Turkish Phishing Email Detection: Cross-Lingual Transfer Learning Approaches

Alkım Gönenç Efe, Atilla Uğur
Department of Computer Engineering
İzmir Katip Çelebi University
İzmir, Turkey

***Abstract*—Phishing attacks remain a critical cybersecurity threat, with detection systems achieving high accuracy in English but leaving low-resource languages like Turkish underserved. This research investigates cross-lingual transfer learning approaches to develop effective Turkish phishing email detection without requiring native Turkish labeled datasets. We leverage a comprehensive English phishing dataset of approximately 39,000 emails and employ two strategies: (1) machine translation using MarianMT to create synthetic Turkish data for training BERTurk, and (2) multilingual transfer learning using XLM-RoBERTa trained on English and evaluated on Turkish through zero-shot and few-shot approaches. Traditional machine learning baselines using TF-IDF with Logistic Regression establish performance benchmarks. Due to the absence of authentic Turkish phishing datasets, all approaches rely on machine-translated data for evaluation. Models are evaluated using precision, recall, F1-score, and AUC-ROC metrics with comprehensive error analysis. This work provides empirical evidence on the viability of cross-lingual transfer for cybersecurity applications in low-resource languages.**

***Index Terms*—phishing detection, cross-lingual transfer learning, Turkish NLP, BERT, XLM-RoBERTa, low-resource languages**

I. Introduction

Phishing attacks represent one of the most pervasive cybersecurity threats, accounting for over 80% of reported security incidents globally. While automated phishing detection has achieved remarkable success in English, supported by extensive labeled datasets, low-resource languages like Turkish lack such resources. With over 80 million native speakers and a growing digital economy, Turkey faces increasing exposure to phishing attacks yet lacks native Turkish phishing datasets necessary for effective detection systems.

The absence of authentic Turkish phishing email datasets creates a fundamental challenge: how can we develop effective detection systems without labeled training data in the target language? This research addresses this challenge by investigating whether phishing detection knowledge can be transferred from English to Turkish through two distinct strategies. Recent advances in neural machine translation [1] and multilingual pre-trained language models [9], [10] offer promising solutions. Machine translation can create synthetic Turkish datasets from English sources, while multilingual models demonstrate impressive cross-lingual transfer capabilities without requiring translation.

Our research explores cross-lingual knowledge transfer for Turkish phishing detection using only English source data and machine-translated evaluation sets. We investigate how effectively phishing detection capabilities can be transferred from English to Turkish when no native Turkish training data exists, and which approach—machine translation with language-specific models or multilingual pre-trained models—provides superior performance.

We implement and compare three approaches. First, we establish traditional machine learning baselines using TF-IDF vectorization with Logistic Regression on translated Turkish text [4], [5]. Second, we train BERTurk [6], a Turkish-specific BERT model [2], on machine-translated Turkish data. Third, we leverage XLM-RoBERTa's [3] multilingual representations through zero-shot and few-shot transfer from English. Due to the unavailability of authentic Turkish phishing datasets, all evaluation is conducted on machine-translated Turkish text.

This research contributes: (1) the first systematic comparison of translation-based versus multilingual transfer approaches for Turkish phishing detection in the absence of native data, (2) a methodology for developing phishing detection systems for languages lacking labeled datasets, (3) empirical evidence on cross-lingual transfer effectiveness for cybersecurity tasks when relying entirely on synthetic data, and (4) practical recommendations for deploying detection systems in low-resource language contexts.

II. Related Work

A. Evolution of Phishing Detection Approaches

Phishing detection has evolved significantly from rule-based systems to sophisticated deep learning architectures. Early approaches relied on blacklists and signature-based techniques, which proved inadequate against sophisticated attacks [8]. Traditional methods using email content analysis, sender verification, and URL examination laid the groundwork for modern detection systems [9]. However, these approaches struggled with zero-day attacks and adaptive phishing techniques.

B. Machine Learning and Traditional Approaches

Classical machine learning approaches have demonstrated considerable success in phishing detection. Yasin and Abuhasan [3] proposed an intelligent classification model using Random Forest and J48 algorithms, achieving 99.1% and 98.4% accuracy respectively. Their work demonstrated

that feature engineering remains crucial even in the era of deep learning. Salloum et al. [4] conducted a comprehensive literature survey on phishing email detection using NLP techniques, identifying that traditional ML methods such as Naive Bayes, SVM, and Random Forest, when combined with TF-IDF features, can achieve competitive performance.

C. Deep Learning for Phishing Detection

The application of deep learning to phishing detection has gained significant momentum. Kyaw et al. [10] presented a systematic review of deep learning techniques, analyzing 33 papers. Their review revealed that CNNs achieved accuracy rates ranging from 93% to 99%, while LSTM models demonstrated accuracy between 95% and 98%. Thakur et al. [11] examined 88 papers published between 2017 and 2023, revealing critical limitations: limited datasets with insufficient diversity, concentration on English-language emails only, and shortage of tools for real-time deployment.

D. BERT-Based Architectures

Transformer-based models have revolutionized phishing detection through contextual understanding. Songailaitė et al. [12] investigated three BERT-based models (DistilBERT, TinyBERT, and RoBERTa) for phishing detection, demonstrating that all models achieved accuracy above 98.5%. The research emphasized trade-offs between model size and performance, with RoBERTa's 270 million parameters providing superior accuracy while DistilBERT's 66 million parameters offered more practical deployment characteristics.

E. Cross-Lingual and Low-Resource Challenges

The challenge of phishing detection in low-resource languages remains largely unaddressed. Gül Taş [13] addressed this gap specifically for Turkish, employing TF-IDF vectorization alongside BERTurk embeddings. Results demonstrated that BERT achieved 97.5% accuracy with F1-score of 0.91, significantly outperforming traditional methods. However, this work relied on a manually constructed dataset of only 2,410 emails and did not explore cross-lingual transfer learning approaches.

F. Research Gaps

Despite Turkish being spoken by over 80 million people, phishing detection research for this language remains extremely limited. The complete absence of publicly available, large-scale authentic Turkish phishing email datasets renders traditional supervised learning approaches infeasible. No prior work has comprehensively evaluated building detection systems entirely on machine-translated datasets or systematically compared machine translation approaches with multilingual model architectures. This research addresses these gaps by providing the first systematic comparison of translation-based versus multilingual transfer approaches for Turkish phishing detection, evaluated entirely on synthetic data.

III. Methodology

A. Dataset

1) Data Source: This study utilizes the CEAS '08 phishing dataset, which provides a comprehensive collection of labeled phishing and legitimate emails. This publicly available dataset offers diverse phishing tactics including financial fraud, account verification scams, and various social engineering approaches.

2) Dataset Statistics and Splitting: The dataset comprises approximately 39,000 email samples with a class distribution of roughly 44% legitimate and 56% phishing emails. Each sample includes subject line, body text, sender information, headers, and binary labels (0: legitimate, 1: phishing). The preprocessing pipeline identified and removed duplicate emails based on identical subject-body combinations to prevent data leakage.

To maintain class distribution across all experimental phases, we employed stratified splitting: 70% for training (approximately 27,407 samples), 15% for validation (approximately 4,110 samples), and 15% for test (approximately 5,874 samples). The validation split represents approximately 12.75% of total data due to sequential splitting.

3) Preprocessing: Data preprocessing involved: (1) removing duplicates to prevent training-test contamination, (2) standardizing UTF-8 encoding, (3) extracting URLs using comprehensive regular expressions to preserve links during translation, (4) text normalization including whitespace removal and control character elimination while preserving essential punctuation, and (5) label validation to ensure consistency.

Due to the absence of authentic Turkish phishing datasets, all Turkish data was generated through machine translation from English sources. This reflects the real-world challenge of developing security systems for low-resource languages.

B. Machine Translation

1) Translation System: We utilized MarianMT [12], specifically the Helsinki-NLP opus-mt-tc-big-en-tr model, a transformer-based neural translation system trained on large-scale parallel corpora. The configuration included maximum sequence length of 128 tokens with greedy decoding for deterministic outputs.

2) Translation Process: The translation process preserved critical features: (1) URL protection by replacing URLs with placeholders (URL_0, URL_1, etc.) before translation, (2) independent translation of subject lines and email bodies using batched processing with batch size 32, (3) URL restoration post-translation, (4) retention of metadata without translation, and (5) quality assurance verification. The entire translated dataset was cached for reuse across experiments.

C. Text Preprocessing

For all models, we implemented a consistent preprocessing pipeline: (1) concatenating subject and body with [SEP] separator, (2) optional lowercase conversion, (3) whitespace normalization, and (4) model-specific tokenization. BERTurk utilizes WordPiece tokenization while XLM-RoBERTa employs

SentencePiece. Both were applied with maximum sequence length of 256 tokens.

D. Algorithms and Models

1) Traditional Machine Learning Baseline: TF-IDF vectorization was configured with 5,000 maximum features, n-gram range (1,2), minimum and maximum document frequencies at 2 and 0.95, and sublinear TF scaling. Logistic Regression served as the classifier with regularization parameter C=1.0, liblinear solver, balanced class weights, and maximum iterations of 1,000.

2) Approach 1: BERTurk on Translated Data: We utilized BERTurk (dbmdz/bert-base-turkish-cased) [11], a 110 million parameter model with 12 transformer layers and 768-dimensional hidden states. The architecture processes text through the BERTurk encoder, extracting the [CLS] token embedding, applying dropout (0.3), and using a linear classification layer for binary prediction. Embeddings and first four encoder layers were frozen. Training used maximum sequence length 256, batch size 16, AdamW optimizer with learning rate 2e-5, and 3-5 epochs with early stopping.

3) Approach 2: XLM-RoBERTa Multilingual Transfer: XLM-RoBERTa-base [10], a 270 million parameter multilingual model pretrained on 100 languages, was implemented with identical architecture to BERTurk but using SentencePiece tokenization. We explored two strategies: (1) **Zero-Shot**: trained on English, evaluated on Turkish without Turkish training examples, and (2) **Few-Shot**: pretrained on English, fine-tuned on small Turkish subsets (50 samples) or mixed configuration (100 Turkish + 200 English samples). Zero-shot used learning rate 2e-5, batch size 32. Few-shot employed learning rate 2e-5, batch size 16, freezing eight encoder layers to prevent overfitting.

E. Evaluation Methodology

1) Performance Metrics: Models were assessed using: accuracy (overall correctness), precision (minimizing false alarms), recall (critical for security—detecting threats), F1-score (primary comparison metric balancing precision and recall), and AUC-ROC (discrimination ability across thresholds). Secondary analysis included confusion matrices and per-class metrics.

2) Implementation: Experiments were implemented using Python 3.8+, PyTorch, and Hugging Face Transformers [7]. Validation relied on single stratified train-validation-test split rather than cross-validation due to computational costs of transformer training.

3) Hyperparameter Selection: Hyperparameters were selected based on preliminary experiments and established best practices from literature. Learning rates of 2e-5, batch sizes 16-32, 3-5 epochs with early stopping, and dropout 0.3 were employed based on standard fine-tuning practices for BERT-family models [9].

IV. Experimental Results and Discussion

A. Overall Model Performance

Table I presents a comprehensive performance comparison of all evaluated approaches on the machine-translated Turkish test set. Among all models, BERTurk achieved the highest performance with 99.73% accuracy and an F1-score of 0.9976, demonstrating near-perfect classification. Its exceptional AUC-ROC value of 0.9999 indicates excellent discrimination across all decision thresholds, validating the effectiveness of training language-specific transformer models on translated data.

TABLE I
Performance Comparison on Turkish Test Set

Approach	Acc.	Prec.	Rec.	F1	AUC
BERTurk	99.73	99.76	99.76	99.76	99.99
TF-IDF Baseline	98.83	99.11	98.78	98.95	99.93
XLM-R Zero-Shot	95.59	97.36	94.66	95.99	99.47
TF-IDF Few-Shot	82.21	93.36	73.33	82.14	94.36
XLM-R Few-Shot	84.35	78.37	99.39	87.64	95.97
XLM-R Mixed	94.02	91.59	98.32	94.83	98.43

The TF-IDF Turkish baseline achieved competitive performance with 98.83% accuracy and an F1-score of 0.9895, demonstrating that traditional machine learning methods remain viable for phishing detection. Its high precision of 99.11% minimizes false alarms, a critical property for production environments where excessive false positives may erode user trust or overwhelm security teams.

XLM-RoBERTa's zero-shot configuration achieved 95.59% accuracy with an F1-score of 0.9599, which is remarkable given that the model was not exposed to any Turkish training examples. This result demonstrates strong cross-lingual transfer capability, where phishing-related patterns learned from English generalize effectively to Turkish. Nevertheless, the approximately four percentage point performance gap relative to BERTurk highlights the benefit of incorporating target-language data when available.

B. Few-Shot Learning Analysis

Few-shot experiments reveal clear differences in how traditional and transformer-based models respond to extreme data scarcity. With only 50 Turkish training samples, TF-IDF achieved 82.21% accuracy, while XLM-RoBERTa achieved 84.35%. Although overall accuracy remains modest, the models exhibit fundamentally different behaviors.

XLM-RoBERTa Few-Shot (50TR) achieved an exceptionally high recall of 99.39%, indicating that the model successfully identified nearly all phishing emails. However, this came at the cost of reduced precision (78.37%), reflecting a strong bias toward classifying ambiguous emails as phishing. This behavior aligns with security-oriented objectives where minimizing missed attacks is prioritized over reducing false alarms.

The mixed training strategy combining 100 Turkish and 200 English samples substantially improved performance. XLM-RoBERTa Mixed achieved 94.02% accuracy and an F1-score of 0.9483, representing a 9.67 percentage point improvement

over the 50-sample few-shot configuration. This result confirms that combining limited target-language data with source-language data allows multilingual models to better balance recall and precision.

C. Confusion Matrix Analysis

Figure 1 shows the confusion matrix for BERTurk. The model produces only 8 false positives and 8 false negatives out of 5,874 test samples, corresponding to a total error rate of just 0.27%. The nearly symmetric error distribution indicates well-calibrated predictions and explains the model's near-perfect precision and recall values.

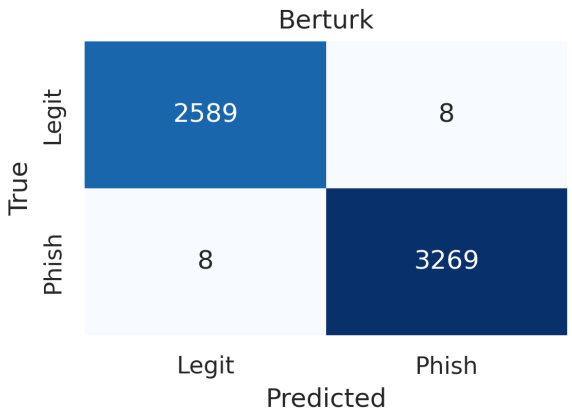

Fig. 1. Confusion Matrix for BERTurk

Figure 2 presents the TF-IDF Turkish baseline confusion matrix. The model generates 40 false positives and 29 false negatives, yielding an error rate of 1.17%. The slightly higher false positive count indicates a mild bias toward recall, which is acceptable in scenarios where flagged emails can be reviewed manually.

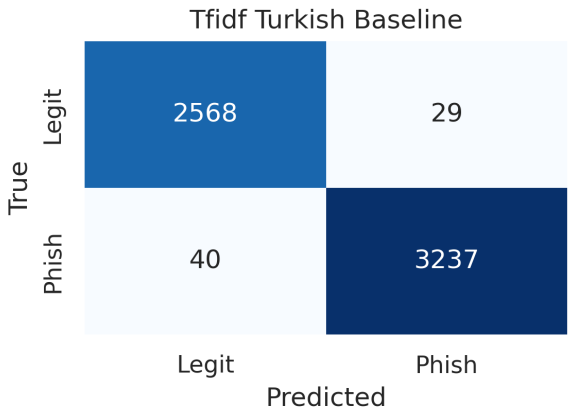

Fig. 2. Confusion Matrix for TF-IDF Turkish Baseline

Figure 3 illustrates the XLM-RoBERTa zero-shot configuration. The model produces 175 false positives and 84 false negatives, indicating a conservative classification tendency when encountering ambiguous Turkish text. This bias reflects the absence of Turkish-specific calibration and explains the observed reduction in precision relative to recall.

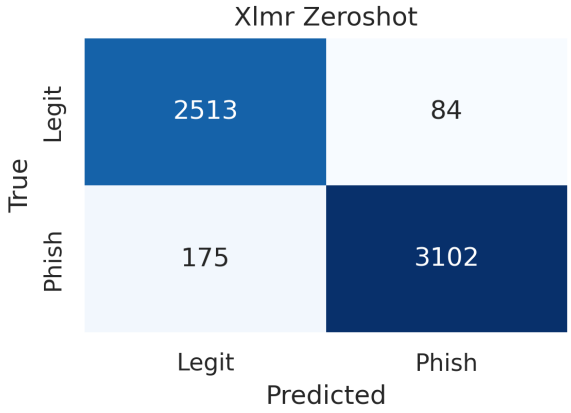

Fig. 3. Confusion Matrix for XLM-R Zero-Shot

Figure 4 highlights the TF-IDF Few-Shot (50TR) configuration. The model produces 874 false negatives, missing approximately 26.7% of phishing emails. This result demonstrates that traditional machine learning approaches are highly sensitive to data scarcity and struggle to generalize when trained on extremely small datasets, making this configuration unsuitable for real-world deployment.

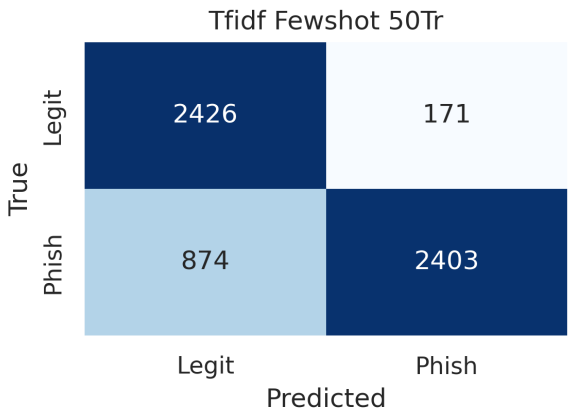

Fig. 4. Confusion Matrix for TF-IDF Few-Shot 50TR

Figure 5 shows the confusion matrix for XLM-RoBERTa Few-Shot (50TR). In contrast to TF-IDF, this model produces only 20 false negatives but generates 899 false positives. This aggressive classification strategy minimizes missed phishing attacks (false negatives below 1%) but introduces substantial operational overhead due to the high false positive rate.

Figure 6 presents the XLM-RoBERTa Mixed (100TR+200EN) configuration. The model achieves a more balanced error profile with 296 false positives and 55

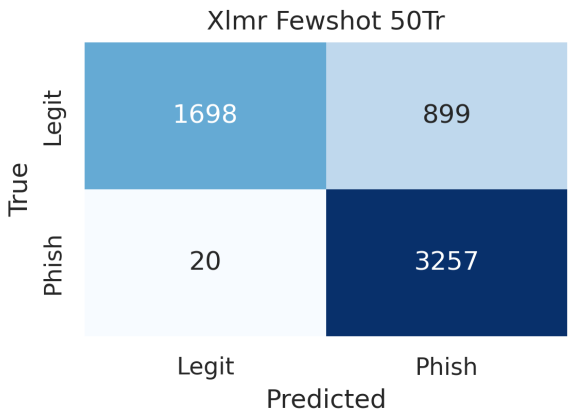

Fig. 5. Confusion Matrix for XLM-R Few-Shot 50TR

false negatives. The false positive rate drops to approximately 11.4%, while the false negative rate remains low at 1.7%, making this configuration a practical compromise between security and usability.

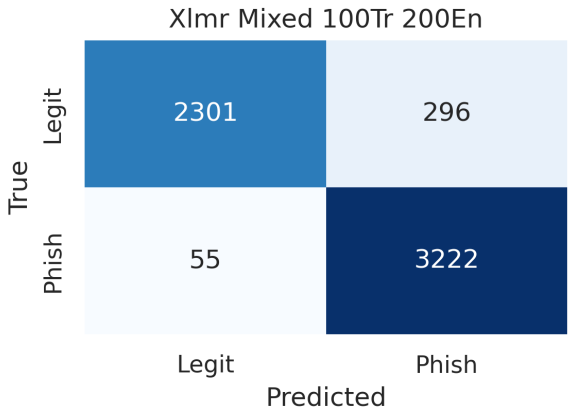

Fig. 6. Confusion Matrix for XLM-R Mixed 100TR+200EN

D. Precision–Recall Tradeoff Analysis

Figure 7 illustrates the precision–recall tradeoff across all approaches. BERTurk and the TF-IDF Turkish baseline occupy the upper-right region of the plot, achieving both high precision ($> 99\%$) and high recall ($> 98\%$), which represents the optimal operating region for phishing detection systems.

XLM-RoBERTa zero-shot maintains a strong balance between precision (97.36%) and recall (94.66%), confirming that zero-shot cross-lingual transfer is viable when no target-language data exists. The few-shot configurations reveal divergent strategies: TF-IDF Few-Shot prioritizes precision (93.36%) at the expense of recall (73.33%), while XLM-RoBERTa Few-Shot prioritizes recall (99.39%) at the expense of precision (78.37%).

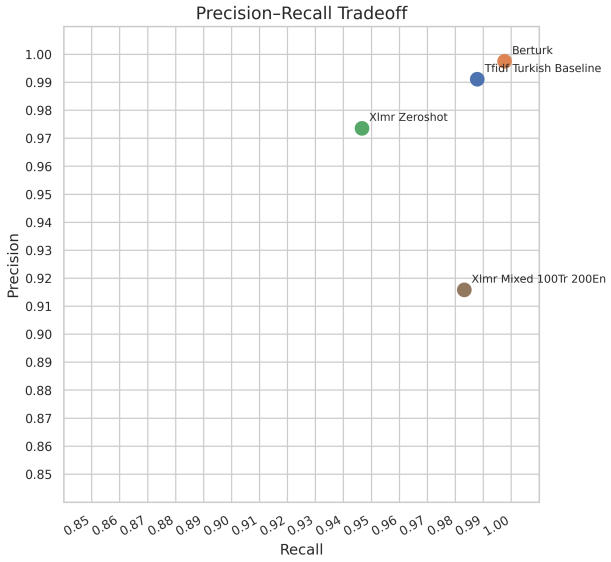

Fig. 7. Precision–Recall Tradeoff Across Approaches

E. Performance Across Metrics

Figure 8 compares accuracy, precision, recall, F1-score, and AUC-ROC across all approaches. BERTurk achieves the highest values across all metrics, establishing it as the strongest performer when translated training data and computational resources are available. The TF-IDF baseline closely follows, reinforcing the relevance of well-engineered traditional methods.

XLM-RoBERTa zero-shot maintains high AUC-ROC (99.47%) despite lower accuracy, indicating strong ranking capability even when classification thresholds are suboptimal. Few-shot models show greater variability, reflecting sensitivity to limited training data and threshold calibration.

Fig. 8. Comprehensive Metric Comparison Across Approaches

F. Discussion

The results demonstrate that phishing detection knowledge transfers effectively across languages due to the largely language-independent nature of phishing indicators, such as urgency cues, financial intent, impersonation patterns, and malicious URLs. BERTurk's superior performance is driven by Turkish-specific pretraining, exposure to the full translated

dataset, and sufficient model capacity to capture complex contextual dependencies.

The strong performance of the TF-IDF baseline suggests that machine-translated text may reduce linguistic variability, inadvertently favoring surface-level lexical models. Consequently, these results should be interpreted as upper bounds, with potential degradation expected on authentic Turkish phishing emails.

Overall, the figure-based analyses confirm that different approaches are optimal under different constraints: BERTurk for maximum accuracy, TF-IDF for efficiency and interpretability, and XLM-RoBERTa for rapid deployment in the absence of labeled target-language data.

G. Limitations

The primary limitation is evaluation on machine-translated rather than authentic Turkish phishing emails. Translation may standardize language, correct grammatical errors, and eliminate cultural nuances, potentially inflating performance estimates. Real-world deployment may reveal degraded performance, particularly for attacks exploiting Turkish-specific cultural knowledge or linguistic creativity.

Dataset limitations include temporal constraints (attacks evolve but evaluation uses static historical data), class distribution effects (44% legitimate, 56% phishing differs from typical email streams), and generalizability questions (results may not extend to other language families or morphologically different languages). Model limitations include BERTurk's computational requirements, transformer black-box nature hindering interpretability, and lack of adversarial robustness testing.

V. Conclusion and Future Work

A. Summary of Findings

This research demonstrated that effective Turkish phishing detection is achievable without native-language labeled datasets through multiple pathways. BERTurk achieved exceptional performance at 99.73% accuracy when trained on machine-translated Turkish data. XLM-RoBERTa zero-shot transfer achieved 95.59% accuracy without any Turkish training examples, proving cross-lingual knowledge transfer viability. Few-shot learning with 50 Turkish samples reached 84.35% accuracy, while mixed approach (100 Turkish + 200 English) achieved 94.02% accuracy, approaching full-data performance with less than 1% of training data. Traditional TF-IDF baseline achieved competitive 98.83% accuracy, demonstrating well-engineered traditional approaches remain relevant.

B. Contributions

This work provides: (1) first systematic comparison of translation-based versus multilingual transfer approaches for Turkish phishing detection, (2) replicable methodology for developing detection systems when labeled data is unavailable, (3) empirical evidence that organizations can deploy effective Turkish phishing detection without expensive annotation efforts, and (4) demonstration that zero-shot enables immediate deployment while few-shot provides improved performance with minimal labeling.

C. Future Work

Critical future directions include: (1) collecting and annotating authentic Turkish phishing emails for validation, (2) assessing advanced translation systems and multimodal approaches analyzing visual elements, (3) developing continual learning frameworks to address concept drift, (4) extending analysis to other low-resource languages, (5) enhancing model trustworthiness through explainability methods, and (6) investigating resource-efficient architectures and adversarial robustness.

The success of cross-lingual transfer provides pathway to democratize security technologies globally, extending protection to billions of speakers of low-resource languages. However, the cybersecurity community must prioritize creating diverse, multilingual, publicly available datasets for rigorous real-world validation before high-stakes deployment.

References

[1] M. Junczys-Dowmunt et al., "Marian: Fast neural machine translation in C++," in *Proc. ACL 2018, System Demonstrations*, 2018, pp. 116–121.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[3] A. Conneau et al., "Unsupervised cross-lingual representation learning at scale," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2020, pp. 8440–8451.

[4] A. Yasin and A. Abuhasan, "An intelligent classification model for phishing email detection," *Int. J. Network Security Its Appl. (IJNSA)*, vol. 8, no. 4, pp. 55–72, Jul. 2016.

[5] S. Salloum, T. Gaber, S. Vadera, and K. Shaalan, "Phishing email detection using natural language processing techniques: A literature survey," *Procedia Comput. Sci.*, vol. 189, pp. 19–28, 2021.

[6] S. Schweter, "BERTurk - BERT models for Turkish," Zenodo, 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3770924.

[7] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in *Proc. EMNLP: System Demonstrations*, 2020, pp. 38–45.

[8] S. Sheng, B. Wardman, G. Warner, L. Cranor, J. Hong, and C. Zhang, "An empirical analysis of phishing blacklists," in *Proc. 6th Conf. Email Anti-Spam (CEAS)*, 2009, pp. 1–10.

[9] M. Khonji, Y. Iraqi, and A. Jones, "Phishing detection: A literature survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 4, pp. 2091–2121, 4th Quart., 2013.

[10] P. H. Kyaw, J. Gutierrez, and A. Ghobakhlou, "A systematic review of deep learning techniques for phishing email detection," *Electronics*, vol. 13, no. 19, p. 3823, Sep. 2024.

[11] K. Thakur, M. L. Ali, M. A. Obaidat, and A. Kamruzzaman, "A systematic review on deep-learning-based phishing email detection," *Electronics*, vol. 12, no. 21, p. 4545, Nov. 2023.

[12] M. Songailaitė, E. Kankevičiūtė, B. Zhyhun, and J. Mandravickaitė, "BERT-based models for phishing detection," in *Proc. 28th Conf. Inf. Soc. Univ. Studies (IVUS'2023)*, Kaunas, Lithuania, May 2023, pp. 1–8.

[13] M. Gül Taş, "Semantic detection of Turkish phishing emails: A natural language processing and deep learning-based approach," *Düzce Üniv. J. Cybersecurity Digit. Econ.*, vol. 1, no. 1, pp. 29–42, 2025.