Detecting Phishing Emails in Nigerian Pidgin English Using a Dialect-Aware and Behavioural NLP Model

Zubaida Muhtar Alhassan ZUBAIDAALHASSAN.01@GMAIL.COM
Department of Information Technology,
Faculty of Computing,
Nile University of Nigeria, Abuja, Nigeria

Editor: Sakinat Folorunso, Roseline Ogundokun, and Francisca Oladipo

Abstract

Phishing attacks that target individuals using culturally specific communication remain one of the most common forms of cybercrime today. In Nigeria, over 100 million people speak Nigerian Pidgin, a language frequently used in informal digital conversations that has also become a conduit for phishing attacks targeting specific regions. However, the linguistic and cultural nuances of Nigerian Pidgin are not detected by existing phishing detection systems, which are predominantly trained on standard English datasets. This study proposes a proof-of-concept dialect-aware and behaviourally informed Natural Language Processing (NLP) model for detecting phishing emails in Nigerian Pidgin. Due to the scarcity of public datasets, a balanced dataset of 870 emails was created using a hybrid translation and generation process, validated by native Nigerian Pidgin speakers. The proposed model combines TF-IDF-based linguistic features with seven behavioural indicators derived from persuasion theory. A Random Forest classifier was utilized for email classification, optimized using a Genetic Algorithm to tune hyperparameters and feature weights. The system achieved 93.89% accuracy, 100.00% precision, and 87.69% recall on the test set, demonstrating strong performance while maintaining a zero false positive rate. Behavioural features associated with reward and urgency were found to significantly improve detection performance. The study demonstrates the importance of integrating behavioural and linguistic analysis for cybersecurity in low-resource language contexts.

Keywords: Phishing Detection, Nigerian Pidgin, Low-Resource NLP, Behavioural Analysis, Random Forest, Genetic Algorithm

1. Introduction

Phishing is a major cybersecurity issue. It entails sending false messages to extract sensitive information such as login credentials and financial details. Today, the use of NLP and machine learning algorithms has greatly enhanced phishing detection. However, most existing phishing detection algorithms are intended for standard English and do not function well in low-resource or informal languages.

In Nigeria, smartphone ownership has grown significantly. Over 200 million Nigerians are currently connected via mobile devices (Statista, 2024). Nigerian Pidgin English is commonly used in online discussions, and attackers utilize it to target victims with limited digital abilities. Nigerian Pidgin uses non-standard orthography, code-switching, and cultural expressions. Current detection systems are not trained to recognize these patterns, limiting their efficiency in such scenarios. Existing phishing systems are primarily trained in Western English. They fail to account for dialect variations and region-specific

manipulation tactics. Moreover, these systems often rely heavily on text features and do not include behavioral manipulation techniques like urgency, authority impersonation, and reward-based persuasion, which are common in phishing attacks. Additionally, optimization methods such as genetic algorithms are not often used to improve model performance in low-resource environments.

In this study, a dialect-aware and behavior-informed NLP model is proposed to detect phishing in emails written in Nigerian Pidgin English. To enhance detection accuracy in low-resource language settings, linguistic and behavioral features are integrated.

The objectives of this study are:

1. To design a dialect-aware machine learning framework that includes linguistic feature extraction and psychologically based behavioral analysis for detecting phishing in Nigerian Pidgin English.

2. To implement the framework using a dialect-aware NLP model integrated with behavioral features that capture psychological manipulation strategies such as urgency, authority impersonation, reward bait, and suspicious URL patterns.

3. To optimize the model using genetic algorithms to improve feature weighting and classification performance.

4. To evaluate the proposed model using precision, recall, F1-score, and false positive rate metrics, and to compare its performance against baseline NLP models.

2. Related Work

2.1. Phishing Detection Techniques

Approaches for phishing detection evolved from rule-based systems to machine learning approaches such as Random Forest, Support Vector Machine (SVM), and Logistic Regression. Kalla and Kuraku (2023) highlighted that Random Forest models using lexical and host-based features detect phishing URLs with good accuracy. Moreover, in a similar paper, Mohammed et al. (2015) found that important factors such as IP address usage, URL length, and domain attributes were powerful predictors of phishing activity.

In recent studies, deep learning and transformer-based models have shown remarkable accuracy. Studies such as Roumeliotis et al. (2024) showed that the transformer architecture can enhance detection by identifying contextual patterns in phishing emails. However, these models require large labelled datasets and are trained on English language corpora, limiting their applicability in low-resource language settings.

2.2. Behavioural Analysis

In many cases, phishing scams frequently utilize psychological manipulation techniques to exploit people's vulnerabilities. According to Cialdini (2006), tactics are consistent with persuasion principles such as urgency, authority, and reward. Adding such behavioural indicators to phishing detection systems improves detection accuracy. Studies such as Mohammed et al. (2015) indicated how scammers use deceptive structures like fake domain names and unusual URLs, while Kalla and Kuraku (2023) argued that features such as URL

shortening and suspicious domains improve detection accuracy. These findings show that behavioural features help indicate attacker intent, depend less on language, and are useful in low-resource and multilingual contexts.

In addition to academic research, recent grassroots initiatives such as the Cyber-Shield project have begun to address phishing detection in local Nigerian languages. Cyber-Shield is an AI-powered system designed to detect scams such as fake bank alerts, Central Bank of Nigeria (CBN) grant scams, OTP harvesting, and lottery fraud across languages including Nigerian Pidgin, Yoruba, Hausa, and Igbo. While this initiative highlights the growing importance of localized cybersecurity solutions, it lacks publicly available technical details and standardized evaluation, reinforcing the need for structured, research-driven approaches.

2.3. NLP in Low-Resource and African Languages

Historically, Natural Language Processing has focused on high-resource languages, with more than 90% of studies concentrating on a limited subset of global languages (Joshi et al., 2020). Due to limited datasets, irregular orthography, and linguistic diversity, African languages including Nigerian Pidgin remain under-represented. Recently, there have been attempts to address this gap. Adelani et al. (2024) contributed towards developing benchmark datasets for African languages and highlighted that transfer learning models such as Afro-XLMR can enhance performance in low-resource settings. However, performance differences remain when compared to English, mainly due to data limitation and linguistic variability.

Emezue (2021) identified resources and infrastructural challenges that limit the creation of NLP systems for African languages. It was also discovered that out of all NLP research on Nigerian languages, only roughly a quarter actually contributes new resources (Inuwa-Dutse, 2025). Similarly, Garba et al. (2024) adapted GPT-2 for Nigerian Pidgin text generation and reported a BLEU score of 0.56, which, though promising, is still below typical English benchmarks. Adelani et al. (2025) assessed large language models like LLaMA2 and BERT on Nigerian Pidgin translation tasks and observed substantial performance degradation compared to English, especially in handling code-switching, cultural references, and non-standard orthography.

Nekoto et al. (2020), through the Masakhane initiative, demonstrated the importance of community-driven dataset creation for African languages but acknowledged persistent resource scarcity and evaluation challenges. Hedderich et al. (2021) further observed that low-resource NLP often relies on data augmentation, transfer learning, and synthetic generation techniques due to insufficient corpora.

Studies like Okoloegbo et al. (2022) demonstrated that using machine learning to detect cyberbullying in Nigerian Pidgin is feasible. While these papers verify that Pidgin can be computationally modelled, they do not address cybersecurity applications such as phishing detection and are limited to text classification tasks or translation.

2.4. Research Gap

There is limited study on phishing detection in Nigerian Pidgin English, despite notable improvements in low-resource NLP and phishing detection. Most existing systems are trained on detecting standard English and do not account for the code-switching, non-standard

spelling, and cultural phrases found in Nigerian Pidgin. Moreover, behavioral analysis is rarely integrated with dialect-aware NLP techniques, even though it has been demonstrated to improve phishing detection. Furthermore, existing phishing detection models underutilize optimization approaches like genetic algorithms, especially in low-resource environments. These drawbacks demonstrate the need for a unifying model that combines dialect-aware NLP with behavioral analysis and optimization techniques. This study fills this gap by proposing an integrated approach specifically designed for Nigerian Pidgin phishing detection.

3. Methodology

3.1. Research Design

This study incorporates an experimental ML strategy which includes dataset creation, feature extraction, model development, optimization, and evaluation.

3.2. Dataset Creation

The study utilized a balanced dataset of 870 emails (435 phishing, 435 legitimate) created through a two-stage hybrid approach to ensure diversity and cultural relevance.

Translation Phase (30%): Approximately 30% of the dataset (261 emails) was translated from a publicly available English phishing dataset on Kaggle (Al-Subaiey et al., 2024) into Nigerian Pidgin using ChatGPT-4.

Custom Generation Phase (70%): The remaining 70% (609 emails) were custom-generated from scratch using ChatGPT-4.

- **Phishing Samples:** Generated based on documented Nigerian scam patterns, such as CBN account verification alerts, bank impersonation (GTBank, Access Bank), telecommunications fraud, and lottery scams.

- **Legitimate Samples:** Generated to mimic natural Nigerian Pidgin communication found in informal digital correspondence.

Two approaches were tested for dataset creation. First, a custom neural machine translation (NMT) model was trained using the Hugging Face English-Nigerian Pidgin parallel corpus (22,000 sentence pairs), BBC Pidgin articles, and Nairaland posts. However, a BLEU score of 0.42 was obtained after analyzing 50 samples. The model struggled with semantic drift and unnatural phrasing. Due to the limited training data (510 samples vs the 10,000+ typically required for NMT), the NMT approach was abandoned in favor of the ChatGPT-4-assisted approach validated by 11 native speakers.

All samples were validated by 11 native Nigerian Pidgin speakers using a 5-point Likert scale based on linguistic naturalness, cultural appropriateness, and preservation of manipulation tactics. The human-validated approach achieved a mean quality rating of 4.17/5.0.

The dataset was split as follows: Training set (609 samples, 70%), Validation set (130 samples, 15%), and Test set (131 samples, 15%).

Table 1: Human Validation Results for Dataset Quality

Dimension	Mean	SD	Interpretation
Linguistic Naturalness	4.14	1.10	Good
Cultural Appropriateness	4.19	1.01	Good
Manipulation Preservation	4.19	1.01	Good
Overall Mean	4.17	1.04	Good

3.3. Feature Extraction

This section combines TF-IDF linguistic features and seven behavioral indicators to improve detection performance. TF-IDF (Term Frequency-Inverse Document Frequency) was utilized to convert the content of the email into numerical representations by capturing the importance of each word in the email in relation to the whole dataset. This enables the system to identify linguistic patterns that are commonly found in phishing emails, like specific keywords and phrasing structures.

However, using TF-IDF alone is limited to surface-level textual patterns and may fail to detect phishing messages that use sophisticated and contextually deceptive language. To address this, seven behavioral features based on Cialdini (2006) persuasion theory were added:

1. Urgency
2. Authority impersonation
3. Reward cues
4. URL count
5. IP Address Detected
6. URL shortener Detected
7. Suspicious TLD Detected (e.g., .tk, .ml, .ga)

Integrating linguistic and behavioral features enables the model to capture both what is being said (text patterns) and how it is being said (manipulation strategies). This combination enhances robustness, particularly in low-resource settings.

3.4. Model Development

The Random Forest classifier was chosen for its reliability, interpretability, and ability to handle high-dimensional feature fields. The final categorization is determined by a majority vote based on the predictions made by each decision tree in the Random Forest ensemble. This technique works well with structured and mixed feature types, such as the TF-IDF and behavioral features utilized in this study, and reduces overfitting when compared to single decision trees.

Furthermore, a Genetic Algorithm (GA) was used to optimize the Random Forest model's hyperparameters and the weights assigned to behavioral traits. The GA iteratively searches for optimal configurations by simulating evolutionary processes such as selection, crossover, and mutation, with the objective of maximizing the F1-score on the validation dataset.

3.5. Evaluation Metrics

The performance of the model was evaluated using standard classification metrics:

- **Accuracy:** The proportion of correctly classified emails. Accuracy = (TP + TN) / (TP + TN + FP + FN)
- **Precision:** The proportion of true phishing emails among those flagged. Precision = TP / (TP + FP)
- **Recall:** The proportion of actual phishing emails that were correctly detected. Recall = TP / (TP + FN)
- **F1-Score:** The harmonic mean of precision and recall. F1 = 2 x (Precision x Recall) / (Precision + Recall)

4. Results

4.1. Baseline Model Performance

The model was tested using the validation dataset with default hyperparameters of 200 trees and a maximum depth of 20.

Table 2: Baseline Model Performance Results

Metric	Score (%)
Accuracy	95.38
Precision	100.00
Recall	90.77
F1-score	95.16

4.2. Genetic Optimization

The Genetic Algorithm was applied to optimize the Random Forest hyperparameters and the weighting of the seven behavioral features. The goal was to maximize the F1-score on the validation dataset.

Optimized Hyperparameters:

- n_estimators: Increased from 200 to 253
- max_depth: Increased from 20 to 21

- min_samples_split: Reduced from 5 to 2

As shown in Table 3, the optimization process increased the F1-score from 95.16% to 96.83%.

Table 3: Genetic Algorithm Optimized Model Performance (Validation Set)

Metric	**Score (%)**
Accuracy	96.92
Precision	100.00
Recall	93.85
F1-score	96.83

Table 4: Optimized Behavioral Feature Weights

Feature	**Optimized Weight**	**Change (%)**
reward_score	2.91	+191.3
url_count	2.51	+151.4
urgency_score	2.15	+115.3
authority_score	0.16	-84.0
has_ip_address	0.72	-28.4
has_suspicious_tld	0.63	-37.3
has_url_shortener	0.63	-36.8

The reward score emerged as the most heavily weighted feature, coinciding with the prevalence of advance-fee fraud and lottery scams in the Nigerian context.

4.3. Optimized Model Performance on Test Set

The optimized Random Forest model was evaluated on the unseen test dataset (131 samples).

The model maintained perfect precision (100%) on the test set, with 57 true positives, 8 false negatives, and 66 true negatives. The recall of 87.69% signifies that the system successfully detected nearly 9 out of 10 phishing attempts. The high ROC-AUC score of 0.9917 confirms excellent overall discriminatory capability.

Table 5: Optimized Model Performance on Test Dataset

Metric	Score (%)
Accuracy	93.89
Precision	100.00
Recall	87.69
F1-score	93.44
ROC-AUC	99.17

4.4. Cross-Validation Results

A 5-fold stratified cross-validation was performed on the training set (609 samples) to assess model stability and generalizability.

Table 6: Cross-Validation Performance Across Folds

Fold	F1-score	ROC-AUC	Precision	Recall
1	96.55	0.9918	98.59	94.59
2	88.72	0.9912	100.00	79.73
3	94.29	0.9909	100.00	89.19
4	95.04	0.9987	100.00	90.54
5	93.53	0.9743	100.00	87.84
Mean	93.62	0.9894	99.72	88.38

The low standard deviation in F1-score (±2.65%) and high mean ROC-AUC confirm that the model is robust, stable, and not overfitting.

5. Discussion

The results demonstrate that integrating linguistic and behavioral features significantly improves phishing detection performance in Nigerian Pidgin. While TF-IDF features effectively capture textual patterns, the inclusion of behavioral indicators enhances the model's ability to detect manipulation strategies that remain consistent across different linguistic expressions.

The Genetic Algorithm optimization further improved model performance by identifying the most relevant features and fine-tuning model parameters. Notably, reward-based and urgency features were found to have the highest impact, reflecting the prevalence of financial and incentive-driven scams in the Nigerian context.

Benchmarking: The proposed model achieved an F1-score of 93.44% on the Nigerian Pidgin test set. This performance is highly competitive when compared to studies on high-resource languages, such as Altwaijry et al. (2024) who reported 94.00% F1-score on English phishing emails using deep learning. Given the low-resource nature of Nigerian Pidgin and the absence of large pre-trained models, achieving comparable accuracy using a Random Forest model with behavioral features validates the effectiveness of the proposed dialect-aware approach.

6. Conclusion

This study illustrates that phishing can be identified efficiently in diverse and low-resource languages through a shift from a universal approach to a more adaptive model. The developed model asserts that the synergistic integration of sociolinguistic intelligence (dialect awareness), psychological insight (behavioral features), and computational efficiency (Random Forest optimized by GA) forms an effective defense strategy.

The model achieved perfect precision, showing that it is possible to build systems that are responsive to threats while adhering to local communication norms. Future work will focus on:

- Collecting larger, real-world datasets.
- Exploring deep learning approaches.
- Implementation in real-world email security systems.

Data Availability Statement

The synthetic Nigerian Pidgin phishing dataset (870 balanced emails) and the complete Python implementation code are available on demand. The dataset was generated using public phishing templates (**?**), ChatGPT-4-assisted translation, and manual validation by 11 native speakers. All samples are de-identified.

Acknowledgments

I express my sincere gratitude to my supervisor, Assoc. Prof. Bilkisu Muhammad-Bello, for her guidance, insight, and encouragement at every stage of this research. I also appreciate the valuable feedback from the faculty and staff of the Department of Information Technology, Nile University of Nigeria. I am grateful to my family for their patience and support throughout this journey.

References

David I. Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoming Shen, Masabata Reid, Michelle Beukman, Jonneke Pfeiffer, Chryzant Emezue, Sebastian Gehrmann, and Sara Hooker. Africanlp: Sentiment analysis and named entity recognition for african languages. *Proceedings of the AfricaNLP Workshop*, 2024.

David I. Adelani, A. Seza Doğruöz, Ifeoluwa Shode, and Ayodele Aremu. Does generative ai speak nigerian-pidgin? issues about representativeness and bias for multilingualism in llms. *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025. URL https://aclanthology.org/2025.findings-naacl.85.pdf.

A. Al-Subaiey, M. Al-Thani, N. A. Alam, K. F. Antora, A. Khandakar, and S. A. U. Zaman. Novel interpretable and robust web-based ai platform for phishing email detection. *arXiv preprint arXiv:2405.11619*, 2024. URL https://arxiv.org/abs/2405.11619.

Norah Altwaijry, Maram A. Al-Turaiki, and Muna Al-Razgan. Advancing phishing email detection: A comparative study of deep learning models. *Sensors*, 24(7):2077, 2024.

Robert B. Cialdini. *Influence: The Psychology of Persuasion*. Harper Business, revised edition, 2006.

Chris Chinenye Emezue. African natural language processing: Challenges and opportunities. *arXiv preprint*, 2021. arXiv:2103.12303.

K. Garba, T. Kolajo, and J. B. Agbogun. A transformer-based approach to nigerian pidgin text generation. *International Journal of Speech Technology*, 27(4):1027–1037, 2024.

Michael A. Hedderich, David Adelani, Dawei Zhu, Jesujoba Alabi, and Dietrich Klakow. A survey on recent approaches for natural language processing in low-resource scenarios. *Proceedings of NAACL*, pages 2545–2568, 2021.

Ibrahim Inuwa-Dutse. Overview of nigerian pidgin and its usage in nigeria. *arXiv preprint*, 2025. arXiv:2502.19784.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, 2020.

S. Kalla and M. Kuraku. Machine learning approaches for phishing detection using url and lexical features. *Journal of Cybersecurity Research*, 2023.

Rami M. Mohammed, Fadi Thabtah, and Lee McCluskey. An assessment of features related to phishing websites using an automated technique. *International Journal of Internet Technology and Secured Transactions*, 5(1):1–15, 2015.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, et al. Participatory research for low-resourced machine translation: A case study in african languages. In *Findings of EMNLP*, pages 2144–2160, 2020.

P. Okoloegbo, C. Nwankwo, and I. Okafor. Multilingual cyberbullying detection for nigerian languages using machine learning techniques. *Proceedings of the African NLP Workshop*, 2022.

K. I. Roumeliotis, N. D. Tselikas, and A. K. Nasiopoulos. Next-generation spam filtering: Comparative fine-tuning of llms, nlps, and cnn models for email spam classification. *Electronics*, 13(5):2034, 2024.

Statista. Number of mobile internet users in nigeria, 2024. URL https://www.statista.com.