Black-box model functionality stealing for Vietnamese sentiment analysis

Pham Xuan Cong[1], Do Viet Binh[1*], Hoang Trung Nguyen[2], Tran Cao Truong[2]

[1]Institute of Information Technology and Electronics, Academy of Military Science and Technology, 17 Hoang Sam, Cau Giay, Hanoi, Vietnam;
[2]Le Quy Don Technical University, 236 Hoang Quoc Viet, Bac Tu Liem, Hanoi, Viet Nam.
[*]Corresponding author: binhdv@gmail.com
Received 8 Mar. 2025; Revised 14 May. 2025; Accepted 10 Jun. 2025; Published 25 Jun. 2025.
DOI: https://doi.org/10.54939/1859-1043.j.mst.104.2025.144-154

ABSTRACT

Black-box deep learning models often keep critical components such as model architecture, hyperparameters, and training data confidential, allowing users to observe only the inputs and outputs without understanding their internal workings. Consequently, there is growing interest in developing "knockoff" models that replicate the behavior of these black-box models without direct access to internal details. We have conducted extensive studies on function extraction attacks targeting English text sentiment analysis models. By employing random or adaptive sampling methods, we have successfully reconstructed knockoff models that achieve functionality equivalent to the original models with high similarity. In this study, we extend our investigation to sentiment analysis datasets in Vietnamese. Experimental results demonstrate that for black-box models in Vietnamese text sentiment analysis, our method remains effective, successfully constructing models with equivalent functionality.

Keywords: Knockoff model; Black-box model functionality extraction; Vietnamese text sentiment analysis.

1. INTRODUCTION

Black-box models are systems where users can observe the inputs and outputs but cannot fully understand their internal information, such as architecture, parameters, or training datasets. This limitation has driven extensive research into the development of knockoff models, which are designed to replicate the decision-making process of the original (victim) models without requiring access to their structure or parameters. This approach, known as black-box model function extraction, involves an attacker continuously feeding input data into the black-box model and observing its corresponding output predictions. The collected input-output pairs are then used to train the knockoff model through knowledge distillation [14].

Most existing research on knockoff models has primarily focused on computer vision (CV), where techniques for replicating model behavior have been extensively studied [14]. However, researchers have given significantly less attention to the field of natural language processing (NLP), particularly in applications such as sentiment analysis. This gap highlights the need for further research to develop more effective methods for mimicking black-box models in NLP.

In Vietnam, there is almost no research on function extraction for Vietnamese language models in general, and specifically for Vietnamese text sentiment analysis. The selection of input data for extracting the functionality of a black-box model is a key research direction. We have studied and proposed both random and adaptive sampling methods for function extraction attacks on black-box models in the context of English text sentiment analysis [11, 12]. Experimental results demonstrate that our approach successfully reconstructs knockoff models with equivalent functionality and high agreement. In this study, we implement and compare the effectiveness of these methods on Vietnamese text sentiment analysis models.

Moreover, the practical implications of functionality stealing attacks remain underexplored in current literature, especially regarding their potential impact on real-world applications. In commercial sentiment analysis systems, such attacks may allow competitors to replicate

proprietary models without incurring the costs of data collection and training, leading to unfair economic advantages. In governmental or public administration contexts, adversaries may use knockoff models to manipulate or intercept public opinion analysis tools, potentially distorting political communication, misinforming decision-making, or compromising national information security. These risks highlight the need to understand and mitigate model stealing threats in both business and policy environments.

The remainder of this paper is structured as follows: section II provides a review of existing research on black-box model extraction, sampling strategies, and sentiment analysis. Section III details the proposed methodology for extracting model functions. Section IV presents the results, and section V concludes with future directions.

2. RELATED WORK

2.1. Black-box model stealing by data sampling

Model functional stealing attacks in machine learning involve an adversary attempting to infer or replicate the functionality of a target (victim) model by analyzing its outputs [14]. We analyze the latest studies on model extraction based on three key aspects: attack targets, sampling strategies, and data sources. Most prior research has focused on extracting black-box models in computer vision [1, 9], particularly deep neural networks. However, our study shifts the focus toward function extraction in pre-trained language models (PLMs) such as [2, 17]. Regarding data sources, some studies have concentrated on online machine learning services (MLaaS) [10, 19], whereas we leverage publicly available Vietnamese datasets [3, 7]. For sampling strategies, methods can include random sampling or adaptive approaches based on reinforcement learning (RL) and active learning (AL) [2, 9, 11, 12].

Using active learning-based sampling techniques, S. Pal et al. [10] introduced a novel attack framework called ActiveThief to extract models from MLaaS. ActiveThief leverages active learning and publicly available unlabeled data to construct a substitute model with functionality equivalent to the original black-box model. This approach exploits various sampling strategies to optimize the querying process for the target model. Later studies built upon the work of S. Pal et al. [10], some sampling techniques have been applied. Further studies by C. Dai et al. [2] and W. Wu et al. [17] have expanded and refined these approaches across additional datasets. C. Dai et al. [2] introduced a novel model extraction attack method called MeaeQ, designed to enhance the efficiency of model extraction attacks in the field of natural language processing. W. Wu et al. [17] proposed the PEEP model extraction attack framework, which extracts models from sentiment analysis APIs using only queries. PEEP leverages publicly available data as substitute training data while employing a new query strategy and a greedy search algorithm to identify the optimal architecture for the extracted model.

T. Orekondy et al. [9] proposed a method for simulating the functionality of image classification models using knowledge distillation in combination with data sampling via reinforcement learning. Their experiments on various datasets demonstrated that knockoff models could effectively replicate the functionality of the victim model with comparable accuracy. In another study, Akshit Jindal et al. [1] trained a group of models with varying levels of complexity to exploit the advantages of collective intelligence. Based on the consensus of the model ensemble, uncertain samples were selected for querying the victim model, while the most stable samples were directly incorporated into the training dataset.

Inspired by the work of T. Orekondy et al. [9] in image classification, our previous studies [11, 12] proposed random and reinforcement learning-based adaptive sampling techniques for functionality extraction attacks on black-box models in English text classification tasks within NLP.

2.2. Vietnamese text sentiment analysis

Sentiment analysis is a crucial task in NLP that involves classifying the sentiment expressed in a given text [15]. In recent years, transformer-based models have revolutionized the field of NLP, including text sentiment analysis. Transformers have achieved state-of-the-art performance by leveraging self-attention mechanisms to capture contextual information from text sequences.

Vietnamese sentiment analysis faces greater challenges than English due to limited annotated data, complex morphology, and ambiguous word segmentation. Unlike English, Vietnamese words often comprise multiple syllables separated by spaces, making preprocessing crucial. Sentiment polarity is highly context-dependent, requiring advanced features and external resources. Recent studies have focused on enhancing performance using pre-trained Transformer models like PhoBERT [4], viBERT [8], and ViSoBERT [6]. Despite progress, handling informal language and data scarcity remains a key research direction.

3. PROBLEM AND METHODS

3.1. Vietnamese text sentiment problem

As highlighted in our recent studies [11, 12], the problem of black-box model function extraction for deep learning models in Vietnamese NLP can be formally stated as follows. The hypothesis considers a black-box model $F_V : X_V \to Y_V$ (referred to as the victim or target model V), designed for Vietnamese text sentiment analysis, with internal details that remain unknown. The function of this victim model is to take a text input and classify it as positive, negative, or neutral. Model V operates on a sentiment analysis dataset X_V, assigning labels to form an annotated dataset $D_V = \{(x_i, y_i) | x_i \in X_V, y_i = F_V(x_i)\}$, which is further divided into a training set D_V^{train} and a test set D_V^{test}. The model F_V is trained on the training set D_V^{train} and evaluated on the test set D_V^{test} [11, 12].

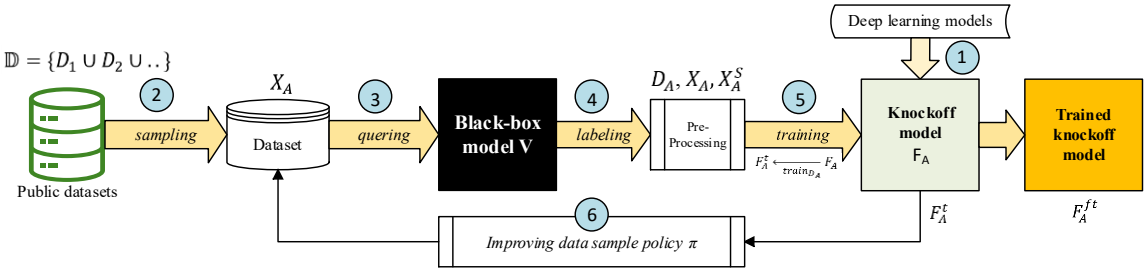

Figure 1. *Overview of text sentiment analysis model extraction process [11].*

The extraction process terminates when a predefined number of iterations is reached or when the data budget B is exhausted. Finally, the knockoff model F_A undergoes a final fine-tuning phase to produce F_A^{ft} (*"ft" stands for final trained)*. The reward scores and sampling policy are derived from the Gradient Bandits algorithm, with minor enhancements to improve efficiency. One such enhancement is the Adaptive Sampling based on Data sources (ASbD) technique, which may help efficiently identify records that were used to train the target model, particularly in scenarios where such records are present in the data.

Without access to the internal details of the victim model V, we aim to develop a method to train a knockoff model F_A that replicates the functionality of V with optimal accuracy and similarity. Using publicly available Vietnamese sentiment analysis datasets and deep learning models, the goal is to select a model F_A trained on a dataset D_A such that F_A meets two key criteria: Achieving accuracy comparable to that of F_V; and maximizing similarity to F_V (specific similarity measures will be detailed below).

3.2. Solutions

3.2.1. Model extraction method

Algorithm 1: Adaptive sampling for model extraction

Input:

$\mathbb{D}$: Vietnamese text sentiment datasets; B : the query budget of data

F_V : Victim model V; $\mathbb{F}$: the set of PLM models for text sentiment

Output: $F_A^{ft} \in \mathbb{F}$: the final trained knockoff model A

Steps:

$F_A \leftarrow \mathbb{F}$ ▷ select a suitable model;

$X_A \overset{B}{\leftarrow} \mathbb{D}$ ▷ randomly select B data from $\mathbb{D}$

Initialize collections and variables;

while $t < B$ do

$x_t \leftarrow$ sample based on policy π on X_A;

$D_A = \{(x_t \in X_A, F_V(x_t))\}, \quad X_A = X_A \setminus \{x_t\}, \quad X_A^S = X_A^S \cup \{x_t\}_t$;

$F_A^t \leftarrow$ train on D_A; $\quad \hat{y}_t = F_A^t(x_t)$;

Calculating rewards: $R^*(\hat{y}_t)$, $R(y_t, \hat{y}_t)$;

Updating learning rate α; Updating policy H_t and π_t;

end

$F_A^{ft} \leftarrow$ final train on X_A^S

return F_A^{ft}

The process of extracting the functionality of a black-box text sentiment analysis model is illustrated in Figure 1 and Algorithm 1, comprising six main steps as follows:

Step 1: Select a suitable deep learning model F_A as the knockoff model.

Step 2: Assemble multiple datasets $D_1, D_2, ...$ for the Vietnamese text sentiment analysis task to form a dataset pool $\mathbb{D} = D_1 \cup D_2 \cup ...$

Step 3: Apply a suitable sampling strategy (by budget B) to extract data from $\mathbb{D}$. The selected data x is continuously queried into the victim model F_V to obtain labeled responses $y = F_V(x)$.

Step 4: Construct the transfer dataset D_A by pairing the obtained labeled data $\{x, y = F_V(x)\}$ after a pre-processing stage, which is then used to train F_A.

Step 5: Train the knockoff model F_A using the dataset D_A.

Step 6: Evaluate F_A and compute reward scores to update the sampling policy for the next iteration, improving the effectiveness of subsequent sampling.

3.2.2. Sampling strategy

To sample the dataset $\mathbb{D}$ for training the knockoff model, we select B samples from $\mathbb{D}$ and pass them through the victim model to construct the transfer set. One approach is random sampling (RS), which is simple but may not be the most effective for imitation. Alternatively, we adopt the Bandit Gradient Algorithm [9, 11] from reinforcement learning to implement an adaptive sampling (AS) strategy.

In the AS method, there are k potential actions, each associated with a preference value $H(a)$. The probability of selecting an action a, denoted as $\pi(a)$, is computed using the softmax function:

$$\pi(a) = \frac{e^{H(a)}}{\sum_{b=1}^{k} e^{H(b)}} \tag{1}$$

Initially, all preference values $H(a)$ are set to 0. At each time step t, an action a_t is selected based on the probability distribution $P(a)$, and a reward R_t is obtained. The average reward $\bar{R}_t$ up to time t is then calculated. Using the learning rate α, the preference values $H(a)$ are updated as follows:

$$H(a) = \begin{cases} H(A_t) + \alpha(R_t - \bar{R}_t)(1 - \pi(A_t)), & a = A_t \\ H(a) - \alpha(R_t - \bar{R}_t)\pi(a), & a \neq A_t \end{cases} \tag{2}$$

To evaluate the quality of sampled data at the time t, we apply specific reward metrics, including confidence, diversity, and similarity. *Confidence* reward is based on the margin metric, which measures the difference in predicted probabilities between the most confident class and the second most confident class in the output. *Diversity* reward ensures that data selection does not overly focus on a single class, preventing bias. *Similarity* reward prioritizes samples where the knockoff model produces outputs similar to those of the victim model. The final total reward score R_t is computed as a balanced combination of these individual metrics [11].

We implement the adaptive algorithm in two directions: ASbD and ASbC techniques [11]. The Adaptive Sampling by Dataset (ASbD) technique prioritizes sample selection based on the origin of datasets within the overall pool, aiming to identify subsets that yield better imitation results. This technique is only applicable in the closed-case scenario, where the dataset may contain samples that were used to train the victim model. In contrast, the Adaptive Sampling by Classification (ASbC) technique prioritizes samples based on the classification outputs of knockoff models that currently align with those of the victim model.

3.2.3. Evaluation metrics

The objective of the attack is to train a knockoff model F_A^{ft} that replicates the behavior of F_V on the same input domain. The performance of F_A^{ft} is evaluated based on two key criteria: similarity and accuracy.

Similarity is measured using the *Agreement* metric [9, 11], which quantifies how closely F_A^{ft} mimics F_V. Since the test set D_V^{test} is unknown, it is only used to assess the effectiveness of F_A^{ft} based on the following similarity measure:

$$Agreement(D_V^{test}) = \frac{1}{\left|D_V^{test}\right|} \sum_{x \in D_V^{test}} I(F_A^{ft}(x) = F_V(x)) \tag{3}$$

where I(·) is the indicator function, which returns 1 if $F_A^{ft}(x) = F_V(x)$ and 0 otherwise, and $\left|D_V^{test}\right|$ is the total number of records in the test set D_V^{test}. A higher *Agreement* score indicates that F_A^{ft} more closely resembles F_V, demonstrating successful imitation of the victim's behavior.

Accuracy is measured using the F1-score, which compares the classification performance of F_A^{ft} with that of the victim model F_V.

3.2.4. Datasets

In our experiments on model extraction attacks for Vietnamese text sentiment analysis, we evaluated several pre-trained language models (PLMs) on the VLSP-2016, AIVIVN-2019, UIT-VSFC, and UIT-ViSFD datasets.

We utilize four widely used Vietnamese sentiment analysis datasets. **VLSP-2016** contains short user reviews on technological devices, labeled as positive, negative, or neutral [7]. **AIVIVN-2019** is the official dataset from the AIVIVN Sentiment Challenge, featuring around 160,000 labeled product reviews and 11,000 unlabeled ones collected from e-commerce platforms. **UIT-VSFC [5]** includes over 16,000 manually annotated sentences for both sentiment and topic classification, split into training, validation, and test sets. Lastly, **UIT-ViSFD** [13] is a benchmark dataset for aspect-based sentiment analysis, consisting of over 11,000 human-annotated smartphone reviews in the mobile e-commerce domain.

3.2.5. Models

PhoBERT [4], based on the RoBERTa architecture, is the first large-scale monolingual model trained exclusively on Vietnamese text. Unlike multilingual BERT-based models, which do not account for the unique word segmentation challenges in Vietnamese, PhoBERT applies word-level tokenization before subword encoding, resulting in significant improvements across various NLP tasks. **viBERT** [8] is a pre-trained BERT model trained on 10GB of Vietnamese text collected from online newspapers, utilizing subword tokenization. This model follows the BERT architecture and is specifically designed for Vietnamese NLP tasks, particularly sequence tagging, part-of-speech tagging, and named entity recognition. **ViSoBERT** [6] is a pre-trained Transformer model developed for processing Vietnamese social media text. Unlike general-purpose Vietnamese language models, ViSoBERT is trained on informal and noisy text from social media platforms, making it highly effective in handling teencode, emojis, abbreviations, and unstructured language.

Table 1. *F1-score of PLM models over Vietnamese text sentiment datasets.*

Model	AIVIVN-2019	UIT-VSFC	UIT-ViSFD	VLSP-2016
PhoBERT	95.29	96.48	92.97	88.75
viBERT	93.69	95.50	92.84	83.61
VisoBERT	95.54	**96.61**	93.55	88.72

These PLM models are directly utilized by the developers or accessed via the Hugging Face library. Table 1 presents the experimental results of PLM models on Vietnamese text sentiment analysis datasets. It can be observed that ViSoBERT achieved the highest performance on the UIT-VSFC dataset, with an accuracy of 96.61%. Therefore, we selected ViSoBERT trained on the UIT-VSFC dataset as the target black-box model, while the other two models were designated as knockoff models for the experiments in this study.

In addition, in our previous studies [11, 12] on English text sentiment model extraction, we experimented with DistilBERT and RoBERTa, achieving effective feature extraction from victim models. In this study, we also implement these models for Vietnamese text data to provide a more comprehensive evaluation. DistilBERT [16] is a lightweight and faster version of BERT, optimized for resource-constrained environments while maintaining competitive performance. RoBERTa [18] improves upon BERT by utilizing additional pre-training data and optimizations, enhancing its effectiveness across various NLP tasks.

3.2.6. Parameters setting

Unlike experiments conducted on English datasets, where large-scale corpora are readily available, Vietnamese public datasets contain relatively fewer records. Therefore, in this study, we use the largest possible sampling budgets from the original datasets, including 1000, 2500, 5000, 10000, and 20000 samples. The sentence length in the datasets does not exceed 256 tokens. We

evaluate two data scenarios: the ***closed-case***, where the sampled dataset contains the training data of the target model, including all four datasets mentioned in section 2.2.4; and the ***open-case***, where the sampled dataset excludes the target model's training data and consists only of AIVIVN, UIT-ViSFD, and VLSP-2016. For training, we employ 4 GeForce GTX 1080 Ti 11GB GPUs. The maximum input length for PhoBERT, viBERT, and ViSoBERT is set to 256, with a training batch size of 16 and 3 epochs. The number of training records is determined by the data budget, with 100 samples per iteration. Meanwhile, RoBERTa and DistilBERT are trained with a maximum sequence length of 512, a batch size of 8.

4. RESULTS AND DISCUSSION

4.1. Result of random sampling

Table 2 presents the performance of 4 knockoff models in extracting features from the Vietnamese text sentiment analysis model using random sampling in both open-case and closed-case scenarios. The results clearly indicate that PhoBERT achieves the highest accuracy and agreement in both cases, closely approaching the accuracy of the victim model (96.61% in table 1). Specifically, PhoBERT consistently demonstrates superior performance, attaining an F1-score of 96.08% at a budget of 20K, with only a 0.53-point difference from the victim model. In comparison, viBERT follows with 94.00%, while RoBERTa and DistilBERT perform considerably lower, reaching only 89.64% and 89.32%, respectively, at the same budget. This trend remains consistent across all budgets, further reinforcing PhoBERT's superiority over the other models.

Table 2. *Performance of knockoff models on random sampling.*

Knockoff model	**Budget**	*Open-case*		*Closed-case*	
		F1-score	**Agreement**	**F1-score**	**Agreement**
PhoBERT	1000	59.94	70.79	88.21	89.46
	2500	90.41	91.73	93.29	94.46
	5000	92.33	93.23	95.48	**96.80**
	10000	93.36	94.30	95.14	96.17
	20000	**94.02**	**95.20**	**96.08**	96.40
viBERT	1000	47.23	62.29	80.47	79.16
	2500	75.10	79.33	87.10	89.00
	5000	80.31	82.93	90.72	92.16
	10000	84.39	85.83	93.33	94.80
	20000	86.07	86.70	94.00	95.27
RoBerta	1000	43.38	61.79	76.29	69.66
	2500	57.55	68.76	84.90	85.26
	5000	70.69	75.29	86.46	88.86
	10000	72.49	77.23	88.48	89.83
	20000	75.41	79.13	89.64	90.43
DistilBert	1000	71.05	73.96	81.03	80.29
	2500	75.29	76.29	85.02	84.83
	5000	76.14	79.49	85.32	86.90
	10000	77.68	80.59	87.68	89.10
	20000	78.28	80.69	89.32	90.83

In the open-case scenario, PhoBERT again leads with an F1-score of 94.02%, followed by viBERT at 86.07%, while RoBERTa and DistilBERT lag behind with maximum scores of 75.41% and 78.28%, respectively. The agreement scores exhibit a similar trend, with PhoBERT achieving

95.20%, while viBERT, RoBERTa, and DistilBERT follow with 86.70%, 79.13%, and 80.69%, respectively. These results suggest that PhoBERT and viBERT provide a significantly better approximation of the victim model, particularly in closed-case settings, whereas DistilBERT and RoBERTa struggle to achieve comparable performance.

4.2. Result of adaptive sampling

Table 3. *Performance of knockoff models on adaptive sampling.*

Knockoff	**Budget**	**Open-case**				**Closed-case**			
		ASbC		*ASbD*		*ASbC*		*ASbD*	
		F1-score	**Agree**	**F1-score**	**Agree**	**F1-score**	**Agree**	**F1-score**	**Agree**
PhoBERT	1000	85.67	86.99	91.20	92.00	89.81	91.41	92.99	94.42
	2500	91.12	91.75	93.36	94.59	95.52	96.57	95.25	96.20
	5000	92.90	93.51	94.21	95.19	95.71	97.01	95.69	96.57
	10000	94.23	95.35	94.36	95.50	96.24	97.55	96.12	97.24
	20000	**94.88**	**95.79**	**94.71**	**95.83**	**96.54**	**97.66**	**96.39**	**97.38**
viBERT	1000	68.18	73.44	77.43	78.70	81.37	81.56	87.07	87.91
	2500	81.67	82.64	82.73	84.10	90.86	93.04	91.29	92.76
	5000	84.89	85.63	84.48	86.51	92.09	93.99	93.43	95.11
	10000	86.64	87.43	86.15	87.91	93.47	95.35	94.62	96.23
	20000	87.00	87.70	86.71	88.18	94.40	96.09	95.17	96.87
RoBerta	1000	66.21	71.67	73.23	73.51	79.99	80.74	78.17	74.32
	2500	71.76	74.69	69.48	73.17	83.73	83.25	84.57	85.02
	5000	73.35	76.66	75.36	76.63	89.23	90.18	87.08	88.01
	10000	76.37	78.94	76.44	78.36	91.27	92.36	89.00	90.66
	20000	75.69	78.46	79.95	80.10	91.80	93.61	90.31	91.78
DistilBert	1000	64.77	69.80	49.95	50.31	83.12	81.90	79.39	79.65
	2500	76.85	76.70	73.90	76.32	88.97	89.37	86.31	86.58
	5000	77.74	77.24	72.86	76.70	89.75	90.93	88.58	89.40
	10000	79.52	80.47	75.77	77.45	91.06	92.87	89.95	91.17
	20000	78.96	79.28	78.06	80.26	91.97	94.29	92.03	93.61

The results in table 3 provide insights into the effectiveness of knockoff models under adaptive attack settings, where the attacker has greater control over the sampling process. When comparing the four models, significant differences in performance are observed across both closed-case and open-case scenarios using the ASbC and ASbD methods.

The results demonstrate that PhoBERT consistently outperforms other models. It achieves the highest F1-scores and Agreement in both open and closed-case settings. At a 20K budget, PhoBERT's F1-score reaches 96.54% on ASbC and 96.39% on ASbD in the closed-case scenario, while maintaining strong open-case performance with 94.88% on ASbC and 94.71% on ASbD. viBERT follows closely, showing slightly lower F1-scores but experiencing a greater drop in open-case performance, suggesting that it depends more on the victim model's training data. In contrast, RoBERTa and DistilBERT perform significantly worse, especially in open-case settings, making them less effective for function extraction attacks.

A comparison between ASbC and ASbD methods indicates that ASbD consistently outperforms ASbC, particularly in closed-case scenarios. Across all models and budgets, F1-score

and Agreement are higher for ASbD, suggesting that it provides a more effective sampling strategy for approximating the victim model. The advantage is more pronounced in weaker models like RoBERTa and DistilBERT, while PhoBERT and viBERT remain strong regardless of the adaptive method used. These results suggest that employing the ASbD technique can lead to more effective knockoff models, especially when the attacker is fortunate enough to sample a portion of the victim's training data (i.e., in the open-case setting).

Comparing these results with random sampling attacks, we observe a clear performance improvement with adaptive attacks. PhoBERT's F1-score in closed-case improves from 96.08% (random) to 96.54% (adaptive), and viBERT experiences a similar improvement. The primary advantage of adaptive attacks is their better approximation of the victim model's decision boundaries, particularly in closed-case scenarios. However, the performance gap between closed-case and open-case settings remains, especially for weaker models, indicating that even with adaptive sampling, knockoff models still struggle to generalize without victim training data.

Overall, the findings confirm that PhoBERT is the best knockoff model, and adaptive attacks significantly outperform random sampling, demonstrating the importance of strategic data selection. ASbD-based attacks show the highest effectiveness, suggesting that different adaptive strategies yield varying results. Future research should focus on further optimizing sampling strategies and reducing the open-case performance gap to enhance the effectiveness of function extraction attacks.

4.3. Analyze the behavior of the Victim and Knockoff models on some specific records.

Table 4. *Some prediction examples by the Victim and Knockoff models.*

Records	True Label	Victim	Knockoff
giảng bài thu hút , dí dỏm.	1	1	0
giữa lý thuyết từ vựng với trò chơi để dễ tiếp thu.	0	1	0
có up bài cho sinh viên làm thêm bài tập.	1	1	0
cách dạy đổi mới.	1	1	0
phương pháp học áp dụng thực tiễn.	1	1	0
chương trình giảng dạy cần được phân bố hợp lý và được cập nhật.	0	0	1
đề cập nhiều thông tin mở rộng ngoài xã hội.	1	0	1
giảng viên nên cân đối thời gian giảng dạy , không để trễ bài.	0	0	1
nâng cao chất lượng giảng viên và trang thiết bị hỗ trợ.	0	0	1
nói tiếng anh lưu loát.	1	1	1
giáo viên rất vui tính.	1	1	1
cô max có tâm.	1	1	1
thầy dạy nhiệt tình và tâm huyết.	1	1	1
cô rất nhiệt tình , dễ thương , dạy dễ hiểu.	1	1	1
trong trường macbook thầy số hai thì không có máy nào số một.	1	0	0
nên đưa ra một vài phương pháp học lập trình hay cho sinh viên.	0	0	0
chưa giỏi chuyên môn cho lắm.	0	0	0
không nhiệt tình chỉ dẫn và luôn gây khó khăn cho sinh viên.	0	0	0
giáo trình chưa có hợp lý.	0	0	0

Table 4 presents a comparison between the predictions of the Victim and Knockoff models on a sentiment classification task. Although it was trained on data unrelated to the Victim, the Knockoff model produced identical predictions on several records during evaluation, suggesting that it can partially imitate the Victim's behavior. These matching cases often involve clear and emotionally charged texts, which likely make the sentiment easier to detect. However, several instances show discrepancies between the two models. For example, in one case, the Victim predicts a positive sentiment while the Knockoff misclassifies it as negative. This suggests that the Knockoff model may struggle with subtle emotional expressions or nuanced language. Additionally, the Knockoff appears to be more sensitive to surface-level keywords, lacking the deeper contextual understanding learned by the Victim. Such divergence highlights limitations in functional mimicry and the challenges of extracting high-fidelity behavior from a black-box model. Overall, the analysis reveals both strengths and weaknesses in Knockoff's ability to replicate the Victim's predictions.

5. CONCLUSIONS

This paper is an extension of our previous research [11, 12], in which we proposed an adaptive sampling technique based on reinforcement learning to build knockoff models that replicate the functionality of black-box models in English text sentiment analysis. The main contribution of this publication is to demonstrate the effectiveness of that technique in Vietnamese text classification tasks. Vietnamese presents unique challenges due to tonal variations and diacritics, making preprocessing and analysis more complex than in English. While the contributions of this paper are modest, we believe it serves as a valuable starting point for exploring black-box functionality extraction in Vietnamese text classification - a research area that, to our knowledge, remains largely unexplored.

For future research, we aim to develop a text data generation model for both English and Vietnamese, creating a synthetic dataset repository to support the sampling process in function extraction attacks on NLP models.

REFERENCES

[1]. Akshit Jindal, Vikram Goyal, Saket Anand et al., *"Army of Thieves: Enhancing Black-Box Model Extraction via Ensemble based sample selection,"* (2023).

[2]. Dai C. W., Lv M. X, Li K. et al., *"MeaeQ: Mount Model Extraction Attacks with Efficient Queries,"* presented at the arXiv:2310.14047, (2023).

[3]. Minh Pham Quang Nhat, *"An Empirical Study of Using Pre-trained BERT Models for Vietnamese Relation Extraction Task at VLSP 2020,"* in Proceedings of the 7th International Workshop on Vietnamese Language and Speech Processing, (2020).

[4]. Nguyen D.Q and Nguyen A.T, *"PhoBERT: Pre-trained language models for Vietnamese,"* arXiv:2003.00744, (2020).

[5]. Nguyen K.V., Nguyen V.D., Nguyen P.X. V. et al., *"UIT-VSFC: Vietnamese Students' Feedback Corpus for Sentiment Analysis,"* presented at the 10th KSE, Vietnam, (2018).

[6]. Nguyen Q.N., Phan T.C., Nguyen D.V. et al., *"ViSoBERT: A pre-trained language model for Vietnamese social media text processing,"* arXiv:2310.11166, (2023).

[7]. Nguyen T.M.H, Nguyen V.H, Ngo T.Q. et al., *"VLSP shared task: sentiment analysis,"* Journal of Computer Science and Cybernetics, vol. 34, no. 4, pp. 295-310, (2018).

[8]. Oanh Tran Thi and Phuong Le Hong, *"Improving sequence tagging for Vietnamese text using transformer-based neural models,"* in Proceedings of the 34th Pacific Asia conference on language, information and computation, pp. 13-20, (2020).

[9]. Orekondy T., Schiele B., and Fritz M., *"Knockoff Nets: Stealing Functionality of Black-Box Models,"* in IEEE/CVF, pp. 4954--4963, (2019).

[10]. Pal Soham, Yash Gupta, Aditya Shukla et al., *"ActiveThief: Model Extraction Using Active Learning and Unannotated Public Data,"* presented at the AAAI-20, (2020).

[11]. Pham X. Cong, Hoang T. Nguyen, Tran C. Truong et al., *"Adaptive Sampling Technique for Building Knockoff Text Sentiment Models,"* in The 18th IEEE-RIVF, Danang, Vietnam, (2024).
[12]. Pham X. Cong, Hoang T. Nguyen, Tran C. Truong et al., *"Textknockoff: Knockoff nets for stealing functionality of text sentiment models,"* Journal of Science and Technique - Section on ICT, vol. 13, no. 1, (2024), doi: 10.56651/lqdtu.jst.v13.n01.821.ict.
[13]. Phan L. L., Pham P. H., Nguyen K.T.T. et al., *"SA2SL: From Aspect-Based Sentiment Analysis to Social Listening System for Business Intelligence,"* arXiv:2105.15079, (2021).
[14]. Rigaki Maria and Garcia Sebastian, *"A Survey of Privacy Attacks in Machine Learning,"* ACM Computing Surveys, vol. 56, no. 4, pp. 1-34, (2020) (arXiv:2007.07646v3 11-2023).
[15]. S. Kumar, P. P. Roy, D. P. Dogra et al., *"A Comprehensive Review on Sentiment Analysis: Tasks, Approaches and Applications,"* arXiv:2311.11250, (2024).
[16]. V. Sanh, L. Debut, J. Chaumond et al., *"DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,"* NeurIPS, (2019).
[17]. W. Wu, J. Zhang, Wei V. J. et al., *"Practical and Efficient Model Extraction of Sentiment Analysis APIs,"* presented at the ICSE 45, (2023).
[18]. Y. Liu, M. Ott, N. Goyal et al., *"RoBERTa: A Robustly Optimized BERT Pretraining Approach,"* arXiv:1907.11692v1, (2019).
[19]. Yuan X., Ding L., Zhang L. et al., *"ES Attack: Model Stealing against Deep Neural Networks without Data Hurdles,"* (2022).

TÓM TẮT

Sao chép chức năng mô hình hộp đen cho bài toán phân tích cảm xúc tiếng Việt

Các mô hình học sâu dạng hộp đen thường giữ bí mật các thành phần quan trọng như kiến trúc mô hình, siêu tham số và dữ liệu huấn luyện, khiến người dùng chỉ có thể quan sát đầu vào và đầu ra mà không hiểu rõ cách hoạt động bên trong. Do đó, ngày càng có nhiều sự quan tâm đến việc phát triển các mô hình "knockoff" có thể tái tạo hành vi của các mô hình hộp đen này mà không cần truy cập trực tiếp vào các chi tiết nội bộ. Chúng tôi đã thực hiện các nghiên cứu chuyên sâu về các cuộc tấn công trích xuất chức năng mô hình hộp đen NLP với dữ liệu văn bản tiếng Anh. Bằng cách sử dụng các phương pháp lấy mẫu ngẫu nhiên hoặc thích nghi, chúng tôi đã tái tạo thành công các mô hình knockoff có chức năng tương đương với mô hình gốc và mức độ tương đồng cao. Bài báo này mở rộng phạm vi nghiên cứu sang các tập dữ liệu văn bản tiếng Việt. Kết quả thực nghiệm cho thấy, đối với các mô hình hộp đen trong phân tích cảm xúc văn bản tiếng Việt, phương pháp của chúng tôi vẫn duy trì hiệu quả, giúp xây dựng thành công các mô hình có chức năng tương đương với mô hình gốc.

Từ khoá: Mô hình nhái; Trích xuất chức năng mô hình hộp đen; Phân tích cảm xúc văn bản tiếng Việt.