

Multi-Label Classification of Indonesian Voice Phishing Conversations: A Comparative Study of XLM-RoBERTa and ELECTRA

Ahmad Hidayat^{1,*}, Sarifuddin Madenda², Hustinawaty³

^{1,2,3}Doctoral Program in Information Technology, Gunadarma University, Depok 16424, Indonesia

(Received: November 29, 2024; Revised: January 5, 2025; Accepted: April 5, 2025; Available online: July 19, 2025)

Abstract

Mobile phones have become a primary means of communication, yet their advancement has also been exploited by cybercriminals, particularly through voice phishing schemes. Voice phishing is a form of social engineering fraud carried out via telephone conversations to illegally obtain personal or financial information. The complexity of voice phishing continues to increase, as a single conversation may involve multiple fraudulent schemes simultaneously, necessitating the application of multi-label classification to comprehensively identify all motives of fraud. Previous studies have predominantly utilized single-label approaches and foreign-language data, making them less relevant to the Indonesian language context and unable to produce speaker segmentation outputs for conversational analysis. This study contributes by developing a multi-label voice phishing classification system specifically for Indonesian telephone conversations to address this gap. Audio data were collected from open sources and simulated recordings, resulting in a total of 300 samples labeled into six categories: five phishing modes and one non-phishing category. The proposed system consists of a preprocessing pipeline that includes noise reduction, speaker segmentation, automatic transcription, and text cleaning to preserve the context of two-way conversations. Two machine learning models based on transformer architectures, XLM-RoBERTa and ELECTRA, are employed to identify various fraud schemes that may occur simultaneously within a single conversation. The dataset was split into training, validation, and testing sets with two division ratios for performance evaluation. Several combinations of hyperparameters were tested to obtain the most optimal model configuration. Evaluation was conducted using a supervised learning approach and various performance metrics. The experimental results show that XLM-RoBERTa achieved the highest average accuracy of $97.04 \pm 1.15\%$ and the highest average F1-score of $92.66 \pm 2.59\%$. These results highlight the novelty of applying multi-label classification in the Indonesian language context for voice phishing detection, contributing to more effective fraud identification in real-world telephony systems.

Keywords: Voice Phishing, Indonesian Language, Speaker Segmentation, Multi-Label Classification, Transformer Model

1. Introduction

Cybercrime continues to evolve alongside the rapid advancement of digital communication technologies. One of the most prevalent forms of cybercrime is digital fraud, which has emerged as a global concern due to its growing complexity and the increasing difficulty in detection [1], [2]. One of the increasingly prevalent forms of attack is voice phishing, a type of social engineering fraud conducted through direct telephone conversations between an attacker and a victim [3], [4]. Within a single conversation, voice phishing can involve multiple fraud schemes simultaneously, such as illegal online loans, fabricated family emergencies, fake investments, fraudulent transactions, and prize scams. The complexity of these attack patterns necessitates the application of a multi-label classification approach to identify all possible fraudulent intents that may arise in a single interaction.

A multinational survey including Indonesia highlights that phishing is a predominant form of digital fraud, with mobile devices serving as the primary vector due to their widespread accessibility and affordability, underscoring the increasing complexity and prevalence of online scams in the Indonesian context [5]. The bidirectional nature of voice phishing enables perpetrators to dynamically manipulate the conversational context. Such interactions are rich in semantic information from both speakers. However, if the transcription of a conversation is performed without speaker segmentation, the resulting text appears as a single block without any indication of who is speaking. This obscures the identification of roles within the dialogue and significantly reduces classification accuracy [6], [7]. Therefore, in this

*Corresponding author: Ahmad Hidayat (ahmad_hidayat@staff.gunadarma.ac.id)

DOI: <https://doi.org/10.47738/jads.v6i3.858>

This is an open access article under the CC-BY license (<https://creativecommons.org/licenses/by/4.0/>).

© Authors retain all copyrights

study, speaker diarization is employed as a crucial preprocessing step to segment utterances based on speaker identity, although the primary focus remains on the development and evaluation of the multi-label classification model. Previous studies have shown that modern diarization methods, such as End-to-End Neural Diarization (EEND) and DiarizationLM, can enhance the quality of speaker segmentation and thereby support the understanding of the structure in bidirectional conversations [8], [9]. Nevertheless, this study focuses on optimizing the performance of the multi-label classification model to detect various voice phishing schemes.

With the advancement of Natural Language Processing (NLP), transformer-based LLMs such as XLM-RoBERTa and ELECTRA have demonstrated strong performance in multilingual text classification and training efficiency. [10] reported that XLM-RoBERTa excels in cross-lingual classification tasks and in handling complex sentence structures. Meanwhile, [11] found that ELECTRA achieves competitive performance with significantly greater training efficiency than BERT, making it a lightweight yet powerful alternative.

Several previous studies have examined voice phishing classification based on conversational; however, most have focused on single-label approaches and the use of non-Indonesian language datasets. [12] utilized the Korean-language KorCCVi v2 dataset with transcription preprocessing and stop-word removal, applying a hybrid Attention 1D CNN-BiLSTM model that achieved an accuracy of 99.32% and an F1-score of 99.31%. [13] Employed the koBigBird-bert-base model within a Retrieval-Augmented Generation (RAG) framework using semi-automatic speaker segmentation, resulting in an accuracy of 95% and an F1-score of 93%. Another study by [14] used KoBERT on the KorCCVi dataset, achieving an accuracy of 99.60% and an F1-score of 99.57%. Classified phishing and non-phishing audio conversations in Indian languages, utilizing audio transcription preprocessing with Microsoft Azure Speech Translation followed by translation into English. The BERT model in this study achieved an accuracy of 94% [15]. However, all of these studies have remained focused on single-label classification, have not integrated preprocessing pipelines involving speaker segmentation, and have yet to utilize Indonesian language voice phishing data.

Based on this background, the present study aims to develop and compare the performance of XLM-RoBERTa and ELECTRA models in multi-label classification of voice phishing based on Indonesian-language telephone conversations. Preprocessing steps such as speaker diarization, automatic transcription, and text cleaning are applied as part of the input pipeline to generate optimal inputs for both models. The evaluation is conducted to identify various fraudulent schemes that may occur simultaneously within a single conversational interaction.

2. Method

This study focuses on the development of a multi-label classification model of voice phishing in Indonesian. The stages of this study are shown in figure 1.

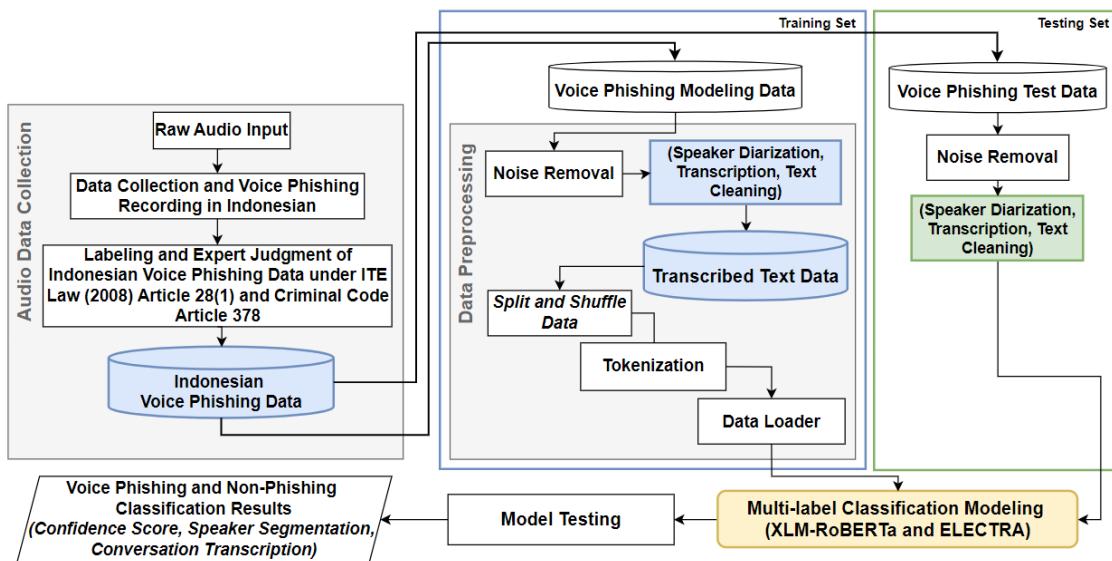


Figure 1. Research Stages

This research is initiated with the collection and recording of Indonesian voice phishing data, sourced from both public platforms and simulated recordings. The collected data is then subjected to labelling and validation through the process of expert judgement, with the objective of constructing an Indonesian voice phishing dataset. The collected data were then divided into two main groups, namely the training set and the testing set. Both sets underwent several preprocessing steps. The initial stage involved noise reduction to clean the audio data from irrelevant sound disturbances. Subsequently, the data were processed using speaker diarization to separate conversations based on speaker identity. Speech-to-text transcription was performed using the Whisper-large-v2 model [16], [17], followed by text cleaning to remove irrelevant characters or elements. The purpose of these steps is to ensure that the data is ready for further processing. The cleaned data is used to build an Indonesian voice phishing corpus dataset. This corpus is then split and shuffled, followed by tokenization and the creation of a data loader. The next stage involves multi-label classification modeling using various Large Language Models (LLMs) such as XLM-RoBERTa and ELECTRA, which are specifically adapted for the Indonesian language. Once the models are trained, they are subjected to a testing phase to produce classification results for voice phishing and non-phishing modes. The implementation and training of the XLM-RoBERTa and ELECTRA models in this study were carried out using the PyTorch deep learning framework and the HuggingFace Transformers library. The noise reduction preprocessing step utilized the DeepFilterNet3 model [18], speaker diarization was performed using pyannote/speaker-diarization-3.1, and transcription was conducted with Whisper-large-v2. The librosa library was used for audio processing, while scikit-learn was employed for data splitting and model performance evaluation.

Overall, this research contributes to the development of a multi-label classification system for Indonesian voice phishing, encompassing corpus dataset construction, integrated preprocessing stages, and the application and evaluation of multiple LLMs to achieve accurate classification across various phishing schemes.

2.1. Data Collection and Voice Phishing Recording in Indonesian

Data collection is sourced from YouTube platform phishing videos and voice phishing simulation recordings. The phishing videos obtained indicate digital fraud. The categories of methods used consist of phishing and non-phishing. Fraudulent videos and voice simulations are in accordance with the interpretation of fraud based on the Indonesian Criminal Code (KUHP) No. 378 [19] and Law Number 19 of 2016 on Electronic Information and Transactions (ITE Law) [20]. **Figure 2** shows the flow of the data collection process. The stages begin by searching for voice phishing videos taken from YouTube.

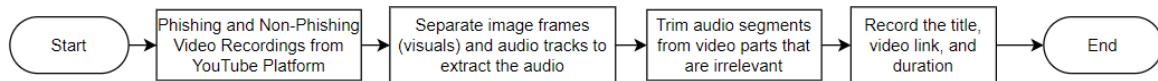


Figure 2. Data Collection Stages

Voice phishing video files consist of two main components, namely image frames (visuals) and audio tracks. Separation is done by extracting audio from video files using an online YouTube video conversion application tool, namely downloader.to (<https://downloaderto.com>). The results of the separation process aim to retrieve audio files. The separated audio is stored in MP3 format. The cutting process aims to remove irrelevant parts in the video, including opening or bumper videos, video clips and irrelevant sound parts in the conversation. The audio cutting process uses an audio editor application, namely audacity (<https://www.audacityteam.org>). The final stage is to document by recording information on the title, video link, and duration. The process of making a voice phishing simulation recording begins with compiling a script shown in **figure 3**.

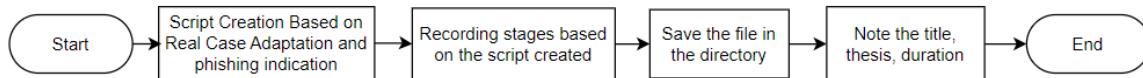


Figure 3. Stages of Creating a Phishing Script and Voice Recording

The preparation of the script adapts voice phishing practices that occur in the real world by analyzing fraud cases that occur with the interpretation of phishing fraud actions based on the Indonesian Criminal Code No. 378 and Law Number 19 of 2016. The script scenario created can describe two or more speakers. The speakers in the conversation consist of two main roles, namely one acting as a victim and the other acting as a phishing fraud perpetrator. The

process of recording the script results is carried out using a voice recording application or video conference application. The results of the voice phishing simulation recording are saved in MP3 format. Files are stored in storage media in the directory. The results of the recording simulation are documented with some title, script and duration information. The duration of each audio varies, adjusting to the length of the conversation context.

2.2. Labeling and Expert Judgement of Indonesian Voice Phishing Data

Labeling of data collected from YouTube data and making of Indonesian language phishing fraud simulation recordings based on expert judgment. This expert judgement process was carried out by a single expert with a background in law and information technology, aiming to ensure that the labeling aligns with the characteristics of fraudulent acts as regulated in Article 378 of the KUHP and Law Number 19 of 2016 on ITE Law. These regulations cover crimes such as phishing, including electronic fraud, identity forgery, and personal data theft. All labels were assigned manually based on the applicable legal criteria. However, the involvement of only one expert is a limitation, as it prevents the measurement of inter-rater agreement. This limitation should be addressed and is expected to be improved in future studies by involving more than one validator to enhance the validity of the data labeling process.

The stages of labeling Indonesian language voice phishing data begin with selecting audio according to the results of data collection. The process of listening to audio conversations to understand the context and meaning of the conversation aims to determine whether the conversation falls into the category of voice phishing or non-phishing. Labeling is carried out according to five voice phishing methods consisting of illegal online loan phishing, phishing under the guise of family crises (accidents, illness, drugs, police tickets), illegal investment phishing, buying and selling phishing (goods/services not arriving, goods/services not as described, money not reaching the seller, etc.), phishing under the guise of gifts and non-phishing. Each audio recording collected in this study was categorized into six main labels. These categories include: (1) illegal online loan voice phishing methods, labeled as 1_p_p_o; (2) family crisis disguised voice phishing methods, such as fraud involving fabricated emergencies, labeled as 2_p_b_k_k; (3) illegal investment voice phishing methods, labeled as 3_p_i_i; (4) buy-sell voice phishing methods related to goods or services, labeled as 4_p_j_b_j; (5) prize disguised voice phishing methods, labeled as 5_p_h; and (6) the non-phishing category, labeled as 6_n_p. These codes were used to facilitate the identification, grouping, and analysis of the data throughout the model training and evaluation process.

2.3. Indonesian Voice Phishing Dataset

The Indonesian voice phishing dataset comprises audio recordings illustrating diverse phishing scam methods and communication patterns in the Indonesian context, including non-phishing categories. Collected by the author from YouTube and simulated vishing recordings, the dataset contains 300 samples evenly distributed across six labels, all expertly annotated. For robust evaluation, the data were split into training, validation, and testing sets with proportions of 70:15:15 and 80:10:10, ensuring effective training and reliable assessment of model generalization.

2.4. Data Pre-processing

The pre-processing stage of voice data derived from voice phishing telephone conversations is carried out to ensure that the data is in a format suitable for model processing. This stage comprises several essential steps: noise reduction, speaker diarization, audio-to-text transcription, text normalization, data shuffling and partitioning, tokenization, and the construction of a data loader. Each of these steps is critical for producing high-quality input that enables the model to achieve optimal performance in voice phishing classification tasks. In this study, the speaker diarization process was conducted prior to transcription using Whisper-large-v2. The audio was first segmented based on speaker identity, and then each diarization segment was transcribed separately with Whisper. This approach allows the transcription results to be clearly mapped to each speaker in the dialogue. To ensure accuracy, the results of diarization and transcription were manually validated and corrected before proceeding to the labeling and modeling stages. To clarify the data processing flow before entering the tokenization and classification stages, the following presents a structured algorithm that describes how speaker diarization, transcription, and text cleaning are performed:

Algorithm 1. Speaker Diarization, Audio Transcription, and Text Cleaning Pipeline

Input: A folder containing several audio files (.wav or .mp3)

Output: Cleaned transcribed text per speaker

1. Start
2. Prepare the directory containing the audio files
3. Initialize the Whisper-large-v2 model for transcription
4. Initialize the pyannote/speaker-diarization version 3 pipeline
5. For each audio file:
 - 5.1 Perform speaker diarization to detect speaker segments
 - 5.2 Segment the audio based on speaker diarization results
 - 5.3 For each segment:
 - 5.3.1 Transcribe with Whisper-large-v2
 - 5.3.2 Save the transcription text along with speaker label and timestamp
 - 5.4 Concatenate transcription results in chronological order
 - 5.5 Clean the transcribed text (remove fillers, normalize)
 - 5.6 Save the final results in .csv format
6. Repeat for all files
7. End

Input: $F = \{f_i \mid i = 1, \dots, N\}$, $f_i \in \{\text{.wav, .mp3}\}$

Output: $C = \{c_i \mid i = 1, \dots, N\}$, c_i = cleaned transcribed text for audio f_i

Initialization:

$M_T \leftarrow$ Whisper-large-v2 transcription model

$M_D \leftarrow$ pyannote speaker diarization pipeline version 3

For each audio file $f_i \in F$:

$S_i = M_D(f_i) = \{s_{i,j} \mid j = 1, \dots, K_i\}$ (speaker segments with labels and timestamps)

For each segment $s_{i,j} \in S_i$:

$T_{i,j} = M_T(s_{i,j})$ (transcribed text of segment)

$L_{i,j} = (sp_{i,j}, t_{start}^{i,j}, t_{end}^{i,j}, T_{i,j})$

$\tilde{c}_i = sort_t(\{L_{i,1}, L_{i,2}, \dots, L_{i,K_i}\})$ (sorted based on $t_{start}^{i,j}$)

$c_i = \text{clean } (\tilde{c}_i)$ (text cleaning: filler removal, normalization)

save (c_i , format=csv)

Algorithm 1 ensures that transcription results are segmented based on speaker identity and systematically cleaned to guarantee consistency and accuracy during the classification stage. The text cleaning stage specifically focuses on removing irrelevant characters or elements without altering the core meaning of the conversation. The first step is to delete any character that appears more than five times consecutively, simplifying patterns such as "AAAAAAA" to "A" in order to avoid unnatural text characteristics. Next, repetitive character patterns that are not relevant to the conversational context, such as "WUHUUHUUHUUH" (which are typically only emotional expressions without clear semantic meaning), are removed. The following step reduces word repetitions that occur more than three times in a row, word repetitions of up to three times are retained as they can represent important expressions of intensity or emotion in communication, such as in the phrase "Very very very good!". In addition, excessive spaces are removed to ensure the text structure is neat and easy to read. This cleaning process does not employ stemming or stopword removal techniques, aiming to preserve the completeness of meaning and context in each conversation.

2.5. Multi-label modeling

In this study, the pre-trained XLM-RoBERTa and ELECTRA models were compared to obtain the best performance. XLM-RoBERTa (Cross-lingual Masked RoBERTa) is a multilingual version of RoBERTa developed by Facebook AI (Meta AI), and trained on more than 100 languages using large data from CommonCrawl [21], [22]. The ELECTRA model used is "google/electra-base-discriminator". This model facilitates the transformation of the dataset into a format

compatible with the model's capabilities. This model is designed for cross-lingual NLP tasks such as text classification, translation, and natural logic inference between languages [23], [24]. The stages of forming a multi-label classification model for Indonesian language voice phishing are shown in [figure 4](#).



Figure 4. Stages of Multi-label Classification Modeling for Indonesian Voice Phishing

The stages consist of model and tokenizer initialization, training and validation, and testing. The process begins with the model and tokenizer initialization stage using the XLM-RoBERTa and ELECTRA models. The model training and validation process begins by creating a training function and a validation function. Each function will be called by the training and validation functions by adjusting the parameters of the number of epochs, learning rate, batch size, and data ratio. Model testing aims to evaluate the performance of the model on data that is not used in the training process to ensure the generalization ability of the model, as well as to detect any tendencies of overfitting or prediction bias. Hyperparameters and model configurations are determined to optimize the performance of the multi-label classification of Indonesian language voice phishing. At the training stage, hyperparameters are determined with a combination of learning rate, batch size, epoch and data ratio parameters.

During the model training phase for multi-label voice phishing classification, several combinations of key hyperparameters were explored to obtain the best configuration. The parameters tested included the learning rate (2e-5 and 5e-5), batch size (8 and 16), number of epochs (5, 25, and 50), and the data split ratios (70:15:15 and 80:10:10 for training, validation, and testing, respectively). The chosen learning rate values are commonly used in training transformer-based architectures. All combinations were evaluated using performance metrics such as accuracy, precision, recall, and F1-score to determine the most optimal settings for both XLM-RoBERTa and ELECTRA models in the multi-label classification of Indonesian voice phishing conversations.

The label encoding process for multi-label classification was carried out using the binary relevance approach. Each conversation that had undergone preprocessing was converted into a six-dimensional one-hot vector, representing five voice phishing categories (illegal online loans, family crises, illegal investments, buying and selling, prizes) and one non-phishing category. A value of 1 in each element of the vector indicates the presence of a particular label in the data, while a value of 0 denotes the absence of the corresponding label. Thus, each conversation can have more than one active label simultaneously, in accordance with the characteristics of real-world voice phishing cases. The binary relevance approach was chosen because it decomposes the multi-label problem into several independent binary classification tasks, so that each label is predicted separately (one-vs-rest). In addition to its simplicity and implementation efficiency, this method is also well-suited for datasets with a limited number of samples and uneven label distribution (sparsity), as is the case with the dataset used in this study. The one-hot vectors resulting from label encoding were then used as targets for training the XLM-RoBERTa and ELECTRA models based on LLMs.

All of these steps were undertaken to ensure the validity and generalizability of the model results, despite the limited data size. To minimize the risk of overfitting due to the limited sample size (300 samples), several strategies were implemented in this study: (1) the data were divided into training, validation, and test sets using two scenarios, namely 70:15:15 and 80:10:10, to ensure a comprehensive evaluation of model performance on unseen data; (2) early stopping was applied by monitoring the validation loss, so that training was halted if there was no improvement in loss over several epochs; (3) hyperparameters such as learning rate, batch size, and number of epochs were systematically tuned to balance model complexity and generalization; and (4) model performance was evaluated on the training, validation, and test data to detect any generalization gaps. This combination of steps effectively reduced the risk of overfitting and ensured that the reported results reflected the actual generalization ability of the models.

Each data point was represented by a six-dimensional one-hot encoded vector, allowing each instance to have more than one active label simultaneously. During model training, the BCEWithLogitsLoss (binary cross-entropy) loss function was used, which calculates the loss for each label independently, making it suitable for handling multi-label classification in voice phishing data.

2.6. Model Testing

The model testing stage is carried out to evaluate the performance of the Indonesian language multi-label voice phishing classification model that has been formed. The model testing stage is shown in figure 5.

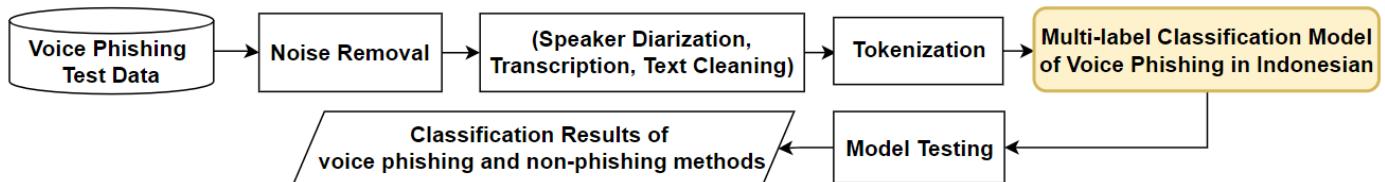


Figure 5. Voice Phishing Model Testing Phase

The testing process begins with preparing voice phishing test data in the form of audio files in WAV or MP3 format including voice recordings in the phishing and non-phishing methods categories. This data goes through a noise removal stage to ensure that the audio quality is clean and not disturbed by external sound interference. Furthermore, the audio data is segmented or separated from the speakers in the voice recording, transcribed audio to text using the whisper-large-v2 model and followed by text cleaning to remove irrelevant characters or elements. The next stage is tokenization, where the cleaned text is processed using the XLM-RoBERTa Tokenization method with the addition of special tokens to capture the context of the conversation in more depth. The results of this tokenization are then entered into the multi-label classification model that has been formed. The model is tested with test data to identify its performance in classifying conversations into several voice phishing categories. The results of this test are the main indicators in evaluating the accuracy, precision, and generalization ability of the model in classifying voice phishing.

2.7. Classification Results

The classification result stage of the model testing process produces output in text form that includes the name of the audio file being tested, so that each classification result can be traced back to the original test data. In addition, the model provides a classification for all categories of voice phishing methods with a confidence score indicating the level of confidence of the model in the predictions given. Other output results include speaker segmentation containing information about the start and end times of each conversation segment, speaker labels. In addition, transcription results are included for each segment, providing a text representation of the voice conversation being tested

3. Results and Discussion

This study developed a multi-label classification system for Indonesian language voice phishing based on the LLM. Based on figure 1, the results of the study are:

3.1. Voice Phishing Data Collection and Recording Results

The results of voice phishing data collection are sourced from the YouTube platform and recording voice phishing simulations. The video data that was successfully collected from the YouTube platform was 102 videos with details of 62 phishing category videos and 40 non-phishing. Part of the results are shown in table 1.

Table 1. Part of the results of collecting phishing video data from the YouTube platform

Title	Link	Duration (minute)	Category
How to Deal with Illegal Online Loan Debt Collectors Recording of Debt Collection by Illegal Online Loan App Go Kredit;	https://www.youtube.com/watch?v=1q5hzPaphU&list=PLNmRHPNj5JLWuLv_7udbZpnqb c50Sc1VW&index=44	6.28	phishing
Scammer Pretending to Be a Police Officer Gets Scammed	https://www.youtube.com/watch?v=jmnm8O92q1c&list=PLlb-kPR4_SQs0O46VO FapauxBPI83qAr&index=26	5.54	phishing
Phone Conversation of Company (Business Communication)	https://www.youtube.com/watch?v=6ALWZr2eb7U	2.08	non-phishing

Practice of Receiving & Making Phone Calls	https://www.youtube.com/watch?v=GfJnWHIC5m0	4	non-phishing
--	---	---	--------------

Table 1 presents data from the results of collecting phishing videos from the YouTube platform including some information on video titles, video links, duration and categories. The title represents various phishing modes. The video source link is used as a reference. Duration describes how long the video is displayed. The category indicates the type of phishing and non-phishing. **Table 2** illustrates the result presented is the development of a script section for simulated voice phishing recordings, consisting of a title, a snippet of the conversation script, and duration.

Table 2. Part of the Results of Script Simulated Voice Phishing Recordings

Title	Portions of conversation	Duration (minute)
Family Debt Scam: Claim Money for Emergency Handling	Scammer: "Thank you for answering my call. I am Haris, an old friend of your father, Mr. Wibowo. I'm calling because he is facing a big problem right now. Your father is currently being held by the lender due to delayed loan payments, and they are threatening to take this matter to court if it's not resolved within 24 hours. "Victim: "What? My father is being held? Why didn't I know about this? Why didn't anyone inform me before?" Scammer: "I understand your confusion, but this issue came up suddenly. I tried to contact other family members, but only you, Tika, were reachable for us to resolve this matter." Victim: "Why would my father borrow money? As far as I know, he never talked about any debts."	3.25

The title represents various phishing schemes. The script snippet contains dialogues to be recorded under two categories: voice phishing and non-phishing. The recording process is carried out using voice recording applications or video conferencing tools. The duration indicates the length of the recorded conversation. A total of 300 audio recordings were successfully collected for this study, comprising 250 voice phishing recordings and 50 non-phishing recordings. Of these, 102 audio files were sourced from the YouTube platform, with 62 labeled as voice phishing and 40 as non-phishing. The remaining 198 audio files were obtained from simulation recordings, consisting of 188 voice phishing and 10 non-phishing samples.

3.2. Data Labeling Results and Expert Judgment

The results of data labeling are carried out by expert judgment as legal and information technology experts with the aim of ensuring that labeling is in accordance with the characteristics of fraudulent acts regulated in Article 378 of the KUHP and Law Number 19 of 2016 ITE Law. **Table 3** presents the results of labeling voice phishing videos taken from YouTube sources. The table contains information about the title, audio file name, and label. The title represents various phishing modes. File naming is done according to a predetermined format such as audio-p-m1-1.mp3.

Table 3. Results of Labeling Videos from YouTube Sources

Title	File Name	Label
How to Deal with Illegal Online Loan Debt Collectors Recording of Debt Collection via Illegal Online Loan App Go Kredit	audio-p-m1-3.mp3	1_p_p_o
Online Cooperative Fraud Miss Bella Hasky	audio-p-m1-4.mp3	1_p_p_o
Scammer Pretending to Be a Police Officer Trying to Scam but Gets Scammed Instead	audio-p-m2-1.mp3	2_p_b_k_k
Beware of Phone Scams, Finally the Scammer Gives Up	audio-p-m2-2.mp3	2_p_b_k_k
Pranking the Scammer on the Phone Until the Scammer Gets Angry and Upset!!	audio-p-m4-2.mp3	4_p_j_b_j

Table 4 presents the results of labeling the voice phishing simulation recording script. The table contains information about the title, audio file name, conversation snippets and labels. The title represents various phishing modes. File naming is done according to a predetermined format such as audio-p-m1-1.mp3. The conversation snippets describe the voice phishing recording conversation script. The first line is a script with the title "Family Fraud Trapped in Debt: Claim Money for Emergency Handling" which is voice phishing that has more than one mode. In the script there are two contexts of the mode, namely phishing under the guise of a family crisis and phishing of illegal online loans. The

perpetrator pretends to be an old friend of the victim's father and informs that the father is being detained due to a debt of Rp15,000,000 to the Loan provider. With an urgent tone, the perpetrator manipulates the victim's emotions and asks the victim to transfer money for repayment, accompanied by a request for personal data, namely KTP number, email, account number, and OTP code. The panicked victim finally provides information and makes a transfer to the perpetrator's account. This phishing fraud piercing illustrates how the perpetrator exploits family relationships and emergency situations convincingly in order to obtain illegal financial gain.

Table 4. Recording Script Labeling Results Section

Title	File Name	Label
Family Debt Trap Scam: Money Claim for Emergency Handling	Scammer: "Thank you for answering my call. I am Haris, an old friend of your father, Mr. Wibowo. I'm calling because he is facing a serious problem right now. Your father is currently being held by the lender due to overdue loan payments, and they are threatening to take this issue to court if it's not resolved within 24 hours." Victim: "What? My father is being held? Why didn't I know about this? Why didn't anyone inform me earlier?" Scammer: "I understand your confusion, but this issue just happened suddenly. I tried to contact other family members, but only you, Tika, are reachable for us to resolve this matter."	2_p_b_k_k, 1_p_p_o

Based on the expert labeling results, a total of 300 audio samples were analyzed in this study. Of these, 180 samples were single-label data, meaning that each audio file was classified into only one category of voice phishing or non-phishing. Meanwhile, 120 samples were multi-label, classified into two or more categories simultaneously within a single conversation. The distribution of label counts for each category shows that there are 74 labels for each voice phishing category: illegal online loans, family crises, illegal investments, buying and selling, and prizes. In the non-phishing category, there are 50 labels. Thus, the total number of labels present in this dataset is 420, which exceeds the number of samples due to the existence of multi-label data, where a single sample may have more than one label as a result of combined schemes within a conversation. To provide a visual overview of this distribution, [figure 6](#) presents a bar chart illustrating the number of labels in each category of voice phishing and non-phishing as determined by expert labeling.

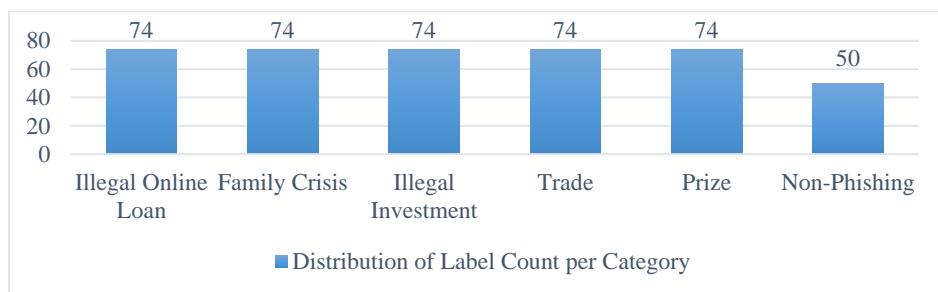


Figure 6. Bar Chart Showing the Distribution of Labels In Each Category of Voice Phishing and Non-Phishing

This visualization demonstrates that each voice phishing category has a balanced number of labels (74 each), while the non-phishing category contains 50 labels. This indicates that the label distribution in the dataset is fairly even across the phishing schemes, highlighting the importance of a multi-label approach in classifying complex voice phishing conversations.

3.3. Data Pre-Processing Results

The pre-processing stage consists of noise removal, speaker diarization, transcription, text cleaning, formation of Indonesian language voice phishing corpus, split shuffle data, tokenization, and data loader creation. The results of noise removal using the DeepFilterNet3 deep learning model. In line with the research objective, which focuses on the development of a multi-label voice phishing classification model based on Indonesian-language conversations, this study did not conduct experiments comparing classification performance before and after the integration of speaker diarization. The evaluation of speaker diarization, transcription, and text cleaning using Bilingual Evaluation

Understudy (BLEU) and Metric for Evaluation of Translation with Explicit ORdering (METEOR) was intended to validate the correspondence of the transcription results with the manual references prior to their use in the classification process. The results of the speaker diarization, transcription and text cleaning processes are carried out for all audio files resulting from noise cleaning. The text formed from the results of speaker diarization, transcription and text cleaning are validated by expert judgment and evaluated using the BLEU and METEOR metrics which are used to evaluate the quality of text translation produced by the transcription model. The data will become a corpus of Indonesian language voice phishing conversations.

Table 5 shows the results of the performance evaluation of the transcription and speaker segmentation models using BLEU and METEOR. Row one of audio-*np-21.mp3* contains a conversation about the online loan voice phishing methods that is transcribed and annotated based on the speaker. The speaker segmentation transcription snippet column shows a comparison of the source text (manual reference) and the results of the transcription model. In the source transcription, the conversation is marked by two main speakers, namely [SPEAKER_00] and [SPEAKER_01], while in the model results there is an additional segment [SPEAKER_02] which indicates additional speaker detection or segmentation errors from the model. The BLEU value of 97.77 and METEOR of 92.76 indicate that the results of the transcription and speaker segmentation of the model are very close to the manual reference text. Based on thresholds established in previous studies, BLEU scores above 40–50 and METEOR scores above 50–60 are considered very good for ASR systems and machine translation [25], [26]. Therefore, the BLEU and METEOR scores obtained in this study indicate that the transcription results are sufficiently reliable to be used in the subsequent classification stage.

Table 5. BLEU METEOR Evaluation Results

File Name	Transcription Chunks and Speaker Segmentation		Grade	
	Source	Model Results	BLEU	METEOR
audio- np- 21.mp3	[SPEAKER_00] Good afternoon, is this Suralinga company? [SPEAKER_01] Good afternoon, yes this is the admin section of Suralinga company. How can I help you? [SPEAKER_00] I want to order the products your company offers. Where can I place an order? [SPEAKER_01] Okay, I will connect you to the ordering sub-department. Please wait a moment.	[SPEAKER_00] Good afternoon, is this Suralinga company? [SPEAKER_01] Good afternoon, yes, this is the admin department of Suralinga company. How can I assist you? [SPEAKER_00] I want to order the products your company offers. Where can I place an order? [SPEAKER_01] Okay, I will connect you to the installation sub-department. Please wait a moment.	97.77	92.76
audio- p-m1- 41.mp3	[SPEAKER_00] Good morning, Sir/Madam! Is this the registered number for our online loan service? [SPEAKER_01] Morning, yes that's correct. Who am I speaking with? [SPEAKER_00] Let me introduce myself, I'm Aldo from Dana Instan, a trusted loan platform that has partnered with many partners across Indonesia. We would like to inform you that you are eligible for a special loan offer up to Rp10 million with very low interest. We offer hassle-free loans, so we want to know if you currently need additional funds for any particular needs?	[SPEAKER_00] Good morning, Ma'am. Is this the registered number for our online loan service? [SPEAKER_02] Hello, good morning, yes that's correct. Who am I speaking with? [SPEAKER_00] Yes, Ma'am. Let me introduce myself, my name is Aldo from Dana Instan, a trusted loan platform that has partnered with many partners across Indonesia. We want to inform you that you are eligible for a special loan offer up to 10 million rupiah with very low interest. We offer hassle-free loans, so we want to know, Ma'am, do you currently need additional funds for any particular needs?	92.67	75.95

Table 6 provides an example of one-hot encoding representation for multi-label classification in voice phishing tasks. Each conversation is represented as a six-dimensional binary vector, where a value of 1 indicates the presence of a particular label and a value of 0 indicates that the label is not present in the conversation.

Table 6. Example of One-Hot Encoding for Multi-label Classification

Data ID	Illegal Online Loans	Family Crisis	Illegal Investment	Buying and selling	Prize	Non-Phishing
Data 1	1	0	0	1	0	0

Data 2	0	1	1	0	0	0
Data 3	0	0	0	0	0	1
Data 4	1	1	0	0	0	0
Data 5	0	1	1	0	0	0

3.4. Multi-label Voice Phishing Modeling Results

Modeling results using XLM-RoBERTa and ELECTRA models. The models are compared to obtain the best performance in performing multi-label classification of Indonesian language voice phishing. The test results are carried out from a dataset of 300 voice phishing data. Testing is carried out using the DGX A100 machine. Each model was tested using the three best hyperparameter scenarios, with each scenario executed in three separate runs using different random data splits. The optimized parameters included learning rate, batch size, number of epochs, and data split ratio.

Based on the test results presented in [table 7](#), the XLM-RoBERTa model consistently achieved higher accuracy and F1-score values compared to the ELECTRA model across all best scenarios. The optimal scenario for XLM-RoBERTa, with a learning rate of 2e-5, batch size of 16, 50 epochs, and a data split ratio of 80:10:10, recorded an average accuracy of $97.04\% \pm 1.15$ and an F1-score of $92.66\% \pm 2.59$. Additionally, XLM-RoBERTa demonstrated good performance stability, as reflected by the relatively small standard deviation in each key metric. On the other hand, the ELECTRA model also showed competitive performance, with its highest average accuracy reaching $93.52\% \pm 0.32$ and F1-score $84.10\% \pm 1.26$ in its best scenario.

Table 7. Summary of the Mean and Standard Deviation of Test Results for the Three Best Scenarios of the Models

Model	Learning Rate	Batch Size	Epoch	Data ratio	Accuracy (%)	Recall (%)	F1-Score (%)
XLM-RoBERTa	2e-5	16	50	80:10:10	97.04 ± 1.15	92.65 ± 4.26	92.66 ± 2.59
XLM-RoBERTa	2e-5	16	25	80:10:10	96.11 ± 2.00	88.37 ± 8.79	90.13 ± 4.77
XLM-RoBERTa	2e-5	16	50	70:15:15	85.19 ± 2.50	74.61 ± 7.50	71.19 ± 8.12
ELECTRA	5e-5	8	50	80:10:10	93.52 ± 0.32	85.96 ± 3.49	84.10 ± 1.26
ELECTRA	2e-5	8	50	80:10:10	92.96 ± 0.64	86.06 ± 0.99	82.96 ± 2.24

Overall, these experimental results confirm that XLM-RoBERTa is more optimal for multi-label voice phishing classification tasks on Indonesian conversational data. The low standard deviation values in most scenarios indicate that both models are capable of producing consistent and stable results, even when tested with different random data splits. To provide a more detailed performance overview for each label as well as the macro/micro averages, the following section presents the classification reports for both models on [table 8](#).

Table 8. Classification Report of XLM-RoBERTa and ELECTRA Models on Test Data Using the Best Hyperparameter

Label / Average	XLM-RoBERTa				ELECTRA			
	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support
1_p_p_o	1.00	1.00	1.00	7	0.71	0.83	0.77	6
2_p_b_k_k	1.00	1.00	1.00	6	1.00	0.67	0.80	3
3_p_i_i	1.00	1.00	1.00	2	0.80	0.80	0.80	5
4_p_j_b_j	1.00	0.86	0.92	7	0.62	0.83	0.71	6
5_p_h	1.00	1.00	1.00	10	0.83	0.83	0.83	6
6_n_p	0.50	1.00	0.67	2	0.80	0.89	0.84	9
Micro avg	0.94	0.97	0.96	34	0.76	0.83	0.79	35
Macro avg	0.92	0.98	0.93	34	0.80	0.81	0.79	35
Weighted avg	0.97	0.97	0.96	34	0.78	0.83	0.80	35
Samples avg	0.97	0.98	0.97	34	0.81	0.88	0.82	35

On [table 8](#) the high and consistent precision, recall, and F1-scores on the test data, particularly the macro average F1-score of 0.93 and micro average recall of 0.97 for the XLM-RoBERTa model, indicate good generalization without overfitting. Overfitting would show a notable performance drop on test data, which is not present here. Balanced metrics across all labels, including those with few samples, further confirm the model's robustness on imbalanced and unseen data. The XLM-RoBERTa model achieved a macro average F1-score of 0.93 and a micro average recall of 0.97, indicating that this model performs very well and consistently in recognizing all labels in multi-label classification tasks. The ELECTRA model achieved a macro average F1-score of 0.79 and a micro average recall of 0.83, which reflects reasonably good performance but less consistency compared to XLM-RoBERTa in recognizing all labels. These results indicate that both models are capable of delivering good performance in multi-label classification scenarios. Further evaluation using the classification report demonstrates that XLM-RoBERTa achieved higher macro and micro average F1-scores than ELECTRA. This suggests that XLM-RoBERTa offers better consistency in recognizing all labels, including those with fewer data samples. Therefore, the XLM-RoBERTa model can be considered more optimal for multi-label voice phishing classification in Indonesian, especially on imbalanced datasets.

The visualization of the confusion matrix in [figure 7](#) illustrates the distribution of correct and incorrect predictions for each voice phishing category for the best-performing model, XLM-RoBERTa. This confusion matrix shows that the model tends to achieve a high rate of correct predictions across most categories. Nevertheless, there are still some cases of misclassification for certain labels, such as 4_p_j_b_j (buying and selling phishing) and 6_n_p (non-phishing), indicating areas that could be improved in future model development. A detailed analysis of the error distribution in this confusion matrix provides additional insights into the strengths and weaknesses of the model in handling multi-label data, particularly for categories with limited data samples.

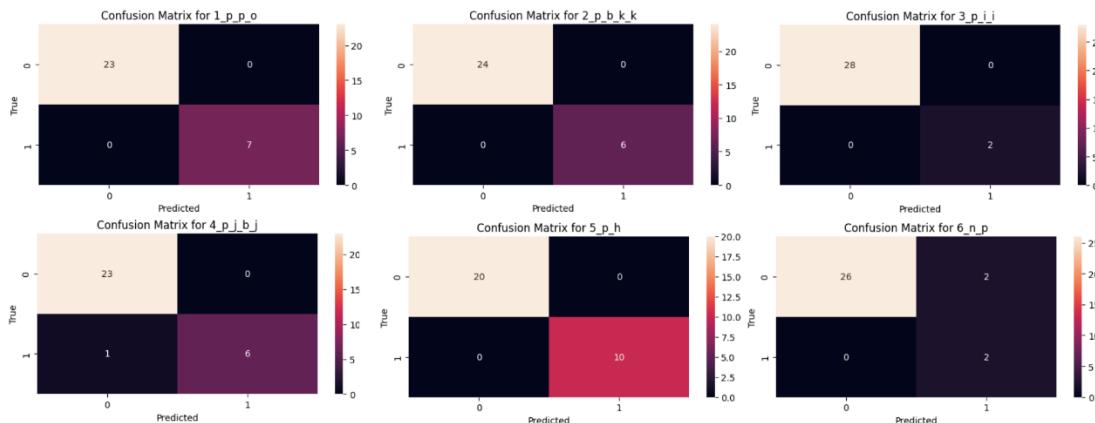


Figure 7. Confusion Matrix the XLM-RoBERTa Model

[Figure 8](#) shows the progression of accuracy and loss for the model on the training and validation data over 50 epochs. The curves indicate that the model's accuracy increases consistently and stably, while the loss for both the training and validation data decreases, suggesting that the training process proceeds well without significant signs of overfitting. This indicates that the training process is optimal and that the model is able to generalize well to the test data.

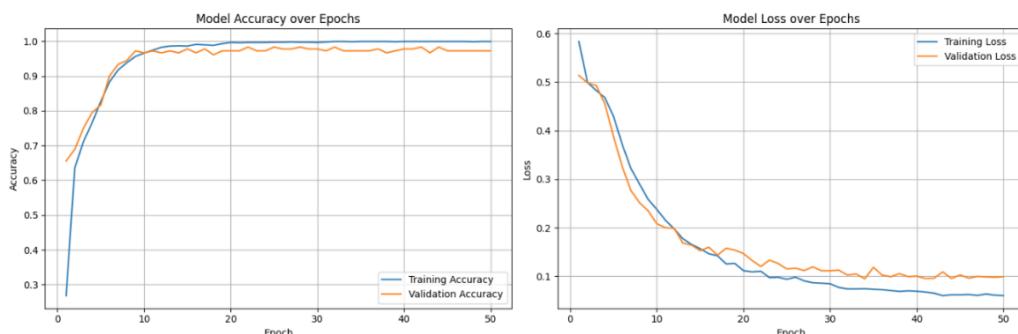


Figure 8. Training and Validation Accuracy and Loss Curves of the XLM-RoBERTa Model

Overall, both XLM-RoBERTa and ELECTRA demonstrate good classification performance across most categories. The XLM-RoBERTa model achieved the highest overall accuracy and F1-score on the test data and exhibited better consistency in recognizing all labels, including minority labels. This makes XLM-RoBERTa particularly suitable for applications requiring high accuracy and stable performance on datasets with imbalanced label distributions. Meanwhile, the ELECTRA model remains a viable alternative, although its performance was not as high as XLM-RoBERTa in this evaluation.

3.5. Evaluation Results of the Best-Performing Model

The best-performing model, XLM-RoBERTa, was evaluated using three test audio files that had not been previously encountered during model training (unseen data). The test set consisted of two voice phishing samples and one non-phishing sample. Each file was processed to obtain predicted labels along with the corresponding confidence scores (the model's level of certainty for each label) within a range of 0–100%. The predicted results were then compared to the initial (manual) labels as a reference for evaluation (see [table 9](#)).

Table 9. Model Evaluation Results on 3 Test Audio Files

No	File Name	Manual Label	Model Output	Manual vs Model Match
1	audio-tp-m1-m5.mp3	1_p_p_o, 5_p_h	1_p_p_o (57.55%), 5_p_h (74.55%)	Match (all manual labels detected)
2	audio-tp-m1-m2-2.mp3	1_p_p_o, 2_p_b_k_k	1_p_p_o (69.84%), 2_p_b_k_k (70.61%)	Match (all manual labels detected)
3	audio-tnp-m6-1.mp3	6_n_p	6_n_p (90.25%)	Match (manual label detected)

Based on the evaluation of the three audio files, all manual labels in the test data were successfully detected by the XLM-RoBERTa model with confidence scores above 50%. This demonstrates the model's strong performance in accurately identifying both voice phishing and non-phishing scenarios in previously unseen data.

4. Conclusion

This study successfully developed a multi-label classification model for voice phishing based on LLMs using Indonesian-language telephone conversations. The resulting model is capable of classifying voice phishing into six main categories: illegal online loan phishing, family crisis, illegal investment, buying and selling, prize, and non-phishing. The dataset, consisting of 300 samples, demonstrates that both XLM-RoBERTa and ELECTRA exhibit competitive performance. The XLM-RoBERTa model achieved the highest average accuracy across all test data at $97.04 \pm 1.15\%$, as well as the highest average F1-score at $92.66 \pm 2.59\%$. According to the classification report, XLM-RoBERTa also showed higher macro and micro average F1-scores compared to ELECTRA (0.93 for XLM-RoBERTa and 0.79 for ELECTRA), indicating better performance consistency across all labels, including minority labels and in imbalanced data scenarios. Therefore, XLM-RoBERTa is recommended for implementation in voice phishing detection systems that require high accuracy and consistency across various categories. Meanwhile, ELECTRA remains a viable alternative, particularly in situations with limited computational resources or where a lighter implementation is needed. The final model selection should be tailored to the specific needs of real-world applications, while taking into account data privacy protection, computational efficiency, and false positive risk mitigation.

5. Declarations

5.1. Author Contributions

Conceptualization: A.H., S.M.; Methodology: A.H., H.; Software: A.H.; Validation: S.M., H.; Formal Analysis: A.H.; Investigation: A.H.; Resources: S.M., H.; Data Curation: A.H.; Writing – Original Draft Preparation: A.H.; Writing – Review and Editing: S.M., H.; Visualization: A.H.; All authors have read and agreed to the published version of the manuscript.

5.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

5.3. Funding

The author would like to express his sincere gratitude to Mr. Andika D. Riyandi, B.L., B.Eng for his invaluable role as a validator, whose expertise and guidance have greatly contributed to the success of this research. The author would also like to express his deepest appreciation to Gunadarma University for its continuous support and conducive academic environment that made this research possible. The facilities, encouragement, and opportunities provided by this institution have been important factors in achieving the research objectives.

5.4. Institutional Review Board Statement

Not applicable.

5.5. Informed Consent Statement

Not applicable.

5.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- [1] M. F. Arroyabe, C. F. A. Arranz, I. F. de Arroyabe, and J. C. F. de Arroyabe, "Revealing the realities of cybercrime in small and medium enterprises: Understanding fear and taxonomic perspectives," *Computers & Security*, vol. 141, Art. no. 103826, pp. 1-12, 2024, doi: 10.1016/j.cose.2024.103826.
- [2] Y. Zhou, M. Tiwari, A. Bernot, and K. Lin, "Metacrime and Cybercrime: Exploring the Convergence and Divergence in Digital Criminality," *Asian Journal of Criminology*, vol. 19, no. 3, pp. 419–439, Sep. 2024, doi: 10.1007/s11417-024-09436-y.
- [3] F. Salahdine and N. Kaabouch, "Social engineering attacks: A survey," *Future Internet*, vol. 11, no. 4, Art. no. 89, pp. 1-12, 2019, doi: 10.3390/fi11040089.
- [4] A. Triantafyllopoulos, A. A. Spiesberger, I. Tsangko, X. Jing, V. Distler, F. Dietz, F. Alt, and B. W. Schuller, "Vishing: Detecting social engineering in spoken communication — A first survey & urgent roadmap to address an emerging societal challenge," *Computer Speech & Language*, vol. 94, Art. no. 101802, pp. 1-12, 2025, doi: 10.1016/j.csl.2025.101802.
- [5] M. Houtti, A. Roy, V. N. R. Gangula, and A. Walker, "A survey of scam exposure, victimization, types, vectors, and reporting in 12 countries," *J. Online Trust Saf.*, vol. 2, no. 4, Art. no. 204, pp. 1-12, Sep. 2024, doi: 10.54501/jots.v2i4.204.
- [6] D. O'Shaughnessy, "Speaker Diarization: A Review of Objectives and Methods," *Applied Sciences*, vol. 15, no. 4, Art. no. 2002, pp. 1-12, 2025, doi: 10.3390/app15042002.
- [7] J. Boyd, M. Fahim, and O. Olukoya, "Voice spoofing detection for multiclass attack classification using deep learning," *Machine Learning with Applications*, vol. 14, Art. no. 100503, pp. 1-12, 2023, doi: 10.1016/j.mlwa.2023.100503.
- [8] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with self-attention," in *Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, vol. 2019, no. 12, pp. 296–303, 2019, doi: 10.1109/ASRU46091.2019.9003959.
- [9] V. Khoma, Y. Khoma, V. Brydinskyi, and A. Konovalov, "Development of Supervised Speaker Diarization System Based on the PyAnnote Audio Processing Library," *Sensors*, vol. 23, no. 4, Art. no. 2082, pp. 1-12, 2023. doi: 10.3390/s23042082.
- [10] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised Cross-lingual Representation Learning at Scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, vol. 2020, no. 4, pp. 8440–8451, Online, Jul. 2020, doi: 10.18653/v1/2020.acl-main.747.
- [11] K. Clark, M. T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators," in *Proc. 8th International Conference on Learning Representations, ICLR*, vol. 2020, no. 3, pp. 1-12, Mar. 2020, doi: 10.48550/arXiv.2003.10555.
- [12] M. K. M. Boussougou and D.-J. Park, "Attention-Based 1D CNN-BiLSTM Hybrid Model Enhanced with FastText Word Embedding for Korean Voice Phishing Detection," *Mathematics*, vol. 11, no. 14, Art. no. 3217, pp. 1-12, 2023, doi: 10.3390/math11143217.

- [13] J.-M. Lee, Y. Baek, M.-S. Baek, H. Park, S. Byon, and E.-S. Jung, "Personalized Response System for Different Voice Phishing Types: Utilizing a Retrieval-Augmented Generation Model," in *Proceedings of the 2024 15th International Conference on Information and Communication Technology Convergence (ICTC)*, vol. 2024, no. 10, pp. 699–702, 2024. doi: 10.1109/ICTC62082.2024.10827466.
- [14] M. K. M. Boussougou and D.-J. Park, "Exploiting Korean Language Model to Improve Korean Voice Phishing Detection," *KIPS Transactions on Software and Data Engineering*, vol. 11, no. 10, pp. 437–446, 2022, doi: 10.3745/KTSDE.2022.11.10.437.
- [15] D. Naidu, "Voice analysis system for detection of vishing using deep learning," *International Journal of Health Sciences*, vol. 6, no. S1, pp. 10457–10466, May 2022, doi: 10.53730/ijhs.v6nS1.7520.
- [16] S. O. Russell, I. Gessinger, A. Krason, G. Vigliocco, and N. Harte, "What automatic speech recognition can and cannot do for conversational speech transcription," *Research Methods in Applied Linguistics*, vol. 3, no. 3, Art. no. 100163, pp. 1-12, 2024, doi: 10.1016/j.rmal.2024.100163.
- [17] L. A. Kumar, D. K. Renuka, and M. C. S. Priya, "Towards robust speech recognition model using deep learning," in *Proceedings of the 2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS)*, vol. 2023, no. 5, pp. 253–256, 2023, doi: 10.1109/ICISCoIS56541.2023.10100390.
- [18] S. S. Hashemi, M. Asadi, and M. Aghabozorgi, "An Audio-Visual System for Sound Noise Reduction Based on Deep Neural Networks," in *Proceedings of the 2021 7th International Conference on Signal Processing and Intelligent Systems (ICSPIS)*, vol. 2021, no. 12, pp. 1–6, 2021, doi: 10.1109/ICSPIS54653.2021.9729351.
- [19] Government of the Republic of Indonesia, *Criminal Code (Kitab Undang-Undang Hukum Pidana)*. Accessed: Apr. 9, 2025. [Online].
- [20] Government of the Republic of Indonesia, *Law of the Republic of Indonesia Number 11 of 2008 on Electronic Information and Transactions*. Accessed: Apr. 9, 2025. [Online].
- [21] A. Gaurav, B. B. Gupta, S. Sharma, R. Bansal, and K. T. Chui, "XLM-RoBERTa based sentiment analysis of tweets on Metaverse and 6G," *Procedia Computer Science*, Proc. 15th Int. Conf. Ambient Syst., Netw. Technol. (ANT) / 7th Int. Conf. Emerg. Data Ind. 4.0 (EDI40), vol. 238, no.4, pp. 902–907, Apr. 23–25, 2024, Hasselt Univ., Belgium. doi: 10.1016/j.procs.2024.06.110.
- [22] V. Mathur, T. Dadu, and S. Aggarwal, "Evaluating neural networks' ability to generalize against adversarial attacks in cross-lingual settings," *Applied Sciences*, vol. 14, no. 13, Art. no. 5440, pp. 1-12, 2024, doi: 10.3390/app14135440.
- [23] H. Fang, G. Xu, Y. Long, and W. Tang, "An Effective ELECTRA-Based Pipeline for Sentiment Analysis of Tourist Attraction Reviews," *Applied Sciences*, vol. 12, no. 21, Art. no. 10881, pp. 1-12, 2022, doi: 10.3390/app122110881.
- [24] C. M. Greco and A. Tagarelli, "Bringing order into the realm of Transformer-based language models for artificial intelligence and law," *Artif. Intell. Law (Dordr)*, vol. 32, no. 4, pp. 863–1010, Dec. 2023, doi: 10.1007/s10506-023-09374-7.
- [25] M. Post, "A Call for Clarity in Reporting BLEU Scores," in *Proc. 3rd Conf. Mach. Transl.: Research Papers*, vol. 2018, no. 10, Brussels, Belgium, Oct. 2018, pp. 186–191, doi: 10.18653/v1/W18-6319.
- [26] M. Freitag, R. Rei, N. Mathur, C.-k. Lo, C. Stewart, E. Avramidis, T. Kocmi, G. Foster, A. Lavie, and A. F. T. Martins, "Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust," in *Proc. 7th Conf. Mach. Transl. (WMT)*, Abu Dhabi, United Arab Emirates (Hybrid), vol. 2022, no.12, pp. 46–68, Dec. 2022.