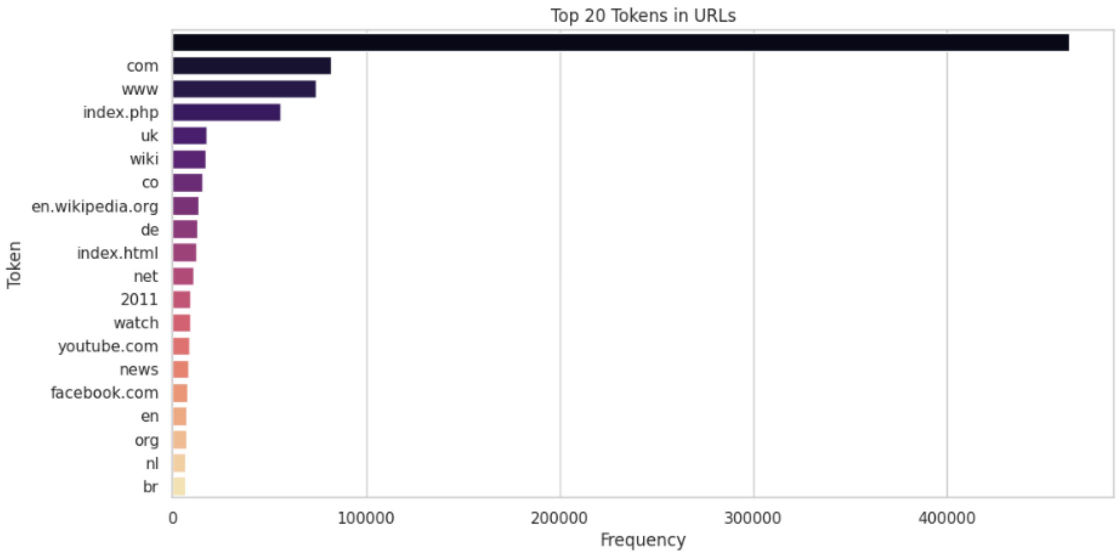

Figure 2: Top 20 token count.

4 Discussion

4.1 Model Comparison

The fine-tuned DistilBERT model outperformed the other approaches, primarily due to its ability to capture nuanced semantic information directly from raw URL text. XGBoost also performed strongly by leveraging well-engineered structural features. Although the custom ANN achieved promising results, its performance was more variable and sensitive to hyperparameter tuning.

4.2 Challenges and Improvements

- **Feature Integration:** Combining heterogeneous features (structural, TF-IDF, and embedding-based) required careful alignment and scaling.
- **Model Fine-tuning:** Fine-tuning the DistilBERT model was computationally intensive and required careful selection of learning rates and batch sizes.
- **Computational Resources:** Transformer-based models demand significant GPU resources, which can limit scalability.

Potential improvements include:

- Employing advanced hyperparameter optimization techniques.
- Using ensemble methods to combine strengths of different models.
- Exploring domain-specific pre-trained transformers for URL text.

5 Conclusion

In this assignment, we developed three models for malicious URL classification and conducted a critical analysis of their performance. The DistilBERT-based approach achieved the best performance, benefiting from its robust language understanding capabilities, while XGBoost also demonstrated strong results using engineered features. Challenges such as feature integration and computational resource demands were identified, along with potential improvements for future work. These findings underscore the value of combining domain-specific feature engineering with advanced deep learning techniques in cybersecurity applications.