Journal of Ambient Intelligence and Humanized Computing (2023) 14:8173–8184
https://doi.org/10.1007/s12652-021-03587-x

ORIGINAL RESEARCH

Real-time Korean voice phishing detection based on machine learning approaches

Minyoung Lee[1] · **Eunil Park**[1]

Received: 9 August 2021 / Accepted: 26 October 2021 / Published online: 12 November 2021

Abstract
Voice phishing, or vishing, is a phishing phone call in which an attacker lures receivers into providing personal their information. Damage from vishing is a serious problem worldwide and is increasing in frequency. Therefore, this study is aimed at detecting vishing in real time. Owing to the absence of research on spam detection using low-resource languages, we detect vishing in the Korean language using basic machine-learning models. We collected actual vishing damage data and converted the voice files into text to achieve spam detection using natural language processing techniques. The focus is on determining whether vishing can be rapidly detected, rather than model development. Based on the results, we suggest that vishing can be detected in real time and requires only a short training time when using machine learning models.

Keywords Voice phishing · Vishing · Spam detection · Machine learning · Natural language processing

1 Introduction

Voice phishing, or vishing, which is a combination of the terms, "voice" and "phishing," is "*an emergent crime in which victims are deceived during phone conversations*" (Choi et al. 2017). Generally, an attacker attempts to lure receivers into revealing their personal/private information and creates an illegitimate advantage through phone conversations (Yeboah-Boateng and Amanor 2014).

With rapid improvements in information and communication technologies, including mobile networks, the problems and social harm caused by vishing have also increased. According to the 2018 Internet Crime Report released by the FBI, damage from phishing in the US was estimated at $48 million with 26,379 victims (Gorham 2019). Moreover, approximately 30% of all global incoming calls are regarded as vishing calls. For instance, approximately 4 billion spam calls were generated in the US per month in 2019, whereas each Brazilian annually received more than 45 vishing calls per month (Cook 2021).

To prevent damage through vishing, several scholars have made notable efforts to use computational approaches to detect such events (Zhang and Gurtov 2009; Tran et al. 2020). For instance, Barraclough et al. (2013) proposed a protection platform by combining several notable features, such as user behavior profiles. As another example, Tran et al. (2020) introduced a vishing detection system, which includes several natural language processing (NLP) techniques that consider white and blacklists. Although the findings of previous studies have provided several cornerstones for vishing detection, some concerns remain. For example, most of these studies have been conducted in English, which is widely used around the world.

However, significant damage is also incurred through vishing in several nations that use low-resource languages, including Kenya (Obuhuma and Zivuku 2020), Japan (Kadoya et al. 2020), and South Korea (Kim et al. 2021). In South Korea, the damage from vishing is continuously increasing (Figs. 1, 2). From 2016 to 2020, 134,112 cumulative cases (27,515 institutional impersonations and 106,597 loan fraud) were reported in South Korea (Korea National Police Agency 2020).

According to Korea Financial Supervisory Service (2021), the cases of fraud have become extremely sophisticated, resulting in serious damage to the victims. Moreover, because several innovative financial technologies have been introduced to provide convenient banking services, the damage incurred from faulty money transfers is significantly increasing. In 2020, more than 75% of the total

✉ Eunil Park
eunilpark@skku.edu

[1] Department of Applied Artificial Intelligence, Sungkyunkwan University, Seoul, Republic of Korea

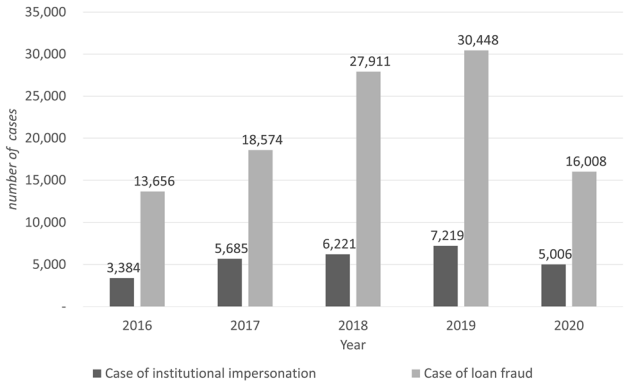

Fig. 1 Case statistics of vishing in South Korea (Korea National Police Agency 2020)

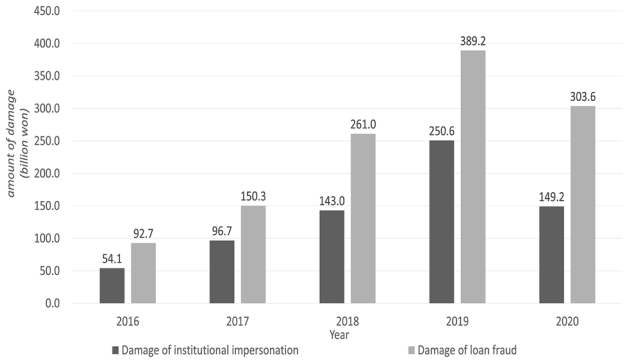

Fig. 2 Damage statistics of vishing in South Korea (Korea National Police Agency 2020)

damage occurred through mobile and Internet banking services (Korea Financial Supervisory Service 2021). This means that advanced financial technologies, aiming at providing greater convenience, are more likely to be exposed to criminal activities.

Because vishing occurs instantaneously (Biswal 2021), it is necessary to investigate in real time whether a specific message (or data) is actually vishing. Therefore, the current study proposes a real-time Korean vishing detection system using machine learning approaches. The main contributions of this paper are summarized as follows:

- The main purpose of this study is to introduce a real-time detection system for vishing in Korean.
- Machine learning techniques with not only fast training procedures, but also rapid detection capabilities, are proposed.
- Importantly, the increased applicability of actual case data under real situations in which vishing has occurred, rather than the use of older and refined spam datasets, is demonstrated.

The remainder of this paper is organized as follows. Section 2 provides an overview of the literature in this field. Related datasets and machine learning models are presented in Sect. 3, followed by performance analyses of the models in Sect. 4. Finally, a discussion and some concluding remarks are presented (Sect. 5).

2 Literature review

Most studies on vishing detection have been conducted in related areas, such as spam detection. Historically, many scholars have considered vishing detection issues as subordinate topics in the field of spam detection (Song et al. 2014). Therefore, we have summarized several spam detection studies using machine learning, neural networks, hybrid approaches, and some notable real-time spam detection studies. In addition, we provide an overview of prior vishing-detection research.

2.1 Spam detection with machine learning approaches

Table 1 provides a summary of prior spam detection research using machine learning approaches. For instance, Drucker et al. (1999) initially introduced a spam classification model using a support vector machine (SVM). They employed spam and nonspam email datasets collected by AT&T and evaluated the performances of several machine learning and rule-based classification models, including SVM, Ripper, and Rocchio algorithms, when considering the recall, precision, error rate, and false alarm rate as their metrics. Moreover, the term frequency and inverse document frequency (TF-IDF), as well as binary forms using TF, were used as the key features. The results indicate that their boosting model recorded the lowest error rate of 0.018, followed by the SVM model at 0.021. Although the error rate of the SVM was not significantly different from that of the boosting model, the proposed SVM model was significantly faster than the other models. As another example, Yan et al. (2020) introduced a robust LDA approach, which can be applied to detect spam information with multiple classifications.

2.2 Spam detection with neural network approaches

A significant limitation of using traditional machine learning approaches is the difficulty in understanding relatively long sentences and contexts. Accordingly, prior research has indicated that the use of a neural network architecture can be helpful for addressing a number of academic issues (Ren et al. 2021). Thus, to contextualize spam detection, some scholars prefer the use of a neural network

Table 1 Summary of prior spam detection research using machine learning approaches

Sources	Type	Method	Datasets	Results
Drucker et al. (1999)	Email spam filtering	Machine learning (supervised; boosting and SVM) and rule-based algorithm approaches (Ripper and Rocchio)	AT&T dataset: 850 spam and 2150 non-spam cases	Error rate: 1.80% (Boosting), 2.13% (SVM)
Koøcz and Alspector (2001)	Email spam filtering	Machine learning approach (supervised; SVM)	Collection of spam-based, ling-spam, and PU1 corpus: 5365 spam and 6043 nonspam cases	Metric: 3.54% of precision
Sasaki and Shinnou (2005)	Email spam filtering	Machine learning approach (unsupervised; spherical k-means clustering)	Ling-spam dataset: 481 spam and 2412 nonspam cases	Precision: 90% (spam), 96% (non-spam)
Gómez et al. (2006)	SMS spam filtering	Machine learning approaches (supervised; Naive Bayes (NB), C4.5 and Partial Decision Trees, and SVM)	Spanish (199 spam and 1157 non-spam) and English datasets (85 spam and 1119 nonspam)	Accuracy: 90% (English) and 99% (Spanish)
Abu-Nimeh et al. (2007)	email spam filtering	Machine learning approaches (supervised; Logistic regression (LR), classification and regression trees, Bayesian additive regression trees, SVM, RF, and neural networks)	Spam email database: 1171 spam and 1718 nonspam cases	Error rate: 7.72% (RF), false positive rate: 4.89% (LR), AUC: 0.9448 (Neural network)
Mccord and Chuah (2011)	Twitter spam filtering	Machine learning approaches (supervised; RF, NB, SVM, K-NN neighbor)	Twitter data: 4435 spam and 7293 non-spam cases in (Follower), and 3535 spam and 1107 nonspam cases (Following)	Precision: 95.7% (RF), F1-score: 95.7% (RF)
Akinyelu and Adewumi (2014)	Email spam filtering	Machine learning approaches (supervised; RF)	spam assassin project & Nazario corpus: 200 spam and 1800 nonspam cases	Accuracy: 99.7%
Trivedi (2016)	Email spam filtering	Machine learning approaches (supervised; Bayesian, SVM, DT, AdaBoost)	Enron email corpus: 3000 spam and 3000 nonspam cases	F1-score: 93.3% (SVM), false positive rate: 6.5% (SVM)

architecture with improved deep learning techniques (Kenton and Toutanova 2019). Accordingly, there are several unique feature extraction approaches for achieving a neural network architecture (Li et al. 2018a, 2018b). Table 2 lists the findings of prior spam detection research using neural network approaches. For instance, Li et al. (2019) proposed an event-adaptive concept for addressing event detection.

As another example, Akinyelu and Adewumi (2014) used random forest (RF) classifiers with a 10-fold cross-validation method. Instead of directly using the original email text, they extracted 15 features for spam detection, including *URLs containing the IP address; disparities between "href" attributes and LINK text; the presence of " Link " "Click," and "Here" in LINK text; the number of dots in the domain name; HTML email; the presence of JavaScript; the number of links; the number of links to a domain from a body match domain check; and word list features*. The results of the RF classifiers showed an accuracy of 99.70% for the 2000 spam and nonspam cases.

As a representative example, Rajet al. (2018) proposed a neural network architecture for spam classification. They employed SVM, decision tree (DT), KNN, RF, and NB machine learning classifiers on SMS messages and obtained an accuracy of 96.17% when using SVM. The LSTM-based model was then applied to determine whether a specific message was spam. The LSTM-based model achieved an accuracy of 97.50%, which is better than that of the machine learning classifiers. Therefore, neural network algorithms can detect spam with longer contexts more accurately than traditional machine-learning approaches.

Roy et al. (2020) demonstrated that neural network based spam detection models are more effective in addressing imbalanced datasets than traditional machine learning models. Experiments were conducted using the SMS Corpus provided by the University of California, Irvine (UCI) repository, which is an imbalanced dataset of 747 spam corpora and 4827 nonspam corpora. Compared to the performance of machine-learning-oriented spam detection models (F1-scores of 18.40% and 88.30% (NB), 0% and 92.50% (LR), 3.60% and 91.60% (RF) for spam and nonspam cases, respectively), convolutional neural network (CNN) oriented models (single layer) achieved F1-scores of 92.20% (spam) and 98.80% (non-spam) with imbalanced datasets. Moreover, three-layer CNN-oriented models with a dropout rate of 0.3 and 10-fold cross-validation procedures achieved an F1-score of 99.80%, whereas the LSTM-oriented models achieved F1-scores of 87.00% (spam) and 98.20% (nonspam).

Table 2 Summary of several prior spam detection research with neural-network and hybrid approaches

Sources	Type	Method	Datasets	Results
Wijaya and Bisri (2016)	Email spam filtering	Hybrid algorithm (DT, LR)	Collection of UCI repository, ling-spam and PU1 corpus: 1813 spam and 2488 nonspam cases	Accuracy: 91.67%
Wu et al. (2017)	Twitter spam filtering	Machine learning (RF, DT, NB), and neural network approaches (Multi-layer perceptron (MLP))	Real-life Twitter dataset: 1376206 spam and 673836 nonspam	Accuracy: 99.34% (MLP)
Rajet al. (2018)	SMS spam filtering	Machine learning (SVM, DT, KNN, RF, and NB), and neural network approaches (Long Short-Term Memory (LSTM))	NUS-SMS Corpus: 1877 spam and 3697 nonspam	Accuracy: 96.17% (SVM), 97.50% (LSTM)
Roy et al. (2020)	SMS spam filtering	Machine learning (NB, RF, GB, and LR), and neural network approaches (CNN and LSTM)	UCI Repository SMS corpus: 747 spam and 4827 nonspam	F1-score: 99% (CNN with three layers), 87% (spam with LSTM), and 98% (nonspam with LSTM)
Ghourabi et al. (2020)	SMS spam filtering	Machine learning (SVM, and RF), neural network (LSTM, CNN), and hybrid approaches (CNN+LSTM)	English (5574) and Arabic cases (2730)	Accuracy: 98.37% (CNN-LSTM synthetic model)

2.3 Hybrid approaches

Several hybrid approaches have also been recently introduced to address spam detection by considering the drawbacks of a traditional system, such as an overfitting. Wijaya and Bisri (2016) presented a novel hybrid model that combines LR with DT to account for noise data. In this case, DT was used for non-noisy data, whereas LR was used for noisy data using the false negative rate threshold levels to increase the true positive results. Considering 4301 cases (1813 spam and 2488 nonspam cases), the authors achieved an accuracy of 91.67% by integrating machine learning and deep learning models, which is greater than that of stand-alone machine learning (90.65%) and deep learning approaches (90.79%).

Ghourabi et al. (2020) attempted to address SMS spam detection and introduced a unique hybrid algorithm, which synthesized two neural network models, i.e., a CNN and an LSTM. In the proposed CNN-LSTM model, the embedding layer was connected to the max-pooling layer through a single convolutional layer with a filter size of 32. The output of the max-pooling layer is linked to the input of the LSTM. The final output was extracted using a sigmoid function to determine whether a specific SMS was spam. With a dataset comprising 5574 English and 2730 Arabic cases, the SVM and RF machine learning models achieved an accuracy of 97.83%, whereas the LSTM and CNN models showed accuracies of 98.13% and 98.19%, respectively. In addition, the hybrid model integrating a CNN and LSTM showed the highest levels of accuracy (98.37%), recall (87.88%), F1-score (91.48%), and ROC-AUC (0.94).

2.4 Real-time spam detection

Because vishing occurs extremely rapidly, real-time spam detection has become one of the most important research topics in our society. Therefore, Gupta et al. (2018) argued the need for real-time spam detection in the social network service, Twitter, where several messages are presented within a short period, making it necessary to quickly detect spam messages and prevent their diffusion and exposure. Although there are several spam-detection tools available on Twitter, such as Google SafeBrowsing and Twitter BotMaker, they have a limited ability to detect real-time spam tweets. Gupta et al. (2018) also introduced a new model that can classify whether specific tweets are spam, considering both user (e.g., account age, number of followers, and number of followings) and tweet (e.g., number of retweets, hashtags, and URLs) information. From the collected 150,000 spam and 250,000 nonspam real-time Twitter data, 100,000 unique words and 13 features were extracted, from which the top-15 indicators of spam and nonspam words were selected, respectively. For the four classifiers considered, accuracy levels of 85.95% (SVM), 91.65% (neural networks), 85.84% (gradient boosting), and 85.25% (RF) were achieved in real-time spam-tweet detection tasks.

Sun et al. (2020) explored the spam-filtering tasks on Twitter. Based on a 30-million tweet dataset including 6.5 million spam cases, they employed each user profile as a key feature of real-time spam detection (e.g., the number of followers and retweets). Then, six types of user-oriented and seven types of content-based information were extracted and employed as features for real-time spam detection. Several machine learning and deep learning models have been applied based on these features, considering the time required for spam detection. According to the results, the LR model showed the fastest performance in exploring real-time spam detection tasks (3 s), whereas the RF model achieved the highest recognition accuracy of 90.58%.

Similar to spam detection in social network services, several approaches have recently been proposed for real-time vishing detection. For example, Zhang and Gurtov (2009) proposed a new vishing detection system using the whitelist of each recipient. If a caller is on the whitelist, the system connects the caller with the recipient; otherwise, the recipient can decide whether to connect. The results of the connection were then employed to improve the vishing detection system (e.g., whitelist, blacklist, and other reputation databases). Tran et al. (2020) also proposed an alert system that was developed using mobile applications and web servers in South Korea.

Although there are several notable cornerstones in real-time vishing detection tasks, there has been a notable lack of research on tasks that consider low-resource languages. Therefore, the current study explores real-time vishing detection tasks using machine learning approaches in a low-resource language, namely, Korean.

3 Method

Figure 3 shows the overall flow of our experiment. After collecting the vishing dataset, the dataset was first preprocessed and then used to train and test the classification models.

3.1 Dataset

We collected a vishing dataset, which was organized as vishing-labeled (47 h) and non-vishing-labeled speech (500 h), from the transcripts. The dataset is open to the public through two South Korean governmental agencies, i.e., the *Korean Financial Supervisory Service*[1] and *National Institute of Korean Language*,[2] and the open-source project.[3]

[1] https://www.fss.or.kr/fss/vstop/avoid/this_voice_l.jsp.

[2] https://corpus.korean.go.kr/.

[3] https://www.anonymous.4open.science/r/vishing-1AF8.

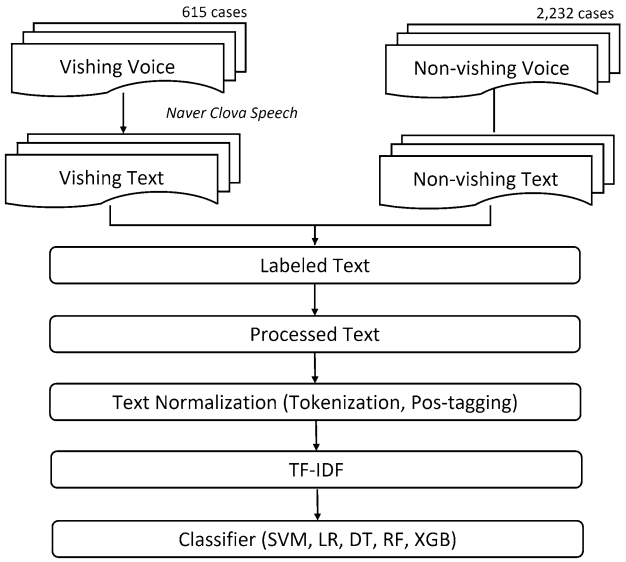

Fig. 3 Workflow of data preprocessing and classification

The vishing-label cases were collected by the *Financial Supervisory Service*. These are actual recorded cases reported by victims in South Korea. Because the cases occurred while talking over mobile phones, there is considerable noise in the data. We employed 615 vishing cases after eliminating personal information and annotating the collected datasets for each vishing case.

Non-vishing labeled cases were collected by the *National Institute of Korean Language*. The cases covered approximately 500 h of daily conversation by 2739 speakers. In each case, the conversation had a duration of approximately 15 min and was on one of 15 topics (*sports, travel, weather, company/school, food, broadcasting, movies, health, gifts, goals, dating, pets, part-time jobs, personality, and family*). After removing any personal information, we employed 2232 non-vishing cases organized using paired speech and text data.

3.2 Preprocessing

We converted the vishing-label cases into text using the *Google Cloud Speech-to-Text API*[4] and *Naver Cloud Platform Clova Speech API*.[5] The results were then compared to validate whether the converted texts matched the original speech cases. We selected the results of the *Naver Cloud Platform Clova Speech API*, which outperformed the *Google Cloud Speech-to-Text* API and was suitable for the Korean language. Table 7 (in Appendix B) presents some examples of the APIs used. An expert in information science reviewed and corrected the converted texts.

Table 3 Summary of the collected dataset

	Vishing (615 cases)	Non-vishing (2,232 cases)
Voice	169,408 s	1,867,555 s
Text	1,040,065 words	10,857,402 Words

We employed a dataset of 2847 cases (full dataset: 615 vishing and 2,232 non-vishing). Table 3 summarizes the collected and applied datasets.

In addition to the full dataset, we created a smaller dataset by excluding cases with more than 5000 words, called the *Top-5000 dataset*, to address the imbalance in the length distributions between vishing and nonvishing cases.

Each word was then tagged by dividing each text into morpheme units. We used open Korean text (OKT) from Korean natural language processing in Python (KoNLPy),[6] and Khaiii (Kakao hangul analyzer),[7] which are widely applied in the Korean NLP process. The nouns, adjectives, and verbs that remained were employed in the analysis.

We then used the TF-IDF vectorizer for feature extraction. Classification experiments were conducted after feature extraction. In addition to the features, the paired POS tags for each word were analyzed to better understand the context of the text.

3.3 Machine learning models

The *full dataset* and *Top-5000 dataset* were randomly divided at a ratio of 8:2 as the training and test datasets, respectively. Table 4 presents the number of cases employed in the training and testing. All experiments were conducted on a single Tesla V100 PCI-E 32 GB GPU and implemented in Python 3.6.

Five machine learning models, namely, SVM, LR, RF, DT, and extreme gradient boosting (XGB) models, were used in this study. A brief description of these approaches is provided below.

- **SVM** is a training algorithm that classifies the labels by maximizing the margin between the examples and class boundary (Boser et al. 1992; Kim et al. 2021a).
- **LR** is one of the most popular classification algorithms based on a statistical method (Dreiseitl and Ohno-Machado 2002; Lee et al. 2021).

[4] https://cloud.google.com/speech-to-text?hl=ko.

[5] https://www.ncloud.com/product/aiService/clovaSpeech.

[6] https://konlpy.org/en/latest/.

[7] https://github.com/kakao/khaiii.

Table 4 Dataset for training and testing

	Full dataset		Top-5000 dataset	
	Vishing	Nonvishing	Vishing	Nonvishing
Average per case (voice)	275.46 s	848.12 s	269.22 s	788.22 s
Average per case (text)	1691 words	4864 words	1030 words	4021 words
Training cases	507 cases	1770 cases	465 cases	881 cases
Testing cases	108 cases	462 cases	104 cases	233 cases

- **DT** is a multistage decision-making approach that breaks a complex decision into a union of several simpler decisions (Safavian and Landgrebe 1991; Kim et al. 2019).
- **RF** is a combination of tree algorithms that independently sample random vectors (Breiman 2001; Hwang et al. 2020).
- **XGB** is a decision-tree-based ensemble algorithm using a gradient boosting framework (Kim et al. 2020; Stein et al. 2019).

Different parameters were used to evaluate each classifier when applying the two datasets. Table 5 lists the optimized classifier parameters selected using the grid search algorithm.[8]

3.4 Evaluation metric

The indexes of the confusion matrix as well as the true positive (TP), false positive (FP), false negative (FN), and true negative (TN) rates were used to calculate the following five performance metrics:

- Precision $= \frac{TP}{TP+FP}$
- Recall $= \frac{TP}{TP+FN}$
- F1-score $= 2 \times \frac{Precision \times Recall}{Precision + Recall}$
- Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$
- Time to training (s)

4 Results

Table 5 shows the performance metrics of the different classification models. Almost all classifiers achieved 100% accuracy when using the *full dataset*. In particular, the SVM classifier showed the highest performance in both morpheme analyses, OKT and Khaiii, with 100% accuracy. The LR classifier reported the fastest training time of 2.223 s, followed by the SVM models.

The SVM classifier again achieved the highest accuracy (100%) with both morpheme analyses when using the *Top-5000* dataset. Other machine learning classifiers, such as RF and XGB, also showed 100% accuracy, unlike the *full dataset*. In addition, the *Top-5000 dataset* required a relatively shorter training time than the *full dataset*. The Khaiii tagging LR classifier was the fastest with a training time of 0.244 s.

The detection time was presented. The fastest case was the LR method with OKT tagging applied to the *Top-5000 dataset*: a speed of 0.03 ms for the testing of each case was achieved.

5 Discussion and conclusions

Because most research into existing spam or vishing detection uses static text datasets, vishing detection in real time with actual Korean voice datasets is limited. Therefore, in the current study, to examine machine learning models for conducting real-time vishing detection tasks in South Korea, actual South Korean voice datasets were collected and features were extracted.

Rather than proposing a new algorithm for vishing detection, we investigated the potential and significance of real-time vishing detection in low-resource languages. We also examined whether the use of traditional machine learning approaches for real-time vishing detection can be an effective and economical solution to preventing potential vishing damage. Accordingly, we have kept our datasets and models open to the public. Although more complex machine learning/deep learning models can provide a better performance than the models employed in this study, they consume more time and resources.

In addition, considering that most prior vishing/spam detection studies conducted in real time have applied high-resource languages, such as English (Roy et al. 2020; Trivedi 2016), the findings of the current study can be a milestone in examining vishing detection tasks in low-resource languages.

Because it is essential to understand voice contexts and realize real-time vishing detection, one of the key characteristics of this study is the use of voice instead of a static text

[8] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html.

Table 5 Classification metrics; *Regularization parameter, **Inverse of regularization strength

			Full dataset							Top5000 dataset						
Tagging	Model	Class	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)	Time to Training (s)	Time to test per case(ms)	Parameter	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)	Time to Training (s)	Time to test per case (ms)	Parameter
OKT	SVM	Vishing	100	100	100	100	5.051	1.51	C* = 10	100	100	100	100	1.973	1.39	C = 10
		non-vishing	100	100	100				kernel = linear	100	100	100				kernel = linear
	LR	Vishing	100	99.21	99.60	99.83	2.223	0.06	C** = 100	100	99.15	99.57	99.70	1.117	0.02	C = 100
									penalty = l2							penalty = l2
		non-Vishing	99.78	100	99.89				max iter = 100	99.55	100	99.77				max iter = 100
									solver = lbfgs							solver = lbfgs
	DT	Vishing	96.15	99.21	97.66	98.95	8.440	0.14	Criteria = entropy	95.76	96.58	96.17	97.33	2.374	0.07	criterion = Gini
		non-vishing	99.77	98.87	99.32				max depth = 10	98.17	97.73	97.95				max depth = 20
	RF	Vishing	100	100	100	100	6.348	0.18	criterion = Gini	100	100	100	100	6.191	0.29	Criterion = Gini
		non-vishing	100	100	100				n estimators = 100	100	100	100				n estimators = 500
									max depth = 20							max depth = 5
	XGB	Vishing	100	100	100	100	72.170	0.11	Max depth = 3	100	100	100	100	27.580	0.10	maximum depth = 7
		non-vishing	100	100	100				n estimators = 1000	100	100	100				n estimators = 1000
									subsample = 0.6							subsample = 1.0
Khaiii	SVM	Vishing	100	100	100	100	5.959	3.80	C = 10	100	100	100	100	2.478	1.38	C = 1.0
		non-vishing	100	100	100				kernel = poly	100	100	100				kernel = poly
	LR	Vishing	100	98.44	99.21	99.65	1.604	0.03	C = 100	100	99.00	99.50	99.70	0.244	0.04	C = 100
									penalty = l2							penalty = l2
		non-Vishing	99.55	100	99.77				max iter = 100	99.58	100	99.79				max iter = 100

Table 5 (continued)

		Full dataset						Top5000 dataset						
								Solver = lbfgs						Solver = liblinear
DT	Vishing	95.42	97.66	96.53	98.42	10.839	0.13	Criterion = Gini	97.06	99.00	98.02	98.81	2.605	0.08	Criterion = Gini
	non-Vishing	99.32	98.64	98.98				max depth = 10	99.57	98.73	99.15				max depth = 5
RF	Vishing	100	99.22	99.61	99.83	23.900	0.36	Criterion = Gini	100	100	100	100	16.885	0.56	Gini
	non-Vishing	99.77	100	99.89				n estimators = 500	100	100	100				n estimators = 1000
								max depth = 10							max depth = 7
XGB	Vishing	100	99.22	99.61	99.86	71.551	0.15	Max depth = 3	100	100	100	100	27.580	0.07	Maximum depth = 7
	non-Vishing	99.77	100	99.89				n estimators = 1000	100	100	100				n estimators = 1000
								subsample = 1.0							subsample = 1.0

Table 6 Data analysis

Vishing	Non−vishing data
본인의, 부분, 그렇죠, 먼저, 없어요, **대출**, 잠시만요, 이번, **조사**, 계속, 건데, 말씀해, 수가, **통장**이, 전화를, 부분에, **은행**, 아니요, 앞으로, 전화, 됩니다, 거고, 있는데, 말씀을, 보니까, 많이, 하면, 고객님께서, 해야, 만약에, 부분이, **불법**, **피해자**, 아니, **조사**를, 네네, **통장**을, 명의로, 되면, 돼요, 하나, 있습니다, 연락을, 거죠, 아니면, 알고, 한번, 바로, **통장**, 저는, 거고요, 직접, 일단은, 전혀, 어떤, **계좌**, 오늘, 그럼, 여보세요, 있어요, 하고, 하는, 다른, 그렇게, 정도, 알겠습니다, 이거, 현재, 본인께서, 이게, 되는, 그냥, 다시, 그러니까, 그런데, 같은, 해서, 사건, 겁니다, 그런, 이런, 대해서, 혹시, 그리고, 그래서, 어떻게, 일단, 고객님, 본인이, 그러면, 이렇게, 있는, 저희, 때문에, 본인, 거예요, 이제, 저희가, 제가, 지금	우리가, 영화, 기억이, 거의, 적이, 거예요, 요즘, 정도, 그래가지고, 없는, 같고, 사람이, 있어서, 하면, 많은, 그거, 나도, 항상, 사람들이, 일단, 엄청, 너는, 거를, 했는데, 뭔가, 있어, 저희, 아니면, 해야, 한번, 그러면, 계속, 되고, 있어요, 가서, 먹고, 하는데, 되는, 싶은, 보면, 거야, 대해서, 사실, 쪼끔, 때는, 그래도, 좋아하는, 거는, 가장, 제일, 보고, 저도, 이게, 생각이, 그때, 좋은, 있는데, 굉장히, 다른, 그러니까, 어떻게, 것도, 같이, 해서, 있고, 약간, 그게, 우리, 같아, 진짜, 보니까, 정말, 생각을, 어떤, 혹시, 때문에, 같아요, 지금, 하는, 그렇게, 나는, 같은, 그리고, 있는, 너무, 내가, 되게, 그냥, 제가, 인제, 저는, 하고, 이런, 이렇게, 근데, 조금, 많이, 그래서, 이제, 그런

Table 7 Comparison of speech-to-text tools

text	Google STT API	Naver Clova Speech
요부분 확인 들어갈 때 제가 전화 드리고 이렇게 진행을 도와드리니까 조금만 기다려주시면 되고, 빠른 진행 도와드릴 수 있도록 하겠습니다.	제가 전화를 드리고 이렇게 지내는 전화 드리니까 조금만 기다려 주시면 되고, 받으시네 도와드릴 수 있도록 하여 드리겠습니다.	그 부분 확인 들어갈 때 제가 전화를 드리고 이렇게 진행을 도와드리니까 조금만 기다려주시면 되고, 아버지네 도와드릴 수 있도록 하겠습니다.
예 안녕하십니까. 저는 ** 은행 **이라고 합니다. 네 그러면 저희 쪽으로 음성 안내 메시지 들으시고 마이너스 통장 생각 문의하시는 분 아니십니까.	예 안녕하십니까. 저희 집으로 음성 안내 메시지 드리시고 마이너스 통장 생각 못하신 분 아니십니까.	예 안녕하십니까. 저희 은행 애기라고 합니다. 네 그러면 저희 쪽으로 음성 안내 메시지 들으시고 마이너스 통장 생각 문의하시는 분 아니십니까.

** Beep processing for personal information

**Beep processing for personal information

dataset. Although several prior studies have presented notable findings on real-time vishing detection (Zhang and Gurtov 2009; Tran et al. 2020), we only employed voice contexts as our features for exploring real-time vishing detection.

Moreover, most of our classifiers have achieved a training time of less than 10 s (except for the XGB classifier). Therefore, because rapid detection is one of the most important criteria in real-time vishing detection, the studied approaches can be practically employed.

5.1 Limitations and future research

Because Korean is one of the low-resource languages, there are certain limitations. For example, the collected datasets were imbalanced and insufficient for examining other sampling techniques. Moreover, the presented approaches and other traditional detection methods should be compared to better understand vishing detection. Although we employed machine learning techniques in this study for rapid vishing detection, there are several notable approaches employing neural network architectures (Wei and Nguyen 2020). According to these latest trends, it will be necessary to consider a lightweight neural network model in the future. Finally, several prior studies have analyzed the characteristics and features of speech cases in machine learning and deep learning models (Shen et al. 2018; Arık et al. 2017). However, we did not consider any characteristics or features of the speech cases, and such a study could be a significant milestone in future vishing detection research. Because we aimed to detect Korean vishing based on South Korean conversation data, our approach can be applied not only to vishing detection, but also to addressing spam messages in the chat programs of a number of social networking services. Moreover, this approach can be used to explore improper information/contents detection in real-time broadcasting environments.

Appendix A. Data analysis

Table 6 shows the top-100 most widely used words in spam and nonspam cases when analyzing spam and nonspam text content, respectively. The nonspam cases mostly included everyday words, such as *us, movies, people, and*

me, whereas the spam cases included words such as *loan, investigation, bank accounts, bank, illegality, and victims* (given in bold). Therefore, understanding the meaning of these words is important in vishing detection.

Appendix B. Speech-to-text tool examples

We converted the collected.mp3 files of voice phishing speech, and the results when using actual voice scripts, *Google speech-to-text API*, and *Naver Clova Speech* speech-to-text conversion tools are shown in Table 7.

Acknowledgements This work was supported by an Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. IITP-2021-0-00358, AI big data-based cyber security orchestration, and automated response technology development). Moreover, this research was supported by National Research Foundation (NRF) of Korea Grant funded by the Korean Government (MSIT) (No. 2021R1A4A3022102).

Author contributions ML and EP designed the study. ML collected and analyzed the data. EP presented the results. ML and EP wrote and revised the manuscript. All authors reviewed the manuscript.

Data availability Statement The datasets collected for the current study are available at https://anonymous.4open.science/r/vishing-1AF8. Additional information used in this study can be obtained from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors have no conflicts or competing interests to declare.

References

Abu-Nimeh S, Nappa D, Wang X, Nair S (2007) A comparison of machine learning techniques for phishing detection. In: Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit, ACM, pp 60–69

Akinyelu AA, Adewumi AO (2014) Classification of phishing email using random forest machine learning technique. J Appl Math 2014:425731

Arık SÖ, Chrzanowski M, Coates A, Diamos G, Gibiansky A, Kang Y, Li X, Miller J, Ng A, Raiman J et al (2017) Deep voice: real-time neural text-to-speech. In: Proceedings of the International Conference on Machine Learning, PMLR, pp 195–204

Barraclough PA, Hossain MA, Tahir M, Sexton G, Aslam N (2013) Intelligent phishing detection and protection scheme for online transactions. Expert Syst Appl 40(11):4697–4706

Biswal S (2021) Real-time intelligent vishing prediction and awareness model (rivpam). In: Proceedings of the 2021 international conference on cyber situational awareness. Data Analytics and Assessment (CyberSA), IEEE, pp 1–2

Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on Computational learning theory, ACM, pp 144–152

Breiman L (2001) Random forests. Mach Learning 45(1):5–32

Choi K, Jl L, Yt C (2017) Voice phishing fraud and its modus operandi. Secur J 30(2):454–466

Cook S (2021) 35+ phone spam stattistics for 2017–2021. https://www.comparitech.com/blog/information-security/phone-spam-statistics/

Dreiseitl S, Ohno-Machado L (2002) Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inform 35(5–6):352–359

Drucker H, Wu D, Vapnik VN (1999) Support vector machines for spam categorization. IEEE Trans Neural Networks 10(5):1048–1054

Ghourabi A, Mahmood MA, Alzubi QM (2020) A hybrid cnn-lstm model for sms spam detection in Arabic and English messages. Future Internet 12(9):156

Gómez Hidalgo JM, Bringas GC, Sánz EP, García FC (2006) Content based sms spam filtering. In: Proceedings of the 2006 ACM symposium on Document engineering, ACM, pp 107–114

Gorham M (2019) 2018 internet crime report. https://www.ic3.gov/Media/PDF/AnnualReport/2018_IC3Report.pdf

Gupta H, Jamal MS, Madisetty S, Desarkar MS (2018) A framework for real-time spam detection in twitter. In: Proceedings of the 2018 10th international conference on communication systems & networks (COMSNETS), IEEE, pp 380–383

Hwang S, Kim J, Park E, Kwon SJ (2020) Who will be your next customer: a machine learning approach to customer return visits in airline services. J Bus Res 121:121–126

Kadoya Y, Khan MSR, Yamane T (2020) The rising phenomenon of financial scams: evidence from Japan. J Financial Crime 27(2):387–396

Kenton JDMWC, Toutanova LK (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies (NAACL-HLT), pp 4171–4186

Kim J, Bae K, Park E, del Pobil AP (2019) Who will subscribe to my streaming channel? The case of twitch. In: Conference companion publication of the 2019 on computer supported cooperative work and social computing (CSCW Companion), pp 247–251

Kim J, Lee J, Park E, Han J (2020) A deep learning model for detecting mental illness from user content on social media. Sci Rep 10(1):1–6

Kim J, Hwang S, Park E (2021a) Can we predict the Oscar winner? A machine learning approach with social network services. Entertain Comput 39:100441

Kim JW, Hong GW, Chang H (2021b) Voice recognition and document classification-based data analysis for voice phishing detection. Human-Centric Comput Info Sci 11:2

Korea Financial Supervisory Service (2021) Analysis of voice phishing status in 2020. https://www.fss.or.kr/fss/kr/promo/bodobbs_view.jsp?seqno=23836

Korea National Police Agency (2020) Voice phishing status. https://www.data.go.kt/data/15063815/fileData.do

Koøcz A, Alspector J (2001) SVM-based Filtering of E-mail Spam with Content-specific Misclassification Costs. In: Proceedings of the workshop on text mining (TEXTDM), Citeseer, pp 1–14

Lee S, Ji H, Kim J, Park E (2021) What books will be your bestseller? A machine learning approach with amazon kindle. Electron Libr 39(1):137–151

Li Z, Nie F, Chang X, Nie L, Zhang H, Yang Y (2018a) Rank-constrained spectral clustering with flexible embedding. IEEE Trans Neural Netw Learning Syst 29(12):6073–6082

Li Z, Nie F, Chang X, Yang Y, Zhang C, Sebe N (2018b) Dynamic affinity graph construction for spectral clustering using multiple features. IEEE Trans Neural Netw Learning Syst 29(12):6323–6332

Li Z, Yao L, Chang X, Zhan K, Sun J, Zhang H (2019) Zero-shot event detection via event-adaptive concept relevance mining. Pattern Recogn 88:595–603

Mccord M, Chuah M (2011) Spam detection on twitter using traditional classifiers. In: Proceedings of the international conference on autonomic and trusted computing (ATC), Springer, pp 175–186

Obuhuma J, Zivuku S (2020) Social engineering based cyber-attacks in kenya. In: Proceedings of the 2020 IST-Africa conference (IST-Africa), IEEE, pp 1–9

Raj H, Weihong Y, Banbhrani SK, Dino SP (2018) Lstm based short message service (sms) modeling for spam classification. In: Proceedings of the 2018 International Conference on Machine Learning Technologies, pp 76–80

Ren P, Xiao Y, Chang X, Huang PY, Li Z, Chen X, Wang X (2021) A comprehensive survey of neural architecture search: challenges and solutions. ACM Comput Surveys (CSUR) 54(4):1–34

Roy PK, Singh JP, Banerjee S (2020) Deep learning to filter sms spam. Futur Gener Comput Syst 102:524–533

Safavian SR, Landgrebe D (1991) A survey of decision tree classifier methodology. IEEE Trans Syst Man Cybern 21(3):660–674

Sasaki M, Shinnou H (2005) Spam detection using text clustering. In: Proceedings of the 2005 international conference on cyberworlds (CW), IEEE, pp 1–4

Shen J, Pang R, Weiss RJ, Schuster M, Jaitly N, Yang Z, Chen Z, Zhang Y, Wang Y, Skerrv-Ryan R et al (2018) Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In: Proceedings of the 2018 IEEE international conference on acoustics. Speech and Signal Processing (ICASSP), IEEE, pp 4779–4783

Song J, Kim H, Gkelias A (2014) ivisher: real-time detection of caller id spoofing. ETRI J 36(5):865–875

Stein RA, Jaques PA, Valiati JF (2019) An analysis of hierarchical text classification using word embeddings. Inf Sci 471:216–232

Sun N, Lin G, Qiu J, Rimba P (2020) Near real-time twitter spam detection with machine learning techniques. Int J Comput Appl. https://doi.org/10.1080/1206212X.2020.1751387

Tran MH, Le Hoai TH, Choo H (2020) A third-party intelligent system for preventing call phishing and message scams. In: Proceedings of the international conference on future data and security engineering (FDSE), Springer, pp 486–492

Trivedi SK (2016) A study of machine learning classifiers for spam detection. In: Proceedings of the 2016 4th international symposium on computational and business intelligence (ISCBI), IEEE, pp 176–180

Wei F, Nguyen T (2020) A lightweight deep neural model for sms spam detection. 2020 International Symposium on Networks. Computers and Communications (ISNCC), IEEE, pp 1–6

Wijaya A, Bisri A (2016) Hybrid decision tree and logistic regression classifier for email spam detection. In: 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE), IEEE, pp 1–4

Wu T, Liu S, Zhang J, Xiang Y (2017) Twitter spam detection based on deep learning. In: Proceedings of the australasian computer science week multiconference (ACSW), ACM, pp 1–8

Yan C, Chang X, Luo M, Zheng Q, Zhang X, Li Z, Nie F (2020) Self-weighted robust lda for multiclass classification with edge classes. ACM Trans Intell Syst Technol (TIST) 12(1):1–19

Yeboah-Boateng EO, Amanor PM (2014) Phishing, smishing & vishing: an assessment of threats against mobile devices. J Emerg Trends Comput Inf Sci 5(4):297–307

Zhang R, Gurtov A (2009) Collaborative reputation-based voice spam filtering. In: Proceedings of the 2009 20th international workshop on database and expert systems application, IEEE, pp 33–37