Detecting SMS Phishing Based on Arabic Text-Content Using Deep Learning

Sadeem Alsufyani and Samah Alajmani

Department of information technology, College of Computer and Information Technology,
Taif University, Taif 21994, Saudi Arabia

Abstract
SMS phishing is a type of cyberattack where fraudsters use text messages to deceive people into giving up confidential information, like their bank details or personal data. While a lot of research has gone into spotting these fake texts in languages like English, there's a real lack of studies on the topic in Arabic. This paper aims to bridge that gap by putting forward a model specifically designed to identify phishing in Arabic SMS messages. The proposed model involves several stages, beginning with the collection of a dataset containing Arabic SMS messages, followed by Arabic SMS dataset cleaning. The linguistic complexities of the Arabic language are then addressed through preprocessing in the textual content of the Arabic SMS messages, such as removing Arabic stop words, diacritics, punctuation, and other irrelevant elements. Since Arabic words can have multiple forms, they are reduced to their Arabic roots using a stemming process. Next, features are extracted using the TF-IDF technique, and the target classes are encoded. Finally, they are passed through the classification process. The model uses three various deep learning techniques: BiGRU, CNN, and GRU, to detect messages as either phishing or legitimate (ham). The study compares the performance of the three models for deep learning utilizing the four main criteria and demonstrates the outcomes that the BiGRU model, with an accuracy rate of 98.71%, outperformed the other models. The GRU model achieved an accuracy rate of 98.32%, and the CNN model achieved an accuracy rate of 97.86%. These outcomes demonstrate BiGRU's capability to Arabic phishing SMS messages detection.

Keywords:
Phishing Threat; Arabic Phishing; Arabic SMS Text-content; Bidirectional GRU (BiGRU); Phone Phishing.

1. Introduction

Phishing represents a significant security threat exploited by attackers to trick them into giving them their information, phishing is a threat that involves the technique of tricking victims into disclosing their private data, such as usernames, passwords, and credit card numbers, this is conducted by impersonating reliable online platforms, such as sources or entities, in order to attract an audience, attackers pretend to be well-known users or companies from social media sites, online payment systems, or any other technologies, phones, emails, bank account information, passwords, and credit card information are usually the targets of phishing, to target the victim, the attacker uses social engineering in order to obtain personal data and account details related to the intended victim [1] [2]. SMS is a component of text connection services in mobile, internet, or phone systems, to enable SMS messaging between mobile or stationary devices, SMS is used with protocols for standardized communication, due to the fact that it's used so extensively, SMS has been a convenient substitute for calls when a call is unwanted or not feasible, furthermore, since texting tends to be significantly less expensive than calling another cellular phone, this has become increasingly common [3]. Therefore, in SMS phishing, the attackers send a message impersonating a trusted bank in an effort to trick the victim, the contents of the message are that the victim's bank account has been closed, and the attacker informs the victim to visit the link in the message to get back his bank account, therefore, the aim of this SMS phishing threat is identity theft and financial loss for the victims [4]. Therefore, with the increase in our use of smart devices, phishing attacks via SMS are increasing. But research that can assist in detecting such a threat is limited as far as the Arabic language is concerned. To bridge this gap, this paper contributes to give a model that detects Arabic SMS messages utilizing three models for deep learning based on Arabic Text-Content and classifies the Arabic SMS message as ham or phishing. The suggested model is beneficial in the sense that it gives a dataset that contains Arabic SMS messages for both ham and phishing categories. As the Arabic SMS dataset is small, we used two translation methods, i.e., machine and human translation, to cope with the problem of the small dataset. The proposed model also seeks to identify the best deep learning model based on the results reported in the form of the accuracy of the detection of Arabic SMS phishing messages based on Arabic Text-Content. The paper has a role to play in proposing a solution to combat this phishing menace via Arabic SMS message based on Arabic Text-Content.

The rest of this paper is organized as follows: Section 2 provides a literature review. Section 3 discusses the proposed model. Section 4 presents the result. Section 5 is a discussion of the proposed model. Finally, Section 6 concludes the paper and future work.

Manuscript received April 5, 2025
Manuscript revised April 20, 2025
https://doi.org/10.22937/IJCSNS.2025.25.4.1

2. Literature Review

In this section, we present some of the associated works in SMS phishing detection.

Saeed, 2023 [5] proposed SMS message models for spam and ham identification messages on the basis of various algorithms of supervised machine learning such as J48, K-Nearest Neighbors, and Decision Tree. The experimentation was performed upon an English dataset of 5574 messages used from the UCI repository (Almeida et al., 2011). The preprocessing of SMS data was done using various methods, such as data cleaning and word embedding techniques. The method was tested using recall, accuracy, and precision. The decision tree's 97.05% accuracy rate was the highest compared with the other machine learning classifiers.

Urmi et al., 2022 [6] the proposed model has used supervised machine learning algorithms—Logistic Regression (LR), Multinomial Naive Bayes (MNB), Support Vector Machine (SVM), Random Forest (RF), and Decision Tree (DT) classifiers—for SMS spam detection. The model was tested on a large dataset of 11,132 SMS messages labeled as spam or ham. The findings of the study showed that the most effective classifier for SMS spam detection is random forest (RF) with an accuracy rate of 99.73%.

Hossain et al., 2022 [7] presented a model that utilized deep learning technologies, i.e., convolutional neural network CNN- and long short-term memory-LSTM, to detect English SMS spam. In addition, they used traditional machine learning algorithms, i.e., MNB and SVM. Their findings showed that the CNN-LSTM model was superior to the model, with an accuracy rate of 98.40%. In addition, the model also contained an AUC of 0.994 and an F1 score of 98%.

Ghourabi et al., 2020 [8] introduced a CNN-LSTM hybrid model for SMS spam detection. The Arabic and English datasets were used to assess the model. Machine learning algorithms such as support vector machine (SVM), decision tree (DT), multinomial Naive Bayes (NB), logistic regression (LR), extra trees, AdaBoost (AB), bagging classifier, random forest (RF), and K-nearest neighbors (KNN) were used for comparison in the assessment. The findings indicated that the proposed model attained a precision rate of 95.39%, accuracy rate of 98.37%, F1 value of 91.48%, recall rate of 87.87%, and AUC value of 93.7%.

Alshahrani, 2021 [9] proposed employing random forest and decision tree to detect SMS spam. A dataset of 4900 ham spam messages and 672 normal spam messages was utilized. The result was that random forest worked better with an accuracy rate of 98.2%.

Abayomi‐Alli et al., 2022 [10] utilized a BiLSTM-based deep learning technique for SMS spam text detection. Two datasets were utilized to validate the proposed technique: the first original data, ExAIS_SMS, and the second is the UCI SMS dataset. Certain machine-learning techniques were utilized to compare the findings. The findings show that the BiLSTM model recorded an accuracy of 98.6% for the UCI dataset and 93.4% for the ExAIS_SMS dataset as well. As for comparing with modern machine learning classifiers, the results showed that the UCI dataset showed SGD 76.02 and Bayes Net 90.92%. while SOM, Naïve Bayes, C4.5, Bayes Net, decision tree, and J48 are 88.24%, 89.64%, 80.24%, 91.11%, 75.76%, and 79.2% ratios for the ExAIS_SMS dataset. Therefore, it was concluded that the proposed method, BiLSTM, showed a significant enhancement compared to traditional machine learning classifiers.

Ghourabi & Alohaly, 2023 [11] developed the SMS spam detector algorithm based on the application of ensemble learning methods as well as the use of GPT-3 transformer in order to generate informative and dense representations in addition to effective classifiers that had been able to categorize messages accurately as either spam or legitimate messages. This algorithm tried to harness advanced natural language processing transformers. Four algorithms for classification were utilized: Support Vector Machine (SVM), K-nearest neighbors (KNN), LightGBM, and CNN. Moreover, the last output of the ensemble learning module was predicted using a weighted vote technique. Based on the results, the proposed model was found to have an accuracy of 99.91%, 99.32% recall, 99.66% F1 score, and 100% precision.

Ibrahim et al., 2024 [12] A technology has been proposed to detect SMS phishing messages that support the Arabic language. Random forest classification and natural language processing were used. AdaBoost, Logistic Regression, and K-Nearest Neighbors (KNN) were among the other machine learning techniques they contrasted it with. The findings proved that the random forest classifier outperformed the other classifiers with a precision of 99.10%, recall of 98.23%, accuracy of 98.66%, and F1 score of 98.67%.

Therefore, there have been some earlier related studies confirming that remediation of this threat from the view of the Arabic language alone would not suffice, which is a research gap because limited mechanisms to date have contributed to alleviating this threat for Arab users. Table 1 illustrates related studies that have proposed solutions for the SMS phishing message detection.

Table 1: Literature review of SMS phishing Detection

Ref	*Model architecture*	*Language*	*Result%*
[5]	J48, Decision Tree, and K-Nearest Neighbors.	English	97.05%
[6]	Logistic Regression, Multinomial Naive Bayes, Support Vector Machine, Random Forest, and Decision Tree classifiers	English	99.73%
[7]	CNN-LSTM, SVM and MNB	English	98.40%
[8]	support vector machine, decision tree, multinomial Naive Bayes, logistic regression, extra trees, AdaBoost, bagging classifier, random forest, K-nearest neighbors, CNN, and LSTM	English and Arabic	98.37%
[9]	Random Forest, Decision Tree	English	98.2%
[10]	SOM, Naïve Bayes, C4.5, Bayes Net, decision tree, and J48, SGD, and BiLSTM	English	98.6% for the UCI dataset 93.4% for the ExAIS_SMS dataset
[11]	GPT-3 transformer and ensemble learning	English	99.91%
[12]	Random forest and NLP	Arabic	98.66%

3. Proposed Model

We show an overview of the methodology used in the deep learning method for Arabic phishing SMS message detection. A model has been suggested that includes the classification of the contents of the Arabic SMS message in terms of Arabic Text-Content. The datasets utilized in this process were sourced from three various sources: i) Kaggle [13], ii) the UCI repository [14], and iii) a paper [12]. The first two sources contain datasets in English, necessitating translation to transform the SMS text from English to Arabic, which supports the proposed model. The TF-IDF feature extraction was used at the word level to identify features related to Arabic SMS phishing message detection. CNN, GRU, and BiGRU models for deep learning have been used to classify Arabic SMS messages and distinguish between Arabic SMS phishing messages and Arabic SMS ham messages. Evaluation criteria are applied to assess the suggested model, such as precision, F1 score, accuracy, and recall. This methodology aims to create a model for Arabic phishing SMS message detection effectively.

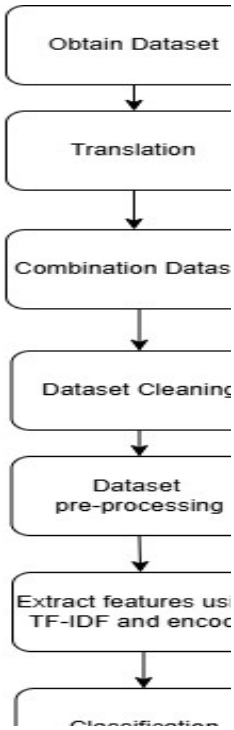

Fig. 1 The suggested model for Arabic SMS messages based on Arabic Text-Content detection.

The aim of the methodology is to provide an effective model for Arabic phishing SMS message detection that will help protect users from the risks of SMS phishing, which can be a dangerous means that will be exploited by attackers to steal personal or financial data. Initially, Fig. 1 illustrates the structure of the suggested model for the process of Arabic SMS messages based on Arabic Text-Content detection. A dataset of SMS messages must be obtained. The importance of collecting the dataset in the suggested model lies in achieving its goal of the model, which is to detect whether Arabic SMS messages are phishing or ham SMS messages. Providing a dataset that includes Arabic SMS messages is necessary, which helps train the proposed model. We collected the dataset from three different sources. The dataset has been modified to meet the requirements of the proposed model, contain two columns: the label of Arabic SMS messages and Arabic SMS text are shown in the first and second columns, respectively.

The following are the steps the proposed model will implement:

Step 1 Obtain Dataset:
The dataset was collected from three various sources. The first dataset from [13] is in English. The second dataset from [14] is in English. The third dataset from [12] is in Arabic.

Step 2 Translation:
Due to the lack of a sufficient dataset that includes Arabic text messages, we translated the following two datasets [13] and [14]. The translation process was performed in two ways: the first is machine translation using Google Translate, and the second is by a human volunteer.

Step 3 Combination Dataset:
This step aims to create a comprehensive dataset and expand the scope of the dataset that the model was trained on, which contains 16521 Arabic SMS messages varying between phishing and ham, i.e. the dataset translated by machine translation—Google Translate—contains 16,521 Arabic SMS messages, and the dataset translated by human volunteer translation contains 16,521 Arabic SMS messages.

Step 4 Data Cleaning:
Dataset cleaning is an important step to make the dataset free from errors and noise, which aids in improving the training and testing of deep learning models such as BiGRU, GRU, and CNN. We have removed null values as well as duplicate values from the dataset.

Step 5 Dataset Preprocessing:
Dataset Preprocessing is a critical step after the dataset cleaning process, which prepares the dataset for analysis. This step includes several steps to address the linguistic complexity of the Arabic language, as follows:

1. Remove process, which includes:
   - Remove Arabic stop words that include: pronouns, prepositions, and conjunctions.
   - Remove URL addresses.
   - Remove email.
   - Remove numbers.
   - Remove English characters.
   - Remove emojis and Flags.
   - Remove special characters.
   - Remove white space.
   - Remove diacritics.
   - Remove punctuation.
2. Normalization of letters such as Alif "ٱأآإ" “converted to "ا “. The letter Ya "ى ي" is converted to "ي", and ha "ۀ ة" is converted to "ه" to become uniform.
3. Tokenization and Stemming Process: This stage tokenizes the Arabic Text-Content message into individual words and then stems each word back to its linguistic origin.

Step 6 TF-IDF Feature Extraction and Encoding:
Here we will carry out two steps: one step is feature extraction using TF-IDF, and the second step is label encoding. Firstly, the process of feature extraction through the TF-IDF technique. The major goal of this technique is to determine the importance of each word in an Arabic SMS message based on its frequency, within the dataset. This technique represents each Arabic SMS text as a matrix in which each entry is the TF-IDF value for each word. Second is the process of encoding. This is accomplished in order to convert labels of the Arabic SMS message type into numerical forms so that they can be interpreted by the model. The type of phishing was converted to value 1, and the other type of ham was converted to value 0.

Step 7 Classification:
In the classification phase, we took three main tasks. Initially, we divided the dataset, allocating 30% to testing and the remaining 70% to train our models. Then, we constructed a deep learning model specifically to categorize Arabic SMS messages. For this task, we implemented three different models:

- BiGRU as part of the GRU model, the BiGRU model consists of putting together two unidirectional GRU units for processing the forward and backward portions of the sequence, the forward and backward GRU units' hidden states are then integrated at each time step, and the hidden state information of the forward and backward GRU units is linearly computed in order to retrieve the outputs of the BiGRU model at each instant [15]. Due to the fact that the input sequence is presented to one network in regular right-to-left chronological order and to the second network in reverse Arabic chronological order, this architecture is able to offer comprehensive contextual data [16].
- GRU is a type of gated recurrent neural network called the gated recurrent unit model (GRU), designed to address the issues of vanishing and exploding gradients that arise when learning long-term dependencies in traditional recurrent neural networks [17]. In contrast to the traditional LSTM model, the GRU model is distinguished by its straightforward structure and limited number of parameters, the GRU model consists of a reset gate, while the input and forget gates are combined into a single update gate [18]. The GRU model's internal state retains the prior state throughout the time step, demonstrating its ability to deal with sequential data and catch long-term dependencies among elements [19].
- CNN has three different kinds of layers—convolution, pooling, and fully connected

layers—that usually make up a CNN model, convolution and pooling are the initial two layers that extract features, in the final output, the extracted characteristics are output by the third layer, which is the fully connected layer [20]. Depending on the data that needs to be processed, the CNN model includes several dimensions: signals and text are processed using the one-dimensional model, images or audio are processed using the two-dimensional model, and video is processed using the three-dimensional model [21]. Our model classifies Arabic text using the one-dimensional model.

-

Finally, in the third task, we evaluated the performance of each model for deep learning using evaluation criteria: precision, F1 score, accuracy, and recall. Additionally, the confusion matrix illustrates the ability of each model for deep learning to distinguish between types of Arabic SMS messages, whether Phishing or Ham.

Assessing models' performance for deep learning requires training and testing. These four metrics are recall, accuracy, F1 score, and precision, as follows:

1. TP-True Positive: The amount of correctly identified positive Arabic SMS phishing detections.
2. TN-True Negative: Refers to how many negative Arabic SMS Ham messages were accurately identified.
3. FP-False Positive: Refers to the amount of negative cases wrongly labeled as positive, such as the number of Arabic SMS ham messages labeled as Arabic SMS phishing messages.
4. FN-False Negative: Refers to the amount of positive cases wrongly labeled as negative; for example, the number of Arabic SMS phishing messages detected as Arabic SMS Ham messages.

The criteria are then explained as follows:

- Accuracy:

It refers to the model's accurate predictions, whether the classifications are identifying an Arabic SMS message as phishing, or identifying an Arabic SMS message as ham. It is following represented by equation (1):

$$\text{Accuracy} = \frac{TN+TP}{TN+FP+FN+TP} \quad (1)$$

- Precision:

It refers to the correct positive predictions, meaning the Arabic SMS messages identified as phishing. The following is represented by equation (2):

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

- Recall:

The percentage of Arabic SMS messages represents the actual instances of Arabic SMS phishing messages properly identified by the model. The following equation represents it as (3):

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

- F1 score:

It is used to measure the performance of our models, which indicates to the average recall and precision. The following equation denotes it as (4):

$$\text{F1 score} = \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 2 \quad (4)$$

4. Results

This presentation showcases the results of using deep learning to sort Arabic SMS messages, highlighting how translation impacts the data. It compares the effects of translating the dataset with machine translation tools like Google Translate versus using a human volunteer for translation.

4.1 Google Translate

Models for deep learning (BiGRU, CNN, and GRU) were utilized in Arabic SMS messages classified as ham or phishing. The findings of the models are shown as follows:

4.1.1 CNN model for classification

Table 2 shows the total outcome of the CNN model, which explains the model's general performance while classifying Arabic SMS messages. It achieved a total accuracy of 97.12%, which shows that the model was able to classify Arabic SMS messages with high accuracy. Recall, the model was able to identify correctly 89.52% of Arabic SMS phishing messages. Precision, the model correctly predicted 85.60% of the Arabic SMS phishing messages as Arabic SMS phishing messages and the F1 score illustrated a best trade-off of 87.52%. Fig. 2. The CNN model confusion matrix shows how it can classify Arabic SMS messages as ham or phishing. The confusion matrix indicates that 2872 Arabic SMS ham messages were accurately identified as Arabic SMS ham messages and 333 Arabic SMS phishing messages were indeed Arabic SMS phishing messages. It misclassified 56 Arabic SMS ham messages as Arabic SMS phishing messages and 39 Arabic SMS phishing messages as Arabic SMS ham messages.

Table 2: Evaluation metrics for CNN model – Google translate

Model	*Metrics of Evaluation*			
	Accuracy	*Precision*	*Recall*	*F1 score*
CNN	97.12%	85.60%	89.52%	87.52%

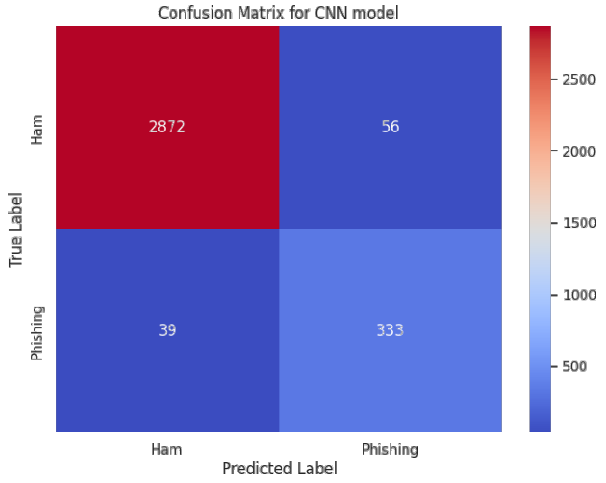

Fig. 2 Confusion matrix of the CNN model.

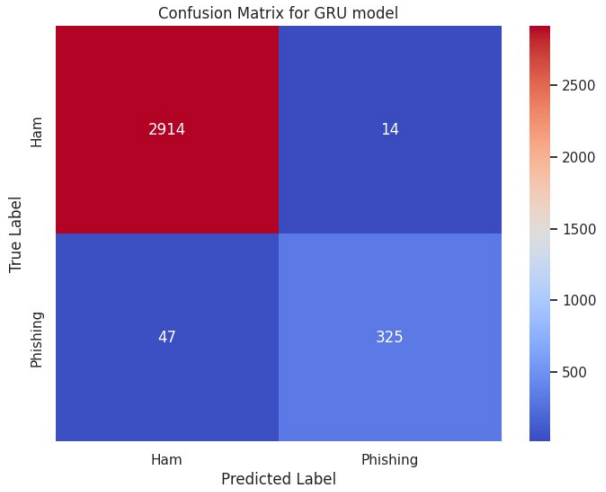

Fig. 3 Confusion matrix of the GRU model.

4.1.2 GRU model for classification

Table 3 demonstrates the overall findings of the GRU model in the Arabic SMS message classification. The findings showed that the model accomplished an overall accuracy in correctly classifying Arabic SMS messages of 98.15%, the model's ability to recall 87.37% of Arabic SMS phishing messages, a precision of 95.87%, and an F1 score of 91.42%. In Fig. 3. the confusion matrix shows the model's ability to classify both ham and phishing messages, classifying 2914 Arabic SMS ham messages as Arabic SMS ham messages and 325 Arabic SMS phishing messages as Arabic SMS phishing messages, while it was unable to classify correctly, classifying 14 Arabic SMS ham messages as Arabic SMS phishing messages and 47 Arabic SMS phishing messages as Arabic SMS ham messages.

Table 3: Evaluation metrics for GRU model – Google translate

Model	*Metrics of Evaluation*			
	Accuracy	*Precision*	*Recall*	*F1 score*
GRU	98.15%	95.87%	87.37%	91.42%

4.1.3 BiGRU model for classification

The BiGRU model's overall performance in distinguishing between Arabic SMS phishing and ham messages is summarized in Table 4. It managed to accurately classify Arabic SMS messages with an overall success rate of 98.06%. When it came to identifying Arabic SMS phishing messages, it had a recall rate of 87.90%. The precision rate, reflecting the accuracy of its classifications, was 94.51%. Finally, the F1 score, which provides a balanced measure, stood at 91.09%. Fig. 4. the confusion matrix, shows how the model distinguishes Arabic SMS phishing and ham messages. The model correctly classified 327 Arabic SMS phishing messages as Arabic SMS phishing messages and 2909 Arabic SMS ham messages as Arabic SMS ham messages, while it incorrectly classified 19 Arabic SMS ham messages as Arabic SMS phishing messages and 45 Arabic SMS phishing messages as Arabic SMS ham messages.

Table 4: Evaluation metrics for BiGRU model – Google translate

Model	*Metrics of Evaluation*			
	Accuracy	*Precision*	*Recall*	*F1 score*
BiGRU	98.06%	94.51%	87.90%	91.09%

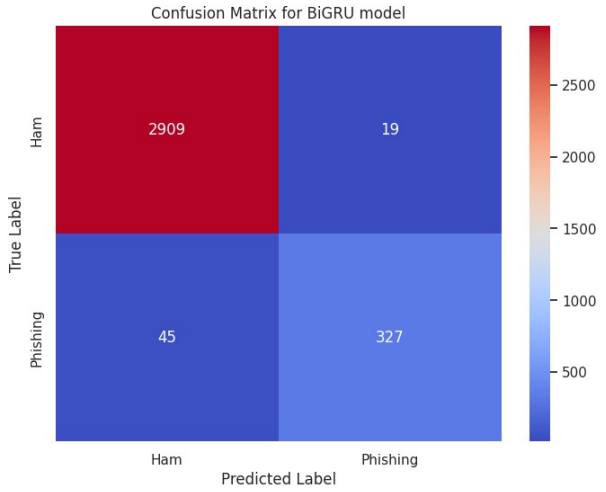

Fig. 4 Confusion matrix of the BiGRU model.

4.1.4 Comparison of the three models on the Google Translated dataset

Table 5 shows a comparison between all the findings of the three models in detecting Arabic SMS messages translated using Google Translate and evaluating the performance of the models using the four main criteria for evaluating the model performance. The findings displayed that in comparison to other models, the GRU model achieved the highest accuracy rate, reaching 98.15% and a precision rate of 95.87%, while the recall rate reached 87.37%, which was less compared to the other models. It achieved a balance between precision and recall with an F1 score of 91.42%, followed by the BiGRU model, whose overall accuracy rate reached 98.06% in classifying Arabic SMS messages and achieved a precision rate of 94.51%, while the recall rate reached 87.90% and the F1 score indicating balance reached 91.09%. The CNN model was the lowest among the other models, achieving the lowest accuracy rate of 97.12%. However, it achieved a precision of 85.60%, which is the lowest among the other models, a recall rate of 89.52%, which is the highest compared to the other models, and the lowest balance between recall and precision, which reached an F1 score of 87.52%. Therefore, based on these findings regarding the dataset that includes Arabic SMS messages translated using Google Translate, we can determine that the GRU model is the one that achieves the highest accuracy and is the best in classifying Arabic SMS messages translated using Google Translate.

Table 5: Evaluation metrics for three models – Google translate

Model	*Metrics of Evaluation*			
	Accuracy	*Precision*	*Recall*	*F1 score*
CNN	97.12%	85.60%	89.52%	87.52%
GRU	98.15%	95.87%	87.37%	91.42%
BiGRU	98.06%	94.51%	87.90%	91.09%

4.2 Human volunteer translation

Deep learning models (CNN, GRU, and BiGRU) were used on Arabic SMS messages classified as phishing or ham and translated by a human volunteer. The findings of the models are shown as follows:

4.2.1 CNN model for classification

Table 6 demonstrates the model's performance in the Arabic SMS message classification process, which accomplished an overall accuracy rate of 97.86% in classifying Arabic SMS messages, a recall rate of 95.28% for Arabic SMS phishing messages, a precision of 88.33%, which indicates correct classification of Arabic SMS phishing messages, and an F1 score achieved a balance of 91.67%. As displayed in Fig. 5. the accuracy of the CNN model in the classification of genuine (ham) and false (phishing) Arabic SMS messages is depicted in the form of a confusion matrix. The matrix demonstrates that the model correctly classified 3993 Arabic SMS messages as ham and 545 Arabic SMS messages as phishing accurately. However, it was incorrect to a certain degree: 72 Arabic SMS ham messages were classified as phishing, and 27 Arabic SMS phishing messages were classified as ham.

Table 6: Evaluation metrics for CNN model – Human Volunteer translate

Model	*Metrics of Evaluation*			
	Accuracy	*Precision*	*Recall*	*F1 score*
CNN	97.86%	88.33%	95.28%	91.67%

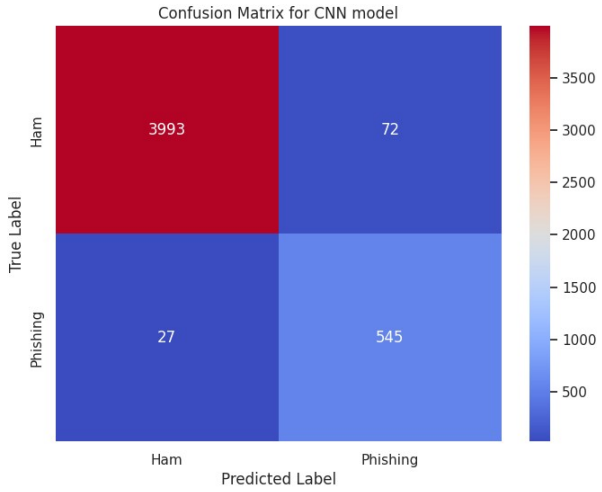

Fig. 5 Confusion matrix of the CNN model.

4.2.2 GRU model for classification

Table 7 shows that the model has the ability to distinguish Arabic SMS messages, as the model accuracy was 98.32%, which indicates the overall accuracy in correct Arabic SMS classification, recall rate was 92.66%, which shows that the model is able to remember phishing messages for Arabic SMS messages, and the precision was 93.64% and achieved an F1 score of 93.15%. The model was able to classify each category (phishing and ham). Fig. 6. appears the confusion matrix of the GRU model, as the model was able to distinguish 4029 Arabic SMS ham messages as Arabic SMS ham messages and 530 Arabic SMS phishing messages as Arabic phishing SMS messages. Nevertheless, the model misclassified the messages as it classified 36 Arabic SMS ham messages as Arabic SMS phishing messages and 42 Arabic SMS phishing messages as Arabic SMS ham messages.

Table 7: Evaluation metrics for GRU model – Human Volunteer translate

Model	*Metrics of Evaluation*			
	Accuracy	*Precision*	*Recall*	*F1 score*
GRU	98.32%	93.64%	92.66%	93.15%

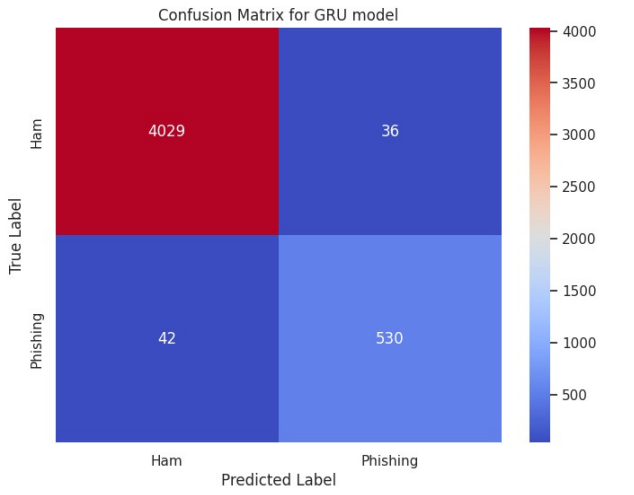

Fig. 6 Confusion matrix of the GRU model.

4.2.3 BiGRU model for classification

The BiGRU model performance is shown in Table 8, where the model's overall accuracy in correctly classifying Arabic SMS messages was 98.71% and the model recall for Arabic SMS phishing messages was 91.96%, reaching a precision of 97.41% and an F1 score balanced of 94.60%. In Fig. 7. the BiGRU model in the confusion matrix shows the model's capability to correctly classify 4051 Arabic SMS ham messages as Arabic SMS ham messages and 526 Arabic SMS phishing messages as Arabic phishing SMS messages, but the model could not correctly classify the category of Arabic SMS ham messages and Arabic SMS phishing messages, as it classified 46 Arabic SMS phishing messages as Arabic SMS ham messages and 14 Arabic SMS ham messages as Arabic SMS phishing messages.

Table 8: Evaluation metrics for BiGRU model – Human Volunteer translate

Model	*Metrics of Evaluation*			
	Accuracy	*Precision*	*Recall*	*F1 score*
BiGRU	98.71%	97.41%	91.96%	94.60%

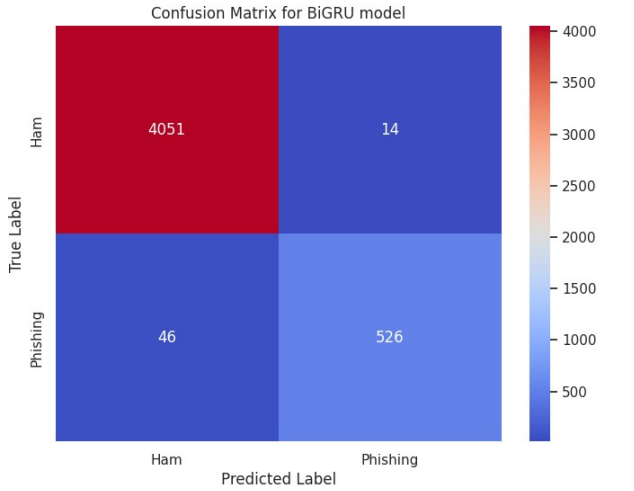

Fig. 7 Confusion matrix of the BiGRU model.

4.2.4 Comparison of the three models on the Human Volunteer Translated dataset

Table 9 shows a comparison between all the findings of the three models in the process of Arabic SMS message detection translated by a human volunteer and evaluating the performance of the models using the four main criteria for model evaluation. The findings explained that the BiGRU model accomplished the supreme accuracy rate compared to the other models, which reached 98.71%, and the precision rate was 97.41%, while the recall rate was the lowest compared to the other models, which reached 91.96%. However, the balance rate between recall and precision in the F1 score criterion was the highest, reaching 94.60%, followed by the GRU model, whose overall accuracy rate reached 98.32%, which is higher than CNN but lower than BiGRU, and a recall rate of 92.66% and precision of 93.64%, and achieved a balance in the F1 score of 93.15%. When compared to other models, the CNN model achieved the lowest accuracy rate, reaching 97.86%. However, in comparison to other models, it achieved the highest recall rate, reaching 95.28%, and the lowest precision compared to other models, reaching 88.33%, and the F1 score, which represents the balance between precision and recall, was 91.67%. Therefore, based on the findings represented, we can identify the BiGRU model as the best in terms of its accuracy in

detecting phishing messages via SMS in the Arabic language.

Table 9: Evaluation metrics for three models – Human Volunteer translate

Model	*Metrics of Evaluation*			
	Accuracy	*Precision*	*Recall*	*F1 score*
CNN	97.86%	88.33%	95.28%	91.67%
GRU	98.32%	93.64%	92.66%	93.15%
BiGRU	98.71%	97.41%	91.96%	94.60%

4.3 Arabic SMS messages Google vs. Human Volunteer translated

As we explained earlier in the proposed model section, there is a problem in finding a dataset that supports Arabic SMS messages. Therefore, in order to provide a dataset that includes Arabic SMS messages, we performed the translation step, which used two methods: machine translation (Google Translate) and human volunteer translation by a volunteer. Table 10 shows a comprehensive comparison of all the results presented in the results section in the previous two sections for classifying Arabic SMS messages as phishing or ham using Google Translate and human volunteer translation. We conduct a thorough comparison of the model's performance for deep learning (BiGRU, GRU, and CNN) in terms of their capability to classify Arabic SMS messages as phishing or ham. As we can see in Table 10, translation by a human volunteer showed higher findings compared to machine translation (Google Translate). In terms of accuracy in classifying Arabic SMS messages, the BiGRU model performed the highest compared to all models in both translations, reaching 98.71%, precision reaching 97.41%, F1 score balance ratio reaching 94.60%, and recall ratio reaching 91.96%. The GRU model followed, achieving an accuracy rate of 98.32% higher than machine translation, a recall rate of 92.66%, and an F1 score balance rate of 93.15%. However, the GRU used in machine translation achieved a higher precision than that used in human translation at 95.87%. The CNN model achieved a higher accuracy rate in human translation than that used in machine translation at 97.86%, a precision of 88.33%, and a recall rate of 95.28% which is the highest compared to all models used in both translations, and an F1 score balance rate of 91.67%. So based on the comparison, we can determine that human translation outperformed machine translation in terms of accuracy, and we can determine that the BiGRU model in human translation is superior among other models in both translations in the process of classifying Arabic SMS messages.

Table 10: Comparative Google vs. Human Volunteer Translated

Model	*Metrics of Evaluation*							
	Google Translate				*Human volunteer Translate*			
	Accurac	*Precision*	*Recall*	*F1 score*	*Accurac*	*Precision*	*Recall*	*F1 score*
CNN	97.12%	85.60%	89.52	87.52%	97.86%	88.33%	95.28%	91.67%
GRU	98.15%	95.87%	87.37	91.42%	98.32%	93.64%	92.66%	93.15%
BiGRU	98.06%	94.51%	87.90	91.09%	98.71%	97.41%	91.96%	94.60%

5. Discussion

The findings showed that the translation process of the dataset affects the performance of the three models (BiGRU, GRU, and CNN) in terms of accuracy in classifying Arabic SMS messages. We also found that the models trained and tested on the dataset of Arabic SMS messages translated by a human volunteer outperformed those translated by Google Translate. Based on the purpose that we had to fulfill in choosing the highest accuracy deep learning model for the classification process, the BiGRU model worked better in the Arabic SMS message classification task out of the three models that were conducted in human translation than in machine translation.

The GRU model came after it, which worked better in human translation compared to machine translation, and finally, the CNN model, which worked worse compared to other models. In contrast, the paper [12] used the same objective: Arabic phishing SMS messages detection based on Arabic Text-Content. But using machine learning algorithms with the TF-IDF technique. Their results showed an accuracy of 98.66%. In our paper, using TF-IDF with deep learning models yielded even higher results, with an accuracy rate of 98.71% for the BiGRU model. Based on these results, it can be concluded that deep learning models have the ability to handle the complexities of the Arabic language. Table 11 summarizes a comparison between our paper and paper [12].

Table 11: Comparison of our paper with the paper [12]

Ref.	*Model architecture*	*TF-IDF Used*	*Detection based on the Arabic text content of the Arabic SMS message*	*Language*	*Accuracy%*
[12]	Random forest and NLP	Yes	Yes	Arabic	98.66%
Our Paper	BiGRU	Yes	Yes	Arabic	98.71%

6. Conclusion

As users are increasingly using smartphones, attackers have taken advantage of this trend to create cyber threats that trick smartphone users by exploiting SMS messages. These attackers seek to trick users into revealing private information, such as banking and personal information. In this paper, we focus on such a threat in Arabic due to the gap in previous studies that have addressed this threat in other languages, such as English. We contributed to presenting a model that Arabic SMS phishing message detection using three deep learning models such as BiGRU, GRU, and CNN. These models help in detecting Arabic SMS messages, whether they are phishing or ham messages. A dataset was collected from different sources and unified to be in one language, which is Arabic, to support the model in achieving the goal. The Arabic SMS Messages dataset cleaning, pre-processing, and feature extraction process were then carried out using the TF-IDF technique and encoding the label of the Arabic SMS messages to be passed to the models for deep learning for training and testing. The findings demonstrated that, when compared to the other models, the BiGRU model had the highest accuracy rate at 98.71%, followed by the GRU model, where the accuracy rate was 98.32%, and finally the CNN model, where the accuracy rate was 97.86%. The shortcomings we faced were that the dataset was insufficient, and this is a gap in our paper because it affects the accuracy of results in the model training and testing process. Therefore, in future work, we will aim to make the Arabic SMS messages dataset larger to conduct more experiments as well as test other models to improve the process of detecting phishing SMS messages via Arabic, can also be conducted to evaluate the proposed model across multiple languages for Phishing SMS detection, as well as enhancing it for real-time applications.

Acknowledgments

The researchers would like to acknowledge Deanship of Scientific Research, Taif University for funding this work and support.

References

[1] F. I. F. Rikzan and M. F. Zolkipli, "A Study of Phishing Attack towards Online Banking," vol. 6, 2023.

[2] Z. Alkhalil, C. Hewage, L. Nawaf, and I. Khan, "Phishing Attacks: A Recent Comprehensive Study and a New Anatomy," *Front. Comput. Sci.*, vol. 3, p. 563060, Mar. 2021, doi: 10.3389/fcomp.2021.563060.

[3] T. A. Almeida, J. M. G. Hidalgo, and T. P. Silva, "Towards SMS Spam Filtering: Results under a New Dataset," *Int. J. Inf. Secur. Sci.*, vol. 2(1), pp. 1–18, 2013.

[4] A. Jain, "A Novel Approach to Detect Spam and Smishing SMS using Machine Learning Techniques," *Int. J. E-Serv. Mob. Appl.*, vol. 12, Nov. 2019, doi: 10.4018/IJESMA.2020010102.

[5] V. A. Saeed, "A Method for SMS Spam Message Detection Using Machine Learning," *Artif. Intell. Robot. Dev. J.*, pp. 214–228, Feb. 2023, doi: 10.52098/airdj.202366.

[6] A. Urmi, Md. T. Ahmed, M. Rahman, and A. Islam, "A Proposal of Systematic SMS Spam Detection Model Using Supervised Machine Learning Classifiers," 2022, pp. 459–471. doi: 10.1007/978-981-16-8225-4_35.

[7] S. Md. M. Hossain *et al.*, "Spam Filtering of Mobile SMS Using CNN–LSTM Based Deep Learning Model," in *Hybrid Intelligent Systems*, vol. 420, A. Abraham, P. Siarry, V. Piuri, N. Gandhi, G. Casalino, O. Castillo, and P. Hung, Eds., in Lecture Notes in Networks and Systems, vol. 420. , Cham: Springer International Publishing, 2022, pp. 106–116. doi: 10.1007/978-3-030-96305-7_10.

[8] A. Ghourabi, M. A. Mahmood, and Q. M. Alzubi, "A Hybrid CNN-LSTM Model for SMS Spam Detection in Arabic and English Messages," *Future Internet*, vol. 12, no. 9, p. 156, Sep. 2020, doi: 10.3390/fi12090156.

[9] A. Alshahrani, "Intelligent Security Schema for SMS Spam Message Based on Machine Learning Algorithms," *Int. J. Interact. Mob. Technol. IJIM*, vol. 15, no. 16, p. 52, Aug. 2021, doi: 10.3991/ijim.v15i16.24197.

[10] O. Abayomi-Alli, S. Misra, and A. Abayomi-Alli, "A deep learning method for automatic SMS spam classification: Performance of learning algorithms on indigenous dataset," *Concurr. Comput. Pract. Exp.*, vol. 34, no. 17, p. e6989, Aug. 2022, doi: 10.1002/cpe.6989.

[11] A. Ghourabi and M. Alohaly, "Enhancing Spam Message Classification and Detection Using Transformer-Based Embedding and Ensemble Learning," *Sensors*, vol. 23, no. 8, p. 3861, Apr. 2023, doi: 10.3390/s23083861.

[12] A. Ibrahim, S. Alyousef, H. Alajmi, R. Aldossari, and F. Masmoudi, "Phishing Detection in Arabic SMS Messages using Natural Language Processing," in *2024 Seventh International Women in Data Science Conference at Prince Sultan University (WiDS PSU)*, Riyadh, Saudi Arabia: IEEE, Mar. 2024, pp. 141–146. doi: 10.1109/WiDS-PSU61003.2024.00040.

[13] "Spam / Ham SMS DataSet." Accessed: Oct. 04, 2024. [Online]. Available: https://www.kaggle.com/datasets/vivekchutke/spam-ham-sms-dataset

[14] J. H. Tiago Almeida, "SMS Spam Collection." UCI Machine Learning Repository, 2011. doi: 10.24432/C5CC84.

[15] Z. Zha *et al.*, "A BiGRU Model Based on the DBO Algorithm for Cloud-Edge Communication Networks," *Appl. Sci.*, vol. 14, no. 22, p. 10155, Nov. 2024, doi: 10.3390/app142210155.

[16] M. M. Abdelgwad, T. H. A. Soliman, A. I. Taloba, and M. F. Farghaly, "Arabic aspect based sentiment analysis using bidirectional GRU based models," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 9, pp. 6652–6662, Oct. 2022, doi: 10.1016/j.jksuci.2021.08.030.

[17] G. Shen, Q. Tan, H. Zhang, P. Zeng, and J. Xu, "Deep Learning with Gated Recurrent Unit Networks for Financial Sequence Predictions," *Procedia Comput. Sci.*, vol. 131, pp. 895–903, 2018, doi: 10.1016/j.procs.2018.04.298.

[18] M. Zulqarnain, R. Ghazali, Y. M. Mohmad Hassim, and M. Rehan, "Text classification based on gated recurrent unit combines with support vector machine," *Int. J. Electr. Comput. Eng. IJECE*, vol. 10, no. 4, p. 3734, Aug. 2020, doi: 10.11591/ijece.v10i4.pp3734-3742.

[19] M. Zulqarnain, S. Abd, R. Ghazali, N. Mohd, M. Aamir, and Y. Mazwin, "An Improved Deep Learning Approach based on Variant Two-State Gated Recurrent Unit and Word Embeddings for Sentiment Classification," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 1, 2020, doi: 10.14569/IJACSA.2020.0110174.

[20] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights Imaging*, vol. 9, no. 4, pp. 611–629, Aug. 2018, doi: 10.1007/s13244-018-0639-9.

[21] D. Alsaleh and S. Larabi-Marie-Sainte, "Arabic Text Classification Using Convolutional Neural Network and Genetic Algorithms," *IEEE Access*, vol. 9, pp. 91670–91685, 2021, doi: 10.1109/ACCESS.2021.3091376.

Sadeem Alsufyani received a Bachelor's degree in Information Technology in 2022 and currently a Post-Graduate Student Master's degree in Cyber Security, both at Taif University, Taif, Saudi Arabia. She is interested in research topics Cybersecurity, Phishing Attacks, AI, and Deep Learning.

Samah Alajmani received the B.Sc. degree in 2004 and Ph.D. degree in 2019 in King Abdulaziz University, Jeddah, Saudi Arabia, both in Computer Science. She earned the M.Sc. degree in Information Technology from the Queensland University of Technology, Brisbane, Australia. She is currently an Assistance Professor at Taif University, Taif, Saudi Arabia. Her research interests include Cyber Security, AI, IoT, Deep Learning and Machine learning.