CERIAS

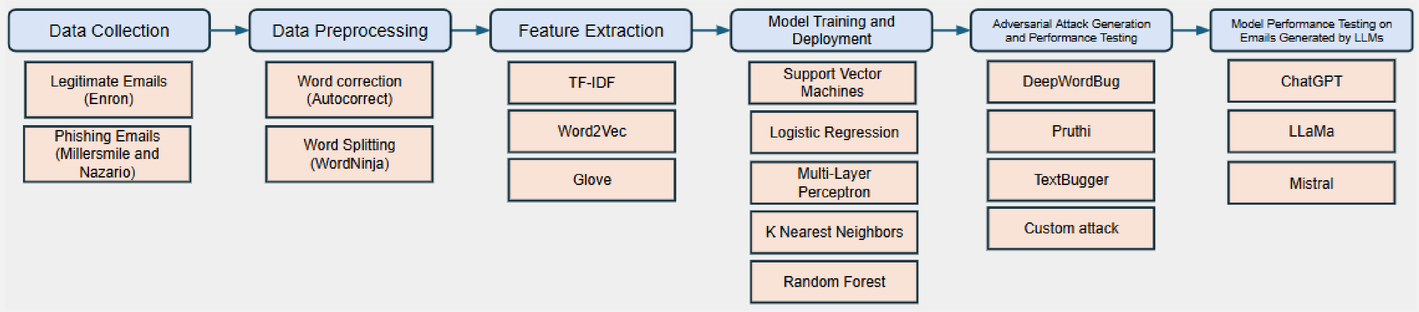
The Center for Education and Research in Information Assurance and Security

Adversarial Attacks Analysis of a Phishing Email Detection System based on Machine Learning and Word Error Correction

Deeksha Hareesha Kulal, and Quamar Niyaz, Electrical and Computer Engineering Department
Purdue University Northwest, Hammond, IN 46323

Motivation

Phishing remains a critical cyber threat, with traditional ML-based detection models relying on grammatical errors and word anomalies as key indicators. However, LLM-generated phishing emails are well-structured, grammatically sound, and highly deceptive, making detection increasingly challenging. This research explores the impact of word correction and splitting techniques in strengthening ML-based phishing detection. We further investigate how these enhancements improve detection accuracy against adversarial attacks and well-crafted LLM-generated phishing emails, paving the way for more resilient and adaptive cybersecurity solutions.



Research Questions

1) Can an ML-based phishing detection model, enhanced with word correction and splitting techniques, effectively identify phishing emails?
2) How does the proposed ML-based phishing detection model perform against adversarial attacks?
3) How effectively does the proposed ML-based phishing detection model detect LLM-generated phishing emails?

Methodology

Dataset Collection - Phishing and legitimate emails gathered from Millersmile, Nazario, and Enron dataset.

	Legitimate emails	Phishing emails
Model Development		
Training dataset	96,000	24,000
Test dataset	24,000	5,500
Model Deployment	22,458	4,492

Data Preprocessing - **Dataset 1** is preprocessed by removing stop words, numbers, special characters, and symbols while identifying email IDs, links, and attachments, whereas **Dataset 2** further improves text quality by applying spelling correction (Autocorrect) and word splitting (WordNinja) along with all the preprocessing steps from Dataset 1.

Original Email	Preprocessed Email with NLTK	Word correct and splitting
Hey! How are you? Can we schedule a meeting to discuss regarding the paper tomorrow?	hey schedule meeting discuss regarding paper tomorrow	hey schedule meeting discuss regarding paper tomorrow
You have won a **lotery** of $1000, click this link https://lotterywin.com and email me your bank **accountdetails**	**lotery** click link [LINK] email bank **accountdetails**	**lottery** click link [LINK] email bank **account details**

Feature Extraction - Conversion of text into numerical representations using TF-IDF, Word2vec and Glove embeddings.

Model Training and Deployment - Trained ML models (SVC, LR, RF, KNN and MLP) with extracted features, optimized performance through hyperparameter tuning and cross-validation. Evaluated models on unseen phishing emails in different timeline.

Adversarial Attack Generation and Performance Evaluation - Generated adversarial phishing emails for the models and validated their robustness against adversarial attacks. These attacks include character-level attacks from the PyTextAttack library, such as DeepWordBug, Pruthi, TextBugger, and one custom attack.

Attack Type	Transformed Email
DeepWordBug	acc0unt locked react1vated replly contact lnik unsubscrbe unsubscrbe
Pruthi	acount locked reactiavted rpelly contact unsibscribe unsibscribe
TextBugger	acc0unt locked react1vated reeply contact lnik unsbscribe unsbscribe
Custom Attack	accounɑ l0ckked reactvated replyy contacct link unsubscr1be unsubscr1be

Model Performance on Emails Generated by LLMs - Evaluated models with 580 phishing emails generated through LLMs (ChatGPT, Mistral, LLaMa) by using prompt engineering techniques.

Results

1) MLP-based models using Word2Vec and GloVe for feature extraction achieved the highest performance, with testing accuracies of 99.77% and 99.67% and F-scores of 0.994 and 0.991, respectively, during the development phase. An SVM-based model with Word2Vec achieved an accuracy of 94.32% and an F-score of 0.843 during the deployment phase.

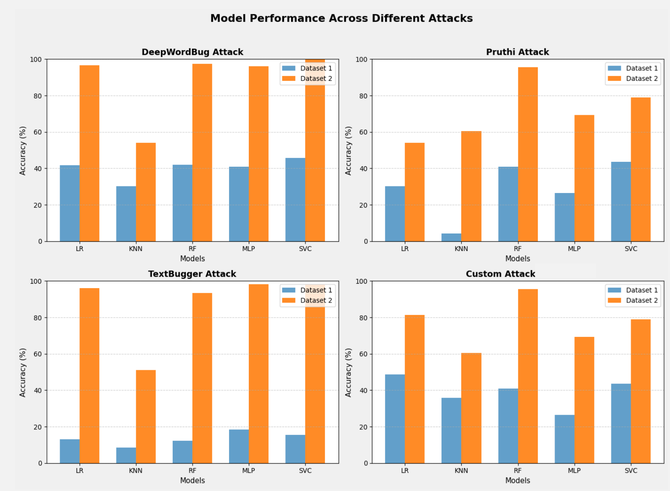
Model Development Phase

TF-IDF					
Model	**Training Accuracy (in %)**	**Testing Accuracy (in %)**	**Precision**	**Recall**	**F-score**
LR	97.67	97.81	0.966	0.915	0.940
SVM	98.78	98.52	0.980	0.940	0.960
RF	98.87	98.58	0.983	0.941	**0.961**
KNN	94.40	93.75	0.758	0.976	0.854
MLP	98.83	98.53	0.979	0.942	0.960
Word2Vec					
Model	**Training Accuracy (in %)**	**Testing Accuracy (in %)**	**Precision**	**Recall**	**F-score**
LR	98.96	98.93	0.972	0.970	0.971
SVM	99.86	99.74	0.996	0.990	0.993
RF	99.99	99.45	0.997	0.973	0.985
KNN	99.99	99.49	0.999	0.973	0.986
MLP	99.99	99.77	0.995	0.992	**0.994**
Glove					
Model	**Training Accuracy (in %)**	**Testing Accuracy (in %)**	**Precision**	**Recall**	**F-score**
LR	98.03	98.08	0.951	0.946	0.949
SVM	99.25	99.25	0.984	0.976	0.980
RF	99.99	99.16	0.993	0.961	0.977
KNN	99.99	99.28	0.997	0.964	0.981
MLP	99.98	99.67	0.994	0.988	**0.991**

Model Deployment Phase

TF-IDF				
Model	**Accuracy (in %)**	**Precision**	**Recall**	**F-score**
LR	94.04	0.905	0.718	0.801
SVM	94.23	0.888	0.748	0.812
RF	94.84	0.934	0.743	**0.828**
KNN	89.53	0.654	0.790	0.716
MLP	94.40	0.897	0.750	0.817
Word2Vec				
Model	**Accuracy (in %)**	**Precision**	**Recall**	**F-score**
LR	94.30	0.809	0.862	0.835
SVM	94.32	0.781	0.915	**0.843**
RF	93.46	0.872	0.713	0.784
KNN	93.53	0.922	0.668	0.775
MLP	93.61	0.747	0.934	0.830
Glove				
Model	**Testing Accuracy (in %)**	**Precision**	**Recall**	**F-score**
LR	92.21	0.751	0.796	0.773
SVM	93.66	0.785	0.854	0.818
RF	92.21	0.751	0.796	0.773
KNN	91.55	0.779	0.689	0.731
MLP	94.11	0.805	0.853	**0.828**

2) A comparative performance analysis of four adversarial attacks using the Word2Vec feature extractor highlights that preprocessing with spelling correction and word splitting significantly improves the classification of adversarial emails.



3) The best-performing model, Word2Vec with SVC, was tested on LLM-generated emails both before and after applying word correction and splitting. The results indicate that the proposed model's performance improved slightly after preprocessing.

Correctly classified	547	548
Incorrectly classified	33	32
Accuracy (in %)	94.31	**94.48**

Conclusion

Enhancing phishing email detection with word correction and splitting techniques improves model accuracy and robustness, especially against adversarial attacks and LLM-generated emails. This approach strengthens cybersecurity defenses and helps create more reliable detection systems.

PURDUE UNIVERSITY

CERIAS