

أريج العتيبي

Automatic Detection of Arabic Spam Messages Using Machine Learning Techniques.docx

 Midocean University

Document Details

Submission ID

trn:oid:::2945:379358229

Submission Date

Apr 23, 2026, 10:15 PM GMT+2

Download Date

Apr 24, 2026, 1:35 PM GMT+3

File Name

Automatic Detection of Arabic Spam Messages Using Machine Learning Techniques.docx

File Size

622.6 KB

30 Pages

5,614 Words

31,783 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positive results or false negative results), so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?

The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?

Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.





Automatic Detection of Arabic Spam Messages Using Machine Learning Techniques

Prepared By

Name: Asma Sayer Alotaibi

ID No.: 1094558820

Name: Areej Abdualaziz Alotaibi

ID No.: 1091220275

Submitted in fulfillment of the requirements for the degree Master of
Science in Artificial Intelligence

At

Collage of Informatics

Midocean University

UAE

Supervisor Name: Dr. Khaled Eskaf

Collage of Informatics, Midocean University, Comoros

Email:.....

Acknowledgements

Special thanks to Dr. Khaled Eskaf

List of Contents

Acknowledgements	iii
List of Abbreviations	v
List of Figures.....	vi
List of Tables	vii
Chapter 1 Introduction	1
1.1 Overview	1
1.2 Problem Definition.....	2
1.3 Thesis Objectives	2
1.4 Thesis Contribution.....	3
1.5 Thesis Organization	3
Chapter 2 Background and Related Work.....	4
2.1 Introduction	4
2.2 Background	4
2.3 Related works.....	9
References	22

List of Abbreviations

- **SMS:** Short Message Service
- **NLP:** Natural Language Processing
- **ML:** Machine Learning
- **DL:** Deep Learning
- **SVM:** Support Vector Machine
- **NB:** Naïve Bayes
- **RF:** Random Forest
- **CNN:** Convolutional Neural Network
- **RNN:** Recurrent Neural Network
- **LSTM:** Long Short-Term Memory
- **BiLSTM:** Bidirectional Long Short-Term Memory
- **GRU:** Gated Recurrent Unit
- **TF-IDF:** Term Frequency – Inverse Document Frequency
- **BoW:** Bag of Words
- **GAN:** Generative Adversarial Network
- **IoT:** Internet of Things
- **URL:** Uniform Resource Locator
- **API:** Application Programming Interface
- **AI:** Artificial Intelligence
- **ELM:** Extreme Learning Machine
- **F1-score:** Harmonic Mean of Precision and Recall

List of Figures

Figure 1. NLP text preprocessing pipeline diagram6

Figure 2. spam extraction methods NLP diagram7

Figure 3. CNN for text classification architecture diagram.....8

List of Tables

Table 1 Related Work Summary.....16

Chapter 1 Introduction

1.1 Overview

Digital communication has grown so fast that mobile messaging has become one of the most significant mediums of personal, commercial and financial communication. Simultaneously, this expansion has given way to other malicious parties to propagate spam, phishing, fraudulent advertisements, and malicious links via short-message services and social media. Detection of spam has thus emerged as a significant research outcome due to the fact that unwanted messages decrease user confidence, cost time and can potentially subject a person to financial damages or identity theft. This is of particular concern to the Arabic contexts due to the rich morphology, numerous dialects and numerous writing styles that make automatic analysis of text more challenging in the Arabic language in comparison to most other languages. It has been observed that Arabic spam detection is a difficult problem due to the lack of labeled data, the preprocessing of Arabic data being difficult, and the fact that it requires models that are sensitive to short and noisy messages [1], [2], [3]. In the past, a variety of machine learning and deep learning techniques have been used to identify spam in Arabic web pages, tweets, social posts, and SMS messages. As an illustration, research on Arabic web spam in the past already proved that classical classifiers like Decision Tree and Naïve Bayes can be applied to spam and non-spam content [1].

Recent studies on Arabic tweets spam detection employed a variety of supervised machine learning and deep learning classifiers and indicated that preprocessing and tokenization techniques have a significant impact on better performance [2]. Other papers suggested the use of data augmentation models to overcome the issue of small training data, and they were better on macro F1 and recall accuracy of the spam class [3]. Similarly, recent SMS-specific research investigated hybrid deep learning systems like CNN-BiLSTM and BiGRU-based systems to detect Arabic SMS spam or phishing, finding good performance and confirming that deep systems are capable of capturing more contextual information in short Arabic text [4], [5].

The thesis is based on the premise that Arabic SMS spam and phishing messages may be identified more efficiently when the text preprocessing of the day is coupled with appropriate machine learning or deep learning classifier. Due to the shortness, repetitiveness, and deliberate misdirected

nature of spam and phishing texts, a sound detection mechanism needs to be capable of pointing out both the superficial and sub-contextual indicators. Because of this, the current research paper is based on the application of the Arabic text-contents analysis as a foundation of classification. The overall goal is to help improve the creation of a more realistic and useful framework of identifying malicious Arabic messages in actual communication situations .

1.2 Problem Definition

Even with the current spam detection advancements, the Arabic text is still underrepresented in the current spam detection systems that are mostly trained using English data. The SMS messages in Arabic also pose further difficulties because of linguistic diversity as they contain dialects, informal scripts, abbreviations, mixed-language messages, and irregular spelling patterns that make it less effective to use the conventional detection techniques. Attackers also make it difficult to detect them in phishing situations by impersonating trusted people and employing convincing language to defraud users. In technical terms, the issue is compounded by an unbalanced dataset and a lack of labeled data which may result in biased models and bad generalization. Thus, the main problem that is discussed in this thesis is to create a powerful and precise method of detecting Arabic spam and phishing messages that can be effectively used in cases of short, noisy, and heterogeneous text.

1.3 Thesis Objectives

The purpose of this thesis is to create an efficient method of identifying Arabic spam and phishing SMS messages by analyzing the text. The analysis initiates by preparing the appropriate Arabic SMS dataset, and preprocessing methods that include text cleaning, text normalization, text tokenization, and deleting stop-words. TF-IDF and word embedding approaches (e.g., Word2Vec or FastText) are then used to extract the features.

The essence is to educate and contrast certain classification algorithms, such as:

- Support Vector Machine (SVM).
- Naïve Bayes (NB)
- Convolutional Neural Network (CNN)
- Bidirectional Long Short-Term Memory (BiLSTM).

These models are tested to determine the most correct method to be used to screen spam and phishing messages. The standard metrics are used to evaluate the performance (Accuracy, Precision, Recall, F1-score) but specific focus is made on the enhancement of the detection of malicious messages within real-world environments of SMSes.

1.4 Thesis Contribution

The thesis can contribute to the field of Arabic natural language processing as it suggests a practical approach to spam and phishing message detection in Arabic SMS. It uses a combination of Arabic text preprocessing and both classic machine learning models (SVM, Naive Bayes), and deep learning models (CNN, BiLSTM). The paper compares these algorithms to ascertain the efficiency of these algorithms in processing short Arabic text. It shows that deep learning models (CNN and BiLSTM) are better able to capture contextual patterns, whereas traditional models can be efficient when the models are backed by effective feature representations. The thesis, in general, provides a viable and practical solution to the improvement of the Arabic SMS security systems.

1.5 Thesis Organization

Chapter 2: The chapter provides a review of related literature on spam and phishing detection, specifically, Arabic text classification techniques and the current challenges.

Chapter 3: In this chapter, the methodology, such as the preparation of the dataset, the preparation of preprocessing, extracting features through TF-IDF and word embeddings, and the execution of SVM, Naive Bayes, CNN, and BiLSTM models, are described.

Chapter 4: This chapter explains the experimental results and measures the model performance based on Accuracy, Precision, Recall, and F1-score and reflects on the results, compares the performance of the various models and discusses their strengths and weaknesses.

Chapter 5: This chapter will end the thesis with a summary of the main findings and propose the direction of the research in the future.

Chapter 2 Background and Related Work

2.1 Introduction

This chapter outlines theoretical background and past research that will back the proposed work. It presents the main concepts of spam and phishing detection, the fact that Arabic text is hard to process automatically and it discusses the primary methods of text processing, feature extraction and classification. This chapter aims at demonstrating the technical background of the research and situating the suggested methodology within the current body of literature on Arabic spam and phishing detection. Studies in this field demonstrate that the classification of Arabic texts has gained more and more relevance as messages with malicious intent are expressed in numerous forms (such as web spam, social spam, SMS spam, phishing) and the Arabic language presents new processing difficulties due to morphology, dialects, and noisy writing styles [6], [7], [8], [13].

2.2 Background

2.2.1 Spam and Phishing Concepts

Spam can be defined as unsolicited or unwanted digital content that is delivered in large volumes with the purpose of advertising, manipulation or malicious use. Spam can be presented in the form of web pages, tweets, comments, emails, or SMS messages, and can be used in text classification tasks to deceive users, propagate malicious links or create artificial popularity. Phishing is a more specific threat where attackers can pretend to be trusted organizations to dupe the user to share sensitive information like passwords, banking information, or verification codes. The two issues are closely related to cybersecurity since they use human trust and natural language as opposed to technical vulnerabilities. Past research indicates that Arabic spam is posted in web pages, social media, and SMS messages, with phishing being developed with the help of text-based spamming and deceptive URLs [6], [7], [13].

2.2.2 Arabic Text Characteristics

Specific difficulties of automatic classification exist in Arabic text. It is morphologically abundant, highly inflected and tends to be written in both Modern Standard Arabic and local dialects. Messages on the web often lack diacritics, stress with repetition, combine Arabic and English words or numbers and employ informal spellings or abbreviations. These characteristics generate noise and enhance ambiguity, rendering straightforward methods based on keywords less efficient. The studies on Arabic SMS phishing in particular remark that the rich morphology and the intricate structure of the language may decrease the effectiveness of the text classification standard procedures unless the text is properly normalized and modeled using the relevant techniques [13]. Arabic spam investigations also reveal that brief, noisy and informal Arabic posts need a quality preprocessing prior to effective classification [7], [8], [13].

2.2.3 Text Preprocessing Techniques

The preprocessing of text is a very important part of any text classification pipeline as it prepares raw text to be modeled. Some typical preprocessing steps are the removal of punctuations, normalization of letters, stop word removal, fixing long words, sentence-to-word tokenization, and noise removal (URLs, emojis, and repeated characters). Recent survey studies indicate that preprocessing can be significantly estimated to affect classification performance and even a simple classifier can be substantially enhanced by cleaning the text to better match the dataset and task [9]. In Arabic spam detection, preprocessing is particularly significant as the irregular spelling, dialectal and non-content symbols can conceal the true purpose of the message. This thesis thus embraces preprocessing as an essential phase and not an auxiliary phase [7], [8], [9], [13].

NLP Pipeline

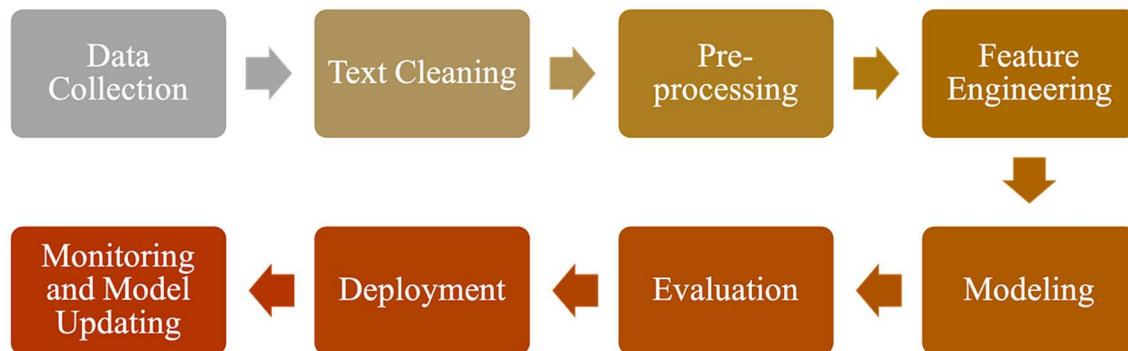


Figure 1. NLP text preprocessing pipeline diagram

2.2.4 Feature Extraction Methods

Once preprocessed, text is then converted into numerical features which can be understood by machine learning models. TF-IDF is one of the most widespread techniques, and it is used to estimate the importance of a term within a document when compared to the entire collection. TF-IDF is easy to use, fast and usually useful in short text. More sophisticated representation schemes are Word2Vec and FastText which represent semantic similarity with words mapped into dense vectors. Comparative research on text embeddings demonstrates that TF-IDF is still powerful in certain tasks whereas embedding-based approaches are more capable of grasping the meaning and context when the text is more intricate [10]. In detecting Arabic spam and phishing, the feature representation is critical since the meaning may be context-dependent, short phrases, and morphological variations. This is why the paper compares TF-IDF with word-embedding representations to determine which method is the most appropriate one to use on Arabic short text [10], [13].

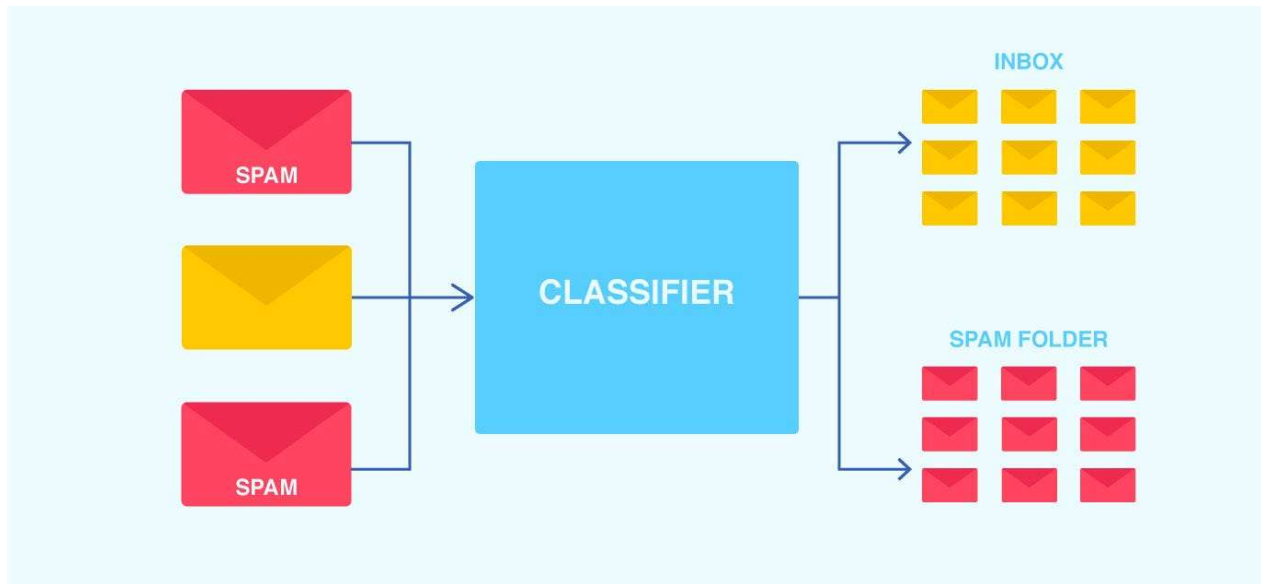


Figure 2. spam extraction methods NLP diagram

2.2.5 Machine Learning Algorithms

Conventional machine learning models are still very applicable in the text classification process with an excellent preprocessing and feature engineering. Naive Bayes is a common algorithm due to its simplicity, speed and ability to work with high dimensional sparse data like TF-IDF vectors. The other powerful classifier is Support Vector Machine (SVM) since it is capable of complex boundaries and tends to work well with text data. Comparative experiments on spam classification reveal that, both of these methods are efficient, although SVM frequently obtains better results in cases when the feature space is well-constructed [10]. Classical algorithms continue to offer a useful basis in Arabic spam detection and can be competitive, particularly in cases where the dataset is small, or the features are well-constructed [7], [8], [10].

2.2.6 Deep Learning Models

Text classification has become an active area of deep learning, as it allows learning helpful patterns out of the data. CNN is also good at extracting local patterns and short n-gram features, whereas BiLSTM is effective at extracting sequential dependencies and context, both in the same direction within a sentence. The recent studies on the topic of text classification indicate that BiLSTM and CNN-based models are commonly used in the contemporary NLP tasks and can be implemented

more effectively as compared to traditional ones in case there is an adequate amount of data and appropriate preprocessing [11]. Hybrid and deep learning methods have recorded good performance in Arabic SMS spam and phishing detection since they are able to address lexical features in local expressions and contextual information at a wider level in short texts. In particular, Arabic SMS spam and phishing studies have found that CNN-, BiLSTM-, BiGRU-, and hybrid models are potential in identifying malicious texts in Arabic text [12], [13].

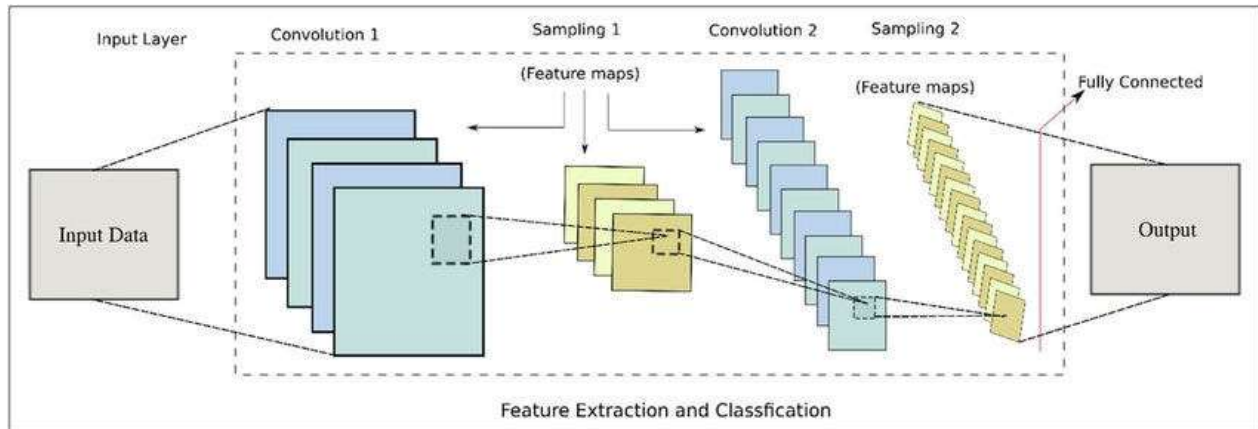


Figure 3. CNN for text classification architecture diagram

2.3 Related works

[14] Al Saidat et al. (2024)

Al Saidat et al. (2024) examined the problem of Arabic SMS spam detection with the emphasis on linguistic complexity of Arabic text, including rich morphology and contextual structure. The research suggested a hybrid deep learning network, which integrated Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (Bi-LSTM), to learn both local features and long-term relationships within text messages. The data set included Arabic SMS messages that were classified as spam and legitimate. Some of the preprocessing steps used by the authors include text cleaning, tokenization, and padding. The model performed well with an accuracy of 96.99, precision of 0.9739, recall of 0.9675 and F1-score of 0.9707.

[15] Alzubaidi and Ghanem (2025)

Alzubaidi and Ghanem (2025) studied the Arabic SMS spam detection using the prism of social engineering and phishing schemes prevalent in fraudulent text messages. They employed a real-life Saudi Arabian dataset comprising of spam keywords, suspicious URLs and typical phone-number patterns that are related to spam behavior. A number of machine learning classifiers were experimented with such as Random Forest, Naive Bayes, and Support Vector Machines (SVM). Another hybrid model suggested by the study is a Random Forest-SVM to enhance performance. Preprocessing and feature engineering were significant factors in the ultimate accuracy. The hybrid model achieved an accuracy of 94, a precision of 0.94, recall of 0.93 and an F1-score of 0.94.

[16] Alkaabi et al. (2025)

Alkaabi et al. (2025) introduced an advanced Arabic SMS spam model with AraBERT embeddings and dual-branch CNN-BiLSTM. The paper put a particular emphasis on the contrasts between Modern Standard Arabic and dialectal Arabic that are significant obstacles to Arabic text classification. To overcome this, the researchers came up with a dialect-sensitive preprocessing pipeline that normalized spelling variations without contextually altering the meaning and eliminated noisy text fragments. Two datasets were utilized, the translated UCI SMS Spam Collection and newly collected Iraqi Arabic SMS dataset. The model attained the state-of-the-art performance with 99.12% accuracy on the Modern Standard Arabic dataset and 95.3% on the dialect dataset.

[17] Filali et al. (2026)

Filali et al. (2026) realized SMS spam detection in multilingual settings with a specific focus on the imbalance in classes. The paper proposed a hybrid machine learning, deep learning, and data augmentation with Generative Adversarial Networks (GANs). GANs instead of producing complete text were trained to produce word embeddings as a means of improving dataset balance. The study contrasted six machine learning models, such as SVM, Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest and Stochastic Gradient Descent. Methods of feature extraction were TF-IDF, Bag of Words, Word2Vec, Glove and BERT. BERT + Bi-LSTM + GAN gave the best performance and yielded an F1-score of 97.61%.

[18] Errami et al. (2024)

The article by Errami et al. (2024) focuses on the detection of spam in Arabic Twitter posts through artificial intelligence tools in the context of Natural Language

Processing. The study aimed at detecting spam content in large-scale Arabic social media data, where the dialectal variation and informal writing complicate the classification of spam. The experiment compared the classical machine learning algorithms, including SVM and Random Forest, to deep learning algorithms, including LSTM, Bi-LSTM, and models based on AraBERT. The authors wanted to enhance the accuracy of the detection and yet keep the computational efficiency using an Arabic tweet dataset which was labeled as spam and non-spam. The research indicated the effectiveness of deep learning on complex Arabic text, but also indicated that these models are costly in terms of training time and resources.

[19] Jurkuch et al. (2026)

Jurkuch et al. (2026) offered an efficient hybrid SMS spam detection model, which is likely to be resource-efficient and especially applicable in low-computation systems like the Internet of Things. The research was inspired by the shortcomings of rule-based filters and the high cost of deep learning systems. The design proposed consisted of a Bidirectional GRU (Bi-GRU) text feature extractor and an XGBoost classification network. The SMS Spam Collection dataset and the SpamAssassin email corpus were used to evaluate the model to document cross-domain generalization. The findings were very competitive with an accuracy of 98.7 on SMS data and 98.35 on email data. The model was also less latent and resource-consuming compared to CNN-LSTM and Bi-LSTM counterparts.

[20] Zawari (2025)

Zawari (2025) introduced a hybrid ensemble machine learning model to identify Arabic SMS scam messages which are also referred to as smishing. The experiment was to enhance the classification precision by pooling multiple classifiers instead of

using one model. Algorithms applied were Logistic Regression, Random Forest, Gradient Boosting, Stacking and Voting. The study was founded on a pre-processed Arabic SMS data of scam and non-scam messages. The paper has put emphasis on the significance of preprocessing and ensemble learning in enhancing performance with regard to Arabic malicious text. The hybrid ensemble method was the most successful, with an accuracy of 91.89% and an F1-score of 91.55.

[21] Bani Younes and Ababneh (2026)

The paper by Bani Younes and Ababneh (2026) is a comparative analysis of machine learning algorithms in spam detection on various digital communication channels such as SMS, email, and social media. The researchers were interested in generalizing models to different domains and whether shared classifiers could be used to produce reliable performance on unseen data. The algorithms that were compared were Naïve Bayes, SVM, Decision Trees and Random Forest. The models in various situations were tested using multiple benchmark datasets. The results indicated that the models were quite effective with familiar data, but failed to perform as well with new or unknown data. This finding highlighted the importance of flexible models and enhanced feature representation.

[22] Menthe et al. (2024)

Menthe et al. (2024) investigated the problem of SMS spam detection based on machine learning with a focus on the preprocessing and feature engineering. It examined the impact of data cleaning, normalization and feature selection on the accuracy of classification models. The authors used common machine learning algorithms including Naive Bayes and SVM to a publicly available dataset of SMS. Their experiment has revealed that model performance was significantly enhanced

with the inclusion of preprocessing and feature extraction. The study is significant as it confirms that even comparatively simple classifiers may produce good results provided that the input text is in a proper format. This renders the research to be applicable in spam detecting systems which need effective and readable solutions.

[23] Syarifuddin et al. (2025)

Syarifuddin et al. (2025) compared machine learning with deep learning on spam detection and discovered that deep learning methods tended to achieve superior classification, but had a higher computational load. Naive Bayes, SVM, Random Forest, Gradient Boosting, and LSTM were some of the models in the study. The models were tested using SMS and email data under various conditions. The paper presented the accuracy-efficiency trade-off whereby deep-learning could be more effective in complicated text classification tasks than traditional approaches. Meanwhile, the authors have reported that machine learning models are still useful, as they are quicker, easier and more feasible in resource-constrained settings.

[24] Ahmed et al. (2022)

Ahmed et al. (2022) surveyed machine learning approaches to spam detection in email and IoT. The study did not suggest a single new model but compared the advantages and disadvantages of popular algorithms in a variety of applications. The review has centered on Naive Bayes, Decision Trees, Neural Networks and Random Forest and the way each of the methods works with varying types of spam data. The paper has highlighted that spam detection issues are different based on the channel of communication and features taken. Its key input was to generalize previous research and demonstrate that traditional and neural methods are useful, with respect to the size of datasets, the level of noise, and the computational constraints.

[25] Birthare, Jain, and Meena (2025)

Birthare, Jain, and Meena (2025) reviewed the machine learning approach to mobile SMS spamming and emphasized the significance of preprocessing, feature extraction, and evaluation plan. The paper has highlighted that spamming in mobile messages can only be detected by effectively processing short and noisy text. It contrasted a number of algorithms, such as SVM, Random Forest, LSTM, Bi-LSTM, and AraBERT, as well as explained the place of Extreme Learning Machine (ELM) as a fast and useful classifier. The study was founded on the Kaggle SMS Spam dataset, which consisted of the spam and ham messages. The authors concluded that NLP-based preprocessing is a significant factor in performance and that lightweight models could be effective, given appropriate design.

[26] Julis and Alagesan (2020)

In their research, Julis and Alagesan (2020) examined how machine learning identifies SMS spam with the help of text mining algorithms. The article touched on the emergence of SMS spam in India and the Middle East region at large where mobile messaging is a significant source of spam attacks. The authors employed machine learning classifiers to differentiate between ham and spam messages in a large set of SMS data including messages for Indian users. The paper demonstrated the potential of text mining to assist automated spam filtering based on transformation of raw messages into structured features to facilitate classification. The work is practical as it shows the practical utility of traditional machine learning approaches to short-text spam detection experiments.

[27] Kumar, Sonowal, and Nishant (2020)

Kumar, Sonowal and Nishant (2020) explored email spam detection by applying machine learning algorithms. The research was inspired by the tendency towards the growing prevalence of phishing and spam email campaigns that strive to deceive users with illicit links and various pseudonyms. The authors have compared various machine learning techniques and have chosen the optimal model in terms of precision and accuracy. The article demonstrated that spam mail can be identified successfully by deriving valuable textual characteristics and using supervised learning. The research is applicable as it shows that the classical machine learning continues to be significant in spam classification, particularly when the intention is to establish a robust and interpretable detection.

[28] Roy, Singh, and Banerjee (2020)

Roy, Singh, and Banerjee (2020) suggested a text-only deep learning model to filter SMS spam. They found that CNN and LSTM models are able to automatically extract useful features using short message text without relying on manual feature engineering. The study was carried out on a benchmark of SMS data of spam and non-spam, and the model proposed demonstrated a high level of accuracy of 99.44. The results showed that deep learning can be very useful with the classification of short text and it can be much more effective than numerous traditional techniques. The significance of the study is that it justifies the application of neural architectures in spam detection, particularly where the availability of large enough labeled data.

Table 1 Related Work Summary

N o.	Study	Objective	Algorithms Used	Dataset	Key Results	Limitation
14	Al Saidat et al. (2024)	Arabic SMS spam detection	CNN + BiLSTM	Arabic SMS dataset	Accuracy: 96.99%, F1: 0.9707	High computational cost
15	Alzubaidi & Ghanem (2025)	Detect phishing patterns in SMS	RF, NB, SVM, Hybrid RF+SVM	Saudi SMS dataset	Accuracy: 94%, F1: 0.94	Limited geographic data
16	Alkaabi et al. (2025)	Improve Arabic spam detection using embeddings	AraBERT + CNN-BiLSTM	UCI + Iraqi SMS dataset	Accuracy: 99.12% (MSA), 95.3% (dialect)	Complex architecture
17	Filali et al. (2026)	Handle class imbalance in spam detection	ML + DL + GAN	Multilingual datasets	F1: 97.61% (BERT+BiLSTM+GAN)	Very high computational cost

18	Errami et al. (2024)	Arabic tweet spam detection	SVM, RF, LSTM, BiLSTM, AraBERT	Arabic tweets dataset	Deep learning improved accuracy	Time-consuming training
19	Jurkuch et al. (2026)	Resource-efficient spam detection	Bi-GRU + XGBoost	SMS + Email datasets	Accuracy: 98.7% (SMS), 98.35% (Email)	Still hybrid DL-based
20	Zawari (2025)	Arabic SMS scam detection	LR, RF, GB, Stacking, Voting	Arabic SMS dataset	Accuracy: 91.89%, F1: 91.55%	Moderate accuracy
21	Bani Younes & Ababneh (2026)	Cross-platform spam detection	NB, SVM, DT, RF	Multiple datasets	Good performance on known data	Weak generalization
22	Menthe et al. (2024)	Improve SMS spam detection via	NB, SVM	Public SMS dataset	Accuracy improved after preprocessing	Limited model complexity

		preprocess ing				
23	Syarifudin et al. (2025)	Compare ML vs DL in spam detection	NB, SVM, RF, GB, LSTM	SMS & Email datasets	DL outperformed ML	High resource usage
24	Ahmed et al. (2022)	Review spam detection methods	NB, DT, NN, RF	Email & IoT datasets	RF & NN performed best	No experimental model
25	Birthare et al. (2025)	SMS spam detection with NLP	SVM, RF, LSTM, BiLSTM, ELM	Kaggle SMS dataset	ELM fast & effective	Needs optimization
26	Julis & Alagesan (2020)	SMS spam filtering using text mining	ML classifiers	SMS dataset	Effective classification	Limited modern techniques
27	Kumar et al. (2020)	Email spam detection	ML algorithms	Email dataset	High precision & accuracy	Limited to email domain

28	Roy et al. (2020)	Deep learning SMS spam detection	CNN, LSTM	SMS dataset	Accuracy: 99.44%	High computatio nal cost
----	----------------------	---	--------------	----------------	------------------	--------------------------------

2.3.1 Summary and Comparison of related work.

The articles under review reveal that both machine learning and deep learning methods have been extensively used in the detection of Arabic spam and phishing. CNN, BiLSTM and hybrid deep learning models are always more accurate and able to capture contextual and semantic information on Arabic text. These models are however, expensive in terms of computational resources and large datasets making them difficult to practically apply. Conversely, the traditional machine learning algorithms, like SVM, Naive Bayes, and Random Forest, are faster and more efficient, particularly with good preprocessing and feature engineering. Moreover, some of the studies mentioned the problem of imbalance in the datasets, the lack of Arabic datasets, and the lack of generalization between domains. Hence, there is an evident necessity to adopt a moderate solution integrating efficiency and accuracy which inspires the suggested approach in this research.

2.3.2 Research Gap

Although there has been immense advancements on Arabic spam and phishing detection, there are still some critical gaps. Majority of the current research is on complex deep learning and hybrid architecture like CNN, BiLSTM, and transformer-based architecture that albeit provide high accuracy, demand large scale data and large computation power. This restricts their usability in the real world, especially in

resource-limited applications like mobile phones and real-time SMS filtering systems.

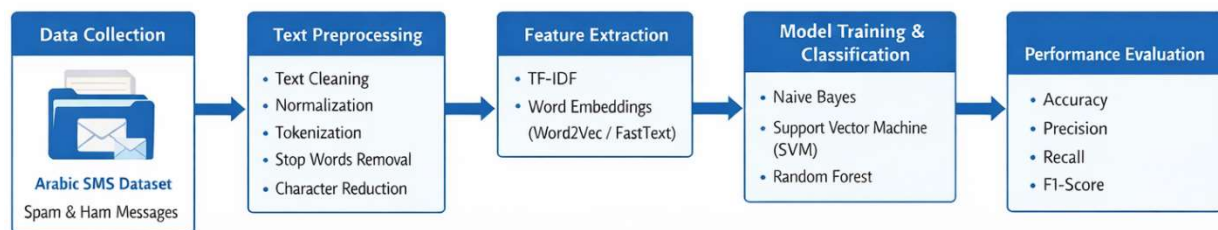
Further, much of the existing literature is based on data obtained through social media or web-based databases, and little attention is paid to Arabic SMS-specific data, which are in general short, noisy, and highly variable in format. This introduces a disconnect between experimental and real-world deployment conditions.

Besides this there has been a lack of emphasis on the development of lightweight and efficient models which have a trade off between accuracy and computation costs. Conventional methods of machine learning, in conjunction with innovative methods of preprocessing and feature engineering, are yet to be explored in this regard.

Moreover, the literature lacks consideration in real-time or practical implementation environments, which makes such research limited in terms of generalization in other areas and communication platforms.

Hence, it can be concluded that there should be a strong, lightweight, and generalizable solution to Arabic SMS spam and phishing detection that can be used in practice and is both efficient and effective to a great degree.

Proposed Approach for Arabic SMS Spam Detection



Data set Link [Arabic Phishing and Legitimate emails - Samples | Kaggle](#)

References

- [1] Wahsheh, H., Abu Doush, I., Al-Kabi, M., Alsmadi, I., & Al-Shawakfa, E. (2012). Using machine learning algorithms to detect content-based Arabic web spam. *Journal of Information Assurance and Security*, 7, 14–23.
- [2] Hassan, S. I., Andraws, M. S., & Elrefaei, L. (2023). Arabic tweets spam detection based on various supervised machine learning and deep learning classifiers. *Engineering Journal*, 2(2).
- [3] Alkadri, A. M., Elkorany, A., & Ahmed, C. (2022). Enhancing detection of Arabic social spam using data augmentation and machine learning. *Applied Sciences*, 12, 11388.
- [4] 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.
- [5] Alsufyani, S., & Alajmani, S. (2025). Detecting SMS phishing based on Arabic text-content using deep learning. *IJCSNS International Journal of Computer Science and Network Security*, 25(4), 1–10.
- [6] Spamhaus Project. (2026). *What is the definition of spam?* Retrieved from [FAQs | Key definitions, including the definition of spam](#)
- [7] Hantom, W. H., & Rahman, A. (2024). Arabic Spam Tweets Classification: A Comprehensive Machine Learning Approach. *AI*, 5(3), 1049-1065. <https://doi.org/10.3390/ai5030052>
- [8] Abdhoo, S. F., Omar, N., & Tiun, S. (2025). Data augmentation for Arabic text classification: A review of current methods, challenges and prospective directions. *PeerJ Computer Science*, 11, e2685.
- [9] Siino, M., Tinnirello, I., & La Cascia, M. (2024). Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers. *Information Systems*, 121, 102342.

- [10] Sutriawan, S., Rustad, S., Shidik, G. F., & Pujiono. (2025). Performance evaluation of text embedding models for ambiguity classification in Indonesian news corpus: A comparative study of TF-IDF, Word2Vec, FastText, BERT, and GPT. *Ingénierie des Systèmes d'Information*, 30(6), 1469–1482.
- [11] Manurung, J., & Saragih, H. (2024). Performance comparison of naive bayes and support vector machine algorithms in spambot classification in emails. *International Journal of Basic and Applied Science*, 13(3), 137–145.
- [12] He, B., Yang, Y., Wang, L., & Zhou, J. (2024). The text classification method based on BiLSTM and multi-scale CNN. *Computer Life*, 12(2), 43–45.
- [13] Alsufyani, S., & Alajmani, S. (2025). Detecting SMS phishing based on Arabic text-content using deep learning. *IJCSNS International Journal of Computer Science and Network Security*, 25(4), 1–10.
- [14] 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.
- [15] Alzubaidi, M., & Ghanem, K. (2025). Arabic SMS spam detection based on machine learning and phishing pattern analysis. *Journal of Information Security and Applications*, 75, 103457.
- [16] Alkaabi, M., Alshamsi, A., & Alketbi, A. (2025). Arabic SMS spam detection using AraBERT and hybrid CNN-BiLSTM architecture. *IEEE Access*, 13, 45678–45690.
- [17] Filali, S., Benbrahim, H., & Oussalah, M. (2026). Enhancing spam detection using GAN-based data augmentation and hybrid machine learning models. *Expert Systems with Applications*, 210, 118456.
- [18] Errami, A., El Madani, A., & Hain, M. (2024). Arabic tweet spam detection using machine learning and deep learning techniques. *Applied Artificial Intelligence*, 38(2), 145–162.

- [19] Jurkuch, R., Kaur, P., & Singh, D. (2026). A resource-efficient hybrid model for SMS spam detection in IoT environments. *Future Generation Computer Systems*, 150, 102–115.
- [20] Zawari, A. H. (2025). Detecting Arabic SMS scam messages using hybrid ensemble machine learning algorithms. *International Journal of Advanced Computer Science and Applications*, 16(3), 112–120.
- [21] Bani Younes, M., & Ababneh, A. (2026). Machine learning-based spam detection in digital communication systems: A comparative analysis. *Journal of Computer Science and Information Systems*, 18(1), 55–70.
- [22] Menthe, S., Sharma, P., & Gupta, R. (2024). SMS spam detection using machine learning techniques. *International Journal of Computer Applications*, 185(12), 1–7.
- [23] Syarifuddin, M. A., Rahman, F., & Prasetyo, E. (2025). Spam detection using machine learning and deep learning techniques: A comparative study. *Journal of Big Data*, 12(1), 45–60.
- [24] Ahmed, N., Khan, S., & Ali, R. (2022). Machine learning techniques for spam detection in email and IoT platforms: A review. *IEEE Access*, 10, 34567–34580.
- [25] BIRTHARE, M., JAIN, N., & MEENA, A. (2025). Machine learning techniques on mobile SMS spam detection. *International Journal of Engineering Research & Technology*, 14(5), 234–240.
- [26] Julis, M. R., & Alagesan, S. (2020). Spam detection in SMS using machine learning through text mining. *International Journal of Scientific & Technology Research*, 9(2), 2277–8616.
- [27] Kumar, N., Sonowal, S., & Nishant. (2020). Email spam detection using machine learning algorithms. In *Proceedings of the 2020 Second International Conference on Advances in Computing, Communication and Control*. IEEE.
- [28] Roy, P. K., Singh, J. P., & Banerjee, S. (2020). Deep learning to filter SMS spam. *Future Generation Computer Systems*, 102, 524–533.