People's Democratic Republic of Algeria

Ministry of Higher Education and Scientific Research

Setif 1 University - Ferhat Abbas

Faculty of sciences

Final Year Project:

Domain: Math and Computer Science
Sector Computer Sciences
Speciality: CS, IDTW

Master Thesis

Detect social engineering attacks: Phishing Email Detection Using Deep learning and transformers for Cybersecurity

Presented by:

MERIR Ishak Abdeldjebar

BENBEKHOUCHE Karim

Supervisor:

M. BERRHAIL Fouaz

2025/2026

Contents

List of Abbreviations 9
List of Abbreviations....10

Chapter 1: Introduction to Social Engineering and Cybersecurity 15
1.1 Social Engineering Attacks....15
1.1.1 Social Engineering Life Cycle....15
1.2 Cybersecurity....16
1.2.1 Foundational Cybersecurity Models....16
1.2.1.1 The CIA Triangle: 16
1.2.1.2 Defense-in-Depth: 16
1.2.1.3 Zero Trust Architecture (ZTA): 17
1.2.1.4 Risk Management Framework (RMF): 17
1.3 Cyberattack 17
1.3.1 Objectives of Cyberattacks....17
1.3.2 Types of Cyberattacks....18
1.3.3 Social Engineering Attack Types....19
1.4 Phishing: The Most Prevalent Social Engineering Attack....19
1.4.1 History of Phishing....19
1.4.2 Types of Phishing....20
1.4.3 Phishing Emails....20
1.4.4 Phishing Email Life Cycle....21
1.4.5 Impact and consequences of successful attacks....21
1.5 Analysis of Phishing Vectors....21
1.5.1 Phishing URL Detection....22
1.5.2 Phishing Webpage Detection....22
1.5.3 Phishing Email Detection....22
1.5.3.1 Structure of a phishing email....22
1.6 Conclusion....25

Chapter 2: Natural Language Processing (NLP) 26
2.1 Introduction:....26
2.2 Natural Language Processing Approaches:....26
2.2.1 Symbolic NLPs:....26

2.2.2 Statistical NLP: . . . 26
2.2.3 Neural NLP: . . . 27
2.3 Natural Language Processing Fundamentals: . . . 27
2.3.1 Text Preprocessing: . . . 27
2.3.2 Text Representation Techniques: . . . 27
2.3.2.1 Traditional/statistical: . . . 27
2.3.2.2 Word Embeddings (Static Vectors): . . . 28
2.3.2.3 Contextual Embeddings: . . . 28
2.4 NLP in Cybersecurity: . . . 28
2.5 Modern NLP for Phishing detection: . . . 29
2.5.1 Some Key Linguistic Features in Phishing Detection: . . . 29
2.6 Conclusion : . . . 29

Chapter 3: State of the Art 30
3.1 Traditional/Non-ML Based Detection Methods: . . . 30
3.1.1 List-based detection methods . . . 30
3.1.1.1 Limitations: . . . 30
3.1.2 Heuristic-based detection methods . . . 31
3.1.2.1 Limitations: . . . 31
3.2 Machine Learning for Phishing Detection . . . 31
3.2.1 Random Forest: . . . 31
3.2.2 SVM: . . . 32
3.2.3 Naïve Bayes: . . . 32
3.3 Deep Learning for Phishing detection . . . 33
3.3.1 Recurrent Neural Networks (RNNs) and LSTMs . . . 33
3.3.1.1 Traditional RNN: . . . 33
3.3.1.1.1 Limitations: . . . 34
3.3.1.2 Long Short-Term Memory (LSTM): . . . 34
3.3.1.3 Bidirectional LSTM (BiLSTM): . . . 34
3.3.2 Convolutional Neural Networks (CNNs) . . . 35
3.4 Transformers for phishing detection . . . 35
3.4.1 Types of Transformer-based Models . . . 36
3.4.2 Applications of Transformers in Phishing Detection: . . . 37

Chapter 4: Methodology 39
4.1 Introduction . . . 39
4.2 Dataset Collection: . . . 40
4.2.1 Dataset Sources . . . 40
4.3 Dataset Preprocessing . . . 41
4.3.1 Phase one: Data Cleaning . . . 41
4.3.2 Phase two: Normalization . . . 43

4.3.3 Phase three: Tokenization....43
4.4 Feature representation:....44
4.4.1 Step 1: Text-to-Sequence Conversion....44
4.4.2 Step 2: Padding....45
4.4.3 Step 3: Embedding....45
4.4.4 Statistical Representation:....46
4.4.4.1 TF-IDF:....46
4.4.5 Semantic Representation: Word Embeddings....46
4.4.5.1 Word2Vec:....47
4.4.5.2 GloVe:....47
4.4.5.3 End to End example :....48
4.4.5.3.1 RAW Email:....48
4.5 Description of the Selected Deep Learning and Transformer Models:....50
4.5.1 Baseline Traditional ML Model (Naïve Bayes):....50
4.5.1.1 Naïve Bayes for Phishing Detection:....51
4.5.1.2 Theoretical Limitations:....51
4.5.2 Baseline Deep Learning Models:....51
4.5.2.1 Long Short-Term Memory (LSTM):....51
4.5.2.1.1 Architectural Overview:....51
4.5.2.1.2 LSTM for phishing:....52
4.5.2.1.3 Theoretical Limitations:....52
4.5.2.2 Bidirectional LSTM(BiLSTM):....52
4.5.2.2.1 Architectural Overview:....53
4.5.2.2.2 Architectural components:....53
4.5.2.3 Convolutional Neural Network (CNN)....54
4.5.2.3.1 Architectural Overview:....54
4.5.2.3.2 Architectural components:....55
4.5.2.3.3 CNN for phishing:....56
4.5.2.3.4 Theoretical Limitations:....56
4.5.3 Transformer model (BERT):....56
4.5.3.1 Bidirectional Encoder Representations from Transformers (BERT):....57
4.5.3.1.1 Overview and architecture:....57
4.5.3.1.2 Architectural components:....58
4.5.3.1.3 BERT for phishing:....59
4.5.3.1.4 Theoretical Limitations:....59
4.6 Description of performance evaluation:....59
4.6.1 Training and Optimization Strategy (Conceptual):....59
4.6.2 Evaluation Metrics and Validation Protocol:....60
4.6.2.1 Confusion Matrix Components:....60

4.6.2.2 Evaluation Metrics:.....61

Chapter 5: Implementation 62

5.1 Experimental Setup.....62
5.1.1 Hardware and software Environment.....62
5.1.2 Libraries and framework versions.....62
5.2 Data collecting.....63
5.3 Preprocessing data.....63
5.3.1 Data cleaning.....63
5.3.2 Data splitting.....66
5.3.3 Tokenization.....66
5.3.3.1 Configuration.....66
5.3.4 Text to Sequence.....66
5.3.4.1 Explanation.....67
5.3.5 Padding Sequences.....67
5.3.5.1 Purpose.....67
5.3.5.2 Explanation.....67
5.3.6 Feature Representaions.....68
5.3.6.1 TF-IDF.....68
5.3.6.1.1 Explanation.....68
5.3.6.2 GloVe Embedding.....68
5.3.6.3 Word2Vec Embedding.....69
5.4 Model Architecture and Results.....70
5.4.1 Naive Bayes Model.....70
5.4.1.1 Implementation.....70
5.4.1.1.1 Explanation:.....70
5.4.1.2 Results(Naive Bayes).....70
5.4.2 CNN Model.....71
5.4.2.1 Model construction.....71
5.4.2.2 Compilation.....72
5.4.2.2.1 Explanation.....72
5.4.2.3 Training.....72
5.4.2.3.1 Explanation.....72
5.4.2.4 Discussion.....73
5.4.2.5 Results CNN with Word2Vec embedding.....73
5.4.2.6 Results CNN with GloVe embedding.....74
5.4.3 BiLSTM Model.....75
5.4.3.1 Model construction.....76
5.4.3.1.1 Explanation.....76
5.4.3.2 Compilation & training.....76

5.4.3.3 Discussion 76
5.4.3.4 Results: BiLSTM with Word2Vec 77
5.4.3.5 Results (BiLSTM + GloVe) 78
5.4.4 BERT Model 79
5.4.4.1 Model Construction 79
5.4.4.2 Optimization Setup: 80
5.4.4.2.1 Explanation: 80
5.4.4.3 Training Strategy: 81
5.4.4.4 Results (BERT) 81
5.5 Training Process and Hyperparameter Tuning 82
5.5.1 Optimization Strategy: 83
5.5.2 Hyperparameter Configuration 83
5.6 Comparison of results and discussion 84
5.6.1 Impact of Architecture and Embedding on Model Performance . . 84

List of Figures

1.1 Example of a phishing email[1].....23
1.2 Example 1[1].....23
1.3 Example 2[1].....24
1.4 Example 3[1].....24
1.5 Example 4[1].....25
1.6 Example 5[1].....25

4.1 The methodology Pipeline.....39
4.2 Text Processing Pipeline for Phishing Email Detection.....49
4.3 Comparison of Text Representation Methods.....50
4.4 BiLSTM architecture.....53
4.5 CNN architecture.....55
4.6 The method of processing an Email with BERT.....58

5.1 Dataset Distribution.....63
5.2 Email type Distribution.....65
5.3 Naive Bayes confusion matrix.....71
5.4 DeepCNN with Word2Vec training history.....73
5.5 DeepCNN Word2Vec confusion matrix.....74
5.6 DeepCNN GloVe training history.....75
5.7 DeepCNN GloVe confusion matrix.....75
5.8 BiLSTM Word2Vec training history.....77
5.9 BiLSTM Word2Vec confusion matrix.....78
5.10 BiLSTM GloVe training history.....78
5.11 BiLSTM GloVe confusion matrix.....79
5.12 BERT training history.....82
5.13 BERT confusion matrix.....82

List of Tables

3.1 Traditional Machine Learning Models for Phishing Email Detection . . . 33
3.2 Deep Learning Models for Phishing Email Detection.....35
3.3 Transformer-Based Models for Phishing Email Detection.....38
3.4 Grouped Models per Study for Phishing Email Detection.....38

4.1 Evaluation Metrics Used in Phishing Email Detection.....61

5.1 Libraries used in our study.....62
5.2 Model Training Workflow.....83
5.3 Unified Optimization Strategy Across Models.....83
5.4 Comparison of Hyperparameters Across Models.....84
5.5 Performance Comparison Across Architectures and Embeddings.....84

List of Abbreviations

AdaGrad	Adaptive Gradient Algorithm
AI	Artificial Intelligence
AOL	America Online
APWG	Anti-Phishing Working Group
AUC	Area Under the Curve
BART	Bidirectional and Auto-Regressive Transformer
BCE	Binary Cross-Entropy
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional Long Short-Term Memory
BOW	Bag of Words
CALO	Cognitive Agent that Learns and Organizes
CBOW	Continuous Bag of Words
CC-News	Common Crawl News
CIA	Confidentiality, Integrity, and
Availability CLS	Classification Token
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CSV	Comma-Separated Values
CTI	Cyber Threat Intelligence
DDos	Distributed Denial of Service
Dim	Dimension
DistilBERT	Distilled Bidirectional Encoder Representations from Transformers
DL	Deep Learning
Email	Electronic mail
ELMo	Embeddings from Language
Models FN	False Negative
FP	False Positive
GB	Giga Byte
GELU	Gaussian Error Linear Unit
GloVe	Global Vectors for Word Representation
GPT	Generative Pre-trained Transformer
GPU	Graphics Processing Unit
GRU	Gated Recurrent Unit
HTML	HyperText Markup Language
ID	Identification
IP	Internet Protocol
LLaMA	Large Language Model Meta AI

List of Abbreviations

LLM	Large Language Model
LSTM	Long Short-Term Memory
ML	Machine Learning
MMS	Multimedia Messaging Service
NB	Naive Bayes
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
OOV	Out Of Vocabulary
RAM	Random-Access Memory
ReLU	Rectified Linear Unit
RMSProp	Root Mean Square Propagation
RNN	Recurrent Neural Network
RoBERTa	Robustly optimized BERT approach
ROC	Receiver Operating Characteristic
ROC-AUC	Receiver Operating Characteristic - Area Under the Curve
SEP	Separator
Seq2Seq	Sequence to Sequence
SMS	Short Message Service
SQL	Structured Query Language
SSL	Secure Sockets Layer
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency
TIA	Threat Intelligence Analytics
TN	True Negative
TP	True Positive
T5	Text-to-Text Transfer Transformer
UCI	University of California Irvine Repository
UNK	Unkown
URL	Uniform Resource Locator
Word2Vec	Word to Vector
WW2	World War 2

Abstract

Phishing emails continue to represent one of the most dangerous cybersecurity threats due to their increasing sophistication and ability to imitate legitimate communication. Traditional detection techniques based on keyword matching and handcrafted features of-ten struggle to identify modern phishing attacks. This thesis investigates the application of deep learning and transformer models for phishing email detection, with a particular focus on BERT architectures, in addition to exploring alternative solutions like hybrid models. Several machine learning and deep learning models were evaluated, including Naive Bayes, CNN, BiLSTM, BERT, and a hybrid architecture combining BERT con-textual embeddings with TF-IDF statistical features. The models were trained and tested on a carefully preprocessed and balanced dataset containing 20,000 email samples. The proposed methodology includes preprocessing, feature representation using Word2Vec, and GloVe. Furthermore, the notion of attention mechanism was explored from imple-mentation to effect on model performance. For evaluation, metrics such as accuracy, precision, recall, and F1-score were used. As a result, the proposed BERT + TF-IDF hy-brid model produced the best overall results, achieving 99.25% accuracy and an F1-score of 99.07%, confirming that combining contextual semantic understanding with statistical keyword representation provides a more robust phishing detection mechanism. The find-ings also highlight the importance of high-quality preprocessing and feature engineering in improving model performance, and the fundamental role that attention provides in enhancing results across the board.

Keywords: Phishing attacks; Phishing email detection; Social engineering; Cybersecurity; Deep learning; transformers; Artificial intelligence.

ملخص

لالا تزال رسائل البريد الإلكتروني الاحتيالية تشكّل أحد أخطر التهديدات الأمنية الإلكترونية نظرًا لتطورها المتزايد وقدرتها على محاكاة الرسائل المشروعة. غالبًا ما تواجه تقنيات الكشف التقليدية، القائمة على مطابقة الكلمات المفتاحية واللصائص المُصممة يدويًا، صعوبة في تحديد نجمات التصيّد الحديثة. تتناول هذه الأطروحة في تطبيق نماذج التعلّم العميق ونماذج اللّغة الكبيرة للكشف عن رسائل البريد الإلكتروني الاحتيالية، مع التركيز بشكل خاص على بنية BERT، بالإضافة إلى استكشاف حلول بديلة مثل النماذج الحجينة. تم تقييم العديد من نماذج التعلّم الآلي والتعلّم العميق، بما في ذلك Naive Bayes وCNN وBiLSTM وBERT، وبنية نجينة تمزج بين تضمينات BERT السياقية واللصائص الإحصائية

TF-IDF. تم تدريب النماذج واختبارها على مجموعة بيانات مُعالجة مسبقًا ومتوازنة بعناية تحتوي على ما يقارب 20,000 عينة بريد إلكتروني. تشمل المنهجية المقترحة المعالجة المسبقة، وتمثيل اللصائص باستخدام word2Vec وGloVe، علاوة على ذلك، تم استكشاف مفهوم آلية الانتباه من حيث التنفيذ وتأثيرها على أداء النموذج. وللتقييم، تم استخدام مقاييس مثل الدقة، والضبط، والاستدعاء، ومقياس F1 ونتيجة لذلك، حقق النموذج الحجين المقترح TF-IDF + BERT أفضل النتائج الإجمالية، حيث بلغت دقته 99.25% ودرجة F1 99.07%، مما يؤكد أن التجميع بين الفهم الدلالي السياقي

والتمثيل الإحصائي للكلمات المفتاحية يوفر آلية أكثر فعالية للكشف عن التصيد الاحتيالي. تُبرز النتائج أهمية المعالجة المسبقة عالية الجودة وهندسة الميزات في تحسين أداء النموذج، والدور الأساسي الذي يلعبه الانتباه في تعزيز النتائج بشكل عام.

الكلمات المفتاحية : نجمات التصيد الاحتيالي؛ كشف رسائل البريد الإلكتروني للتصيد الاحتيالي؛ الحندسة الاجتماعية؛ الأمن السيبراني؛ التعلم العميق؛ اللولالات؛ الذكاء الاصطناعي

Résumé

Les courriels d'hameçonnage représentent toujours l'une des menaces les plus dangereuses en matière de cybersécurité, en raison de leur sophistication croissante et de leur capacité à imiter les communications légitimes. Les techniques de détection traditionnelles, basées sur la correspondance de mots-clés et l'utilisation de caractéristiques extraites manuellement, peinent à identifier les attaques d'hameçonnage modernes. Cette thèse étudie la détection des courriels d'hameçonnage à l'aide de modèles de type Transformer, en se concentrant particulièrement sur les architectures BERT et l'apprentissage profond. Plusieurs modèles d'apprentissage profond ont été évalués, notamment CNN, BiLSTM, Naive Bayes, ainsi que des modèles hybrides qui combinent les plongements contextuels BERT avec les caractéristiques automatiques et statistiques (TF-IDF). Les modèles ont été entraînés et testés sur un ensemble de données soigneusement prétraité, contenant environ 20 000 exemples de courriels équilibrés. La méthodologie proposée comprend le prétraitement, la représentation des caractéristiques à l'aide de Word2Vec, GloVe et BERT. De plus, la notion de mécanisme d'attention a été explorée pour son impact sur la mise en œuvre. Pour l'évaluation des performances du modèle, des métriques telles que l'exactitude, la précision, le rappel et le F1-score ont été utilisées. En conséquence, le modèle hybride BERT + TF-IDF a produit les meilleurs résultats globaux, atteignant une précision de 99,25 % et un F1-score de 99,07 %, confirmant ainsi que la combinaison d'une compréhension sémantique contextuelle et d'une représentation statis-tique des mots-clés offre un mécanisme de détection plus robuste. Les résultats soulignent également l'importance d'un prétraitement de haute qualité et de l'ingénierie des carac-téristiques pour améliorer les performances du phishing, ainsi que le rôle fondamental de l'attention dans l'amélioration globale des résultats.

Mots-clés : Cybersécurité ; Attaques de phishing ; Détection des e-mails de phishing ; Apprentissage profond ; Transformers ; Intelligence artificielle ; Ingénierie sociale