Exploring Cross-Lingual Transfer in Transformer-Based Fraud Detection Models

Iván Martínez Murillo and **Robiert Sepúlveda-Torres** and **Juan Pablo Consuegra-Ayala**
Language Processing and Information Systems Group, University of Alicante, Spain
{ivan.martinezmurillo, robiert.sepulveda, juan.consuegra}@ua.es

Abstract

Digital communication channels have become major vectors for large-scale cyber fraud, including spam, phishing, and recruitment scams, causing significant financial losses and eroding user trust. While transformer-based models have improved text classification, most systems remain English-centric, limiting their effectiveness in multilingual environments and across diverse fraud typologies. In this work, we hypothesize that fraud detection models trained in a single language can exhibit cross-lingual generalization capabilities. We investigate the use of Natural Language Processing (NLP) for detecting multiple forms of textual fraud by developing and evaluating three transformer-based discriminative models targeting spam, phishing, and fake job postings. Specifically, we compare two architectures—MrBERT (308M parameters) and mROBERTa (283M parameters)—across English, Spanish, and Valencian data. Our results demonstrate that while models trained exclusively on English achieve near-perfect performance in-language ($F1 \approx 0.98$), they exhibit limited cross-lingual generalization in other languages such as Spanish ($F1 \approx 0.44$ for phishing detection, and $F1 \approx 0.85$ for spam detection). However, incorporating multilingual training data dramatically improves performance in target languages ($F1 \approx 0.95$–0.99) while maintaining or even enhancing English accuracy. We further observe that cross-lingual transfer is most effective when datasets are parallel or closely aligned. These findings underscore the critical importance of multilingual data inclusion for building robust, scalable cybersecurity frameworks in diverse linguistic settings.

1 Introduction

Digital communication channels such as Short Message Service (SMS), email, and advertisements platforms have become major vectors for large-scale cyber fraud campaigns (Halim et al., 2025). The widespread adoption of online services and mobile devices has significantly expanded the attack surface for malicious actors, enabling scalable spam campaigns, phishing attacks, and recruitment scams (Osamor et al., 2025). These threats cause substantial financial losses and erode user trust in digital communication systems (Akbar et al., 2025). In particular, fraudulent job postings have emerged as a growing form of online deception, exploiting the increasing reliance on digital job platforms to obtain sensitive information or extract fraudulent payments from job seekers (Kumar et al., 2025).

While often conflated in non-technical contexts, spam, phishing, and fake job postings represent distinct forms of cyber fraud. Spam typically consists of unsolicited bulk messages intended for advertising or traffic generation, whereas phishing is a targeted social engineering attack designed to extract sensitive information such as login credentials or financial data (Saka et al., 2025). Recruitment scams, in contrast, involve impersonation of organizations or recruiters to deceive job seekers (B P et al., 2025). Despite sharing superficial linguistic similarities, these fraud types differ in intent, structure, and risk profile, making their automatic detection a challenging NLP task.

Another important challenge is linguistic diversity in digital communications. Most existing detection systems are trained primarily on English-language datasets, leaving fewer resources for other widely used languages such as Spanish (Herrera-Semenets et al., 2024). This imbalance limits the effectiveness of fraud detection systems in multilingual environments. Furthermore, many existing approaches focus on a single fraud type—typically spam—without addressing other related threats such as phishing or recruitment scams, which often exhibit more subtle and manipulative language patterns (Opara et al., 2025). These limitations highlight the need for research addressing multiple fraud typologies and multilingual communication settings.

Proceedings of the Second International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security (NLPAICS'2026)
11–12 June, 2026.

Recent advances in Natural Language Processing (NLP), particularly transformer-based language models have substantially improved text classification performance across a wide range of NLP tasks (Sepúlveda-Torres et al., 2025). Pre-trained models such as RoBERTa and its multilingual variants capture contextual semantic representations that can be effectively adapted to domain-specific classification problems with limited fine-tuning data (Babu et al., 2025). Their ability to model nuanced linguistic cues makes them well suited for detecting deceptive language commonly found in cyber fraud messages.

The objectives of this paper are as follows. We hypothesize that training fraud detection models in one language may enable them to generalize and learn the underlying fraud detection tasks in other languages, thereby highlighting the potential of cross-lingual knowledge transfer in multilingual NLP applications. In this paper, we present a study that investigates the use of NLP techniques for detecting multiple forms of textual cyber fraud. We develop three transformer-based discriminative models targeting different fraud scenarios: spam detection in SMS and email messages, phishing detection in textual communications, and fake job posting detection in online recruitment content. Together, these models provide a broader perspective on how modern NLP architectures can be applied to diverse fraud detection tasks across communication channels and languages.

The main contributions of this paper are as follows:

- We investigate multilingual fraud detection by incorporating English, Valencian, and Spanish communication data, and study how training in one language can enable the models to generalize to other languages.
- We develop and evaluate three transformer-based discriminative models for spam detection, phishing detection, and fake job posting detection.
- We provide an empirical evaluation across fraud types and communication channels, highlighting differences in linguistic characteristics and model performance.
- We present the ALIA cybersecurity framework for studying NLP-based detection of multiple textual cyber fraud types.

2 Related Work

This section reviews prior literature on textual fraud detection, from traditional machine learning approaches based on handcrafted features to recent transformer-based methods. We focus on evidence from spam, phishing, and fake job posting detection, with special attention to multilingual and cross-lingual settings that motivate the methodology adopted in this work.

2.1 Traditional Fraud Detection

Early fraud detection systems relied on manually engineered features combined with classical machine learning classifiers (Al Saidat et al., 2024). For email spam filtering, models such as Naïve Bayes and logistic regression leveraged lexical features (e.g., n-grams and TF–IDF), stylometric cues, and handcrafted metadata, while Support Vector Machines and tree-based models captured more complex decision boundaries (Thakur et al., 2023; Salloum et al., 2022).

Phishing detection extended these approaches with domain-specific signals, including URL structure, domain age, and HTML/JavaScript characteristics (Thakur et al., 2023).

Regarding fake-jobs posting detection, early systems relied on TF-IDF and BoW techniques (Keerthana et al., 2021) along with classifiers such as Random Forest, Support Vector Machines, Logistic Regression, and Multi-Layer Perceptrons (Habiba et al., 2021), achieving accuracies ranging from 71% to 90% (Srikanth et al., 2022). DL-based methods using word embeddings (e.g., GloVe) and sequential neural networks further improved performance, reaching up to 97% accuracy (Ranparia et al., 2020).

Although effective in controlled settings, these methods required substantial feature engineering and struggled to generalize across languages and evolving attack strategies, limiting scalability in multilingual and dynamic environments (Utturkar et al., 2026). These limitations motivated the shift toward representation learning and neural approaches that reduce reliance on manual feature design and improve cross-domain transfer (Utturkar et al., 2026).

2.2 Deep Learning Approaches

The advent of Transformer-based Pre-trained Language Models (PLMs), such as MrBERT (Tamayo et al., 2026) and RoBERTa (Liu et al., 2019),

has catalyzed a paradigm shift in fraud detection. By leveraging self-attention mechanisms to encode nuanced semantic and syntactic dependencies, these models produce high-dimensional contextual representations that outperform traditional n-gram baselines. Standard practises typically fine-tune these architectures on domain-specific datasets, yielding state-of-the-art results in tasks such as phishing detection and fake job identification (Taneja et al., 2025). Furthermore, recent research has moved toward hybrid frameworks; by fusing PLM embeddings with graph neural networks (GNNs) or metadata-derived features, these systems achieve greater robustness against adversarial attacks and sophisticated evasion strategies (Uddin et al., 2026).

However, a critical limitation persists: the prevailing literature remains largely monolingual, with an overwhelming bias toward English-centric benchmarks. This imbalance obscures the impact of linguistic typology on fraudulent discourse and model generalizability. For instance, the lexical and morphological markers that signal deceptive intent in one language may not translate across different language families, creating a leakage that adversaries can exploit via language-specific nuances (Utturkar et al., 2026). Despite the global nature of cybercrime, the interplay between linguistic diversity and fraud remains significantly underinvestigated, particularly regarding the efficacy of cross-lingual transfer in adversarial contexts.

3 Method

This section describes our end-to-end methodology for developing the ALIA fraud detection framework, including how datasets are selected, how multilingual resources were prepared, and how transformer models were configured and trained for the three fraud detection tasks.

3.1 Data Selection Rationale

Our dataset construction followed a license-aware and task-driven methodology designed to support reproducible multilingual fraud detection. We first performed a systematic search on open data repositories, primarily Hugging Face[1] and Kaggle[2], and complemented this search with datasets referenced in recent scientific papers on spam, phishing, and fake job posting detection. Candidate corpora were screened according to four criteria: (i) relevance to the target fraud typology, (ii) availability of raw textual content with reliable labels, (iii) sufficient sample size and class balance for robust fine-tuning, and (iv) legal compatibility with research use and redistribution.

To ensure legal and ethical reuse, we prioritized datasets distributed under permissive licenses such as Creative Commons Zero (CC0), Creative Commons Attribution (CC-BY), and Apache. We also accepted datasets released under Massachusetts Institute of Technology (MIT) when the license explicitly covered the distributed resource and allowed our intended academic use. Although MIT is primarily used for software rather than data, it is still permissive and can be valid for dataset releases when clearly specified by the provider. Datasets with restrictive, unclear, or non-redistributable terms were excluded.

After identifying eligible resources, we harmonized schema and label spaces across corpora (e.g., phishing vs. non-phishing, spam vs. ham) and retained only the textual fields needed for model training. For the Valencian setting, where publicly available labeled corpora are limited, selected source data were translated to Valencian and then reviewed by a language expert to preserve semantic intent, fraud-related cues, and annotation consistency. This process enabled the creation of parallel or aligned multilingual subsets and supported controlled experiments on cross-lingual transfer.

3.2 Base Architecture

To evaluate the impact of the underlying multilingual encoder on fraud detection performance, we experimented with two transformer-based architectures: MrBERT (Tamayo et al., 2026) and mROBERTa based on the RoBERTa architecture (Liu et al., 2019).

MrBERT is a $308M$-parameter multilingual encoder based on the ModernBERT architecture. It is pretrained with a masked language modeling objective on 6.1 trillion tokens covering 35 European languages and code, and supports contexts up to $8,192$ tokens.

mROBERTa is a $283M$-parameter multilingual encoder based on the RoBERTa architecture. It is pretrained from scratch with a masked language modeling objective on $12.8TB$ of data covering 35 European languages and 92 programming languages, using a $256K$ SentencePiece vocabulary

[1] https://huggingface.co/
[2] https://www.kaggle.com/

and a maximum context length of 512 tokens.

Both architectures follow the transformer encoder design composed of stacked self-attention layers and position-wise feed-forward networks. For the downstream binary classification tasks (spam detection and phishing detection), we appended a task-specific linear classification head on top of the contextualized representation. The models were fine-tuned end-to-end using supervised learning.

The motivation for selecting these two encoders is twofold: (1) their strong empirical performance in multilingual NLP tasks and (2) their ability to capture contextual semantic patterns critical for detecting deceptive language. By comparing MrBERT and mROBERTa under identical fine-tuning conditions, we assess the influence of pre-training strategy and architectural optimization on fraud detection performance.

3.3 Training Architecture Setup

To ensure a consistent and reproducible experimental pipeline across tasks and languages, we developed an internal framework on top of the Hugging Face `Trainer` API. The framework standardizes data loading, tokenization, model instantiation, training, validation, and test-time evaluation for all experiments, while keeping configuration modular across fraud typologies and language settings.

In practice, this implementation allows us to run monolingual and multilingual experiments under a unified interface, controlling hyperparameters, random seeds, and split definitions in a comparable way across MrBERT and mROBERTa. The codebase used in this work is publicly available at https://github.com/gplsi/transformer_model_classifier.

3.4 Training Setup

The summary of training setup configuration can be seen in Table 1. We trained all models on the official training split described in Section 4.1 and evaluated them on the corresponding test split. From the training data, we further held out 10% as a validation set for model selection and early evaluation.

Input texts were tokenized with a maximum sequence length of 512 tokens. The prediction target was the `label` column. No additional feature columns were used, and we did not apply class weighting.

We optimized model parameters using AdamW with a learning rate of 2×10^{-5}. Training was conducted for 2 epochs with a per-device batch size of 8 for both training and evaluation.

Hyperparameter	Value
Training split	Official training split
Validation split	10% of training data
Test split	Official test split
Maximum sequence length	512
Prediction target	`label` column
Additional features	None
Class weighting	Not applied
Optimizer	AdamW
Learning rate	2×10^{-5}
Number of epochs	2
Training batch size	8
Evaluation batch size	8
Mixed precision training	FP16 enabled
Random seed	852

Table 1: Training hyperparameters and experimental settings.

To improve computational efficiency, we enabled mixed-precision (FP16) training. All experiments were run with a 852 fixed random seed to ensure reproducibility.

Experiments were conducted on a single NVIDIA A100 40GB GPU within an NVIDIA DGX A100 system.

4 Experimental Setup

This section describes the experimental setup designed to test our methodology in a controlled and comparable way across tasks and languages. It defines the selected datasets, language-specific splits, and training/evaluation conditions used to assess the robustness of our multilingual fraud detection approach.

4.1 Data

Following our methodology for constructing reliable datasets, we compiled the resources detailed in Table 2. It is important to note, that although many of the datasets are balanced across labels, some of them are not, containing a higher proportion of non-fraud instances than fraud ones. This distribution reflects real-world conditions, where legitimate messages significantly outnumber harmful or fraudulent ones.

Phishing datasets: We used four publicly available datasets for phishing detection:

- **SMS Phishing Dataset for Machine Learning and Pattern Recognition** (Mishra and Soni, 2023): This dataset contains $5,971$ English SMS messages labeled as ham $(4,844)$

Dataset	Task	Language	Ham	Fraud	Total
SMS Phishing Dataset (Mishra and Soni, 2023)	Phishing	English	4,777	1,194	5,971
Phishing Email Dataset (Alam, 2024)	Phishing	English	39,595	42,891	82,500
SpearPhishMX (Diaz et al., 2026)	Phishing	Spanish	1,115	1,891	3,006
SpaPhish (Bustio-Martinez et al., 2026)	Phishing	Spanish	664	731	1,395
Spamspa (Zuñiga, 2024)	Spam	Spanish	4,825	747	5,572
(softecapps, 2024)	Spam	Spanish	586	621	1,207
SMS Spam Collection Dataset (Almeida et al., 2011)	Spam	English	4,827	747	5,574
Real or Fake: Fake Job Description Prediction (Bansal)	Fake Job Detection	English / Valencian	865	865	1,730

Table 2: Summary of datasets used in the experiments.

or phishing ($1,127$). Each instance includes the raw message text and a set of engineered features derived primarily from malicious samples. For this study, we only used the raw text and the labels to train and evaluate text-based models.

- **Phishing Email Dataset** (Alam, 2024): This corpus consists of approximately $82,500$ English emails collected from multiple sources, including Enron, Ling, CEAS, Nazario, Nigerian Fraud, and SpamAssassin. It contains $42,891$ phishing emails and $39,595$ legitimate emails. Each email includes the subject, body text, and contextual metadata (e.g., sender, recipient, date); for our experiments, we only used the text and labels.

- **SpearPhishMX** (Diaz et al., 2026): SpearPhishMX is a Spanish spear-phishing email dataset containing $3,006$ messages labeled for binary classification ($1,891$ phishing mails and $1,115$ ham). It includes subject and a research-ready body representation, along with derived attributes for personalization and URL analysis, and provides a sanitized public version that anonymizes sensitive information and defangs links to support responsible data sharing. For the end of this study, only the text and label was used.

- **SpaPhish** (Bustio-Martinez et al., 2026): SpaPhish is a Spanish-native phishing email dataset containing $1,395$ messages. It enables research beyond binary phishing detection by providing authentic Spanish emails scraped between 2014 and 2025. The dataset is balanced, including 664 ham emails and 731 phishing emails.

For each dataset, we applied an 80/20 train–test split. We conducted two training setups: (i) a monolingual setting using only the English training data, and (ii) a multilingual setting in which the English and Spanish training partitions were combined to form a unified training dataset. The Spanish training data was only used within this multilingual configuration and was not used independently. Evaluation was performed on four test sets: (i) English-only test instances, (ii) Spanish-only test instances, (iii) Spanish-only instances from the training distribution, and (iv) a combined multilingual test set.

Spam datasets: We conducted experiments using three publicly available SMS spam datasets in Spanish and English:

- **Spamspa** (Zuñiga, 2024): This dataset contains $5,572$ Spanish SMS messages annotated for spam detection with binary labels (spam vs. non-spam). Each instance consists of the raw message text and its corresponding label. The dataset is designed to support the development and evaluation of text classification models for detecting unwanted messages in SMS and related communication platforms.

- **Spam_ham_spanish** (softecapps, 2024): This dataset comprises $1,207$ Spanish-language SMS messages labeled as spam or ham. Each entry includes the raw message and its associated label, enabling supervised training and evaluation of automatic spam detection systems.

- **SMS Spam Collection Dataset** (Almeida et al., 2011): This dataset contains $5,574$ English SMS messages labeled as ham ($4,827$) or spam (747). Each message includes a label and raw text, organized with one message per

line. The corpus was compiled from multiple publicly available sources, and is suitable for training and evaluating machine learning models for spam detection in SMS communications.

Because the spam datasets contain fewer instances than the phishing datasets in our study, we applied a 90/10 train–test split. We conducted two training configurations: (i) a monolingual setup using only the English training data, and (ii) a multilingual setup in which the English and Spanish training partitions were combined into a single training dataset. As in the phishing experiments, the Spanish training data was only used as part of this multilingual configuration and not as a standalone training set. For evaluation, we constructed four distinct test sets: (i) English-only test instances, (ii) Spanish-only test instances, (iii) Spanish-only instances from the training distribution, and (iv) a combined multilingual test set.

Fake-jobs posting datasets: For studying fake job posting detection, we used the [Real or Fake]: Fake Job Description Prediction dataset (Bansal). This dataset contains approximately $18,000$ job descriptions in English, of which around 800 are fraudulent. It includes both textual information and additional metadata about the job postings. The dataset can be used to develop classification models that distinguish fraudulent job listings from legitimate ones. Since the dataset is highly imbalanced, we created a balanced subset. Specifically, we included all fraudulent items (865) and randomly selected an equal number of legitimate items (865), resulting in a dataset of $1,730$ job postings.

To study the impact of training on a widely represented language (English) on generalizability to other languages, we asked a professional native Valencian translator to translate the subset of $1,730$ job postings instances into Valencian, obtaining a parallel dataset in English and Valencian. The translations were performed by a native speaker to ensure high linguistic fidelity and naturalness in the target language, while carefully preserving the original meaning and semantic content of the English source texts.

Finally, we split both datasets into 85% training and 15% testing, ensuring parallel splits across languages, such that each English training instance had a corresponding Valencian counterpart. We conducted two training configurations: (i) a monolingual setup using only the English training data, and (ii) a multilingual setup in which the English and Valencian training partitions were combined into a single training dataset. The Valencian training data was only used within this multilingual configuration and was not used independently as a separate training set. We evaluated the models under four setups: (i) English-only test set, (ii) Valencian-only test set, (iii) Valencian instances from the training distribution, and (iv) combined multilingual test set.

4.2 Metrics

We evaluate model performance using the macro-averaged F1-score, which is widely adopted in NLP benchmarks. The macro-F1 computes the unweighted mean of the F1-score across all classes, treating each class equally regardless of its support. This is particularly suitable for our binary task, where the dataset exhibits moderate class imbalance (Section 4.1), as it ensures that the evaluation reflects performance on both the minority and majority classes rather than being dominated by the majority class.

5 Results

This section reports the experimental outcomes for the three fraud detection tasks and compares monolingual and multilingual training settings to evaluate cross-lingual generalization.

5.1 Phishing Detection Results

Table 3 presents phishing detection performance for two transformer-based models, MrBERT and mROBERTa, under two training regimes: (i) trained only on English data, and (ii) trained on both English and Spanish data.

Model trained only on English				
Model	EN_test	ES_test	ES_train	Test full
MrBERT	0.982	0.445	0.444	0.875
mROBERTa	0.983	0.440	0.426	0.878
Model trained on English and Spanish				
Model	EN_test	ES_test	ES_train	Test full
MrBERT	0.983	0.950	0.986	0.975
mROBERTa	0.975	0.955	0.982	0.972

Table 3: Phishing prediction results

When trained exclusively on English, both models achieve near-perfect performance on the English test set (EN_test $\approx$ 0.98), but their performance on Spanish examples (ES_test) is poor

($\approx 0.44 - 0.45$), indicating a lack of cross-lingual generalization. Including Spanish data during training dramatically improves performance on Spanish examples (ES_test ≈ 0.95), while maintaining strong English performance, demonstrating that multilingual training enables effective cross-lingual phishing detection. Overall, MrBERT and mROBERTa show comparable trends, with minor differences in performance across metrics, suggesting that model choice has less impact than the language coverage of the training data. One aspect to remark, is that while for MrBERT, training on both languages enhances the performance in the English data, in the case of mROBERTa, the model trained on both languages obtains worse results in English than the model trained only on English. However, in both cases, the difference is very small.

5.2 Spam Detection Results

Table 4 summarizes spam detection performance for MrBERT and mROBERTa under monolingual (English-only) and multilingual (English+Spanish) training.

Model trained only on English				
Model	EN_test	ES_test	ES_train	Test full
MrBERT	0.994	0.851	0.837	0.921
mROBERTa	0.991	0.846	0.832	0.917
Model trained on English and Spanish				
Model	EN_test	ES_test	ES_train	Test full
MrBERT	0.997	0.987	0.995	0.992
mROBERTa	0.994	0.974	0.988	0.983

Table 4: Spam prediction results

When trained solely on English, both models achieve near-perfect performance on the English test set (EN_test ≈ 0.99), while performance on Spanish test data (ES_test ≈ 0.85) is relatively high, indicating certain cross-lingual transfer. Adding Spanish training data substantially improves Spanish test performance (ES_test $\approx 0.97 - 0.99$) and slightly enhances English performance, highlighting the benefit of multilingual training for robust spam detection. Across both training regimes, MrBERT consistently achieves slightly higher scores than mROBERTa, though differences are minor.

5.3 Fake Jobs Posting Results

Table 5 reports model performance on detecting fake job postings in English (EN) and Valencian (VA). When trained exclusively on English, both MrBERT and mROBERTa achieve strong English test performance (EN_test $\approx 0.92 - 0.93$). Interestingly, for this task, where the English and Valencian data are parallel, mROBERTa exhibits strong zero-shot performance on the Valencian test set (VA_test $\approx 0.89 - 0.90$), demonstrating effective cross-lingual generalization. MrBERT, while acceptable (VA_test $\approx 0.75 - 0.79$), performs worse on Valencian, indicating more limited cross-lingual transfer.

Model trained only on English				
Model	EN_test	VA_test	VA_train	Test full
MrBERT	0.922	0.747	0.791	0.838
mROBERTa	0.926	0.899	0.902	0.913
Model trained on English and Valencian				
Model	EN_test	VA_test	VA_train	Test full
MrBERT	0.930	0.934	0.962	0.931
mROBERTa	0.919	0.923	0.953	0.921

Table 5: Fake-jobs prediction results

Including Valencian in the training set further improves VA_test scores ($\approx 0.92 - 0.93$) while maintaining robust English performance, showing that multilingual training is effective for low-resource languages. In this setting, MrBERT slightly outperforms mROBERTa, suggesting that model architecture can interact more efficiently with the distribution of training data across languages. Overall, these results highlight the importance of multilingual training for reliable detection in underrepresented languages.

6 Discussion

Our experiments demonstrate the critical role of multilingual training in enhancing the performance of transformer-based models across different fraud detection tasks. In the phishing and spam detection tasks, models trained exclusively on English achieve near-perfect performance on English test sets, but their performance on Spanish test sets is substantially lower. This highlights the limited cross-lingual generalization of monolingual models, particularly when the target language is underrepresented or differs structurally from English. Including Spanish data in training substantially improves performance on Spanish examples while enhancing also the English accuracy, confirming that multilingual exposure enables robust cross-lingual

transfer. MrBERT and mROBERTa show comparable trends across these tasks, with only minor downgrade of performance of mROBERTa in the multilingual setup, when predicting English texts.

The fake jobs posting task introduces an interesting contrast. Here, the English and Valencian data are parallel, and mROBERTa exhibits surprisingly strong cross-lingual performance even without training on Valencian (although it saw the same instances in English during training). MrBERT, in contrast, performs worse under the same conditions, illustrating that model-specific factors can interact with cross-lingual transfer when data are aligned across languages. As with the other tasks, adding Valencian data further boosts performance while preserving strong English performance, and even enhancing it, underscoring the importance of multilingual training for low-resource languages.

Across all three tasks, several patterns emerge. First, multilingual training consistently improves performance on target languages while maintaining English performance, validating the effectiveness of training on multiple languages. Second, cross-lingual generalization is stronger when data across languages are parallel or closely aligned, as seen in the fake jobs task. Third, while MrBERT and mROBERTa generally perform similarly, architecture can slightly influence in certain settings. Overall, these results highlight that careful multilingual data inclusion is essential for building robust models capable of handling multiple languages and low-resource scenarios.

7 Conclusions & Future Work

In this work, we hypothesize that fraud detection models trained in a single language can exhibit cross-lingual generalization capabilities. To this end, we investigate the performance of transformer-based models, MrBERT and mROBERTa, on phishing, spam, and fake job posting detection across English, Spanish, and Valencian. Our results show that models trained exclusively on English achieve high performance on English test sets but exhibit limited cross-lingual generalization, particularly for underrepresented languages. Incorporating additional language data significantly improves performance on non-English test sets while maintaining strong English accuracy, demonstrating that multilingual training is crucial for robust detection across languages. Moreover, our experiments highlight that cross-lingual transfer is more effective when the datasets are parallel or closely aligned, as observed in the fake jobs task.

The main outcomes of this work are summarised as follow:

- We analyzed cross-lingual transfer behavior and showed its dependence on language alignment and dataset parallelism.
- We demonstrated that multilingual training data substantially improves lower-resourced language performance while preserving strong English results.
- We provide a unified empirical comparison of two multilingual transformer encoders (MrBERT and mROBERTa) under consistent training conditions.
- We introduce the ALIA multilingual fraud detection framework covering phishing, spam, and fake job posting scenarios.

For future work, several directions are promising. First, extending our study to additional low-resource languages would provide insights into the scalability of these approaches. Second, evaluating robustness to adversarially crafted phishing, spam, or fake job content could help develop models resilient to real-world attacks. Finally, investigating data augmentation techniques or synthetic multilingual corpora may further enhance cross-lingual generalization without requiring extensive annotated datasets in each target language.

Acknowledgements

This research is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA (Resol. SEDIA 19.08.2024). Moreover, this work is also partiall funded by the project "SAFEWORDS: Language Anonymization with Ethical and Legal Safeguards through NLP" (AIA2025-163322-C63) funded by MICIU/AEI/10.13039/501100011033. Furthermore, this work is also partiall funded by the R&D project "QUMLAUDE: Mecánica cuántica para comprensión y generación del lenguaje" (PID2024-160791OB-I00), funded by MCIN/AEI/10.13039/501100011033/ and by "ERDF A way of making Europe, and by project HEART-NLP-UA (PID2024-156263OB-C22).

References

Faisal Akbar, Junaid Hussain, Muhammad Babar Usman, and Jamil Afzal. 2025. The impact of financial scams on consumer trust in the banking sector: A qualitative analysis. *International Journal of Discovery in Social Sciences*, 1(1).

Mohammed Rasol Al Saidat, Suleiman Y Yerima, and Khaled Shaalan. 2024. Advancements of sms spam detection: A comprehensive survey of nlp and ml techniques. *Procedia Computer Science*, 244:248–259.

Naser Abdullah Alam. 2024. Phishing email dataset.

Tiago A. Almeida, José María G. Hidalgo, and Akebo Yamakami. 2011. Contributions to the study of sms spam filtering: new collection and results. In *Proceedings of the 11th ACM Symposium on Document Engineering*, DocEng '11, page 259–262, New York, NY, USA. Association for Computing Machinery.

Pradeep Kumar B P, Bhagya D J, Gagana G, and Mani N. 2025. Fake job post detection using machine learning. In *2025 International Conference on Computing for Sustainability and Intelligent Future (COMP-SIF)*, pages 1–6.

B Eswar Babu, Sathya Prakash Racharla, Pasupuleti Pavani, Yadaiah Balagoni, Mohan Ajmeera, et al. 2025. A performance evaluation of transformer models and recurrent neural networks models in efficient text classification tasks. In *2025 8th International Conference on Computing Methodologies and Communication (ICCMC)*, pages 1171–1177. IEEE.

Shivam Bansal. Real / Fake Job Posting Prediction — kaggle.com. https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction/data. [Accessed 16-03-2026].

Lazaro Bustio-Martinez, Viviana Fuentes-Fuentes, Luisa Fernanda Agudelo-Fuentes, Vitali Herrera-Semenets, Darian Llanes-Guilarte, Felipe Antonio Trujillo-Fernández, Carlos Antonio Cardeña-Matamoros, Carlos Francisco Betancourt-Moreno, Joshua Ismael Haase-Hernández, Andrés Guillermo Molano-Jiménez, and Jan van den Berg. 2026. SpaPhish: A spanish dataset for phishing and psychological pattern detection.

Juan Manuel Diaz, Alfonso Martínez Cruz, and Lazaro Bustio Martinez. 2026. Spear-Phishing en español: dataset de correos dirigidos con señales de personalización (SpearPhishMX).

Sultana Umme Habiba, Md Khairul Islam, and Farzana Tasnim. 2021. A comparative study on fake job post prediction using different data mining techniques. In *2021 2nd international conference on robotics, electrical and signal processing techniques (ICREST)*, pages 543–546. IEEE.

Mehraj Ibne Halim, Mohammad Zibran Hasan, Md Humayun Kabir, Mohammad Nadib Hasan, Hassan Jaki, Ahmad, and Hasibul Hasan. 2025. Enhancing phishing detection: A machine learning approach to predicting malicious emails, urls, and sms messages. *Applied Computational Intelligence and Soft Computing*, 2025(1):6633979.

Vitali Herrera-Semenets, Lázaro Bustio-Martínez, Yamel Pérez-Guadarramas, Jorge Ángel González-Ordiano, and Jan van den Berg. 2024. Unmasking phishing attempts: A study on detection in spanish emails. In *Iberoamerican Congress on Pattern Recognition*, pages 1–15. Springer.

Bodduru Keerthana, Anumala Reethika Reddy, and Avantika Tiwari. 2021. Accurate prediction of fake job offers using machine learning. In *Machine Intelligence and Soft Computing: Proceedings of ICMISC 2020*, pages 101–112. Springer.

Shashikant Kumar, Sadiya Yasmeen, Prabhat Kumar, et al. 2025. The growing threat of fake job postings: A review of machine learning and nlp-based detection approaches. *Revolutionary Advances in Computing and Electronics: An International Journal*, pages 41–51.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Sandhya Mishra and Devpriya Soni. 2023. Sms phishing dataset for machine learning and pattern recognition. In *Proceedings of the 14th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2022)*, pages 597–604, Cham. Springer Nature Switzerland.

Chidimma Opara, Paolo Modesti, and Lewis Golightly. 2025. Evaluating spam filters and stylometric detection of ai-generated phishing emails. *Expert systems with applications*, 276:127044.

Jude Osamor, Moses Ashawa, Alireza Shahrabi, Anand Phillip, and Celestine Iwend. 2025. The evolution of phishing and future directions: A review. In *International Conference on Cyber Warfare and Security*, pages 361–368. Academic Conferences International Limited.

Devsmit Ranparia, Shaily Kumari, and Ashish Sahani. 2020. Fake job prediction using sequential network. In *2020 IEEE 15th international conference on industrial and information systems (ICIIS)*, pages 339–343. IEEE.

Tarini Saka, Kami Vaniea, and Nadin Kökciyan. 2025. Sok: Grouping spam and phishing email threats for smarter security. *IEEE Access*.

Said Salloum, Tarek Gaber, Sunil Vadera, and Khaled Shaalan. 2022. A systematic literature review on phishing email detection using natural language processing techniques. *Ieee Access*, 10:65703–65727.

Robiert Sepúlveda-Torres, Iván Martínez-Murillo, Estela Saquete, Elena Lloret, and Manuel Palomar. 2025. To write or not to write as a machine? that's the question. *IEEE Transactions on Big Data*.

softecapps. 2024. spam_ham_spanish.

Cheekati Srikanth, M Rashmi, S Ramu, and Ram Mohana Reddy Guddeti. 2022. A novel fake job posting detection: an empirical study and performance evaluation using ml and ensemble techniques. In *International Conference on Security, Privacy and Data Analytics*, pages 219–234. Springer.

Daniel Tamayo, Iñaki Lacunza, Paula Rivera-Hidalgo, Severino Da Dalt, Javier Aula-Blasco, Aitor Gonzalez-Agirre, and Marta Villegas. 2026. Mrbert: Modern multilingual encoders via vocabulary, domain, and dimensional adaptation. *arXiv preprint arXiv:2602.21379*.

Khushboo Taneja, Jyoti Vashishtha, and Saroj Ratnoo. 2025. Fraud-bert: transformer based context aware online recruitment fraud detection. *Discover Computing*, 28(1):9.

Kutub Thakur, Md Liakat Ali, Muath A Obaidat, and Abu Kamruzzaman. 2023. A systematic review on deep-learning-based phishing email detection. *Electronics*, 12(21):4545.

Mohammad Amaz Uddin, Md Mahiuddin, and Iqbal H Sarker. 2026. An explainable transformer-based model for phishing email detection: A large language model approach. *Computer Networks*, page 112061.

Soumitra Utturkar, Shravani Sonawane, Aditya Shah, Shruti Palange, Rupesh Jaiswal, and Mousami Munot. 2026. Phishing detection using machine learning for english and multilingual. In *Proceedings of International Conference on Computational Intelligence and Information Retrieval: ICCIIR 2025, Volume 2*, volume 2, page 43. Springer Nature.

Gabriela Zuñiga. 2024. Spam detection messages dataset.