Title: ISmishU: Smishing Detection System for Code-mixed Messages Using Transfer Learning of XLM-RoBERTa

Authors: Capulong, Mark Daniel M.
Digang, Cyrel P.
Morales, Mark Jerico P.
Recio, Paolo Luigi G.
Soriano, Stephanie Elenn V.

Adviser: Prof. Aleta C. Fabregas

I. Introduction

In 2023, phishing and smishing attacks accounted for nearly half of all reported fraud cases in the Philippines, driven by the widespread use of mobile technology. Smishing, which is a combination of short message services or ‘SMS’ and ‘phishing’ in which the phishing attack is initiated through text messaging, remains a significant issue despite government efforts like the SIM Registration Act and blocking malicious URLs.

Existing smishing detection systems rely primarily on language models trained in high-resource languages like English, leaving low-resource languages like Taglish underrepresented. This gap poses unique challenges, as Taglish exhibits distinct linguistic patterns that existing systems fail to address.

This study proposes a smishing detection system tailored to Taglish, leveraging XLM-RoBERTa, a multilingual language model known for its ability to analyze message content. By incorporating both text analysis of the fine-tuned XLM-RoBERTa and URL verification through the VirusTotal API, the system aims to detect smishing messages, or "ham" (non-smishing) messages, with improved accuracy.

Training large models like XLM-RoBERTa on low-resource datasets introduces challenges such as the possibility of overfitting, where the model becomes too tailored to training data and performs poorly on new data. To mitigate this, the study explores techniques like layer freezing, where only certain layers of the model are updated, improving generalization and performance. By combining advanced techniques and

addressing the unique linguistic characteristics of Taglish, this research seeks to strengthen cybersecurity efforts and mitigate the threat of smishing in the Philippines.

II. Methodology

This study adopts an experimental quantitative approach to develop a smishing detection system for Taglish SMS using a fine-tuned XLM-RoBERTa model. Data is collected over an eight-week period through a Google Form survey distributed via social media platforms, targeting Filipino residents who have received spam or suspicious messages within the past two years. The collected dataset consists of smishing and legitimate ("ham") messages in Tagalog, English, and Taglish. To ensure accuracy and reliability, all messages are validated by a linguist's expert. Preprocessing steps include standardizing the text by correcting common misspellings (e.g., “anu” to “ano,” “22o” to “totoo”), removing special characters and punctuation, and converting all text to lowercase for consistency.

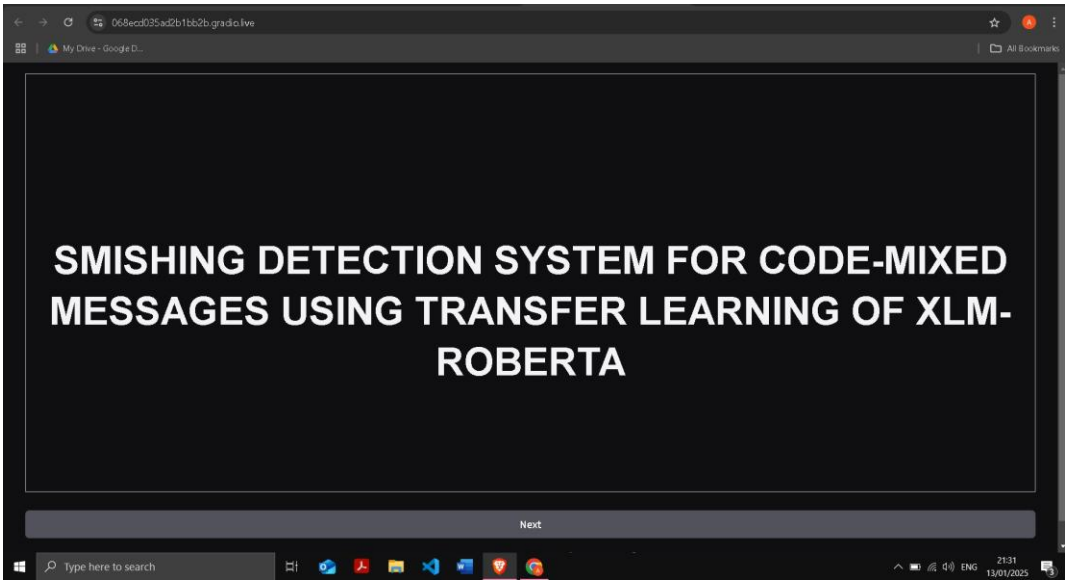
The annotated dataset is used to train and test the XLM-RoBERTa model. The Experiment was conducted using various training dataset ratios and layer freezing configurations are implemented during the training to address possible overfitting by selectively updating the weights of certain layers while freezing others. The model’s performance is assessed using accuracy, precision, recall, and F1 score, with testing results validated against evaluations from expert linguists to ensure reliability and robustness. Finally, a one-way Analysis of Variance (ANOVA) test is conducted to assess whether significant differences exist in the model’s performance across the different training dataset ratios.

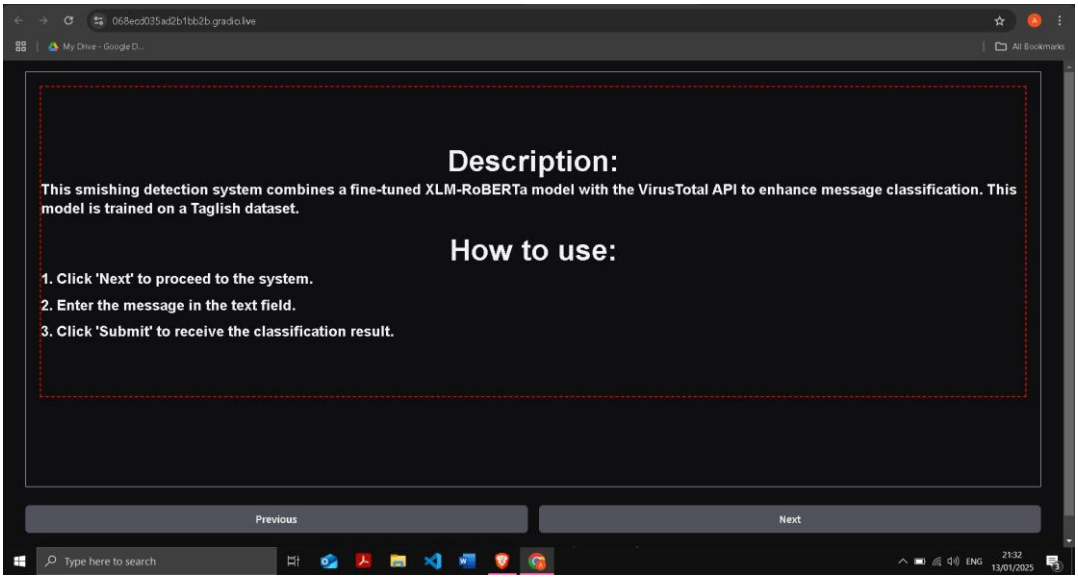
The system will have a two-layer analysis. If a URL is detected in the message, it is sent to VirusTotal through its API for URL analysis. If the URL is identified as malicious, the user interface immediately displays the result indicating that the message is a smishing attempt. If no URL is detected or the URL is not malicious, the text body of the message proceeds to content analysis of the most optimized model configuration, determined through experiments with varying training dataset ratios and layer freezing setups. The combination of text analysis and URL verification ensures a more comprehensive and accurate smishing detection system.of the system's effectiveness.

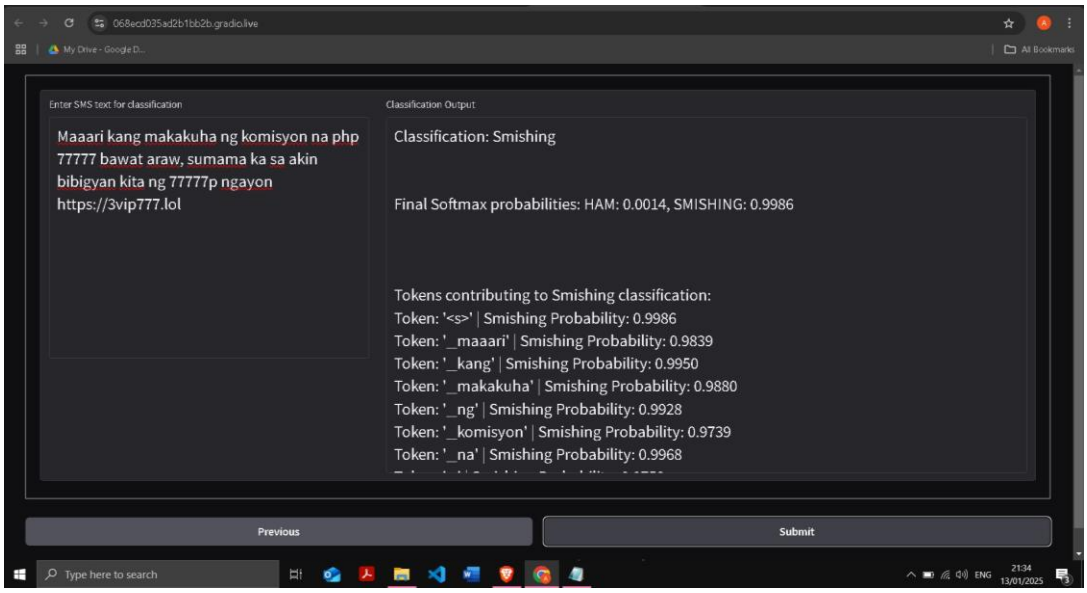
III. Expected Results/Outcomes

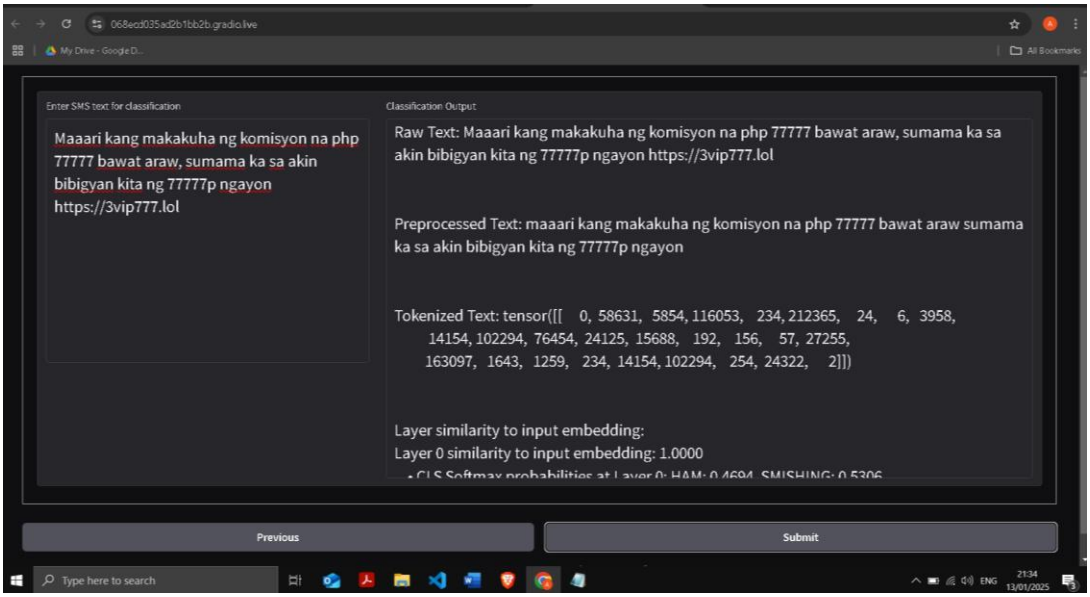
The system is expected to accurately detect whether a message, based on its content or the URL it contains, is classified as smishing or not. When the system performs content analysis using the fine-tuned XLM-RoBERTa model, it will provide a detailed output that includes the classification results, showing the percentage likelihood of the message being "ham" or "smishing" to help explain its decision. Additionally, the system will display a cosine similarity to input embeddings across all layers, offering insights into how the model processes and understands the message. For messages classified as smishing, the system will highlight specific tokens that contributed most to the classification, providing transparency and interpretability of the model's decision-making process.

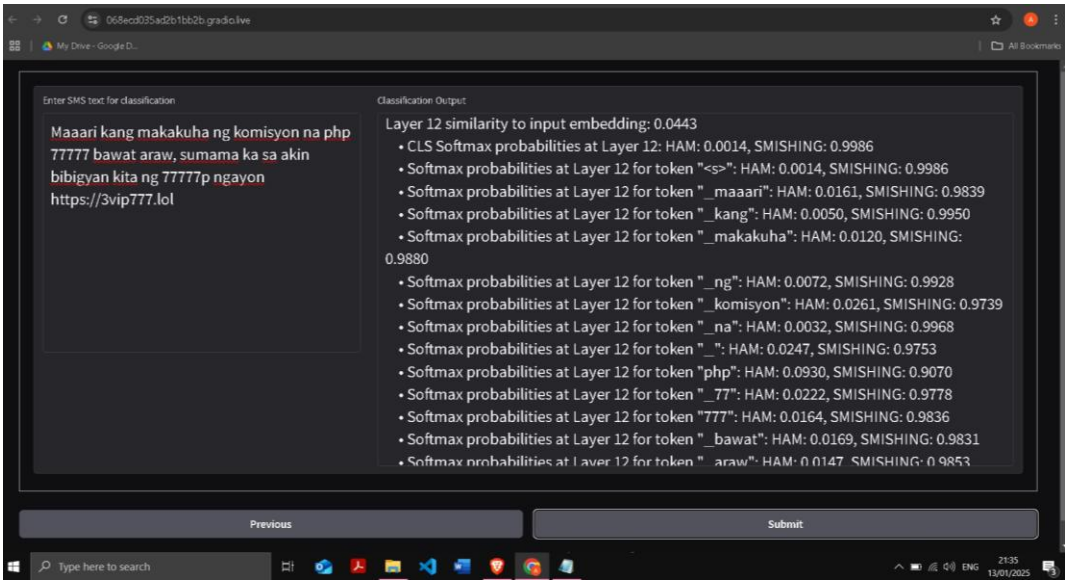
IV. System

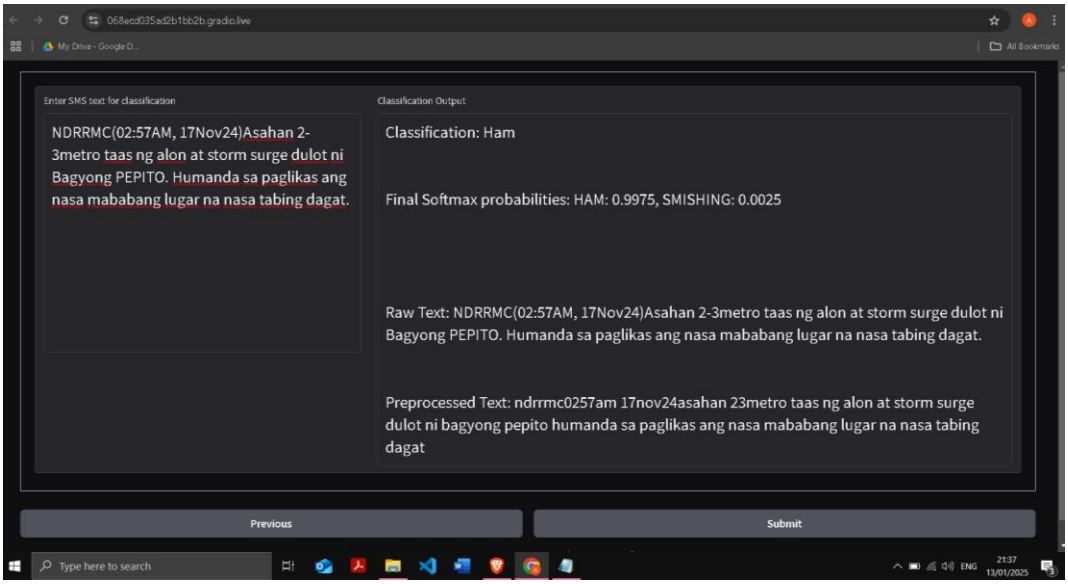


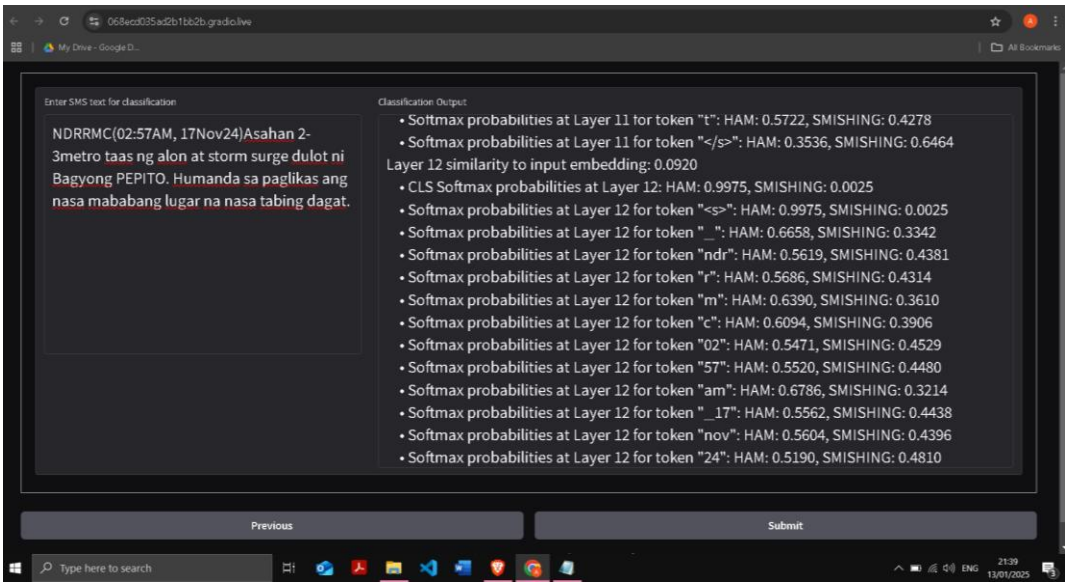

068ecd035ad2b1bb2b.gradio.live
Description:
This smishing detection system combines a fine-tuned XLM-RoBERTa model with the VirusTotal API to enhance message classification. This model is trained on a Taglish dataset.
How to use:
1. Click 'Next' to proceed to the system.
2. Enter the message in the text field.
3. Click 'Submit' to receive the classification result.
Previous
Next


068ecd035ad2b1bb2b.gradio.live
Enter SMS text for classification
Maaari kang makakuha ng komisyon na php 77777 bawat araw, sumama ka sa akin bibigyan kita ng 77777p ngayon https://3vip777.lol
Classification Output
Classification: Smishing
Final Softmax probabilities: HAM: 0.0014, SMISHING: 0.9986
Tokens contributing to Smishing classification:
Token: '<s>' | Smishing Probability: 0.9986
Token: '_maaari' | Smishing Probability: 0.9839
Token: '_kang' | Smishing Probability: 0.9950
Token: '_makakuha' | Smishing Probability: 0.9880
Token: '_ng' | Smishing Probability: 0.9928
Token: '_komisyon' | Smishing Probability: 0.9739
Token: '_na' | Smishing Probability: 0.9968
Previous
Submit


068ecd035ad2b1bb2b.gradio.live
Enter SMS text for classification
Maaari kang makakuha ng komisyon na php 77777 bawat araw, sumama ka sa akin bibigyan kita ng 77777p ngayon https://3vip777.lol
Classification Output
Raw Text: Maaari kang makakuha ng komisyon na php 77777 bawat araw, sumama ka sa akin bibigyan kita ng 77777p ngayon https://3vip777.lol
Preprocessed Text: maaari kang makakuha ng komisyon na php 77777 bawat araw sumama ka sa akin bibigyan kita ng 77777p ngayon
Tokenized Text: tensor([[0, 58631, 5854, 116053, 234, 212365, 24, 6, 3958, 14154, 102294, 76454, 24125, 15688, 192, 156, 57, 27255, 163097, 1643, 1259, 234, 14154, 102294, 254, 24322, 2]])
Layer similarity to input embedding:
Layer 0 similarity to input embedding: 1.0000
Previous
Submit


Enter SMS text for classification
Maaari kang makakuha ng komisyon na php 77777 bawat araw, sumama ka sa akin bibigyan kita ng 77777p ngayon https://3vip777.lol
Classification Output
Layer 12 similarity to input embedding: 0.0443
• CLS Softmax probabilities at Layer 12: HAM: 0.0014, SMISHING: 0.9986
• Softmax probabilities at Layer 12 for token "<s>": HAM: 0.0014, SMISHING: 0.9986
• Softmax probabilities at Layer 12 for token "_maaari": HAM: 0.0161, SMISHING: 0.9839
• Softmax probabilities at Layer 12 for token "_kang": HAM: 0.0050, SMISHING: 0.9950
• Softmax probabilities at Layer 12 for token "_makakuha": HAM: 0.0120, SMISHING: 0.9880
• Softmax probabilities at Layer 12 for token "_ng": HAM: 0.0072, SMISHING: 0.9928
• Softmax probabilities at Layer 12 for token "_komisyon": HAM: 0.0261, SMISHING: 0.9739
• Softmax probabilities at Layer 12 for token "_na": HAM: 0.0032, SMISHING: 0.9968
• Softmax probabilities at Layer 12 for token "_": HAM: 0.0247, SMISHING: 0.9753
• Softmax probabilities at Layer 12 for token "php": HAM: 0.0930, SMISHING: 0.9070
• Softmax probabilities at Layer 12 for token "_77": HAM: 0.0222, SMISHING: 0.9778
• Softmax probabilities at Layer 12 for token "777": HAM: 0.0164, SMISHING: 0.9836
• Softmax probabilities at Layer 12 for token "_bawat": HAM: 0.0169, SMISHING: 0.9831
Previous
Submit


Enter SMS text for classification
NDRRMC(02:57AM, 17Nov24)Asahan 2-3metro taas ng alon at storm surge dulot ni Bagyong PEPITO. Humanda sa paglikas ang nasa mababang lugar na nasa tabing dagat.
Classification Output
Classification: Ham
Final Softmax probabilities: HAM: 0.9975, SMISHING: 0.0025
Raw Text: NDRRMC(02:57AM, 17Nov24)Asahan 2-3metro taas ng alon at storm surge dulot ni Bagyong PEPITO. Humanda sa paglikas ang nasa mababang lugar na nasa tabing dagat.
Preprocessed Text: ndrrmc0257am 17nov24asahan 23metro taas ng alon at storm surge dulot ni bagyong pepito humanda sa paglikas ang nasa mababang lugar na nasa tabing dagat
Previous
Submit


Enter SMS text for classification
NDRRMC(02:57AM, 17Nov24)Asahan 2-3metro taas ng alon at storm surge dulot ni Bagyong PEPITO. Humanda sa paglikas ang nasa mababang lugar na nasa tabing dagat.
Classification Output
• Softmax probabilities at Layer 11 for token "t": HAM: 0.5722, SMISHING: 0.4278
• Softmax probabilities at Layer 11 for token "</s>": HAM: 0.3536, SMISHING: 0.6464
Layer 12 similarity to input embedding: 0.0920
• CLS Softmax probabilities at Layer 12: HAM: 0.9975, SMISHING: 0.0025
• Softmax probabilities at Layer 12 for token "<s>": HAM: 0.9975, SMISHING: 0.0025
• Softmax probabilities at Layer 12 for token "_": HAM: 0.6658, SMISHING: 0.3342
• Softmax probabilities at Layer 12 for token "ndr": HAM: 0.5619, SMISHING: 0.4381
• Softmax probabilities at Layer 12 for token "r": HAM: 0.5686, SMISHING: 0.4314
• Softmax probabilities at Layer 12 for token "m": HAM: 0.6390, SMISHING: 0.3610
• Softmax probabilities at Layer 12 for token "c": HAM: 0.6094, SMISHING: 0.3906
• Softmax probabilities at Layer 12 for token "02": HAM: 0.5471, SMISHING: 0.4529
• Softmax probabilities at Layer 12 for token "57": HAM: 0.5520, SMISHING: 0.4480
• Softmax probabilities at Layer 12 for token "am": HAM: 0.6786, SMISHING: 0.3214
• Softmax probabilities at Layer 12 for token "_17": HAM: 0.5562, SMISHING: 0.4438
• Softmax probabilities at Layer 12 for token "nov": HAM: 0.5604, SMISHING: 0.4396
• Softmax probabilities at Layer 12 for token "24": HAM: 0.5190, SMISHING: 0.4810
Previous
Submit