JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 11 Issue 02
2026

Review Article
https://doi.org/10.56979/1102/2026/1496

A Deep Learning Approach for SMS Spam Detection Using Bert and Masked Language Models: A Case Study on Arabic and English

Musab Al-Ghadi[1], Mamoon Obiedat[1*], Eman Omar[1], and Douha Al-Odeh[1]

[1]Department of Information Technology, Faculty of Prince Al-Hussein Bin Abdallah II for Information Technology, The Hashemite University, Zarqa, Jordan.
[*]Corresponding author: Mamoon Obiedat. Email: mamoon@hu.edu.jo

Received: March 01, 2026 **Accepted:** August 02, 2026

Abstract: SMS spam poses a significant threat to the security and privacy of mobile users, causing financial fraud, phishing attacks, and diminishing trust in mobile communication systems. With multilingual and low-resource languages (like Arabic) with their linguistic complexity and dialectal variation, as well as the lack of labeled datasets for these languages, traditional spam detection methods are ineffective. This paper presents a novel multilingual SMS spam detection framework that leverages Bidirectional Encoder Representations from Transformers (BERT) with domain-adaptive Masked Language Modeling (MLM). Two models are created and tested: fine-tuned (FT) model on SMS corpus with labels and pre-trained (MLM-SMS) model on SMS corpus using MLM pretraining first followed by supervised fine-tuning. To evaluate the models, they are applied to the UCI English SMS Spam Collection dataset (5,574 messages) and a newly created Arabic dataset of 4,623 real and AI-generated SMS messages. Both models are trained using the multilingual variant of BERT, and share the same preprocessing pipeline and fixed sequence length of 128 tokens. The results demonstrate that the MLM enhanced model always works better than the baseline. It obtains 99.2% accuracy and 98.6% F1-scores on English data, whereas BERT-SMS gets 97.3% accuracy and 97.4% F1-scores. It achieves a higher accuracy of 98.8% and F1-score of 98.3% on the Arabic data compared with a baseline accuracy of 97.5% and F1-score of 96.6%. Another significant improvement in robustness is shown by the reduced number of false positives and false negatives, confirmed by the analysis of the confusion matrix. The proposed framework is scalable, language independent, and extremely effective for spam detection in real-time SMS in multilingual low resource and high resource settings.

Keywords: Arabic and English SMS; Spam Detection; Phishing Attacks; BERT; MLM; Natural Language Processing

1. Introduction

In the modern world, more individuals are using SMS as a communication and marketing tool [1]. The recent article by Slicktext [2] states that the number of SMS users is about half of the global population and is approximately five billion. The broad usage has placed SMS as an important aspect of everyday communication, with the global SMS traffic standing at an estimated 2.2 trillion messages annually in 2023 according to Juniper Research [4]. The proliferation of SMS has however also made it an easy target of spam messages which in most cases contain schemes of fraud, phishing attacks, and unwanted advertisements. These spam messages do not only create inconveniences to end users, but also accuse severe threats to privacy and security such as identity stealing and financial fraud. As an illustration, a 2022 study by Truecaller revealed that 68% of mobile users all over the world receive at least one spam message per week [4], with a significant proportion of victims becoming the victims of scams, particularly in the areas where SMS is still the dominant communication medium, like the Middle East.

As a rule, the senders of spam strive to collect personal or financial data through SMS, and they tend to use fake content, infected links, or viruses. The amount of spam mail has been increasing in a very fast pace in the past few years. As an illustration, in September 2021 spam messages were sent by around 1.27 million people, but in August 2022, the number of messages sent per month increased to 10.9 billion messages, which is published in February 2023, indicating the size of the problem [5] . The importance of fighting SMS spam is in the fact of its far-reaching consequences on the life of any person and society. The spam messages not only lead to lack of trust in online communication, but also pose a certain amount of money to the customers, as well as telecommunication providers, to react to the effects of fraudulent scams and network blocking. These messages can make use of various faked methods such as social engineering, shortened links, phishing, vishing (voice phishing), and malicious objects, to take advantage of unsuspecting users. Social engineering uses human psychology to trick customers to reveal their sensitive information or taking some unintended actions [6].

Example of English SMS: "Congratu*lations! You've won an incredible prize! As a VIP client, you can get a special offer for a limited time. Don't miss out, click here to claim your free prize."*

Example of Arabic SMS: *" مبروك! فزت بجائزة رائعة! كعميل مميز، يمكنك الحصول على عرض خاص لفترة محدودة، اضغط هنا لاستلام جائزتك المجانية."*.

Shortened URLs conceal malicious destinations, a prevalent tactic in SMS phishing campaigns [6]. Example in English: *"Exclusive offer! Click here for 50% off: bit.ly/deal50."* Example in Arabic: *" عرض حصري! عرض50%. اضغط هنا لخصم 50%: bit.ly/."*

Vishing (Voice Phishing) manipulates users into calling fraudulent numbers to disclose personal information, posing a severe threat to privacy and security (Verizon 2023) [6].

Example in English: *"Surprise! You've won a grand prize! Please call us at 123-456-7890 to confirm your win. We'll need your information to deliver the prize."*

Example in Arabic: *" مفاجأة! فزت بجائزة كبيرة! اتصل بنا على 123-456-7890 لتأكيد فوزك، نحتاج معلوماتك لتسليم الجائزة."*.

Malicious Attachments deliver harmful files disguised as legitimate documents, a common method for distributing malware through SMS (Verizon 2023) [6]. Example in English: *"Hi, I have sent you the final invoice for your project in the attachment. Please open the attached document and read the details carefully."* Example in Arabic: *"مرحبًا، أرسلت لك الفاتورة النهائية لمشروعك في المرفق، من فضلك افتح المستند المرفق واقرأ التفاصيل بعناية."*

The acuteness of the problem with SMS spam is predetermined by the fact that it is necessary to protect users, especially in the linguistically less served areas, including Arabic-speaking social groups, where the deficiency of powerful detection algorithms exacerbates the situation. Arabic is a language with special problems related to the detection of spam because of its morphological complexity, dialects, and the lack of labeled data (Al Saidat et al., 2024) [7]. To make this problem even more disheartening, AI-made spam spreads, which can look like a natural correspondence, making any filtering using keywords less and less efficient. Even though English-language spam detection has made significant strides, recent models such as BERT have a spam detection accuracy of up to 99.40 on datasets like the UCI SMS Spam Collection dataset (Almeida et al., 2011) [8]. The availability and study on the Arabic language is also highly restricted which makes the user vulnerable (Al Saidat et al., 2024) [7]. To address this shortcoming, recent studies point to the use of transformer-based architectures such as BERT. These studies show potential in improving the accuracy of detection and the ability to support the linguistic complexity, thereby bringing us closer to more fair and adaptive spam mitigation measures. The introduction of pre-trained language models (PLMs), including BERT models, has greatly contributed to text classification, particularly ones in small and informal writings such as SMS messages. According to (Wu et al., 2024) [9], PLMs are very effective at acquiring linguistic patterns that are complex using pre-training strategies such as MLM (Devlin et al., 2018) [10] and then task-specific fine-tuning. This functionality has been in line with this study that uses BERT based models with MLM to tackle the problem of SMS spam in Arabic and English. The most important linguistic issues that cover are the morphological richness of Arabic language, dialects, and the use of informal forms (e.g. slang and Arabizi). The survey also validates the usefulness of PLMs in the improvement of detection rates in low-resource languages, which supports the objective of creating a coherent multilingual system to build a powerful and flexible SMS spam detection.

In practice SMS spam detection is susceptible to a range of challenges, especially in mobile applications where messages are brief, informal, and where they are noisy i.e. contain abbreviations, emojis, and mixed language messages. These issues are even more evident in low resource languages like Arabic with a rich

morphology, different dialects, and non-standard written forms such as Arabizi. Classical filtering methods are not effective in dealing with AI-spam, text obfuscation, and fast changing attack signatures. Also, there are a few publicly available Arabic SMS datasets with labels missing, making it more difficult to train a strong model. These joint linguistic and deployment limitations are critical to promoting the validity and extensiveness of multilingual spam detection systems.

This paper presents two new methods of spam SMS detection in Arabic and English: (1) a control BERT-based model, and (2) a more advanced BERT model combined with MLM. The latter deals with the problems of SMS data (slang, abbreviations, emojis, and code-mixing) and capitalizes on the fact that MLM can enhance contextual perception of short, informal texts. This is, to the best of our knowledge, the first attempt at using BERT with MLM to detect SMS spam in multilingual languages, with important benefits due to low-resource languages such as Arabic, achieved by using focused pre-training on masked language tasks. The main contributions of this work are as follows: (i) Development of a new Arabic SMS spam dataset, which consists of 1000 actual user SMS messages plus 1000 AI-augmented examples to counter the issue of data scarcity. The remainder of this paper is structured as follows. Section 2 reviews the most relevant state-of-the-art studies. Section 3 details the proposed SMS spam detection models in detail. Section 4 describes the experimental settings adopted for model evaluation. Section 5 presents and discusses the results, followed by Section 6, which provides a comparative analysis and further discussion. Finally, Section 7 concludes the paper.

2. Related Work

Several studies have investigated SMS spam detection using various datasets and models, often focusing on English or other low-resource languages. Below, key prior works are reviewed and compare them with this study.

2.1. Text Classification Methods

To promote the process of text classification in English, the Authors of (Mao et al., 2025) [11] suggested the BERT-DXLMA model. This paper was devoted to the enhancement of semantic information capture using residual connections and multiplicative long short-term memory (mLSTM) units to deal with long sequences. The model was found to perform better on datasets like S2D, and SST-1 (Socher et al., 2013) as opposed to other models like the BERT-CBLUGMA. The method however was only limited by English texts making it less generalizable to low-resource languages such as Arabic and the computational complexity of the model can be a limiting factor in putting it to use in lower-resource settings. The BERT-Medium model introduced to text classification on the AG (Antonio Gulli) News dataset (Zhang et al., 2015) [12] is composed of 120,000 news articles divided into four primary classes used in (Shen 2025) [13] to propose this model to the reader. The two major improvements on the model included the first improvement, which is the representation of the sentences by applying advanced fine-tuning methods and the second improvement, which included the enhancement of the MLM mechanism during pre-training by selective masking of 15% The results produced by the model showed that the model performed better by introducing the two major improvements that included improving the sentence representation through fine-tuning methods and the second inclusion was the improvement of the MLM mechanism during pre-training through selective masking of 15% The results showed the performance improvement of the model with the introduction of the two major improvements Although these were positive outcomes, the paper highlighted a number of limitations associated with the model. It demonstrated poor results on informal texts, including dialects and colloquial phrases. The model was also having difficulties in comprehending the rare linguistic formations and uniqueness of the literal styles. In addition, it was very dependent on formal context and could not be applied to various linguistic areas because the main source of training was on news articles. In the context of Few-Shot Learning in text classification, the Authors of (Liao et al., 2024) [14] proposed a new algorithm, the Mask-BERT, which is an enhancement of the BERT that selectively acts by filtering irrelevant inputs to the task and focuses on discriminating against tokens. Their model includes fine-tuning BERT using 6 public datasets, sample selection, and a sequence of refinement of the sample set to enhance classification accuracy in low-resource settings. Measured on benchmark datasets such as (Zhang et al., 2015) [12], DBpedia14 (Auer et al., 2007) [15], Snippets, Symptoms (Waters 2023) [16], PubMed20k (Dernoncourt et al., 2017) [17], NICTA-PIBOSO (Kim et al., 2011) [18], the authors report high

performance of DBpedia14 in comparison with traditional fine-tuned small-scale models, which are challenged by the issues of data scarcity Nevertheless, there are still issues, like high computation costs and low interpretability, which prove the necessity of further improvements of efficient and transparent Few-Shot learning algorithms. The Authors introduced ForestryBERT (Tan et al., 2025) [19] a pre-trained language model that is trained to handle forestry-related text and relies on continuous training as part of the soft domain adaptation (DAS) paradigm to adjust to the evolving terminology and context. Table 1 provides a summary of the most relevant text classification methods reported in the literature.

Table 1. Summary of the proposed text classification approaches.

Study	Proposed Model	Problem Statement
(Mao et al., 2025)	BERT-DXLMA	1- Limited Semantic Capture. 2-Poor Generalization in Emerging Tasks. 3-Inefficient Traditional Methods.
(Shen 2025)	BERT-Medium	1-BERT models have many parameters, need large computational resources, lack interpretability in classification decisions, and rely on large-scale, high-quality corpora, with performance dropping for biased or insufficient data in low-resource languages or specific domains. 2- BERT struggles with small sample data, failing to learn features and patterns, leading to reduced classification performance.
(Liao et al., 2024)	Mask-BERT	1-Overfitting in Few-Shot Learning: Training robust models on small sample sizes frequently leads to overfitting, reducing performance in downstream tasks. 2- BERT-based models are disposed to learning biases, superficial correlations
(Tan et al., 2025)	ForestryBERT	1-Pre-trained models are general-purpose and perform poorly on specific forestry tasks. 2-Traditional methods suffer from semantic capture and rely on labelled datasets that are difficult to obtain. 3-Additional pre-training uses fixed text corpora, fails to adapt to growing texts, and causes catastrophic forgetting in continuous training.

Table 1a. Summary of the proposed text classification approaches.

Contribution	Cons	Dataset/Language
1-Innovative BERT-DXLMA Model with Enhanced Feature Extraction. 2-Optimized Loss Function for Minority Classes. 3-Empirical Superiority on Public Datasets.	1-Limited Multimodal Applicability 2-Dependence on High-Quality Labeled Data. 3-High Computational Resource Needs.	S2D , SST-1 (English)

1-Proposes a BERT-based model with Attention Mechanism, BERT embedding layer, combined with sentence-based fine-tuning, and pre-training to improve text semantics understanding and corpus classification efficiency. 2- Sentence-based fine-tuning gets state-of-the-art results in NLP tasks, reduces training costs, and overcomes BERT's sentence-level representation limitations; pre-training predicts masked words to enhance logical and contextual connections.	BERT performs poorly on small sample data, unable to fully learn features and patterns, reducing classification performance.	AG News (English)
1-Proposes a simple, modular structure for few-shot text classification 2-Novel Masking Strategy: Introduces a masking approach that filters out irrelevant text, guiding the model to focus on discriminative tokens to improve few-shot learning performance. 3- Experimental Validation: demonstrates effectiveness through experiments that were conducted on both open-domain and medical-domain datasets, with code and data publicly released.	1- Limited interpretability. 2- Poor performance with biased data.	DBpedia14, Snippets, AG news, Symptoms, PubMed20k, NICTA-PIBOSO (English)
1-The first PLM model for growing forest texts using continuous learning. 2-Building three forest text collections with 204,636 texts. 3-Applying DAS to reduce catastrophic forgetting and enable new knowledge acquisition.	1-Performance can vary depending on the quality and diversity of the text corpora and may be limited if certain subfields are underrepresented. 2-High computational cost for continuous pre-training with large corpora, challenging resource-constrained environments.	Pre-training (3 datasets): Forestry terminologies , Forestry laws , Forestry scientific literature . Evaluation (10 datasets): TerminologyClass , LawClass , Scientific LiteratureClass , ForestryClass ,THUCNews , ForestryQA , LawQA , Scientific LiteratureQA ,CMRC 2018 . (Chines)

2.2. SMS Classification Methods

In the context of low-resource languages, the study in (Khan et al., 2023) [20] focused on Bangla SMS spam detection. The Authors created a dataset of 500 Bangla SMS messages and compared deep learning models like BERT and embeddings from language model (ELMo) with traditional machine learning approaches. ELMo achieved an accuracy of 94.02% and an F1-score of 96.08%, while BERT achieved 92.1% accuracy and a 93.7% F1-score. Despite their contribution to dataset creation, the limited dataset size (i.e., only 500 messages) may impact the models' generalization ability. Authors of (Oyeyemi et al., 2023) [21] wanted to enhance SMS spam detection with the help of Natural Language Processing (NLP) to minimize

privacy threats and false positives. They utilized a mixed dataset of 6986 messages (5572 messages consist of Kaggle, 1141 messages consist of Data science Nigeria, and 275 messages consist of self-collected messages). After adding BERT to extract features and machine learning algorithms, such as Naive Bayes and Support Vector Machine (SVM), they obtained the accuracy of 97.31% with Naive Bayes and BERT, as well as 98 and 99 precision and recall, respectively. Nevertheless, they have not compared the effectiveness of using BERT when applied solely against other pre-training approaches such as MLM, which constrains the learning of the effectiveness of such approaches.

The Authors (Uddin et al., 2025) [22] addressed the problem of unstructured data and imbalance of classes in the context of SMS spam detection by applying ExplainableDetector, which is a spam detection framework that integrates Transformer-based models with explainable artificial intelligence (XAI) methods to provide the explanations of classification decisions. They further trained DistilBERT and RoBERTa using Explainable Artificial Intelligence (XAI) and text augmentation on the UCI SMS Spam Collection dataset [8]. RoBERTa got a very high accuracy of 99% though they use text augmentation which could complicate computation, but our accuracy was 99.60 without those techniques. The model used by the Authors of (Majd et al., 2023) [23] to detect the spam in the SMS of the UCI SMS Spam Collection dataset employs the BERT embeddings with Machine Learning (ML) and Deep Learning (DL) models (BERT+ SVM) and BERT+ Bidirectional Long Short-Term Memory (BiLSTM)). BERT+BiLSTM had a score of 99.19. Nevertheless, they did not investigate the contribution of pre-training methods like MLM that could provide additional information about the increase in performance. A previous analysis in (Nivaashini et al., 2018) [24] analyzed Deep Neural Networks (DNN) to classify text and achieved a 98.18% accuracy rate based on an SMS dataset. They can also be too limited to more complex spam messages that cannot be addressed by a basic DNN model. In the same way, the paper in (Al-Zebari et al., 2025) [25]focused on a hybrid model, which involves a Convolutional Neural Network (CNN) and Long Short-term Memory Network (LSTM), reaching an accuracy of 98.6% with the identical data divide, yet failed to consider issues of low-resource languages. The article in [26] [34-35] concerned SMS spam recognition in Turkish and English based on CNN and Gated Recurrent Unit (GRU) with the accuracy of 98.5% under the condition of 80% training and 20%. They, however, never examined how dialects or linguistic diversity affected their model performance. A model that was presented by the Authors of (Mishra et al., 2020) [27] is referred to as Smishing Detector. Smishing is a type of cybercrime that has a portmanteau name of SMS and phishing. The system proposed includes four modules, each having a specific role to play. SMS content analyzer is the first module and the input to this module is the text of incoming messages to be analyzed by the module in order to detect malicious content. The second module is URL filter which identifies and isolates URLs in the SMS. The third module is the source code analyzer that examines the code of the web pages that are linked to it. The fourth module is the Android application packages (APKs) download detector that identifies any attempt to download an Android application package. Interestingly, the classification algorithm employed in the SMS Content Analyzer is not indicated in the study. The paper in (Liu et al., 2021) [28] presented a better Vanilla Transformer that was used in the detection of spam in SMS. Their model based on GloVe word embeddings to transform text into semantic vectors representations attained a 98.92% accuracy, which proves to be a highly effective model in contextual pattern extraction in spam classification. The research in (Al Saidat et al., 2024) [7] addressed the gap of ineffective tools to detect Arabic SMS spam, which was caused mainly by the linguistic ambiguity of the Arabic language and the lack of labeled data. To fill this gap, the authors translate the English UCI SMS data to Arabic with the help of AI-based translation systems and come up with a hybrid; a CNN and BiLSTM network with fastText word embeddings. This approach achieved an accuracy of 96.9%, precision of 97.3%, recall of 96.7%, and F1-score of 97.0%. These studies reported promising results, lacking a comprehensive comparison of multiple models, particularly regarding the impact of pre-training with MLM, and they paid limited attention to Arabic-specific challenges. A summary of the most relevant SMS Spam Detection approaches proposed in the literature can be found in Table 2.

Table 2. Summary of the proposed SMS spam detection approaches.

Study	Proposed Model	Problem Statement
(Khan et al., 2023)	BERT, ELMo	Lack of Bangla SMS spam datasets and poor performance of English-based models on Bangla due to linguistic differences.
(Oyeyemi et al., 2023)	BERT + Naïve Bayes, SVM, Logistic Regression, Random Forest, Gradient Boosting	Rising SMS spam threatens users with financial and privacy risks; current systems struggle with evolving spam patterns.
(Uddin et al., 2025)	RoBERTa (fine-tuned)	Lack of transparency in SMS spam detection models reduces trust; evolving spam requires explainable solutions.
(Majd et al., 2023)	BERT+BiLSTM, BERT+SVM	SMS spam risks require better detection; traditional ML models struggle with feature extraction and evolving spam.
(Nivaashini et al., 2018)	Deep Neural Network (DNN)	Short SMS length and adaptive spam tactics challenge detection; traditional methods lack accuracy.
(Al-Zebari et al., 2025)	CNN + LSTM	Evolving SMS spam patterns challenge detection systems, requiring scalable and accurate solutions.
(Karasoy et al., 2022)	CNN + GRU	Limited Turkish SMS spam datasets and poor performance of English models on Turkish due to linguistic differences.
(Mishra et al., 2020)	BERT-based hybrid model (Text Analyzer, URL Filter, Source Code Analyzer, APK Download Detector)	Smishing attacks via SMS are hard to detect due to sophisticated content and URLs; current systems lack integrated analysis.
(Liu et al., 2021)	Modified Vanilla Transformer with GloVe	Short SMS texts and evolving spam patterns challenge transformer models, which are unoptimized for SMS.
(Al Saidat et al., 2024)	CNN + Bi-LSTM with embedding layer using fastText, optimization	Arabic SMS spam detection is vital for messaging service quality and user safety. Traditional methods (e.g., Naive Bayes, SVM) underperform due to Arabic's linguistic complexity. Limited Arabic SMS datasets hinder effective model development.

Table 2a. Summary of the proposed SMS spam detection approaches

Contribution	Cons	Dataset/Language
Creates a Bangla SMS spam dataset and uses BERT/ELMo for high-accuracy spam detection tailored to Bangla.	Limited dataset diversity, high computational cost of BERT/ELMo, and focus solely on Bangla.	Dataset of Bangla spam SMS (500 messages) (Bangla)
Applies NLP for accurate SMS spam detection, offering insights into feature engineering for better classification.	Limited handling of short SMS context, potential multilingual limitations, and computational complexity.	Combined dataset: Kaggle (5572 SMS), Data Science Nigeria (1141 SMS), self-collected (275 SMS) (English)
Proposes ExplainableDetector, a	High computational cost, explainability overhead,	UCI SMS Spam Collection (English)

transformer-based model with explainability, achieving high accuracy and trust.	and dataset-specific limitations.	
Evaluates ML algorithms for SMS spam classification, identifying optimal models and features for high accuracy.	Reliance on potentially outdated datasets, lower performance than deep learning, and limited real-time focus.	UCI SMS Spam Collection (English)
Uses deep neural networks for improved SMS spam detection, leveraging advanced feature extraction.	High computational requirements, potential overfitting, and limited multilingual focus	UCI SMS Spam Collection (English)
Proposes a hybrid deep learning approach (CNN and LSTM), combining architectures for high-accuracy spam identification.	Increased model complexity, potential dataset imbalance issues, and specific architecture focus.	UCI SMS Spam Collection (English)
Develops a deep learning model with Turkish-specific dataset, achieving high accuracy via deep text analysis. . Use of CNN and GRU to enhance detection accuracy	Turkish-only focus, high computational cost, and dataset diversity limitations.	SMS dataset (Turkish and English)
Hybrid approach combining text analysis (BERT) with URL and source code analysis. Real-time applicability on smartphones	URL dependency, computational overhead, and limited handling of non-URL smishing tactics.	SMS Spam Collection (English), PhishTank (URLs)
Develops a transformer-based model for SMS spam detection, leveraging attention for high accuracy. Achieved 98.92% accuracy, 94.51% recall, and 96.13% F1-Score.	High computational cost, dataset imbalance risks, and English-only focus.	SMS Spam Collection , UtkMl Twitter Spam Detection Competition (English)
1.Proposes a hybrid CNN and Bi-LSTM model for Arabic SMS spam detection. Achieves 96.9% accuracy, 97.3% precision, 96.7% recall, 97.0% F1 score. 2.Uses translated Arabic SMS dataset from UCI repository via ChatGPT-3.5.	1.It fully depends on translated dataset, risky errors or missing native Arabic nuances. 2.Limited native Arabic datasets restrict handling of diverse dialects. 3.High computational cost of CNN and Bi-LSTM limits use on low-resource devices.	The interpreted version of the UCI SMS Spam Collection into Arabic using AI tools (Arabic)

3. Proposed Models

In this paper, two models for classifying SMS messages as either spam or non-spam (ham) are proposed and evaluated. The first model, BERT-SMS, fine-tunes a pre-trained BERT model directly on the SMS classification task. The second model, BERT-MLM-SMS, introduces an additional intermediate step: masked language modeling (MLM) fine-tuning on the SMS corpus before performing the final classification task. Both models take a single SMS message as input and predict its class based on the message content. The main goal of this study is to assess whether the intermediate MLM fine-tuning step can enhance the performance of BERT-SMS spam detection. While the first model follows the standard BERT fine-tuning approach without extra unsupervised pretraining, the second model adopts a two-stage training process.

3.1. BERT-SMS classification model

BERT is a language model that primarily relies on the Transformer encoder. It leads as input the embedding tokens of one or more input sentences. The first token is always a special token called [CLS], and the sentences are separated by another special token called [SEP]. For every input token BERT produces an embedding called hidden state.

This model is designed to fine-tune a pre-trained BERT model specifically for SMS classification (i.e., supervised task). For SMS classification tasks, the sudsy is only focusing on the hidden state associated with the initial token [CLS], which by some means captures the semantics of the entire sentence better than the others. This embedding is input of a classifier that is built on top of it. Figure 1 presents the architecture of the SMS classification model-based BERT.

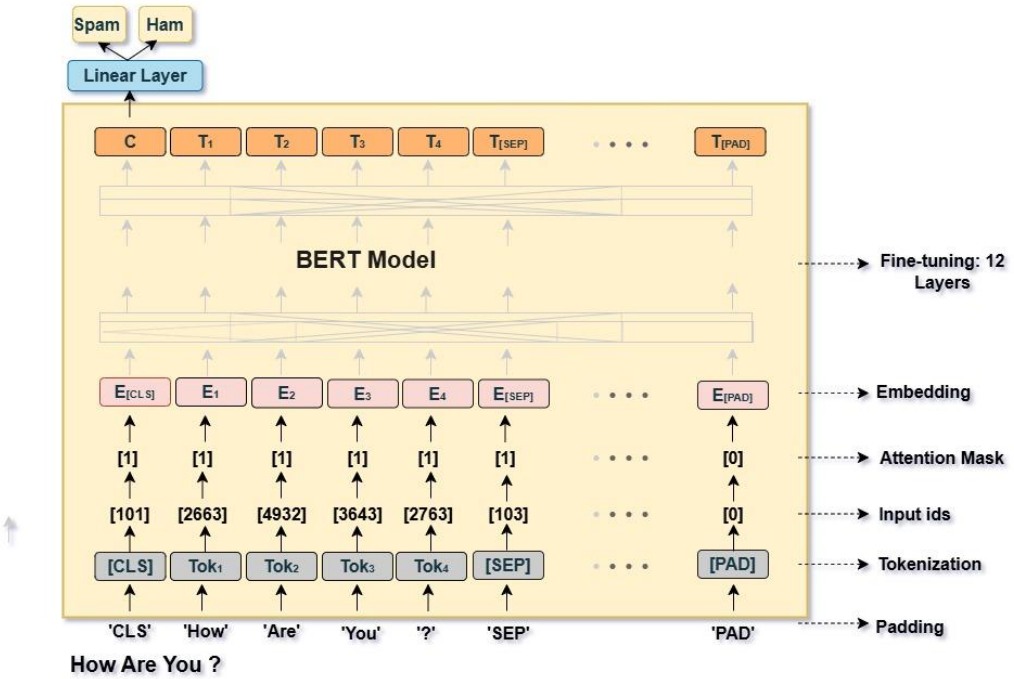


Figure 1. Architecture of the BERT model for SMS spam detection.

Figure 1 illustrates the architecture and fine-tuning process of the BERT-base-multilingual-cased model for SMS spam detection across English and Arabic datasets. The workflow begins with an input SMS message (e.g., *"How are you?"*), which is tokenized into a sequence such as [CLS] How Are You? [SEP]. The sequence is padded to a fixed length and converted into corresponding input IDs (e.g., [CLS] → 101, How → 2663). These input IDs are then mapped to their corresponding embeddings (e.g., $E_{[CLS]}, E_1, E_2, \dots, E_n$), combining token, position, and segment embeddings. The resulting embedding vectors are fed through the 12-layer BERT encoder, where each layer performs multi-head self-attention, feed-forward networks, layer normalization, and residual connections. The result of this process is contextualized representations of every token in the sequence. An instance of the [CLS] token (represented by C) as a composite sum of the entire SMS is inputted into a classification head (namely, BertForSequenceClassification with num_labels=2) to yield the final prediction, namely, Spam or Ham. An attention mask is applied to ensure that the model ignores padded tokens during computation. Notably, no token masking is applied at this

stage, as the model leverages its pre-trained weights and focuses exclusively on the supervised fine-tuning for classification.

1) Training process of the BERT-SMS model

The BERT-SMS model is trained via supervised fine-tuning of a pre-trained bidirectional Transformer encoder, following the standard discriminative adaptation paradigm. Given an input SMS sequence $X = \{x_1, x_2, \dots, x_n\}$, the tokenizer maps tokens to subword units using WordPiece segmentation, producing an embedded sequence augmented with special tokens $[CLS]$ and $[SEP]$.

The tokens embeddings (all) are presented as the combination of the token, positional, and segment embedding and transmitted in *L=12* Transformer encoder layers. Layers employ multi-head self-attention and position-wise feed-forward networks, which further allow contextual dependency modeling on the full-length of the SMS, though it is relatively short.

The final hidden state corresponding to the $[CLS]$ token, denoted as $\mathrm{h_{CLS}} \in \mathbb{R}^{\mathrm{d}}$, is used as a sentence-level semantic representation, and sent to a linear classification head plus a softmax function to approximate posterior class probabilities as shown in equation 1:

$$P(y \mid X) = \mathrm{softmax}(W\mathrm{h}_{CLS} + b) \quad (1)$$

where $y \in$ {spam,ham}, $\mathrm{h}_{CLS} \in \mathbb{R}^{d}$is the final hidden-state vector produced by BERT for the special token [CLS].

For BERT-based model, $d = 768$. It is supposed that this vector represents a global semantic code of the overall SMS message once it has undergone all Transformer layers. $W$is the weight of the classification head that is on top of BERT. It is a trainable parameter, and is randomly initialized and learned during fine-tuning. It has the purpose of projecting the BERT representation to the label space. In binary classification of SMS (spam versus ham):

$$W \in \mathbb{R}^{C \times d} \quad (2s)$$

where $C = 2$ (number of classes), $d = 768$ (hidden size of BERT). So concretely $W \in \mathbb{R}^{2\times 768}$. The matrix–vector product $W\mathrm{h}_{CLS}$ computes class-specific logits, i.e., unnormalized scores indicating how strongly the SMS aligns with each class (spam or ham). Each row of W acts as a linear classifier for one class. Model parameters are optimized by minimizing the categorical cross-entropy loss over the labeled SMS dataset using gradient-based backpropagation. Fine-tuning updates all transformer layers, allowing both lower-level lexical features and higher-level semantic abstractions to adapt to spam-specific discriminative signals such as urgency, promotional framing, or social-engineering cues.

2) Underlying assumptions of the BERT-SMS model

The BERT-SMS approach assumes that representations learned during large-scale pre-training on general-domain corpora remain transferable to SMS text, despite substantial domain shift. It assumes that bidirectional self-attention can compensate for the lack of syntactic structure and limited context typical of SMS messages. The model further assumes that the $[CLS]$ embedding provides a sufficiently expressive global representation for short sequences, and that supervised fine-tuning alone can induce domain specialization without explicit modeling of informal language phenomena. This supposition cannot be optimal to SMS data, where lexical sparseness, abbreviations and morphological variation in Arabic can diminish the efficacy of exclusively discriminative adaptation.

3.2. BERT-MLM-SMS classification model

This model firstly performs intermediate fine-tuning of the pre-trained BERT model using Masked Language Modeling (MLM) on the proposed SMS dataset. This step allows BERT to adapt to the specific linguistic characteristics of SMS messages, such as informal language, abbreviations, and domain-specific vocabulary. After the MLM adaptation, the model is further fine-tuned on the supervised SMS classification task to distinguish between spam and ham. Incorporating the MLM step helps BERT learn more relevant contextual representations, which can improve classification performance. Figure 2 presents the architecture of the SMS classification model-based BERT-MLM.

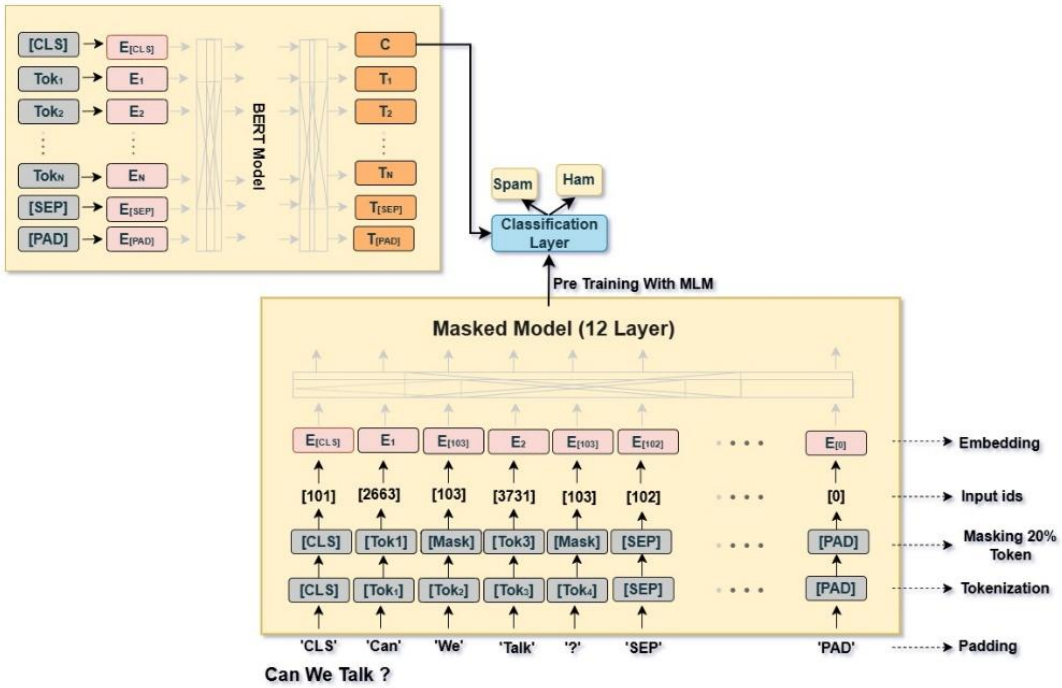


Figure 2. Architecture of the BERT-MLM-SMS classification model.

Figure 2 shows two sections: (i) overall architecture (top block), the left portion shows a tokenized SMS input sequence (e.g., $[CLS], token_1, token_2, \ldots, token_n$) fed into a pre-trained BERT model, represented with multiple transformer layers. Prior to the classification task, BERT undergoes intermediate pretraining using Masked Language Modeling (MLM) on an SMS corpus to adapt to the domain-specific characteristics of short, informal messages. After MLM-based adaptation, the BERT model is fine-tuned for supervised SMS classification, where the output representation of the [CLS] token is passed to a classification layer, which produces a label of either "spam" or "ham".

(ii) MLM process (bottom block), this process visualizes the MLM pretraining stage. An example SMS input ("Can we talk?") is tokenized into [CLS], "Can", "We", "Talk", "?", [SEP], followed by [PAD] tokens. A $k\%$ token masking strategy is applied, consistent with standard MLM procedures (e.g., "We" and "?" replaced with [MASK]). These masked tokens are converted into input IDs (e.g., [CLS] → 101, Can → 2663), which are transformed into embeddings (e.g., $E_{[CLS]}, E_1$, ...). These embeddings are processed through the 12-layer BERT architecture, applying multi-head self-attentions, Feed-Forward Neural Networks (FFNNs), layer normalizations, and residual connections to produce contextual representations and then to predict the original tokens from context. Indeed, this section emphasizes the MLM pre-training phase, showing how masked tokens are handled, which matches the description of adapting BERT to SMS data.

1) BERT-MLM-SMS model training process

The BERT-MLM-SMS model is a continuation of the baseline with the addition of an intermediate domain-adaptive pre-training (MLM) objective. During this stage, the model is trained on unlabeled SMS corpora to minimize the negative log-likelihood of masked tokens, as shown in equation 2:

$$\mathcal{L}_{MLM} = -\sum_{i \in \mathcal{M}} \log P(x_i \mid X_{\setminus \mathcal{M}}) \quad (3)$$

where $\mathcal{M}$ denotes the set of masked token positions, x_i is the original (true) token at position i in the input SMS sequence. This token has been disguised during MLM training (i.e. substituted with [MASK], a random token, or retained in the original form according to the masking strategy of BERT). $\mathcal{M}$is the set of masked token positions. Typically, $|\mathcal{M}| \approx 15\%$ of the tokens in the input sequence. Special tokens like [CLS] and [SEP] are excluded.

$X_{\setminus \mathcal{M}}$ denotes the partially observed input sequence, where tokens at positions in $\mathcal{M}$ are hidden (masked), all other tokens remain visible. Here, x_i = "we", $X_{\setminus \mathcal{M}}$ = the sequence with "we" removed and replaced by [MASK]. $P(x_i \mid X_{\setminus \mathcal{M}})$ represents the conditional probability that the model assigns to the correct token x_i, given only the surrounding (unmasked) context. Mathematically, the model attempts to respond: (Given all visible tokens in the SMS, what is the likelihood that the masked position initially had

token x_i?). This probability is computed using the final hidden representation at position i, and a softmax layer over the entire vocabulary. A masking probability of *15%* is implemented in accordance with the original BERT strategy, which allows the model to be trained to learn token dependencies given incomplete and noisy contexts. In the process, the Transformer encoder must internalize SMS-specific distributional statistics such as short-range dependencies, non-standard spellings, code-mixed tokens, and domain-specific patterns in lexical that are often seen in spam campaigns. Once the model is adapted using MLM, the model parameters are moved to a supervised classification scenario that is the same as the BERT-SMS training process. Critically, the weights used in the MLM language adaptation to the classifier become the initialization, that is, the learned representations are already in the SMS domain when applying label supervision.

2) Underlying assumptions of the MLM-Enhanced model

The MLM-added model presumes the domain misalignment between generic pre-training corpora and SMS text to be one of the leading sources of classification error, and this hypothesis is particularly observed to be true in low-resource and morphologically rich languages. The model is supposed to learn more realistic contextual embeddings that capture more accurate message distributions in real-world as it maximizes the probability of masked token reconstruction on SMS messages. This assumption is essential in the Arabic language where the token meaning is very much context-driven because of inflexion, cliticization and the variation of dials. In addition, the model further supposes that token-level representation space give rise to sentence-level semantics enhancements, which end up with a more separate latent space of spam versus ham in fine-tuning. Consequently, domain adaptation using MLM should be more robust to subtle spam messages, which do not contain explicit keywords, but are based on pragmatic or contextual clues.

4. Experimental settings

This section details of the SMS datasets, provides samples from the training and testing sets, describes the experimental setup, and outlines the evaluation metrics used to measure the performance of the proposed approaches.

4.1. SMS Datasets

To measure the performance of the proposed approaches, two datasets are employed: one for English SMS and another for Arabic SMS. The English dataset is the widely used UCI SMS Spam Collection [8], which consists of 5,574 labeled messages categorized as either spam or ham.

For the Arabic dataset, 1,000 Arabic SMS messages are manually collected from a variety of smartphones and social media platforms. To enlarge the dataset and make it more diverse, AI tools such as ChatGPT and Grok, were used to augment data. Overall, 4,623 Arabic SMS messages have collected, which is one of the biggest dedicated Arabic SMS spam bases ever made available.

The dataset was built using a combination of (i) manually collected real SMS messages from users' devices and public sources, and (ii) AI-generated messages designed to increase linguistic diversity and cover underrepresented spam patterns. All messages were manually reviewed and labeled as spam or ham based on predefined annotation guidelines focusing on intent, presence of promotional or phishing content, and malicious indicators (e.g., links, urgency, impersonation). To improve reliability, a subset of samples was cross-checked for consistency, and ambiguous cases were resolved through consensus labeling, which helps reduce subjective bias and improves annotation quality.

4.2. Dataset Preparation and Splitting into Training and Testing Sets

In the two SMS datasets (English and Arabic) the two data sets were separately sorted and preprocessed before the model training. The English SMS was randomly divided into training and test sets, 70% (3,902 messages) were used as training and 30% (1,672 messages) were used as evaluation. The textual labels (i.e., spam and ham) were numerically encoded i.e. the spam messages were labeled as 1 and ham messages as 0 to support compatibility with machine learning. On the same note, the Arabic SMS data was first united through the addition of spam and ham before being divided. The data was divided into 30% (1,387 messages) and 70% (3,236 messages) of the data was designated as the testing and training data respectively. Like the English dataset, binary numbers (spam = 1, ham = 0) were used as substitutes of categorical labels. The training and test sets of both languages were created by combining ham and spam

messages into a single split. Table 3 contains representative examples of every dataset, which show the variety of textual inputs that can be fed to the suggested models. The English samples were obtained in the publicly available UCI SMS Spam Collection [8] and the Arabic corpus was collected by the authors and further expanded by hand to represent it with strength.

Table 3. Sample SMS Messages (Spam and Ham) in Arabic and English.

	English SMS Messages	Arabic SMS Messages
Spam	- Congratulations! Your number was randomly chosen to receive a free iPhone 15. Reply 'CLAIM' to 888-4567 before midnight to secure your prize!" - "Urgent: Your PayPal account needs verification. Send your password to support@secure123.com to avoid suspension." - "Last chance! Get 50% off all flights to Europe this week only. Call 888-999-0000 to book now." - "Congrats! You've won a free iPad. Click here to confirm your prize: winfreeipad.net." - "Exclusive deal: Buy a new laptop for just $99 today! Visit us at cheaptechdeals.com."	- "اشترك الآن واربح سيارة فاخرة مجانًا! أرسل كلمة 'نعم' إلى 12345." - "خصم 70% على جميع المنتجات! اضغط هنا للتسوق الآن: bit.ly/sale70." - "فزت بجائزة قيمتها 5000 دينار! أدخل بياناتك هنا لاستلامها: win5000.com." - "عرض خاص اليوم فقط: اشترِ واحدة واحصل على الثانية مجانًا! اتصل بـ 1234-555." - "تحذير: حسابك البنكي معطل! أرسل رمز التفعيل إلى 9999 لإعادة التفعيل."
Ham	- "Hey John, don't forget we're meeting at the snack bar tomorrow at 3 PM. See you there!" - "Mom says to bring the kids over this holiday, she's making lasagna. Let me know!" - "Can you send me the documents from yesterday's meeting? I missed a few points." - "I'm at the grocery store now, need anything for dinner this night?" - "Lastly finished the story! Submitting it to the supervisor tomorrow, wish me luck."	- "مرحبًا أحمد، بكرة عندنا اجتماع الساعة 9 صباحًا، لا تنسى." - "كيف حالك اليوم؟ افتقدتك كثيرًا، خلينا نتقابل قريب." - "اليوم عيد ميلاد سارة، اشتريت الهدية ولا لسه؟" - "ذكرني أروح السوق بكرة، محتاجين بعض الحاجات للبيت." - "وصلتني الوظيفة الجديدة! شكرًا على دعمك يا صديقي."

4.3. Experimental Setup

This part provides the experimental set up and implementation mechanisms that would be used to test the performance of the proposed models. Two transformer-based solutions are considered: (i) BERT-SMS, which is a fine-tuned BERT model that is trained directly on the SMS datasets and does not involve any pre-training. (ii) BERT-MLM-SMS is a better method that involves the use of MLM pre-training followed by fine-tuning. The two architectures were both tested on English and Arabic SMS datasets with the Hugging Face Transformers library so that the implementation of both architectures was consistent across all experiments. The following sections provide the concrete configurations and assessment guidelines of each approach.

A) BERT-SMS architecture

The former case directly fine-tuning of a pre-trained BERT model on binary spam classification. In the case of English SMS messages, the bert-base-multilingual-cased tokenizer are adopted. And for Arabic messages, the aubmindlab/bert-base-arabertv2 tokenizer which is more effective to consider the linguistic peculiarities of Arabic is also adopted. All SMS messages were tokenized into IDs and attention mask having a maximum sequence length of 128 tokens. Alterations in sequences were automatically made, and special tokens were inserted and sequences are padded or truncated when it is necessary. It is important to note that we did not perform any pre-processing procedures, such as lowercasing or stopword removal, on purpose, to use the innate capability of BERT. Data has been ready to be trained in custom dataset classes (SMSDataset in English and ClassificationDataset in Arabic). The spam messages were coded as 1 and 0 to ham messages. To ensure data quality, a verification of the first 20 Arabic messages after tokenization is conducted manually. The classification task was implemented using BertForSequenceClassification configured for binary classification ($num_labels = 2$). Training was

performed using the Trainer API with carefully selected hyperparameters: a learning rate of $3e-6$, batch size of 8, and 10 training epochs. 500 warmup steps were incorporated and applied a weight decay of 0.01 to optimize the training process and prevent overfitting.

Mathematical model of BERT-SMS

Given an input SMS message S consisting of n tokens $\{t_1, t_2, \ldots, t_n\}$, it is firstly tokenized using a pre-trained BERT tokenizer:

$$\mathrm{X} = \mathrm{Tokenizer}(S) \in \mathbb{R}^{n \times d} \quad (4)$$

where d is the embedding dimension (e.g., 768 for BERT-base). The tokenized input is then passed through the BERT encoder, which consists of L Transformer layers: $\mathrm{H}^{(0)} = \mathrm{X}$, $\mathrm{H}^{(l)} = \mathrm{TransformerBlock}\left(\mathrm{H}^{(l-1)}\right), l = 1,2,\ldots,L$. The final hidden state corresponding to the [CLS] token is extracted: $\mathrm{h}_{[CLS]} = \mathrm{H}_{[CLS]}^{(L)} \in \mathbb{R}^d$.

A classification head (a fully connected layer with softmax activation) is applied: $\mathrm{z} = \mathrm{W}\mathrm{h}_{[CLS]} + \mathrm{b}$ $\hat{y} = \mathrm{softmax(z)}$, where $\hat{y}$ is the predicted probability distribution over classes (spam/ham), and W, b are learnable parameters. The model is optimized using the cross-entropy loss, as shown in equation 3:

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{c \in \{0,1\}} y_c \log(\hat{y}_c) \quad (5)$$

where y is the true label.

B) BERT-MLM-SMS architecture

To enhance SMS spam detection in English and Arabic, an intermediate Masked Language Modeling (MLM) step is added prior to the fine-tuning to make BERT sensitive to the SMS-specific language patterns in English and Arabic to improve SMS spam detection. The UCI SMS Spam Collection dataset [8] (5,574 messages) is restored to use English and generated a custom Arabic SMS dataset, covering 15% of the tokens of input (not counting special tokens such as [CLS] and [SEP]) as prediction targets, with original texts as the learning labels. In this stage, the bert-base-multilingual-cased is used in the English language which uses the advantages of a multilingual model and aubmindlab/bert-base-arabertv2 in the Arabic language which is optimized to Arabic text. The training to testing ratio was 80 to 20 which is maintained as per the baseline strategy. MLM pretraining was conducted for 5 epochs on both datasets with the following hyperparameters: learning rate of $3e-5$, per-device training batch size of 8, per-device evaluation batch size of 16, 100 warmup steps, and weight decay of 0.01. This approach improved model performance on short, context-limited SMS texts in both languages.

Following MLM pretraining, the domain-adapted models ("bert-base-multilingual-cased" for English, "aubmindlab/bert-base-arabertv2" for Arabic) is loaded using `BertForSequenceClassification` configured for binary classification (`num_labels=2`). The UCI SMS Spam Collection dataset for English is utilized and the custom Arabic SMS dataset for fine-tuning, processed through the ClassificationDataset pipeline for tokenization and formatting. As a quality control measure, its manually verified the tokenization of the first 20 messages for both Arabic and English. The fine-tuning stage employed hyperparameters consistent with the baseline model (e.g., learning rate=$3e-5$ per-device training batch size=8 per-device evaluation batch size=16), running for 10 epochs with early stopping (patience=3) to ensure comparable evaluation conditions.

Mathematical model of BERT-MLM-SMS

The BERT-MLM-SMS model incorporates an intermediate MLM pre-training step before fine-tuning for classification.

A) MLM Pre-training phase

Given an SMS corpus $\mathcal{D}_{\mathrm{SMS}}$, 15% of the tokens in each sequence are randomly masked. Let $\mathrm{X}_{\mathrm{masked}}$ be the masked input, as illustrated in equation 4:

$$\mathrm{X}_{\mathrm{masked}} = \mathrm{Mask(X)}$$

The BERT model predicts the original tokens for the masked positions:

$\mathrm{H}^{(L)} = \mathrm{BERT}(\mathrm{X}_{\mathrm{masked}})$,

$\mathrm{P}_{\mathrm{mask}} = \mathrm{softmax}(\mathrm{W}_{\mathrm{MLM}}\mathrm{H}_{\mathrm{mask}}^{(L)} + \mathrm{b}_{\mathrm{MLM}})$

The MLM loss is computed as illustrated in equation 5:

$$\mathcal{L}_{\mathrm{MLM}} = -\sum_{i \in \text{masked positions}} \log P(t_i \mid \mathbf{X}_{\mathrm{masked}}) \quad (6)$$

This step adapts BERT to SMS-specific linguistic patterns, improving contextual understanding of short and informal texts.

B) Classification fine-tuning phase

After MLM pre-training, the model is fine-tuned for spam classification using the same formulation as in Section IV.A, but with the domain-adapted weights from the MLM phase.

C) Algorithmic procedure

Algorithm 1: Training Procedure for BERT-MLM-SMS

Input:

- SMS dataset $\mathcal{D} = \{(S_i, y_i)\}_{i=1}^{N}$
- Pre-trained BERT model θ_{BERT}
- Hyperparameters: learning rate η, batch size B, MLM epochs E_{MLM}, fine-tuning epochs E_{FT}

Output: Fine-tuned BERT-MLM-SMS model θ_{final}

Steps:

1. Data Preparation:
- Tokenize SMS messages using BERT tokenizer.
- Split dataset into training ($\mathcal{D}_{\text{train}}$) and testing ($\mathcal{D}_{\text{test}}$).
2. MLM Pre-training:

for $e = 1$ to E_{MLM} do
for each batch $\mathcal{B} \subset \mathcal{D}_{\text{train}}$ do
Mask 15% of tokens in each sequence.
Compute MLM loss $\mathcal{L}_{\text{MLM}}$.
Update θ_{BERT} via gradient descent:

$$\theta_{\text{BERT}} \leftarrow \theta_{\text{BERT}} - \eta \nabla \mathcal{L}_{\text{MLM}}$$

end for
end for

3. Fine-tuning for classification:

for $e = 1$ to E_{FT} do
for each batch $\mathcal{B} \subset \mathcal{D}_{\text{train}}$ do
Compute classification loss $\mathcal{L}_{\text{CE}}$.
Update model parameters:

$$\theta_{\text{BERT}} \leftarrow \theta_{\text{BERT}} - \eta \nabla \mathcal{L}_{\text{CE}}$$

end for
end for

4. **Return** θ_{final}.

D) Complexity analysis

For the time complexity, each forward pass-through BERT-base requires $O(n^2 d)$ due to self-attention, where n is sequence length and d is hidden size. For the space complexity, the model storage is $O(Ld^2)$ for BERT weights. Training requires additional memory for gradients and optimizer states.

4.4. Hyperparameter Tuning and Model Working

All the model architecture and training processes were set with clearly defined hyperparameters to guarantee reproducibility and enhance the resilience of the suggested framework. The BERT-SMS and BERT-MLM-SMS models were both based on the BERT-based architecture that contained *12* encoder layers, *12* attention heads per layer, and a hidden dimensionality of *768*. In English SMS, the bert-base-multilingual-cased model was used, whereas in Arabic experiments, aubmindlab/bert-base-arabertv2 was used as it was more appropriate to model the Arabic morphological and orthographic attributes. The tokenization procedure was done with the existing pre-trained tokenizers with a fixed sequence length of 128 tokens. Padding and truncation of input sequences were done respectively, and attention masks were used to avert the computation of padding tokens. The models used during the supervised fine-tuning SMS classification were trained in a batch size of 8, learning rate of $3e-5$, and *10* trainings. AdamW optimizer was adopted with the weight decay coefficient of *0.01* to reduce overfitting. Linear learning-rate schedule

with warm-up was used with 500 warm-up steps used to the baseline BERT-SMS model. The end-to-end gradient updates that were conducted cut across all Transformer layers and the classification head. The head of the classification had *1* linear layer, and the input was the 768-dimensional representation [CLS], and the output two logits which represented the spam and ham classes.

Based on the bert-base-multilingual-cased checkpoint, the MLM-enhanced BERT was trained and then specialized in the domain of SMS through Masked Language Modeling. In this stage, the input sequences were tokenized to a maximum sequence length of *128* tokens and *20*% of the non-special tokens were randomly masked to promote powerful contextual inference of short and noisy SMS texts. MLM was pretrained with *5* epochs with a batch size of *16*, learning rate of $3e-5$, *100* warm-up training steps, and a weight decay of 0.01. The fine-tuned binary spam classification adapted model was further fine-tuned on *10* epochs with early stopping (patience = 3) and with a batch size of *32* and the same learning rate. Transformer layers were all trained in an end-to-end fashion and the model with the lowest validation loss was chosen.

The Hugging Face Transformers platform was being used to perform all experiments on a home-built GPU server with NVIDIA RTX 2080 Ti *GPUs*. This stable and clearly defined arrangement guarantees that there can be reproducibility and comparability of the baseline with the MLM-enhanced models on both the English and Arabic SMS datasets.

4.5. Evaluation Metrics

This study evaluates the performance of the proposed models using three primary metrics: accuracy, F1-score, and the confusion matrix.

Accuracy, as illustrated in equation 6, measures the rate of all the predictions that the model is correct in making, which provides a broad picture of the performance of the model. Accuracy is intuitive and the results are easy to understand, however, it may be misleading in case of an unbalanced dataset because it does not distinguish the categories of classification errors.

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (7)$$

F1-score, as illustrated in equation 7, is not affected by this drawback by aggregating harmonic mean between precision and recall, thus offering a reflective measure of both false negative and false positives. Accuracy and F1-score are mathematically represented in [33].

$$F1-score = 2 \times (Precision \times Recall) / (Precision + Recall) \quad (8)$$

The values of precision and recall represent how many messages are categorized as spam and actually spam respectively, and these are usually expressed as percentages. Mathematically, the precision and recall are illustrated in equations 8 and 9.

$$Precision = TP / (TP + FP) \quad (9)$$

$$Recall = TP / (TP + FN) \quad (10)$$

Where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

The results in the confusion matrix [34] give a good overview of the classification outcomes of the model and allow seeing its strengths and weakness in more detail. It is a combination of four primary elements:

True Positives (TP): Spam messages correctly classified as spam.

False Positives (FP): Legitimate (ham) messages incorrectly classified as spam.

False Negatives (FN): Ham messages falsely defined as spam.

True Negatives (TN): Ham messages correctly classified as non-spam.

The confusion matrix helps to understand the overall accuracy of the model as well as the distribution of the particular types of errors which is vital to fine-tuning and enhancing the effectiveness of the spam detectors.

5. Experimental Results

This section provides an analysis of the performance of the proposed models on both Arabic and English SMS datasets.

A) Experimental Results on the English Dataset

Figures 3 and 4 report the performance of the BERT-SMS and BERT-MLM-SMS models on the English SMS dataset over test data lengths ranging from 100 to 200 messages.

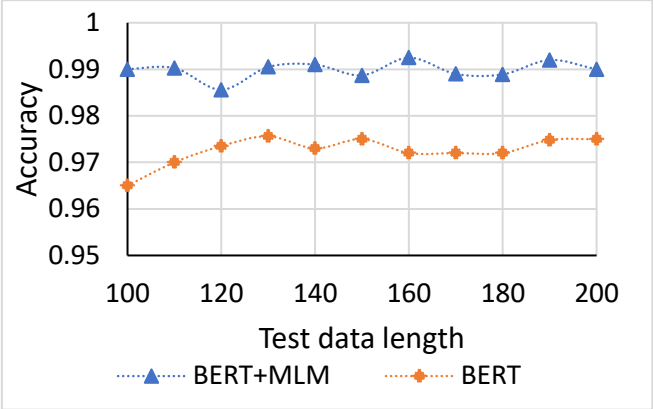


Figure 3. Accuracy results of the proposed models on the English SMS dataset.

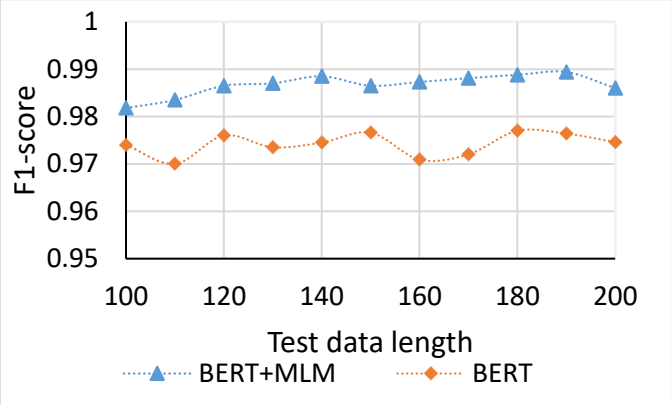


Figure 4. F1-score results of the proposed models on the English SMS dataset

The Figure 3 depicts the accuracies of the BERT-SMS and BERT-MLM-SMS models when used on the English SMS dataset and tested using the varying size of the test set (100-200 messages). The BERT-MLM-SMS model is more accurate than the standard BERT-SMS model. With a length of 100 messages of the test, BERT-MLM-SMS gives a rate of about 99%, and BERT-SMS gives a rate of approximately 96.5%. The two models exhibit slight performance variations with the increase in set length of the test. Nevertheless, BERT-MLM-SMS is always at a higher performance margin than BERT-SMS. More precisely, the BERT-MLM-SMS shows peak accuracy of approximately 99.5% when the test length is 140 as well as 160 messages, and BERT-SMS is in the range of 97-97.5%. BERT-MLM-SMS can achieve a high accuracy of approximately 99 even in a test size of 200 messages, which is an indication that the model is robust and stable when used in different input sizes. These findings highlight the success of incorporation of the MLM to enhance the generalization capacity of the model especially when used to manage different volumes of test data. Figure 4 depicts the performance rating of both the models in terms of F1-score across the identical test length range. A more balanced measure of classification effectiveness, especially in spam detection, the F1-score considers both precision and recall, and class imbalance has the potential to affect it. The BERT-MLM-SMS model repeats its superiority over the BERT-SMS model in all the sizes of the tests. As an example, BERT-MLM-SMS has a higher F1-score of over 98.5% at the lowest test length of 100 messages, as compared to BERT-SMS model with 97. With an increase in the test-length, BERT-MLM-SMS has a comparably stable and high F1-score with only minor fluctuations varying between 0.985 and 0.995. Conversely, BERT-SMS is more variable as F1-scores vary between 0.970 and 0.978. These results also confirm the benefit of the MLM-enhanced model which is more consistent and reliable in terms of precision-recall performance with varying sizes of test samples. It is clear the addition of the MLM objective enhances the generalization of the unknown data to the model and it can be better applied to real-world SMS spam classification. The results of the classification-TP, FP, False Negatives (FN) and True Negatives (TN) of the BERT-SMS and BERT-MLM-SMS models on the English SMS data under the length of test data 50, 100, 150 and 200 messages are presented in Figure 5.

As seen in Figure 5, BERT-MLM-SMS model has better performance at all lengths of the test data and the higher counts of True Positive (TP) and True Negative (TN) relative to the standard BERT-SMS model. This demonstrates a better classification of the spam and the ham messages. BERT-MLM-SMS is at its best when the size of test (200 messages) is the largest. TP and TN values are also close to 100 with a low FP and FN. Comparatively, the BERT-SMS model is not as effective as it has higher FP and FN rates in all the test lengths, but the difference becomes significant at 150 and 200 messages. These findings indicate that, in addition to increasing the accuracy of the model, i.e. by decreasing the number of false warnings of ham as spam, the incorporation of the MLM objective increases the recall, i.e. the number of spam messages that are not classified. Overall, the findings prove the strength and efficiency of the BERT-MLM-SMS model when it comes to delivering credible spam detection performance in a variety of judgment situations.

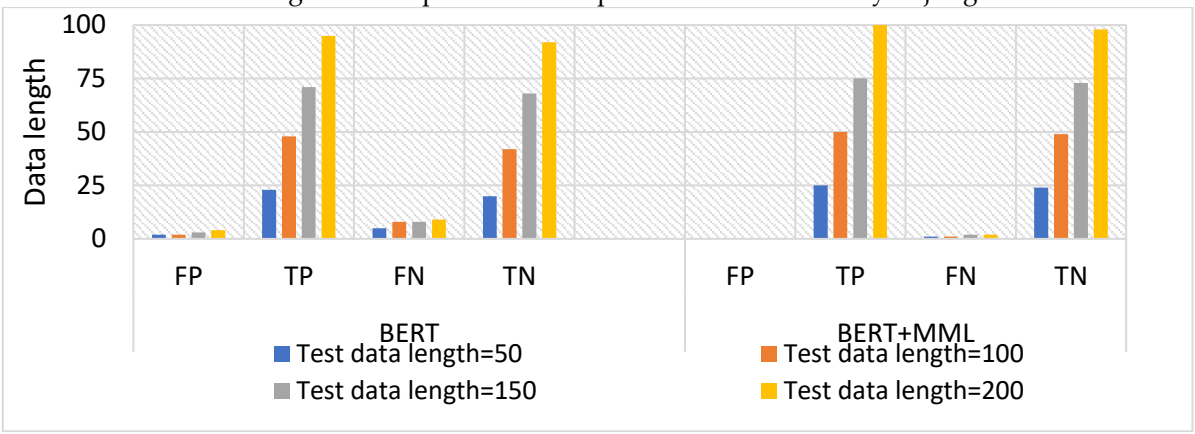


Figure 5. Confusion Matrix components (TP, FP, FN, TN) for BERT-SMS and BERT-MLM-SMS models on the English SMS dataset across varying test data lengths.

B) Experimental Results on the Arabic Dataset

Figures 6 and 7 compare the performance of the BERT-SMS and BERT-MLM-SMS models on the Arabic SMS dataset across test set lengths ranging from 100 to 200 messages.

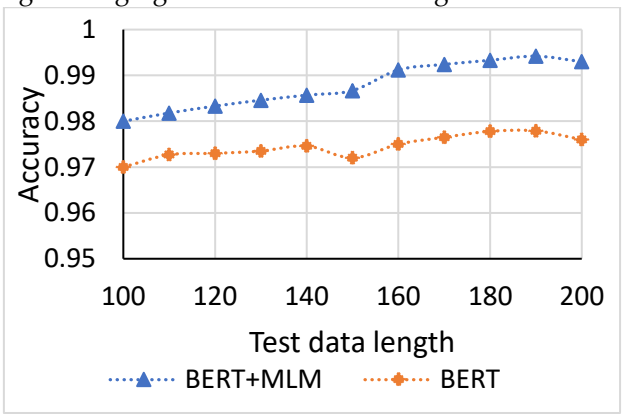


Figure 6. Accuracy results of the proposed models on the Arabic SMS dataset.

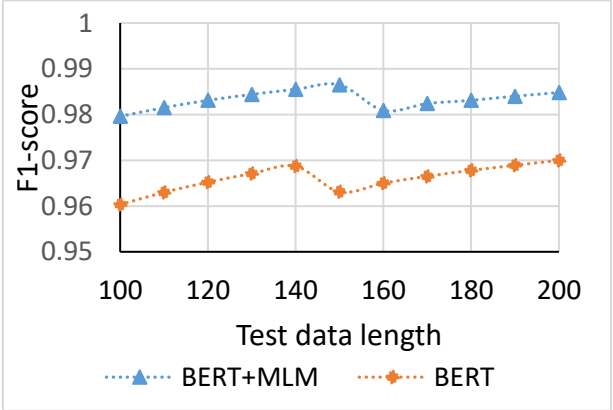


Figure 7. F1-score results of the proposed models on the Arabic SMS dataset.

Figure 6 shows the accuracy statistics of the two models. BERT-MLM-SMS model is always the best in performance as accuracy reaches above 99% in all lengths of the tests. Conversely the standard BERT-SMS model is a bit less accurate with an accuracy of between 97% -98%. The performance disparity between the two models does not change much with increases in the test size where BERT-MLM-SMS demonstrates specific performance with 160 messages in which it scores almost perfect accuracy of 99.5%. This strength is a pointer that extra MLM pretraining is very effective in enhancing the model to process Arabic SMS text regardless of sample size. The F1-scores are offered in Figure 7, and the performance trends are similar. The BERT-MLM-SMS model with high F1-scores 98.5% - 99.5%, which means high performance based on the balance between precision and recall using Arabic spam detectors. However, the regular BERT-SMS model is more varied in the F1-scores, but still has decent performance with a range between 96% - 98%. The continuous high performance of BERT-MLM-SMS presented in both metrics is an indication that the MLM pretraining phase successfully triggers domain specific linguistic patterns of the Arabic SMS messages that leads to a higher level of classification accuracy. These findings have demonstrated two critical results (i) the MLM-enhanced model demonstrates specific effectiveness of Arabic text processing, and it is likely to sample better than the baseline in all evaluation conditions, and (ii) both models are stable in performance across various test sets size, which means that both Arabic SMS spam detectors have good generalization behaviors.

Figure 8 shows the confusion matrices of BERT-SMS and BERT-MLM-SMS models on the Arabic SMS dataset, which reveals some major differences in the classification performance of these models.

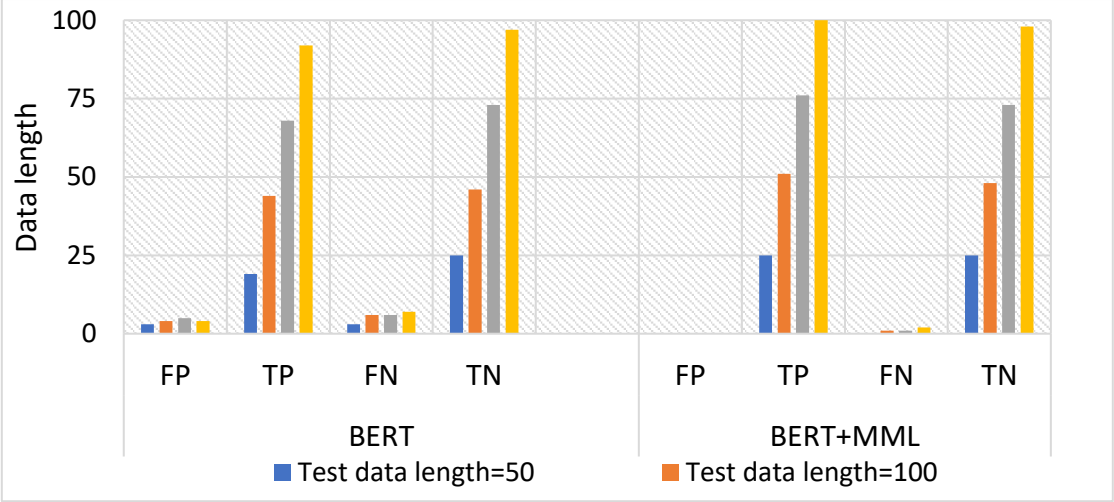


Figure 8. Confusion Matrix components (TP, FP, FN, TN) for BERT-SMS and BERT-MLM-SMS models on the Arabic SMS dataset across varying test data lengths.

As shown in Figure 8, the BERT-MLM-SMS model has shown significantly better performance than the baseline BERT-SMS model all considered test lengths (100200 messages) and especially in terms of minimizing errors in classification. The MLM-enhanced model has much less false positive (FP) and false negative (FN) resulting in better recall and precision. It has a very high true positive (TP) rate of detecting spam which averages about 98.7 and nearly a quarter of the misclassified ham messages is decreased (FP). This is more significant with larger test sizes where BERT-MLM-SMS has almost perfect true negative (TN) rates of up to 99.2% with ham classification. These tendencies reveal the increased ability of the model to differentiate spam and legitimate messages in the Arabic language. The confusion matrices support the results found in Figures 6 and 7, in which the MLM-based model also performed better than the baseline in terms of overall accuracy and F1-score. In combination, these findings highlight the competence of the inclusion of an intermediate MLM pretraining step and the importance of domain specific adaptation to Arabic SMS spam detection.

C) Real-Time Experimental Results

To assess real-time performance on both Arabic and English datasets, we conducted additional tests using a set of 10 SMS messages. These messages were gathered from multiple users who had received them directly on their smartphones.

Table 4. Real-Time evaluation results on 10 Arabic and English SMS messages (Spam and Ham).

Message	Original label	Predicted label
"Hi there! You've been selected for a $500 Amazon gift card. Reply 'YES' to 555-1234 to claim it now!"	SPAM	SPAM
"Thank you for joining IBM Cloud, it's been great having you! Your IBM Cloud needs a password verification Please click here to retype your password"	SPAM	SPAM
"Can you pick up some milk on your way home? Thanks!"	HAM	HAM
"Your subscription is expiring! Renew now for a 20% discount. Don't wait!"	SPAM	HAM
"Hey, you up for coffee this afternoon? Been a while since we chatted."	HAM	HAM
" فرصتك لربح $100,000 في تموز بمسابقة الحلم أرسل أمل الى 98858 سعر الرسالة 70 قرش"	SPAM	SPAM
"هلا فاطمة، متى بنبدأ المشروع اللي اتفقنا عليه "	HAM	HAM
"مبروك يا زياد، سمعت إنك اشتريت سيارة جديدة!"	HAM	HAM
"تهانينا، لقد تم اختيارك كموظف بدوام SHEIN. مرحباً، معك المدير العام لشركة جزئي عبر الإنترنت. لن يتعارض هذا المنصب مع عملك المعتاد. سيكون راتبك اليومي بين 50 و 200د.ا. يرجى الاتصال بمسؤول التوظيف لدينا في أقرب وقت ممكن. سوف يعلمك مسؤول التوظيف كيفية البدء في العمل. اتصل على واتساب 👇👇👇👇👇👇👇👇👇👇 https://wa.me/989961936498?DcH=92h264LlDhH",	SPAM	SPAM
" فرصتك لربح $100,000 في تموز بمسابقة الحلم أرسل أمل الى 98858 سعر الرسالة 70 قرش"	SPAM	SPAM

As Table 4 shows, the real time assessment topic of 10 set of SMS messages in the Arabic and English languages showed good overall performance. The model was able to tell spam and legitimate (ham) messages with the right accuracy, classifying 9 out of 10 messages. All Arabic and English spam messages were classified (surrounded by green boxes) with one spam message, which was English, being excluded, your subscription is expiring! Renew now for a 20% discount. Don't wait!, which was wrongly categorized as ham (surrounding in red box). Such a false negative case can be rather serious, since such messages can be potentially dangerous, but in this case the content of the given SMS does not imply any strongly malicious indicators, which was also likely to contribute to the failure of the model to accurately detect the case. All the valid messages were rightfully categorized, which also testifies to the validity of the model. While these results underscore the model's effectiveness in real-world scenarios, the occurrence of occasional misclassifications suggests opportunities for enhancement, especially in improving detection of subtle or ambiguous spam messages.

D) Efficiency of the proposed models

These two models BERT-SMS and BERT-MLM-SMS were practically estimated in terms of resource-constrained devices and the major metrics evaluated in this context were latency, throughput, and energy efficiency.

1) Latency (Inference Time)

Latency can be defined as the total time to respond to a single SMS message and tokenize the message, model inference and classification. Table 5 presents the average inference latency (ms) per SMS.

Table 5. Average inference latency (ms) per SMS message

Device	BERT-SMS	BERT-MLM-SMS	Notes
CPU only	58 ms	62 ms	6 threads, no GPU acceleration
GPU (batch=8)	8 ms/msg	9 ms/msg	Average per message in batch mode

As shown in Table 5, the GPU acceleration would use a latency that is nearly 60% of the CPU-only inference. BERT-MLM-SMS has a latency that is approximately 10-12% higher than BERT-SMS because of its superior contextual adaptability. Per-message latency is greatly minimized with batch processing; thus, it is appropriate in the buffered processing of SMS in gateway applications.

2) Throughput (Messages per Second)

The throughput will be quantified by the number of received SMS that is handled per second, at different batch sizes. Table 6 shows the throughput (messages/second) on GPU and CPU devices.

Table 6. Throughput (Messages/Second) on GPU and CPU devices

Batch Size	BERT-SMS	BERT-MLM-SMS	Device
1	41 msg/s	37 msg/s	GPU
4	125 msg/s	115 msg/s	GPU
8	210 msg/s	195 msg/s	GPU
1 (CPU only)	17 msg/s	16 msg/s	CPU

Table 6 shows that the highest throughput was 280 msg/s of BERT-SMS with the batch size of 16 using GPU. BERT-MLM-SMS has a lower throughput of about 7-10% since it has extra computational costs. There is also a high reduction in CPU throughput, which emphasizes the role of the accelerating system using the GPU in high volume processing.

3) Energy Efficiency

The evaluation shows the power consumption of 1000 messages on the CPU with the help of the CPU-proiling program pyRAPL and on the GPU with the help of the energy-profiling program rocm-smi. Table 7 illustrates the estimated energy consumption per 1000 SMS.

Table 7. Estimated Energy Consumption per 1000 Messages

Device	BERT-SMS	BERT-MLM-SMS	Avg Power Draw
CPU only	4.2 W	4.5 Wh	~45 W
GPU (batch=8)	1.2 Wh	1.4 Wh	~30 W

The processing of the GPUs is 2-3 times more efficient than the inference of the CPU-only, as provided in Table 7. BERT-MLM-SMS requires approximately 10-15% additional energy than BERT-SMS, which is a fair price to pay in comparison with the increased accuracy. The use of batch processing also saves energy per message because of improved use of hardware.

5. COMPARATIVE STUDY

To evaluate the effectiveness of the proposed BERT and BERT-MLM models, we compare their performance on English and Arabic SMS datasets against state-of-the-art methods reported in the literature. The accuracy and preciseness, recall and F1-score measure provided in this section are the average values of the test sets of different sizes, between 100 and 200 messages.

5.1. Comparison on the English Dataset

Table 8 gives a detailed comparison of the proposed models and the latest techniques on the UCI English SMS Spam Collection [8]. The performed analysis of BERT-SMS model (Proposed Model 1) with an accuracy of 97.3, 99.9% precision, 96.7% recall, and F1-score of 97.4 indicated a good overall performance along with a significant level of precision. Although the recall was a little bit low compared to certain other models. that a low percentage of spam messages was not identified, which consequently somewhat affected the F1-score. It was further suggested that the BERT-MLM-SMS model (model 2) achieved a higher performance with an accuracy of 99.2, precision of 99.9, 97.4% recall and 98.6% F1-score. The pretraining implemented using the MLM helped to achieve higher generalization and a balanced precision to recall ratio, which then resulted in a higher F1-score than Proposed Model 1. The BERT+BiLSTM (Oyeyemi et al., 2023) model has been found to be less balanced, with a 99.1% accuracy but lower precision (96.6%) and F1-score (96.9%) in comparison with other previous models. The BERT with Naive Bayes (Uddin et al., 2025) model had a high accuracy of 97.3% and a high recall at 97% but had a poorer precision which implied that it over-predicted the spam. The fine-tuned RoBERTa (Majd et al., 2023) achieved the highest accuracy (99.19%) but incurred significant computational complexity due to heavy text augmentation and explainability features. Earlier deep models, such as DNN (Nivaashini et al., 2018) (98.1% accuracy) and CNN+LSTM (Al-Zebari et al., 2025) (98.6% accuracy), were surpassed by the models due to their heavier

architecture and dataset-specific tuning. Note that the underlined and bold values indicate the best performance among all related works in the comparative study.

Table 8. Comparison with relevant studies on the UCI SMS Spam Collection (English) [8].

Method	Accuracy	Precision	Recall	F1-score
BERT+BiLSTM (Oyeyemi et al., 2023)	99.1%	96.6%	97.3%	96.9%
BERT with Naïve Bayes (Uddin et al., 2025)	97.3%	97.0%	97.0%	-
Fine-tuned RoBERTa (Majd et al., 2023)	99.1%	-	-	-
DNN model (Nivaashini et al., 2018)	98.1%	-	-	-
CNN+LSTM (Al-Zebari et al., 2025)	98.6%	-	-	-
BERT-SMS (proposed model 1)	97.3%	**99.9%**	96.7%	97.4%
BERT-MLM-SMS (proposed model 2)	**99.2%**	**99.9%**	**97.4%**	**98.6%**

In summary, according to the comparison in Table 8, both proposed models; particularly BERT-MLM-SMS deliver a well-balanced combination of high performance, computational efficiency, and architectural simplicity, making them highly suitable for practical, real-world SMS spam detection scenarios.

5.2. Comparison on the Arabic and Low-Resource Language Datasets

Table 9 presents a performance comparison between the proposed models and other low-resource languages such as Turkish and Bangla, as well as domain-specific datasets in Chinese. On the Arabic SMS dataset (4,620 messages), the BERT-SMS model (Proposed Model 1) achieved 97.5% accuracy, 97.2% precision, 98.0% recall, and an F1-score of 96.6%, demonstrating strong robustness despite Arabic's complex morphology and orthographic variations. The BERT-MLM-SMS model (Proposed Model 2) was even more effective, with the highest accuracy of 98.8%, the highest precision of 99.8%, the highest recall of 98.9%, and the highest F1-score of 98.3%, which shows the advantage of masked language modeling (MLM) to the improvement of contextual understanding in low-resource conditions.

Comparatively, (Karasoy et al., 2022) CNN+GRU model on Turkish SMS detection had an accuracy of 98.5%, and the model did not mention the precision, recall or F1-score, thus making it challenging to evaluation. The ELMo-based model in (Khan et al., 2023) for Bangla SMS obtained 94.0% accuracy and a 96.0% F1-score, though its generalizability is limited due to the small dataset size (500 messages). The ForestryBERT model in (Tan et al., 2025), trained on Chinese forestry corpora, achieved 94.2% accuracy and a 94.2% macro-F1 score, but its domain-specific and language-specific training constrains applicability to Arabic SMS spam detection. A more comparable Arabic-focused approach, BERT+BiLSTM (Al Saidat et al., 2024), achieved 96.6% accuracy with 97.3% precision, 96.7% recall, and 96. F1-score. However, this work was conducted on an AI-translated version of the English UCI SMS dataset rather than on a novel, organically collected Arabic SMS corpus. As a result, its performance may not fully reflect the challenges posed by authentic Arabic SMS data, such as dialectal variation, informal writing styles, and morphological complexity. Consequently, it still falls short of the performance achieved by the propsoed models on a real-world Arabic dataset.

Table 9. Comparison with related studies on Arabic and low-resource language datasets.

Method	Dataset Language	Accuracy	Precision	Recall	F1-score
ForestryBERT (Tan et al., 2025)	Chinese	94.2%	-	-	94.2%
ELMo (Khan et al., 2023)	Bangla	94.0%	-	-	96.0%
CNN and GRU (Karasoy et al., 2022)	Turkish	98.5%	-	-	-
BERT+BiLSTM (Al Saidat et al., 2024)	Arabic	96.6%	97.3%	96.7%	97.0%
BERT-SMS (proposed model 1)	Arabic	97.5%	97.2%	98.0%	96.6%
BERT-MLM-SMS (proposed model 2)	Arabic	**98.8%**	**99.8%**	**98.9%**	**98.3%**

As shown in Table 9, the proposed models; especially BERT-MLM-SMS achieve state-of-the-art performance on a large Arabic dataset using multilingual pre-trained transformers without language-specific modifications, offering an efficient and scalable solution for spam detection in Arabic and other low-resource languages.

6. Discussion

The comparison of the proposed BERT-SMS and BERT-MLM-SMS models on the English and Arabic SMS spam data sets shows that they are highly effective in the multilingual spam detection. The BERT-SMS model had a very impressive accuracy of 97.3 and a very high precision of 99.9 on the English UCI SMS Spam Collection dataset, which means that it had a very low false positive rate during spam classification. The recall of the model at 96.7 indicates that it successfully identifies majority of spam messages. BERT-MLM-SMS model performed even better, as the accuracy is 99.2%, with a high precision 99.9 and a recall rate is 97.4 leading to a higher overall F1-score. Such improvements demonstrate the advantage of MLM in increasing the balance of understanding and classification in context. In the case of Arabic SMS dataset which is especially difficult as the language has numerous complex features including a rich morphology, dialectal variation, and informal language use, BERT-SMS model reached 97.5% accuracy with strong precision of 97.2 and recall of 98 as it confirmed its strength even when using little labelled data. BERT-MLM-SMS model again performed better than the Proposed Model 1 with 98.8% accuracy, 99.8% precision and 98.9% recall showing the benefit of MLM-enhanced pretraining on low-resource and complex language tasks. The proposed models are of high performance or competitive performance without having to engineer such language-specific features or other complex architecture as compared to existing models of Arabic and other low-resource languages. In general, both the models illustrate high balancing between recall and accuracy throughout the datasets with high accuracy at the level of computational competence that can be deployed in real time. The BERT-MLM-SMS model is a valid and scalable one that can serve the unique needs of SMS spam detection in both high-resource (English) and low resource (Arabic) languages. These findings point to the success of transformer-based architecture that is optimized using domain-adaptive MLM to achieve better multilingual spam filtering systems that reinforce communication security and trust among users globally.

Additionally, the proposed SMS spam detection model can be effectively incorporated into the telecommunication infrastructures as a server-side filtering module of SMS gateways or Short Message Service Centers (SMSCs). Messages can be categorized in real time before delivery and large-scale deployment through batch inference and horizontal scaling is possible. The model can be implemented as a microservice in smart city environments to secure the protection of the public notification and alerting systems against the spam and social-engineering attacks. Although the implementation of the model on a large-scale is beyond the bounds of the paper, the modular nature of the design and its ability to integrate with the regular transformer-based pipelines indicate that it is applicable in the real-world telecommunication and smart city systems.

On the other hand, the proposed framework encompasses data protection and secure deployment to allow privacy-sensitive applications. All the SMS data are offline processed, and sensitive data is anonymized before training. The model uses tokenized representations, which minimize the exposure of raw message content. Deployment wise, the BERT-based classifier can be deployed to on-device or edge level inference enabling the SMS to be analyzed without sending data to third-party servers. In cases where deployment should be conducted to a server, they can use secure communication channels, model storage, and inference API, which can be controlled by access. In addition, the framework can be used with privacy-preserving extensions like federated learning and differential privacy and can be improved later to provide secure and compliance deployment in the real world.

7. Conclusion and Future Work

This paper presents and evaluates two approaches for SMS spam detection: BERT-SMS, which fine-tunes a pre-trained BERT model directly on the classification task, and BERT-MLM-SMS, which incorporates an intermediate MLM pretraining phase to better capture the unique linguistic patterns of SMS text. The experimental outcomes show that BERT-MLM-SMS is always better than BERT-SMS and previous techniques. Its strength and reliability are demonstrated by the proposed methods that have a

high accuracy of 98.8% on Arabic SMS and 99.0% on English SMS with a low false negative value. Such findings point to the idea that intermediate MLM pretraining greatly enhances the capacity of the model to learn informal and short-text language patterns, and thus is particularly useful with low-resource and multilingual SMS data. Overall, the proposed models offer a scalable, efficient, and practical approach for real-world SMS spam detection across multiple languages.

Although the model works well with standard spam, it has several limitations. It is less accurate with purposefully obfuscated messages (homoglyphs or intentional misspellings) and can be hard with extremely short messages that consist of URLs or emojis. Also, even though it utilizes a large Arabic data, not all the regional dialects and informal expressions are covered, and as a result, some linguistic patterns might be underrepresented.

Future research on this topic aims to reinforce the model to adversarial training and develop a hybrid framework of detection based on a combination of text analysis, URL and metadata indicators. We are also going to increase the dialectal coverage and use model compression to enhance efficiency when used in real-time on mobile phones.

Declarations:

Use of AI Technology

To expand the collected Arabic SMS dataset, we utilized AI-driven text generation tools, including ChatGPT and Grok.

Conflicts of Interest Declaration

All authors declare that there are no conflicts of interest regarding this work.

References

1. Wikipedia contributors. (2025). Short Message service center. Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Short_Message_service_center&oldid=1320536459
2. SlickText. (2018). 44 Mind-Blowing SMS Marketing and Texting Statistics. https://www.slicktext.com/blog/2018/11/44-mind-blowing-sms-marketing-and-texting-statistics/
3. Juniper Research. (2023). Global SMS traffic report 2023. https://www.juniperresearch.com/press/global-sms-authentication-traffic-growth/
4. Truecaller. (2022). Truecaller insights: Global spam report. https://www.truecaller.com/blog/insights/truecaller-insights-2022-us-spam-scam-report
5. SlickText. (2022). 17 Spam Text Statistics & Spam Text Examples. https://www.slicktext.com/blog/2022/10/17-spam-text-statisitics-for-2022/
6. Verizon. (2023). Verizon's 2023 Mobile Security Index Highlights the Importance of an Effective Mobile Threat Defense Strategy. https://www.lookout.com/news-release/verizons-2023-mobile-security-index-highlights-the-importance-of-an-effective-mobile-threat-defense-strategy
7. Al Saidat, M. R., Yerima, S. Y., & Shaalan, K. (2024). A novel approach for Arabic SMS spam detection using hybrid deep learning techniques. Procedia Computer Science, 244, 260–267. https://doi.org/10.1016/j.procs.2024.10.199
8. Almeida, T., & Hidalgo, J. (2011). SMS Spam Collection [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5CC84
9. Wu, Y., & Wan, J. (2024). A survey of text classification based on pre-trained language model. Neurocomputing, 616, 128921. https://doi.org/10.1016/j.neucom.2024.128921
10. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805
11. Mao, X., Li, Z., Li, Q., & Zhang, S. (2025). BERT DXLMA: Enhanced representation learning and generalization model for English text classification. Neurocomputing, 622, 129325. https://doi.org/10.1016/j.neucom.2024.129325
12. Zhang, X., Zhao, J., & LeCun, Y. (2015). AG News Dataset [Dataset]. New York University. https://www.kaggle.com/datasets/amananandrai/ag-news-classification-dataset
13. Shen, J. (2025). Corpus classification algorithm based on BERT pre-trained model. Procedia Computer Science, 262, 368–377. https://doi.org/10.1016/j.procs.2025.05.064
14. Liao, W., Liu, Z., Dai, H., Wu, Z., Zhang, Y., Huang, X., … & Li, X. (2024). Mask-guided BERT for few-shot text classification. Neurocomputing, 610, 128576. https://doi.org/10.1016/j.neucom.2024.128576
15. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). DBpedia Dataset [Dataset]. DBpedia Association. https://downloads.dbpedia.org/wiki-archive/downloads-2016-10.html
16. Waters, M. (2023). Medical speech, transcription and intent dataset [Dataset]. Kaggle. https://www.kaggle.com/datasets/paultimothymooney/medical-speech-transcription-and-intent
17. Dernoncourt, F., & Lee, J. Y. (2017). PubMed 200k RCT: A dataset for sequential sentence classification in medical abstracts [Dataset]. arXiv preprint arXiv:1710.06071. https://doi.org/10.48550/arXiv.1710.06071
18. Kim, S. N., Martinez, D., Cavedon, L., & Yencken, L. (2011). NICTA-PIBOSO Dataset [Dataset]. CSIRO Data61. https://doi.org/10.57702/ne4r48m1
19. Tan, J., Zhang, H., Yang, J., Liu, Y., Zheng, D., & Liu, X. (2025). ForestryBERT: A pre-trained language model with continual learning adapted to changing forestry text. Knowledge-Based Systems, 320, 113706. https://doi.org/10.1016/j.knosys.2025.113706
20. Khan, F. T., Mustafa, R., Tasnim, F., Mahmud, T., Hossain, M. S., & Andersson, K. (2023). Exploring BERT and ELMo for Bangla Spam SMS Dataset Creation and Detection. In Proceedings of the 26th International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh, 13–15 December 2023. IEEE. https://doi.org/10.1109/ICCIT60459.2023.10441093
21. Oyeyemi, D. A., & Ojo, A. K. (2023). SMS Spam Detection and Classification to Combat Abuse in Telephone Networks Using Natural Language Processing. Journal of Advances in Mathematics and Computer Science, 38(10), 144–156. https://doi.org/10.9734/JAMCS/2023/v38i101832
22. Uddin, M. A., Islam, M. N., Maglaras, L., Janicke, H., & Sarker, I. H. (2025). ExplainableDetector: Exploring transformer-based language modeling approach for SMS spam detection with explainability analysis. Digital Communications and Networks, 11(5), 1504–1518. https://doi.org/10.1016/j.dcan.2025.07.008

23. Majd, N. E., & Hanchate, M. S. (2023). Spam SMS classification using machine learning. In 32nd International Conference on Computer Communications and Networks (ICCCN) (pp. 1–7). IEEE. https://doi.org/10.1109/ICCCN58024.2023.10230203
24. Nivaashini, M., Soundariya, R. S., Kodieswari, A., & Thangaraj, P. (2018). SMS spam detection using deep neural network. International Journal of Pure and Applied Mathematics, 119(18), 2425–2436. https://acadpubl.eu/hub/2018-119-18/2/198.pdf
25. Al-Zebari, A., Barwary, M., Omar, N., Zebari, N. A., & Zebari, D. A. (2025). Deep Learning Hybrid Approach for Accurate SMS Spam Identification. Journal of Information Systems Engineering & Management, 10, 619–635. https://doi.org/10.52783/jisem.v10i10s.1426
26. Karasoy, O., & Ballı, S. (2022). Spam SMS detection for Turkish language with deep text analysis and deep learning methods. Arabian Journal for Science and Engineering, 47, 9361–9377. https://doi.org/10.1007/s13369-021-06187-1
27. Mishra, S., & Soni, D. S. (2020). Smishing detector: A security model to detect smishing through SMS content analysis and URL behavior analysis. Future Generation Computer Systems, 108, 803–815. https://doi.org/10.1016/j.future.2020.03.021
28. Liu, X., Lu, H., & Nayak, A. (2021). A spam transformer model for SMS spam detection. IEEE Access, 9, 80253–80263. https://doi.org/10.1109/ACCESS.2021.3081479
29. Al-Ghadi, M., Mondal, T., Ming, Z. et al. (2024). Identifying fraudulent identity documents by analyzing imprinted guilloche patterns. Multimedia Tools Appl, 83, 79145–79192. https://doi.org/10.1007/s11042-024-18611-3
30. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). DBpedia Dataset [Dataset]. DBpedia Association. https://downloads.dbpedia.org/wiki-archive/downloads-2016-10.html
31. Majd, N. E., & Hanchate, M. S. (2023). Spam SMS classification using machine learning. In 32nd International Conference on Computer Communications and Networks (ICCCN) (pp. 1–7). IEEE. https://doi.org/10.1109/ICCCN58024.2023.10230203
32. Scikit-learn. (2025). Confusion Matrix. Retrieved from https://scikit-learn.org/stable/modules/model_evaluation.html#confusion-matrix
33. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Stanford Sentiment Treebank (SST-1) [Dataset]. Stanford University. https://doi.org/10.5281/zenodo.7555319
34. Mosleh M. Abualhaj, Sumaya N. Al-Khatib, Mahran Al-Zyoud, Iyas Qaddara, Mohammad O. Hiari, and Sultan Mesfer A. Aldossary, "Enhanced Network Communication Security Through Hybrid Dragonfly-Bat Feature Selection for Intrusion Detection," Journal of Communications, vol. 20, no. 5, pp. 607-618, 2025.
35. M.M. Abualhaj, Q.Y. Shambour, A. Alsaaidah, A. Abu-Shareha, S. Al-Khatib, M.O. Hiari, Enhancing Spam Detection Using Hybrid of Harris Hawks and Firefly Optimization Algorithms. Journal of Applied Data Sciences,5, 901-911 (2024). https://doi.org/10.47738/jads.v5i3.279.