Detection of Malicious URLs Using Machine Learning and LLM-Based Models

Zarmeen Tauseef

21I-0312

Isaam Ansari

21I-0299

1. Introduction

The internet has become a primary medium for both legitimate and illicit activities. Malicious URLs, used for phishing, malware distribution, and spam, pose significant cybersecurity risks. Traditional blacklisting methods are ineffective against newly generated malicious URLs. This study aims to classify URLs into five categories—**benign, defacement, phishing, malware, and spam**—using:

1. **Traditional Machine Learning (Random Forest)**
2. **Deep Learning (LSTMs)**
3. **LLM-Based Approaches (Fine-Tuned BERT)**

We explore different feature extraction techniques and evaluate model performance using confusion matrices and ROC curves.

2. Data Merging & Preprocessing

2.1 Dataset Merging

Two datasets were merged to ensure all five categories were present. The **main dataset** contained four classes, and an additional dataset was used to incorporate the missing class.

2.2 Handling Missing Data

- Removed duplicate entries.
- Filled missing values where necessary.
- Standardized the URL formatting.

2.3 Encoding Categorical Variables

The **target labels** (benign, phishing, malware, spam, defacement) were **encoded** using **Label Encoding** for machine learning models.

3. Balancing the Dataset

Class imbalance was addressed using **SMOTE (Synthetic Minority Over-sampling Technique)** to ensure fair representation across all five categories. This prevented models from being biased toward the majority class.

4. Exploratory Data Analysis (EDA)

EDA helped uncover patterns and relationships between **URL structure and malicious behavior**. Key insights:

1. **Benign URLs tend to be shorter**, while **phishing URLs** have excessive subdomains.
2. **Malware and spam URLs** frequently contain **random alphanumeric strings**.
3. **Phishing URLs** often imitate legitimate domains but contain typos.

4.1 Graphs and Visualizations

- **Distribution of URL lengths across categories** (Histogram)
- **Frequency of special characters in malicious vs. benign URLs** (Bar Chart)
- **Word cloud of common URL tokens in phishing attacks**
- **Correlation heatmap of extracted features**
- **Box plot comparing subdomain length across categories**

5. Feature Extraction

Three feature extraction techniques were applied:

1. **Structural Features (Manually Extracted):**

    - URL Length
    - Number of Special Characters
    - Subdomain Count
    - Path Length
    - Number of Digits

2. **TF-IDF (Text-Based Features):**

    - Converted URLs into **numerical representations** based on character frequency.

3. **Transformer-Based Embeddings (BERT Tokenization):**

    - URLs were tokenized using **BERT tokenizer** to capture contextual meaning.

6. Machine Learning & LLM-Based Models

6.1 Traditional Machine Learning: Random Forest

- **Random Forest** was trained using structural and TF-IDF features.
- Achieved an **accuracy of ~87%** (below the 90% requirement).

6.2 Deep Learning: LSTMs for Sequential Data

- **Long Short-Term Memory (LSTM) networks** were used for URL sequence classification.
- Trained on tokenized URLs to capture sequential dependencies.
- **Accuracy: ~92%** (better than traditional ML).

6.3 LLM-Based Approach: Fine-Tuned BERT

- **BERT (Bidirectional Encoder Representations from Transformers)** was fine-tuned for classification.
- Tokenized URLs and applied **transformer-based embeddings**.

- **Achieved the highest accuracy (~96%)**, outperforming other models.

7. Results & Visualization

7.1 Confusion Matrices for Model Comparison

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	87%	0.85	0.83	0.84
LSTM	92%	0.91	0.90	0.90
BERT	**96%**	**0.95**	**0.95**	**0.95**

- **BERT had the best precision and recall**, correctly identifying malicious URLs with the least false positives.

7.2 ROC Curves

- **Random Forest had the lowest AUC score (~0.88)**.
- **LSTMs performed better (~0.92 AUC)**.
- **BERT achieved the highest AUC (~0.98)**, proving superior classification capability.

8. Critical Analysis & Conclusion

8.1 Model Performance Comparison

- **Traditional ML models (Random Forest)** struggled with contextual understanding of URLs.
- **LSTM-based deep learning models** performed better but were computationally expensive.
- **LLM-based BERT outperformed all other models**, demonstrating strong generalization.

8.2 Challenges Faced

- Handling URLs as text sequences (BERT required special preprocessing).
- Training deep learning models required **high GPU resources**.
- Some URL categories were **difficult to distinguish** due to subtle differences.

8.3 Potential Improvements

- **Experimenting with other transformer models (e.g., GPT, DistilBERT)**.
- **Using ensemble methods** combining traditional ML, LSTMs, and transformers.
- **Integrating metadata** (e.g., WHOIS data, SSL certificate info) for better predictions.

9. Conclusion

This study successfully classified URLs into five categories using **three different approaches**. Fine-tuned **BERT achieved the highest accuracy (96%)**, proving the effectiveness of LLM-based models in cybersecurity. **Future improvements** can focus on **hybrid models** and **real-time threat detection**.