Received: 16 December 2025 | Revised: 3 February 2026 | Accepted: 25 February 2026 | Published online: 10 March 2026

RESEARCH ARTICLE

Artificial Intelligence and Applications
2026, Vol. 00(00) 1–9
DOI: 10.47852/bonviewAIA62028760

BON VIEW PUBLISHING

Transformer Attention-Driven Concept Extraction for Efficient Smishing Detection

Zahriya Lawal Hassan[1,2], Nor Fazlida Mohd Sani[1,*], Muhammad Daniel Hafiz Abdullah[1] and Norwati Mustapha[1]

[1] *Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Malaysia*

[2] *Faculty of Computing, Sokoto State University, Nigeria*

Abstract: Short Message Service (SMS) phishing (smishing) is a form of phishing attack that uses mobile messaging as its delivery medium. Conventional detection methods for smishing attacks fall short at identifying such messages because they resemble spam. The conventional methods rely on static rules or shallow linguistic features. In this study, we propose a transformer-based attention-driven framework for detecting smishing. We develop a concept-level representation to improve both the accuracy and the explainability of our model. The approach extracts and sorts the message signatures uncovered in smishing attacks into three conceptual categories—textual, structural, and behavioral—each of which encompasses a different aspect of how smishing attacks attempt to accomplish. We utilize pretrained Bidirectional Encoder Representations from Transformers (BERT) to construct conceptual representations from the SMS messages. By leveraging BERT's attention weights on the smishing concept categories, informative tokens and patterns that distinguish smishing from benign messages were identified. The final classification is performed using both a fully connected neural network layer and three classical machine learning baseline models, trained on the same features. The result demonstrates that our model achieves performance of (F1-score: 98.71%, accuracy: 99.32%) outperforming the baseline models. Ablation studies further confirmed that each concept category makes a meaningful contribution to the classification performance, with behavioral concept features having the highest impact. This work highlights the potential of attention-driven concept modeling for robust and explainable smishing detection.

Keywords: concept representation, smishing, SMS phishing, BERT, explainable AI

1. Introduction

Mobile technology is an essential component of modern business environments [1], allowing essential activities such as collaboration, communication, and identity verification. The increased reliance on mobile technology as an endpoint has recognized the need for effective security systems that can identify and remediate mobile threats such as smishing and phishing. Despite the convenience brought about by the pervasive use of mobile communications, users are now also facing new and innovative forms of cybersecurity threats [2]. Smishing, a form of phishing attack that makes use of Short Message Service (SMS), is now a sophisticated form of a cybersecurity attack that silently tries to trick victims by masquerading as a legitimate entity, such as a bank or a retailer, and aims to obtain personal and vital user details, perform destructive clicks, and initiate downloads of malicious payloads. With the increase of these sophisticated attacks, simple means of identification based on predefined rules or keyword filters, such as those discussed by Mahmood and Hameed [3], will no longer keep pace with these threats.

An important, but currently underappreciated, aspect for improving smishing detection techniques is involving message signatures; these could be specific characteristics or markers to convey the sender identity, purpose, or authenticity of the received messages [4]. These specific characteristics could be particular structures within SMS or email communications, sender identity, or malicious Universal Resource Locator (URL) markers [5, 6]. Furthermore, although developing such message signatures could improve understanding and detection rates, this task would be complicated by short, ambiguous smishing messages.

There have been studies on detecting smishing using various techniques. In heuristic approaches, messages are identified by manually selected attributes. These approaches ignore important details, such as URLs and their redirects, which are strong indicators of fraudulent intent [7]. Some methods rely on blacklists. Since attackers and domains continually change, it is quite easy for attackers to bypass them. Traditional methods, such as Bag-of-Words, Term Frequency–Inverse Document Frequency (TF–IDF), and Word2vec, were widely used to represent the content of SMSs numerically for classification tasks. Even though traditional models produce decent results, their inability to incorporate subtle, context-aware information from messages limits their generalizability.

The development of models such as the transformer-based Bidirectional Encoder Representations from Transformers (BERT) has led to many improvements in natural language understanding [8]. This is due to the application of the self-attention

***Corresponding author:** Nor Fazlida Mohd Sani, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Malaysia. Email: fazlida@upm.edu.my

mechanism, which assigns each word a weight based on its importance and models short- and long-range dependencies. This makes the transformer model very useful for smishing detection tasks, as it can detect potential triggers such as urgency, deceptive URLs, and language that poses as trusted sources.

The study proposes a novel technique for smishing detection that identifies SMS features derived from message signatures using regular expression (RE) pattern matching, maps the features into concepts, and applies an attention mechanism to the resulting concept representations before classification. Pattern matching is an essential task for recognizing the underlying patterns in SMSs, particularly for identifying smishing patterns and classifying them into three conceptual themes: textual, structural, and behavioral. The three themes are used to identify different attributes of the smishing attack. Textual attributes identify linguistic patterns and word use; structural attributes identify the use of embedded URLs or phone numbers; and behavioral attributes emphasize the use of psychological tactics such as urgency, fear, and impersonation.

BERT is employed to generate contextual embeddings for the defined conceptual representations, which are fed into a classifier to categorize each message as either ham or smish. In contrast to conventional detection approaches that rely primarily on statistical weights or static embeddings, the proposed method further analyzes the BERT model's token-level attention weights. These attention weights reflect the relative importance assigned to individual tokens before the classification process and indicate which parts of the message are considered more relevant or potentially suspicious. By leveraging this information, the model can capture fine-grained message signatures that traditional feature-based methods often overlook, thereby improving both detection performance and interpretability. By jointly exploiting contextual embeddings and attention mechanisms, this technique provides a deeper understanding of structural and conceptual differences between ham and smishing messages. All three concept categories—textual, structural, behavioral—and the ham class are combined to produce an overall representation for each message. The combined embeddings are used to train and test several models to classify messages as ham or smishing. The aim of this study goes beyond the development of an efficient model for classifying messages with high precision accurately; rather, it aims to involve the creation of an interpretable and dynamic model that will respond to the adapting nature of the smishing attack patterns and provide valuable insight concerning the behavioral choices of an attack attempt.

In a recent study by Ludan et al. [9], concept-based learning in text classification was pursued through the lens of the concept bottleneck model and interpretable AI techniques. In image classification, concept learning was used to map low-level image features to meaningful and interpretable concepts such as "wing" and "beak." [10]. To apply the approach to smishing detection, this study identifies and defines high-level concepts, accounting for patterns and structures, as well as triggers for action. It identifies the different dimensions wherein smishing differs from other forms of communication. Subsequently, incorporating these conceptual dimensions into the transformer model will both improve performance, yielding interpretable results for each classification decision.

In focusing on attention-driven concept extraction, this work connects cutting-edge deep learning approaches with the essential needs of cybersecurity: interpretability, flexibility, and robustness. The findings show that attention weights can improve the system's interpretability, detection capability, and the ability to notice both usual and novel smishing messages. With the growing magnitude and complexity of mobile-related threats, integrating concept reasoning and attention mechanisms from the transformer approach appears to offer a promising solution for securing communication.

To the best of our knowledge, no previous smishing detection method has attempted to integrate multi-concept extraction using text, structural, and behavioral concepts. Furthermore, while attention is considered an integral part of transformer models, the application of attention organization into conceptual categories has not been sufficiently investigated in the context of deceptive short messages, such as smishing attacks. This is remedied by our method, which allows attention outputs to be used for concept alignment.

The main contributions from this research are:

1) Message signature-based concept modeling: The concept-based methodology for smishing attacks is presented by defining message signatures. The message signatures are categorized into textual, structural, and behavioral concepts, providing a better understanding of the purpose and structure of SMS messages.
2) Attention-guided embedding and interpretability: Using the attention weights in the BERT model to guide the extraction of the most important tokens within messages, the model's performance can improve while increasing the interpretability of the specific attention zones during classification.
3) Fusion of conceptual embeddings for robust detection: In this study, an approach has been developed to fuse attention-informed embeddings of textual, structural, behavioral, and ham conceptual aspects, aiming to improve the model's generalization to new smishing attacks.

2. Literature Review

Adversaries resort to social engineering techniques to induce fear or urgency in their victims. SMS spam campaigns and smishing appear to share the same surface-level characteristics since both are delivered via the SMS messaging channel [11]. In most cases, SMS spam is used to promote products and services, whereas smishing targets recipients to deceive them into disclosing sensitive information [12].

Early studies on smishing and SMS spam detection primarily relied on handcrafted, low-level features derived from message content, as the currently available datasets lacked metadata. The detection approaches leveraged rule-based and machine learning (ML) classifiers, including naïve Bayes (NB), logistic regression (LR), random forests (RF), and support vector machines (SVM). Commonly used features include lexical cues such as Bag-of-Words and n-grams, statistical properties such as message length, presence of special symbols and numbers, and structural indicators such as URLs, and leet words. A dual-phase detection method was proposed by Jain et al. [12], which utilizes 11 basic features in the first phase to distinguish general spam from ham messages, followed by 4 features related to phishing to further differentiate smishing from the spam category. Text normalization techniques were used to convert nonstandard text to its standardized form [13], thereby significantly enhancing the efficacy of the NB ML classifier. The study shows that normalization improves the discriminative capability of key terms, for example, the probability of the term "claim" appearing in malicious contexts rose to 0.25 after normalization. The standard preprocessing steps included in the framework include stemming, tokenization, and TF–IDF. To improve computational efficiency and reduce the dimensionality of the feature space, the Pearson Correlation

Coefficient-Principal Component Analysis (PCC-PCA) framework [14] was introduced for more efficient feature extraction. Similarly, Ulfath et al. [15] used analysis of variance to select statistically significant features after extracting 10,000 TF–IDF n-gram features. By retaining features $p - values \leq 0.05 - 0.001$, they reduced dimensionality to about $4100 - 4300$ features, thereby improving performance.

Recent advancements involve deep learning models that bypass manual feature engineering through automated extraction. A hybrid CNN-LSTM architecture [16] was developed to automatically extract significant features directly from the text messages and to identify whether they are smishing or ham. Goel et al. [17] utilized enhanced convolutional neural networks (CNNs) that incorporate parameterized Rectified Linear Unit (ReLU) to minimize architectural complexity and manage gradients while extracting meaningful features from unstructured data. These automated approaches address the limitations of manual rules and blacklists.

Transformer models have significantly advanced text classification tasks by enabling context awareness [18]. The BERT contextual embeddings were utilized in the work of Jain et al. [19], where messages were tokenized and passed through a pretrained BERT model to obtain a 768-dimensional [CLS] embedding that contextually represented the entire message and then captured the semantic relationships, word order, and contextual dependencies that are often missed by the conventional feature extraction techniques.

Smishing attempts to resemble legitimate messages but with harmful intent. Effective detection requires models that can identify nuanced linguistic, structural, and behavioral indicators within the message content. The reviewed literature highlights the evolving sophistication of smishing attacks and the growing need for advanced detection mechanisms that surpass simple rule-based filtering and keyword matching. Also, challenges remain, particularly in generating reliable message signatures and detecting attacks with limited prior patterns. These limitations emphasize the need for dynamic feature extraction, interpretable models, and holistic concept representations that can adapt to the subtlety and diversity of smishing attempts in real-world scenarios. This evolving landscape provides a strong foundation for developing a context-aware, attention-guided smishing detection framework that balances predictive accuracy with model transparency.

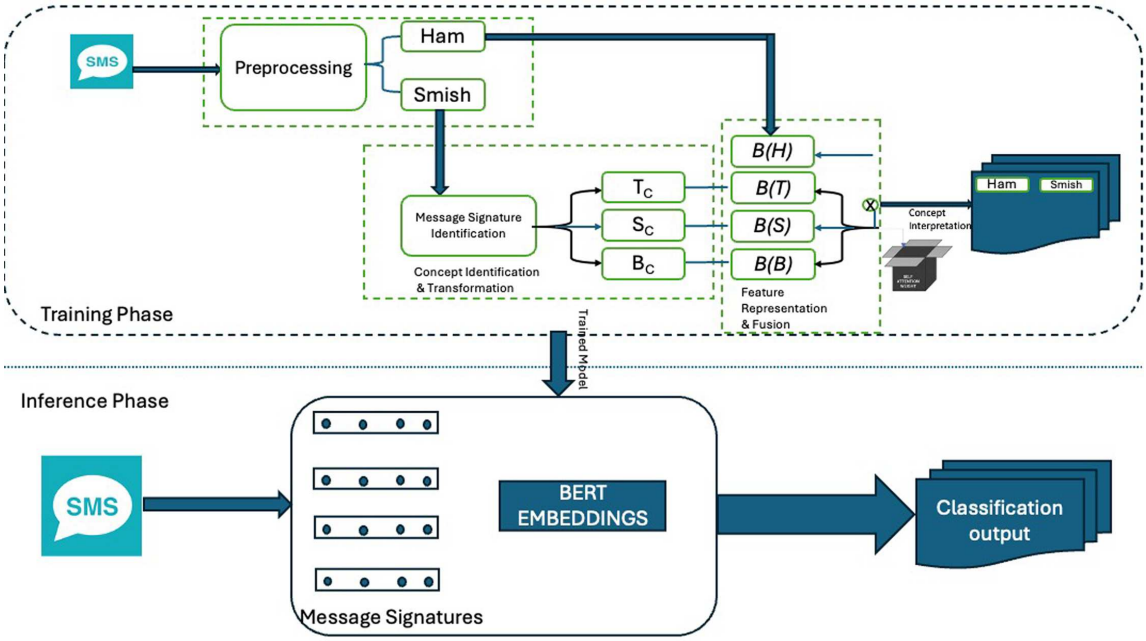
3. Methodology

The framework is illustrated in Figure 1, with two different phases. Phase 1 is the training phase, which consists of five steps: preprocessing, message signature identification, concept representation, attention-based interpretation, and classification. Phase 2, which is the inference phase, shows the trained model being used as an input, with the learned parameters applied to unseen messages to generate model predictions. The training phase is explained in detail.

Preprocessing: SMS messages were retrieved from data sources and prepared for analysis through text preprocessing; standard preprocessing techniques were applied to ensure the data is suitable for natural language processing tasks. These steps included stemming, lemmatization, and tokenization, which collectively reduce noise, and stop-words, numbers, and characters were reserved as they are frequently used in smishing attempts. The preprocessing procedures were implemented using the Natural Language Toolkit library to standardize and optimize the textual data for subsequent feature extraction and modeling.

Message signature identification: Message signatures provide critical insight into the structure and characteristics of malicious messages, yet they are often overlooked by researchers [20]. Signatures consist of distinctive patterns and attributes that serve as unique identifiers for smishing attempts. Signature-based attack detection approaches traditionally extract features from traffic by comparing incoming traffic to already known signatures [21]. This approach yields exceptional accuracy and very few false-positive results. But the signature database must be updated regularly due to the dynamic, zero-day nature of smishing attacks, to enable it to capture new threats. These signatures are usually digital fingerprints of smishing messages that reveal specific patterns, keywords, or other elements commonly found in smishing messages.

Figure 1
Concept extraction and transformer-based classification smishing detection architecture



From the adopted dataset, cues indicative of smishing were extracted by analyzing linguistic features, semantic relationships, and the presence of adversary-oriented vocabulary. To identify the signatures based on these cues, a pattern-matching approach was used to extract the discriminative features. RE-based technique was applied to detect potentially suspicious elements and structural patterns within the messages, adopted from Sharaff et al. [22]. This pattern-matching strategy supports the development of a smishing detection model with a lower false-positive rate. Guided by characteristics of smishing content documented in prior research [20, 23, 24], specific RE patterns were formulated to flag messages containing indicative linguistic or structural markers.

These patterns are then applied to examine smishing messages systematically. Our approach builds on the pattern-matching technique introduced by Alzahrani and Ghorbani [21], chosen for its efficiency in narrowing the search space. The pattern matching is formulated as in Equations (1)–(3):

Let S denote a dataset of SMS messages. Each message is represented as

$$s \in S \tag{1}$$

Equation (2) defines the RE pattern, where each pattern returns 0 or 1

$$P = \{p_1 \ldots\ldots p_n\} \tag{2}$$

The smishing indicator $Sm(s)$ is defined as

$$S_m(s) = \max_{i \in \{1 \ldots n\}} p_{i(s)} \tag{3}$$

The message is flagged as potentially smish if three RE patterns match its contents, helping to reduce messages that do not conform to the smishing identification characteristics. This approach helps develop a model with fewer false-positive predictions. Instead of defining a new set of features, pattern matching evaluates smishing messages against different feature categories that indicate smishing. This includes smishing keywords such as "verify," "click," or "subscribe," as well as patterns indicating URL obfuscation or the presence of a phone number or email address.

Concept extraction/representation: Concept representation is a well-recognized approach in image classification tasks [25], typically involves models identifying high-level semantic attributes, such as texture, color, or object category, to provide interpretability. By linking the abstract features to human-recognized concepts, the model's predictions become more transparent, fostering greater user trust and understanding. Inspired by the paradigm and extending it to the text classification task by Ludan et al. [9], we adapt the concept representation framework to the domain of smishing detection. In this approach, three categories of interpretable concepts are defined and extracted from the dataset: textual-based, structural-based, and behavioral-based. We named it a textual, structural, and behavioral concept. The textual concepts encompass linguistic indicators such as suspicious keywords, informal expressions, and contextually inconsistent terms; the structural concepts capture message-format characteristics, including the presence of URLs, message length, numerical patterns, or special symbols; and behavioral concepts represent cues intended to provoke user action, such as urgency, fear, or requests for sensitive information. These concept categories form a structured and interpretable intermediate representation that enables transparent model predictions and provides clear explanations for smishing classification outcomes.

To derive high-level semantic representations from the input data, a BERT-pretrained transformer model is employed. The three distinct concept categories—textual, structural, and behavioral features—were independently processed by the model through a sequence of transformations, mathematically represented in Equations (4)–(6). The numerical vectors E_i, generated by the tokenizer, are passed to the model as input, as shown in Equation (4).

$$h_i = B(E_i) \tag{4}$$

where $i \in \{t, s, b\}$ for the textual (t), structural (s), and behavioral (b) concepts.

Figure 2 presents the overlap analysis of the three extracted concept dimensions—textual, structural, and behavioral—within the smishing label in the corpus. The plot provides a structured visualization of the intersections and exclusions among these conceptual categories, enabling a deeper understanding of the relationships between linguistic, semantic, and behavioral cues that co-occur in smishing content.

The contextual representation of each concept is obtained from the final hidden layer of the BERT model using the classification token, as shown in Equation (5):

$$r_i = h_i[\mathrm{CLS}] \tag{5}$$

The concept-level representations derived from textual, structural, and behavioral inputs are then fused to form a unified smishing representation through concatenation:

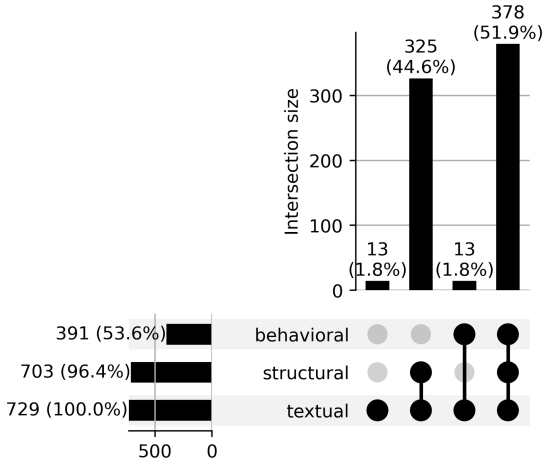
$$r_{(t,s,b)} = r_t \| r_s \| r_b \tag{6}$$

For messages labeled as ham, the tokenizer generated vectors E_h are similarly passed through the BERT model to obtain their contextual representation:

$$h_h = B(E_h) \tag{7}$$

$$r_h = r_h[CLS] \tag{8}$$

Figure 2
Overlap of textual, structural, and behavioral concepts in smishing label messages



The final representation used for classification is formed by concatenating the smishing and ham representations:

$$r_{final} = r_{(t,s,b)} \| r_h \tag{9}$$

The fused representation is then provided to the classifier to predict the message label:

$$y' = C(r_{final}) \tag{10}$$

Signature-to-concept transformation: To demonstrate how the proposed framework addresses the inherent challenges of smishing detection, a message-level example is presented in Table 1. The objective is to show how raw messages are transformed into interpretable concept representations through signature extraction and attention-driven modeling. During the preprocessing step, messages were tokenized, and numerical values, URLs, special symbols, and uppercase characters were preserved as these elements frequently occur in smishing messages. In the message signature identification stage, predefined RE patterns are activated. The presence of obfuscated URL, short numeric code, and urgency-related keywords (e.g., "CHAT" and "contact") is detected. These matched patterns collectively form a compact message signature that characterizes the message's structural and textual properties. The activation of multiple signature patterns allows the framework to filter benign messages while retaining messages that exhibit smishing-related traits.

The extracted message signature is subsequently mapped to the three predefined concept categories, with the features mapped to concepts as shown in Figure 3. This concept-level representation abstracts low-level lexical features into high-level semantic cues, reducing reliance on keyword matching and enabling better generalization to previously unseen smishing variants.

The concept representations are then encoded using the BERT model to generate contextual embeddings. Through the self-attention mechanism, the model assigns a higher attention weight to tokens associated with the extracted concepts. These attention-weighted representations provide an interpretable view of how the model prioritizes smishing-relevant information within short SMS content.

Attention-based interpretation: To interpret the model's output for each concept, the attention weights generated by the transformer's self-attention mechanism were utilized. In BERT, attention is computed internally using the scaled dot-product formulation, as in Equation (11) [26]. Attention is not redefined computationally in the study; rather, the attention weights produced by the BERT model are extracted and analyzed to assess token-level importance within each concept.

$$Attention = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{11}$$

where Q as query, K as key, and V as value obtained from linear transformations of the input sequence.

Figure 4 shows an example of token-level attention distribution for a smishing label message.

From Table 1, it can be observed that tokens, such as "chat" and "86688," receive relatively higher attention values. These tokens are commonly associated with suspicious message characteristics, including unsolicited communication prompts, short codes, and domain-like or service-related terms. Such elements are frequently exploited in smishing messages to attract user interaction or create a sense of legitimacy. The attention distribution indicates that the model does not rely solely on the overall message

Table 1
Smishing instance for attention weight analysis

Label	SMS content
Smishing	Dear U've been invited to XCHAT. This is our final attempt to contact u! Txt CHAT to 86688 150p/MsgrcvdHG/Suite342/2Lands/Row/W1J6HL LDN 18yrs

Figure 3
Mapping smishing-related features to conceptual representation

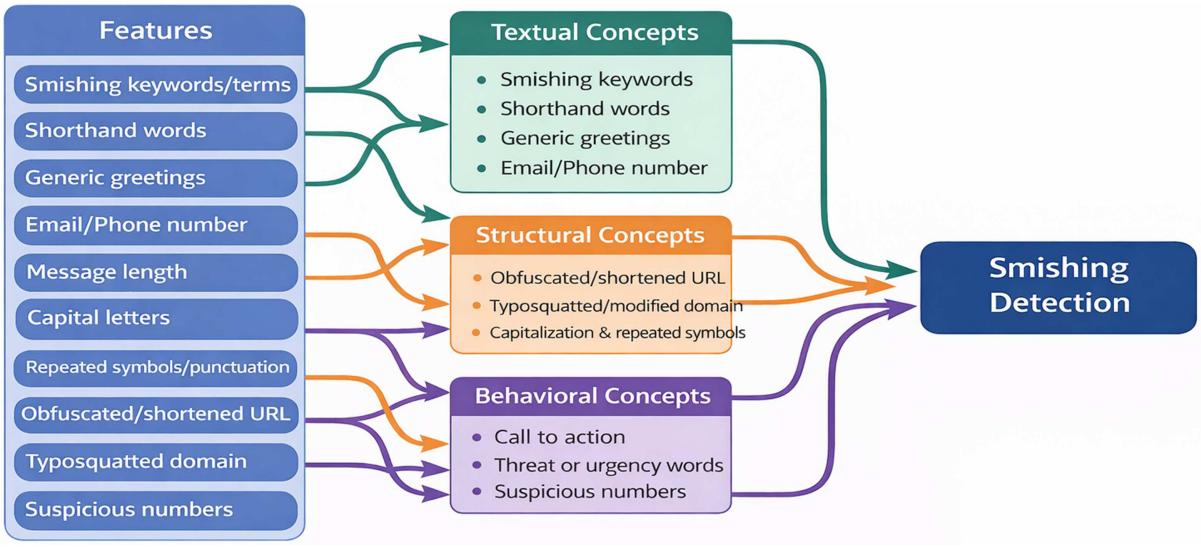
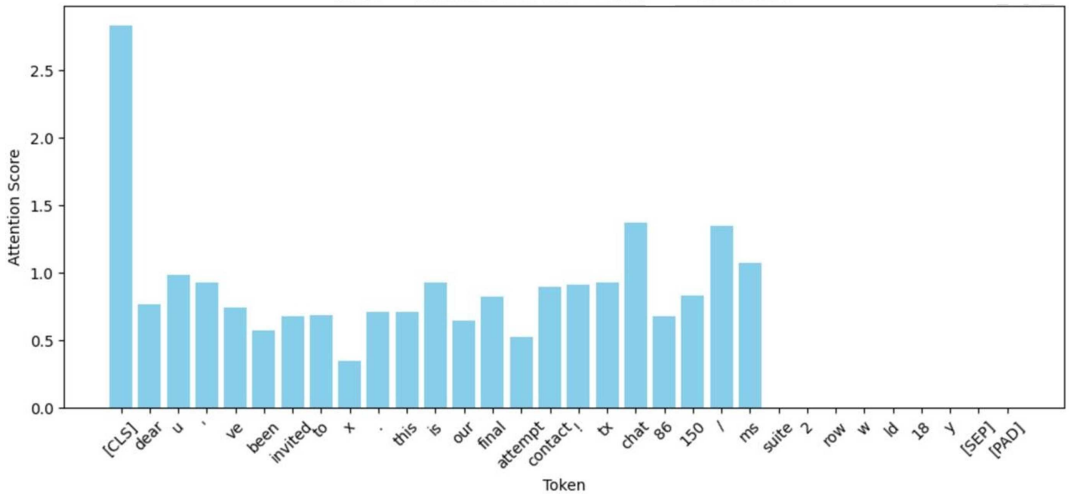


Figure 4
Attention score distribution for the smishing message shown in Table 1



semantics but focuses on specific tokens that may indicate smishing. Therefore, the figure provides an interpretable view of how the attention mechanism captures message signatures and contributes to the final classification decision. For comparison, a legitimate message, "here is my new address -apples&pairs&all that malarky" activates none of the structural or behavioral smishing concepts, resulting in low attention weights for benign tokens.

Classification: The final layer of the model consists of a dense layer that uses attention-based representations for binary classification. A single neuron with a function, sigmoid activation, produces a probability score with a range between 0 and 1, where lower values indicate ham and higher values indicate smish messages. Alongside this deep learning method, the effectiveness of three classical ML models—RF, LR, and SVM—was examined as baseline comparators. Each of these models is evaluated using identical input features to maintain a fair comparison. The complete workflow of the model is outlined in Algorithm 1.

Algorithm 1: Transformer-based attention driven concept extraction for smishing detection

Input: SMS dataset $D = \{(x_i, y_i)\}_{i=1}^{n}$
Output: Trained classifier $\mathscr{M}$ and predicted labels y'
1: Preprocessing
2: **for** each SMS message x_i **do**
3: Tokenize and lemmatize x_i
4: Obtain preprocessed message x_i'
5: end for
6: Message Signature Extraction
7: **for** each message x_i' **do**
8: Apply regular expressions to extract message signatures
9: end for
10: Concept Definition
11: Define concept pattern sets P_t, P_s, P_b
12: **for** each message x_i' **do**
13: Assign message to textual, structural, and/or behavioral concepts
14: end for
15: Concept Representation
16: **for** each concept input **do**
17: Encode using pretrained BERT
18: Extract $[CLS]$ representation
19: end for
20: Attention-Based Interpretation
21: Extract self-attention weights from BERT layers (Equation 11)
22: Fusion and Classification
23: Concatenate representations: $r_{final} = r_t || r_s || r_b$
24: Train classifier $\mathscr{M}$
25: Predict labels $y' \leftarrow \mathscr{M}(r_{final})$
return $\mathscr{M}$ *and* y'

4. Experimental Result

This section presents the experimental results, beginning with an overview of the dataset used and the experimental setup, followed by the results of the evaluation metrics applied, and concluding with a comparative analysis of the proposed models' performance against baseline classifiers and an ablation study of the extracted concepts.

4.1. Dataset and metrics

The experiments in the study were performed using the UCI Machine Learning Repository SMS Spam dataset [27], a labeled dataset consisting of binary classes: legitimate (ham) and smishing messages. Initially, the dataset was used for SMS spam, but later, the spam label was changed to smish due to the resemblance of the two labels. The dataset comprises approximately 5572 labeled text messages, with 4825 labeled as ham and 747 labeled as smishing. The training and testing datasets used a stratified 80:20 split to preserve class proportions. The model's performance was evaluated using standard classification metrics, such as accuracy, precision, recall, F1-score, and receiver operating characteristic-area under the curve (ROC-AUC), to provide a balanced assessment of both predictive effectiveness and robustness, particularly under class-imbalanced conditions.

4.2. Results

This section presents the results of experiments validating the effectiveness of the transformer-based attention-driven concept extraction model for smishing detection. The analysis began with

an overview of the models' overall performance, which integrates BERT embeddings with concept-driven features, followed by a comparative evaluation against classical ML classifiers trained on the same conceptual representations. An ablation study is also included to assess the individual contributions of the different conceptual dimensions to the overall model performance.

4.2.1. Model performance

The proposed model employs BERT-based embeddings combined with a fully connected classification layer to identify smishing messages. This architecture achieved outstanding performance across all evaluation metrics, obtaining an accuracy of 99.32%, precision of 98.01%, recall of 99.42%, F1-score of 98.71%, and a ROC-AUC of 99.95%. These results demonstrate the model's effectiveness in detecting smishing attempts while minimizing false positives.

4.2.2. Baseline model comparison

To establish a performance benchmark, the proposed model was compared with three classical ML models—LR, RF, and SVM—each trained on BERT embeddings. RF achieved the highest result, with an accuracy of 0.9890 and a macro F1-score of 98.59%, closely followed by SVM, which attained an accuracy of 98.70%. LR performed slightly lower, achieving an accuracy of 98.55%. Despite the strong results, none of the baseline models surpassed the transformer-based architecture, particularly in terms of recall and ROC-AUC, underscoring the advantages of contextualized embeddings and attention-based learning for smishing detection. Table 2 summarizes the models' performance and compares it with the three classical ML models.

4.3. Ablation study on concept-based features

To assess the contribution of the individual concept representations, an ablation study across the three concept categories was conducted: textual, structural, and behavioral. Each category was evaluated independently using a BERT-based classifier trained on the corresponding features. Table 3 shows the evaluation results using four metrics: precision, recall, F1-score, and accuracy.

5. Discussion

Message signatures are distinct patterns embedded in a message that reveal its underlying intent and help distinguish malicious messages from ham [20]. Concepts are semantic abstractions that aggregate related signatures into an interpretable unit, enabling the models to reason about the intent rather than isolated surface features.

The integration of textual, structural, and behavioral concepts allows the model to capture the underlying aspects of smishing attacks. The fusion of the concept categories enables a holistic understanding of message intent, which explains the observed performance gains over traditional ML baselines.

The study demonstrates that transforming low-level message signatures into high-level representations significantly enhances smishing detection performance and interpretability. It is a shift from feature-based learning to concept-based reasoning that allows the model to capture patterns that are more robust, generalized, and aligned with the real-world smishing behavior.

The BERT model outperforms traditional ML approaches by a wide margin, thanks to its architectural depth and the ability to process longer text sequences through its transformer backbone. However, smishing generally occurs on mobile devices, where resources are limited. Researchers acknowledge that there is a trade-off between top-tier performance and a computationally intensive structure and that this is a necessary consideration for practical mobile deployments of the proposed model.

We compared our approach with that of Sharaff et al. [22], which combined RE feature engineering with a deep learning model. Their focus was on enhancing performance through regex-informed lexical refinement and a stacked Bidirectional Long Short-Term Memory (BiLSTM) architecture, and Jain et al. [19] built a model by using BERT embeddings that recognized the contextual association between the words. Table 4 compares the methods and performance between the two studies and our proposed model.

The proposed approach extends smishing detection beyond their work by explicitly modeling attack intent through concept-driven feature extraction. By transforming low-level smishing cues into high-level semantic concepts, the model reveals the underlying mechanics of smishing rather than relying on superficial patterns

Table 2
Performance comparison

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC-AUC (%)
Ours (BERT + FC)	99.32	98.01	99.42	98.71	99.95
BERT + LR	98.55	98.29	98.00	98.14	99.91
BERT + RF	98.90	98.89	98.30	98.59	99.71
BERT + SVM	98.70	98.45	98.22	98.34	99.94

Note: BERT + FC stands for BERT with fully connected layers.

Table 3
Evaluation results of the ablation study on the concepts

Concept type	Precision (smish) (%)	Recall (smish) (%)	F1-score (smish) (%)	Accuracy (%)
Textual concepts	97.74	94.89	96.30	99.10
Structural concepts	99.23	94.85	96.99	99.28
Behavioral concepts	98.72	96.25	97.47	99.62

Table 4
Comparison of the proposed model with existing models

Author	Feature extraction	Accuracy
Jain et al. [19]	BERT + Low-level features	94.00 (%)
Sharaff et al. [22]	RE	99.01 (%)
Proposed model	RE + Smishing Concepts + BERT	99.32 (%)

either at the lexical or sequential levels. This modest accuracy gain indicates an added layer of discriminative power provided by concept abstractions beyond raw regex-driven features.

6. Conclusion

This study introduces an attention-driven, concept-based framework for detecting smishing in SMS messages. Message features are first extracted from domain-specific message signatures and then grouped into textual, structural, and behavioral concept categories. These conceptual representations form the basis for an interpretable and robust smishing detection model built on a BERT architecture. By combining contextual embeddings with token-level attention information, the model is able to focus on the most informative parts of each message, thereby enhancing both interpretability and detection accuracy. Experimental results demonstrate that the proposed framework consistently outperforms traditional classifiers, with the BERT-based model achieving the best performance across all evaluation metrics.

In practical deployment environments, efficiency, efficiency and scalability are critical, particularly in mobile environments that rely on resource-constrained devices, where smishing attacks typically occur. Although the standard BERT model delivers the highest accuracy, its large parameter size and computational cost limit its suitability for real-time mobile or edge deployment. To address this limitation, the framework also integrates MobileBERT, a compact transformer model optimized to reduce memory usage and faster inference. While MobileBERT exhibits a slight reduction in accuracy compared to BERT, it offers substantial gains in efficiency, making it more suitable for deployment in resource-constrained environments. In addition, MobileBERT can be fine-tuned on multilingual datasets, enabling cross-language adaptability, while the attention-driven concept extraction mechanism supports continuous adaptation to evolving smishing strategies.

The ablation analysis further revealed that each conceptual component contributed uniquely to performance, with behavioral features proving especially effective in identifying manipulative and deceptive linguistic patterns. Overall, the proposed approach achieves a balance between state-of-the-art deep learning performance and the transparency required for practical cybersecurity deployment. By integrating concept extraction with attention analysis, this framework lays the foundation for transparent, adaptive, and generalizable smishing detection systems capable of addressing both existing and emerging mobile threats. Future work will focus on extending the framework to multilingual datasets and incorporating dynamic behavioral feedback derived from real-time user interactions.

Acknowledgment

The authors would like to acknowledge the support of the Faculty of Computer Science and Information Technology, Universiti Putra Malaysia.

Funding Support

This research is supported by Universiti Putra Malaysia, under the Fundamental Research Grant Scheme UPM.RMC.800-2/1/2024/GP-IPS/9808600.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in the UCI Machine Learning Repository at https://doi.org/10.1145/2034691.2034742, reference number [27].

Author Contribution Statement

Zahriya Lawal Hassan: Conceptualization, Methodology, Resources, Writing – original draft, Visualization. Nor **Fazlida Mohd Sani:** Validation, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Muhammad Daniel Hafiz Abdullah:** Validation, Writing – review & editing, Supervision, Project administration. **Norwati Mustapha:** Writing – review & editing, Supervision, Project administration.

References

[1] Padilla-Vega, R., Sénquiz-Díaz, C., & Ojeda-Castro, A. (2022). Toward a conceptual framework of technology adoption: Factors impacting the acceptance of the mobile technology in the international business growth. In S. A. U. Cavalcanti (Ed.), *Impactos de las tecnologías en las ciencias sociales aplicadas Atena Editora* (*4*, pp. 132–142). https://doi.org/10.22533/at.ed.32722201011

[2] Stanfield, M. (2024). Mobile technologies at risk: A literature review on the evolving challenges and solutions in mobile technology security. *Scientific Bulletin*, *29*(1), 151–162. https://doi.org/10.2478/bsaft-2024-0016

[3] Mahmood, A. R., & Hameed, S. M. (2023). A smishing detection method based on SMS contents analysis and URL inspection using Google engine and VirusTotal. *Iraqi Journal of Science*, *64*(10), 6276–6291. https://doi.org/10.24996/ijs.2023.64.10.41

[4] Shoaib, M., & Umar, M. S. (2023). An investigation in detection and mitigation of smishing using machine learning techniques. *Social Network Analysis and Mining*, *13*(1), 135. https://doi.org/10.1007/s13278-023-01142-4

[5] Maqsood, U., Ur Rehman, S., Ali, T., Mahmood, K., Alsaedi, T., & Kundi, M. (2023). An intelligent framework based on deep learning for SMS and e-mail spam detection. *Applied Computational Intelligence and Soft Computing*, *2023*(1), 6648970. https://doi.org/10.1155/2023/6648970

[6] Abdullahi, A. A., & Kaya, M. (2021). A deep learning based method to detect email and SMS spams. In *2021 International Conference on Decision Aid Sciences and Application*, 430–435. https://doi.org/10.1109/DASA53625.2021.9681921

[7] Mahmud, T., Prince, M. A. H., Ali, M. H., Hossain, M. S., & Andersson, K. (2024). Enhancing cybersecurity: Hybrid deep learning approaches to smishing attack detection. *Systems*, *12*(11), 490. https://doi.org/10.3390/systems12110490

[8] Wu, T., Wang, Y., & Quach, N. (2025). Advancements in natural language processing: Exploring transformer-based architectures for text understanding. In *International Conference on Artificial Intelligence and Industrial Technology Applications*, 1384–1388.

[9] Ludan, J. M., Lyu, Q., Yang, Y., Dugan, L., Yatskar, M., & Callison-Burch, C. (2023). Interpretable-by-design text classification with iteratively generated concept bottleneck, *arXiv Preprint: 2310.19660*.

[10] Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. (2020). Concept bottleneck models. *arXiv Preprint:2007.04612*.

[11] Akande, O. N., Akande, H. B., Kayode, A. A., Adeyinka, A. A., Olaiya, F., & Oluwadara, G. (2022). Development of a real time smishing detection mobile application using rule based techniques. *Procedia Computer Science*, *199*, 95–102. https://doi.org/10.1016/j.procs.2022.01.012

[12] Jain, A. K., Yadav, S. K., & Choudhary, N. (2020). A novel approach to detect spam and smishing SMS using machine learning techniques. *International Journal of E-Services and Mobile Applications*, *12*(1), 21–38. https://doi.org/10.4018/IJESMA.2020010102

[13] Jain, H., & Mahadev, M. (2022). An analysis of SMS spam detection using machine learning model. In *International Conference on Computational Intelligence and Communication Technologies*, 151–156. https://doi.org/10.1109/CCiCT56684.2022.00038

[14] Hapase, D. S., & Patil, L. V. (2024). Telecommunication fraud resilient framework for efficient and accurate detection of SMS phishing using artificial intelligence techniques. *Multimedia Tools and Applications*, *83*(41), 89111–89133. https://doi.org/10.1007/S11042-024-19020-2

[15] Ulfath, R. E., Sarker, I. H., Chowdhury, M. J. M., & Hammoudeh, M. (2021). Detecting smishing attacks using feature extraction and classification techniques. In *International Conference on Big Data, IoT, and Machine Learning*, *677–689*. https://doi.org/10.1007/978-981-16-6636-0_51

[16] Mehmood, M. K., Arshad, H., Alawida, M., & Mehmood, A. (2024). Enhancing smishing detection: A deep learning approach for improved accuracy and reduced false positives. *IEEE Access*, *12*, 137176–137193. https://doi.org/10.1109/ACCESS.2024.3463871

[17] Goel, D., Ahmad, H., Jain, A. K., & Goel, N. K. (2024). Machine learning driven smishing detection framework for mobile security. *arXiv Preprint: 2412.09641*

[18] Kumarasiri, W. L. T. T. N., Siriwardhana, M. K. J. C., Suraweera, S. A. D. S. L., Senarathne, A. N., & Harshanath, S. M. B. (2023). Cybersmish: A proactive approach for smishing detection and prevention using machine learning. In *2023 7th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)*, 210–217. https://doi.org/10.1109/I-SMAC58438.2023.10290228

[19] Jain, A. K., Kaur, K., Gupta, N. K., & Khare, A. (2025). Detecting smishing messages using BERT and advanced NLP techniques. *SN Computer Science*, *6*(2), 109. https://doi.org/10.1007/s42979-024-03532-7

[20] Mishra, S., & Soni, D. (2023). DSmishSMS-A system to detect smishing SMS. *Neural Computing and Applications*, *35*(7), 4975–4992. https://doi.org/10.1007/s00521-021-06305-y

[21] Alzahrani, A. J., & Ghorbani, A. A. (2015). Real-time signature-based detection approach for SMS botnet. In *2015 13th Annual Conference on Privacy, Security and Trust*, 157–164. https://doi.org/10.1109/PST.2015.7232968

[22] Sharaff, A., Pathak, V., & Paul, S. S. (2023). Deep learning-based smishing message identification using regular expression feature generation. *Expert Systems*, *40*(4), e13153. https://doi.org/10.1111/exsy.13153

[23] Martínez-Mendoza, A., Jáñez-Martino, F., Carofilis, A., Fernández-Robles, L., Alegre, E., & Fidalgo, E. (2024). Towards multi-class smishing detection: A novel feature vector approach and the smishing-4C dataset. In *Poster Proceedings of the 40th Annual Conference of the Spanish Association for Natural Language Processing*, *3846*, 58–68.

[24] Akande, O. N., Gbenlea, O., Abikoyec, O. C., Jimoh, R. G., Akande, H. B., Balogun, A. O., & Fatokun, A. (2023). SMSPROTECT: An automatic smishing detection mobile application. *ICT Express*, *9*(2), 168–176. https://doi.org/10.1016/j.icte.2022.05.009

[25] Togo, R., Nakagawa, N., Ogawa, T., & Haseyama, M. (2025). ConcVAE: Conceptual representation learning. *IEEE Transactions on Neural Networks and Learning Systems*, *36*(4), 7529–7541. https://doi.org/10.1109/TNNLS.2024.3404496

[26] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . , & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, *30*, 6000–6010.

[27] Almeida, T. A., Hidalgo, J. M. G., & Yamakami, A. (2011). Contributions to the study of SMS spam filtering: New collection and results. In *Proceedings of the 11th ACM symposium on Document engineering*, 259–262. https://doi.org/10.1145/2034691.2034742

How to Cite: Hassan, Z. L., Sani, N. F. M., Abdullah, M. D. H., & Mustapha, N. (2026). Transformer Attention-Driven Concept Extraction for Efficient Smishing Detection. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA62028760