SMS SPAM DETECTION AND CLASSIFICATION TO COMBAT ABUSE IN TELEPHONE NETWORK USING NATURAL LANGUAGE PROCESSING

ABSTRACT

In the modern era, mobile phones have become ubiquitous, and Short Message Service (SMS) has grown to become a multi-million-dollar service due to the widespread adoption of mobile devices and the millions of people who use SMS daily. However, SMS spam has also become a pervasive problem that endangers users' privacy and security through phishing and fraud. Despite numerous spam filtering techniques, there is still a need for a more effective solution to address this problem[1].This research addresses the pervasive issue of SMS spam, which poses threats to users' privacy and security. Despite existing spam filtering techniques, the high false-positive rate persists as a challenge. The study introduces a novel approach utilizing Natural Language Processing (NLP) and machine learning models, particularly BERT (Bidirectional Encoder Representations from Transformers), for SMS spam detection and classification. Data preprocessing techniques, such as stop words removal and tokenization, are applied, along with feature extraction using BERT.

Machine learning models, including SVM, Logistic Regression, Naive Bayes, Gradient Boosting, and Random Forest, are integrated with BERT for differentiating spam from ham messages. Evaluation results revealed that Naïve Bayes classifier + BERT model achieves the highest accuracy at 97.31% with fastest execution time of 0.3 seconds on the test dataset. This approach demonstrates a notable enhancement in spam detection efficiency and low false-positive rate.

The developed model presents a valuable solution to combat SMS spam, ensuring faster and accurate detection. This model not only safeguards users' privacy but also assists network providers in effectively identifying and blocking SMS spam messages.

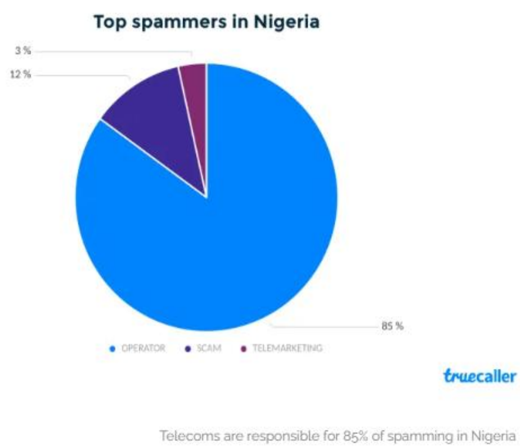
Keywords: SMS Spam, Logistic Regression, Natural Language Processing, Naïve Bayes, Machine Learning, SVM Classifier, BERT model, Gradient Boosting, Random Forest Classifier.

1 INTRODUCTION

Mobile phones are now considered to be a kind of devoted companion to user. The widespread adoption of mobile phones, particularly for SMS communication, has become an integral part of modern life. SMS, a valuable service in the telecommunications industry, contributes significantly to the Gross National Income (GNI) of developing countries [2]. This facility is used by millions of users daily due to its simplicity, accessibility, instant delivery, and its low-price rates as compared to calls. However, the ubiquity of SMS has also led to an increase in unwanted spam messages, including advertisements and scams.Nigeria, in particular, faces a significant SMS spam problem which has endangered mobile users' privacy with phishing and fraud. Nigeria is ranked 3rd among top 10 Africa countries affected by SMS spam according to report by Truecaller. According to the report, an average mobile user in Nigeria received 35 spam messages per month. *Figure 1*shows the statistics of SMS spam received by Truecaller users in Africa and *Figure.2* shows the statistics of the category of SMS spammers in Nigeria.

Spam detection has traditionally relied on keyword filters to differentiate between spam and legitimate messages for the past two decades [3]. More recently, advanced methods like Statistical Learning Theory, Artificial Neural Networks (ANNs), and Support Vector Machines (SVM) have emerged. However, these newer techniques exhibit inconsistent performance across different training datasets without logical or apparent explanation.There are numerous spam filtering techniques, however, because each of these techniques has strengths and drawbacks, no single spam filtering strategy can be guaranteed to be 100% effective at eradicating spam issues. The application of text mining techniques to SMS will improve the effectiveness of detecting and classifying spam messages [4], which will reduce telephone network abuse. There are now a staggering number of different forms of SMS, a different technique that can effectively classify SMS with low latency must be proposed.

This study proposed the use of contextualized embedding, an NLP approach, to enhance SMS spam detection and classification. Contextualized embedding represents words based not only on their static definitions but also on the context in which they appear. By applying BERT to create contextual sentence embeddings for SMS messages, combined with machine learning algorithms like Naive Bayes, Random Forest, Gradient Boosting, Logical Regression, and SVM, this research aims to improve efficiency and accuracy in identifying spam messages in lowest possible time.



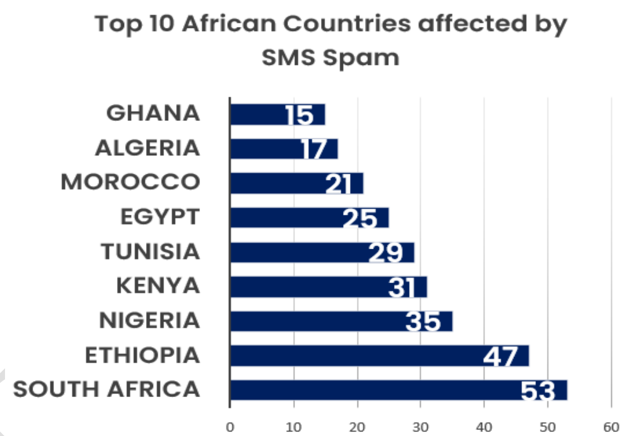


Figure 1Truecaller Average spam sms per user/month in Africa (https://pmnewsnigeria.com/1029/03_02)

Figure 2Category and Percentage (%) of SMS Spammers in Nigeria (www.technext24.com/2019/12/04/truecaller-report)

2 Problem Statement

SMS is a text messaging platform that allows users to exchange short text messages using standardized communication protocols, [5]. These messages are succinct, filled with abbreviations, and less formal, posing a challenge in spam classification. Service providers show less concern about SMS spam due to its low occurrence and their use of SMS for promotions.

In Nigeria, there's an existing solution called Do Not Disturb (DND) offered by network providers through the Nigeria Communication Commission, but it's not entirely effective because it can also block legitimate messages. Moreover, the unavailability of large SMS spam datasets used for training and testing spam detection models has led to issues of overgeneralization and overfitting, reducing the accuracy of classification predictions [6].

Given the limited character length of SMS and the evolving strategies employed by spammers, a key question emerges: Are the current models sufficiently effective and efficient in quickly distinguishingbetween SMS spam and legitimate messages? Addressing this question is pivotal in enhancing the accuracy and efficiency of SMS spam detection and classification.

3 Research Objectives

1. To pre-processing the dataset using Natural Language Processing techniques using BERT (Bidirectional Encoder Representations from Transformers) model.
2. To perform feature extraction and selection using Document Frequency Matrix and BERT preprocessor.
3. To perform vectorization in order to convert the processed dataset to numerical data.
4. To evaluate and compare the outcomes of the utilized machine learning models, which are Naive Bayes, Random Forest, Gradient Boosting, Logical Regression and SVM.

4Scope and Limitation(s) of Study

The study aims to develop an effective SMS spam detection and classification that leverages machine learning algorithms to combat abuse in the telephone network. It involves gathering a diverse dataset of SMS messages from various sources such as Kaggle, UCI and Data Science Nigeria and self-sourced data, preprocessing the datasets it to remove noise and extract relevant features. These features encompass text-based and metadata-based attributes.

To identify spam or legitimate messages, the study employs machine learning algorithms, including fine-tuning the pre-trained BERT model for SMS data dataset to improve its contextual understanding of sentences of the SMS domain and generate highly relevant contextualized sentence embeddings. Machine language algorithms such as Naive Bayes, Random Forest, Gradient Boosting, Logical Regression, and Support Vector Machine are used, and their performances are compared using evaluation metrics

One challenge is the dataset's class imbalance, with far fewer spam records than legitimate ones, requiring down-sampling technique to the ham class, ensuring parity between ham and spam records. The limited dataset for training due to the down-sampling poses a potential performance challenge and it might affect model training as the dataset size increases. Additionally, the study is limited to the English language due to dataset composition, even though it can identify some pidgin words. The models may not effectively recognize slangs or abbreviations, potentially affecting results when non-English words are used.

5Literature Review

5.1Related works

There have been several studies about SMS spam detection and classification in recent years. Researchers have proposed various techniques ranging from machine learning tohybrid approachestodetect and classify spam messages. Keyword Filtering is most common but this approach has limitations, as spammers easily modify words they use or intentionally misspell to evade detection.

5.1.1Machine learning approaches

In the study done by (Baaqeel, Hind, and Rachid Zagrouba, 2020), they tested several supervised machine learning models including an unsupervised model K-means, Decision Trees, Naïve Bayes, SVM, and K-Nearest Neighbor on the UCI Machine Learning Repository dataset. They then developed a hybrid model that combines unsupervised and supervised algorithms to classify SMS messages as ham or spam. Among the combinations tested (Kmeans-SVM, Kmeans-NB, and Kmeans-LR), the Kmeans-SVM combination achieved the best accuracy of 98.8%. They did not explore other combinations of unsupervised and supervised algorithms to create a hybrid model.

[7] proposed a message topic model for spam SMS detection in 2021. Utilizing the KNN algorithm to address message sparsity, the model considered symbol terms and background terms, outperforming the standard LDA model. [7] undertook a similar investigation employing two datasets: one sourced from UCI machine learning, mirroring Kaggle's corpus, featuring 5,574 spam and ham messages; the other encompassing 2,000 messages of both types. Leveraging TF-IDF matrices, a comprehensive suite of machine learning algorithms, including CNN, Naive Bayes, SVM, Random Forest, ANN, and Decision Tree, were applied to the two datasets. CNN achieved state-of-the-art results with 99.10% accuracy, followed by Naive Bayes and SVM.

From the above research, it becomes evident that traditional algorithms like Naive Bayes and SVM consistently outperform other methods. Moreover, considering message length as an additional feature could potentially contribute to the model's overall performance.

5.1.2Deep learning approaches

[8]adopted a deep learning approach called BiLSTM (Bidirectional Long Short-Term Memory), which is a type of Recurrent Neural Network (RNN), for detecting spam messages. The outcome of their study indicate that this model performed better than alternative machine learning algorithms, including BayesNet, J48, Naïve Bayes, SVM, K-nearest neighbor, and decision tree. Notably, the model attains an accuracy rate of 98.6%. They used the Word2vec algorithm for word embedding. However, the model had some limitations such as high preprocessing time due to the manual removal of unstandardized abbreviations.Deep learning techniques have shown promising results in SMS spam detection and classification. [9]titled their paper, "Deep learning to filter SMS Spam". They utilized Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM), the models were based solely on text data only and achieved a high accuracyof 99.44%.

5.1.3Ensemble learning andFeature engineering approaches

Ensemble learning involves combining multiple models to improve accuracy and reduce overfitting. This techniques have also been applied to the domain of SMS spam detection and classification. [10] titled their research "Enhancing Spam Message Classification and Detection Using Transformer-Based Embedding and Ensemble Learning". In this work, the authors used an ensemble learning approach for SMS spam detection by combining four machine learning models into one model.The model performed better than its separate constituent parts achieving a high accuracy of 99.91%.

Feature engineering techniques have been applied to improve the performance of SMS spam detection and classification models. [11]introduced an approach rooted in feature engineering, leveraging semantic analysis to extract features from SMS messages. They created a dictionary using the TF-IDF Vectorizer algorithm, whichincludes all the features of words of a spam SMS. The system classifies SMS referring to this dictionary and based on the message content.

6METHODOLOGY

6.1Research Design and Approach

The research design and approach for this study uses a cross-sectional research method that involves collecting and analyzing data at a specific period of time. The research design is structured into: data collection, data preprocessing, feature extraction, and classification. The study aims to develop a model for SMS classification and detection that can be used to combat abuse in a telephone network.

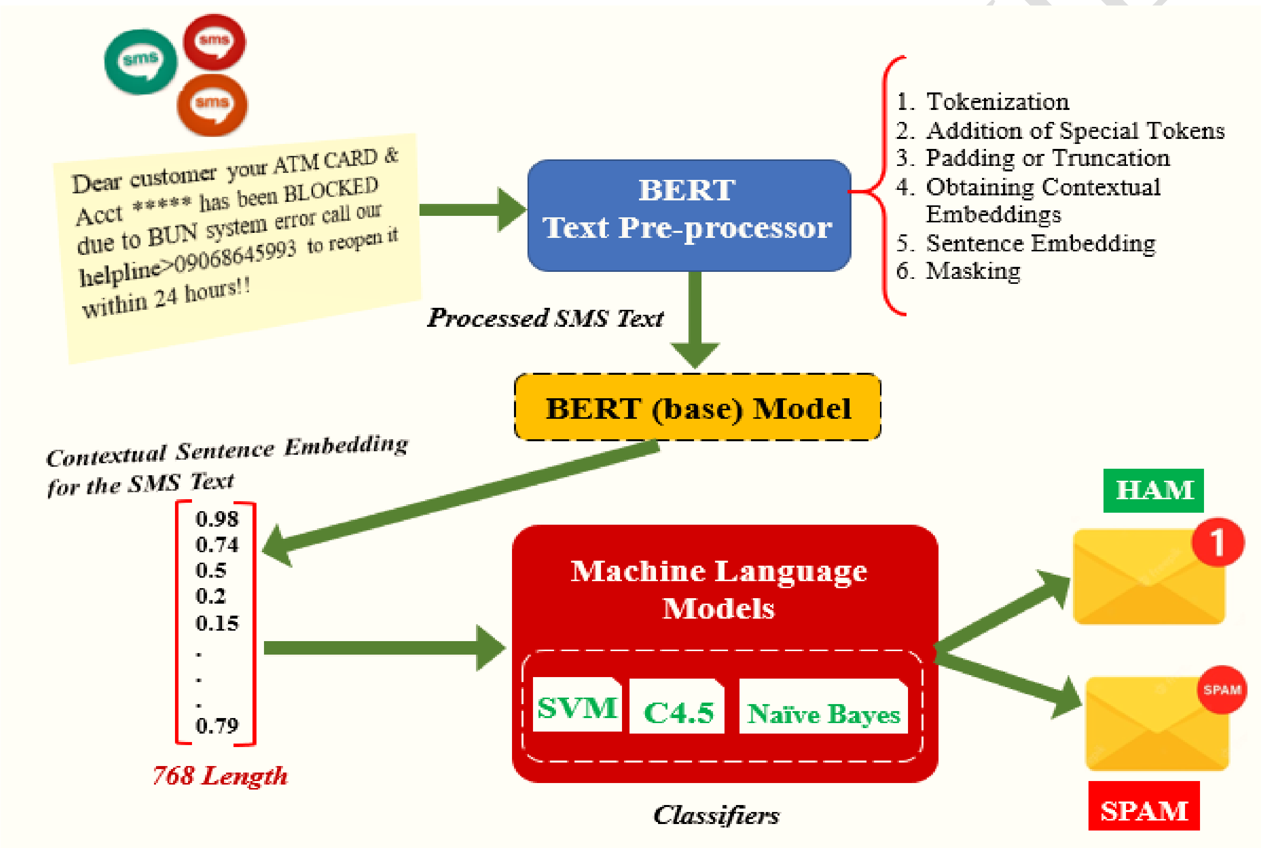


Figure 3 Proposed Methodology for the study

6.1.1Cross-sectional Study

For this research, a cross-sectional study was used to collect data on SMS messages from various online repository. A large number of mobile users in Nigeria fall victim of SMS spam during COVID-19 period (2020/21) compared to SMS spam data gathered from previous years. The spam messages data collected represent a snapshot of the SMS messages at that specific point in time, which was used to detect and classify spam messages in the telephone network.The advantage of a cross-sectional study is that it makes it possible for the collection of data from a large population sample at a specific point in time. This approach allows the testing of hypotheses and the drawing of statistical inferences, providing insights into the effectiveness of SMS spam detection and classification systems.

6.1.2 Data Collection phase

At this primitive level, data is gathered locally andglobally from numerous sources to create a respectable dataset of spam and ham text messages, which will be utilized as the model's input (SMS messages). The spam dataset used for this research contained a total of 6,986 rows.

1. Kaggle: The dataset contains 5,572 rows of SMS classified into ham (nonspam) and spam.
2. Data Science Nigeria (DSN): the dataset contains 1,141 text messages. The dataset comprises of fraudulent messages in the financial and labor sectors received in any location in the country.
3. Self-Data: Google form to collect local spam messages mobile user received (275 spam messages)

The main reason for choosing the dataset is to combined them together by randomly sampling from different resources to gives exposure to different scenarios of ham and spam SMSs.

6.1.3 Data Cleaning and Preprocessing

The three datasets were cleaned to have uniform classes and data before we perform preprocessing on the SMS spam dataset utilizing in python and its libraries such as NLTK library and others. Preprocessing is the initial stage of turning the collected dataset (unstructured data) to more structured data and converting of the input text into numerical representations that can be used for various downstream natural language processing (NLP) tasks. Forthe SMS Spam dataset used in this study, preprocessing was done to remove unnecessary elements that doesn't contribute to the classification task such as stop words, special characters, and punctuations before applying the classification models. Stemming was also be performed to minimize the word count within the messages and converting the text into lowercase to avoid case sensitivity issues, while still preserving their meaning.

BERT model preprocessor and encoder aided the conversion of thedataset into numerical format used for the various downstream natural language processing tasks in this study.The preprocessor prepared the input data for the BERT model by performing tokenization, adding special tokenssuch as [CLS] and [SEP],segmenting the input data into separate sequences, masking of the tokens, and padding the input sequence to a fixed length. Figure 4 and 5 shows how the pre-trained BERT model process text data.

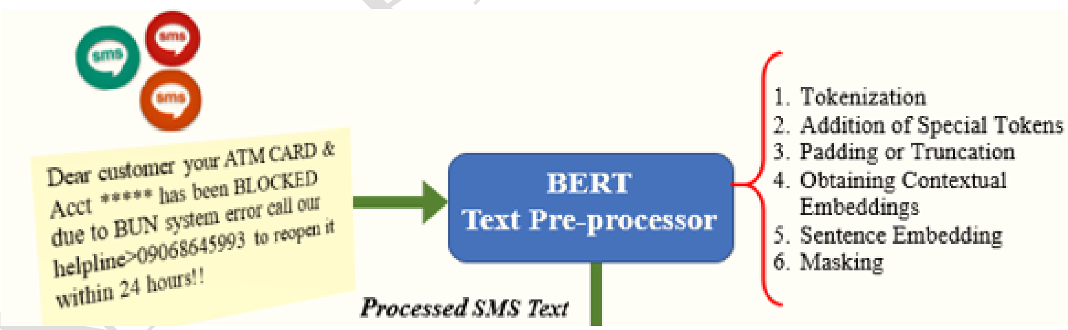


Figure 4 BERT Model Text Pre-Processor

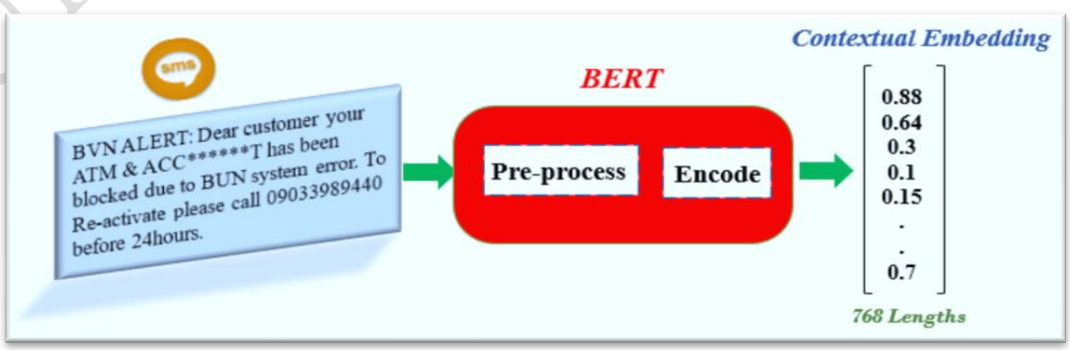


Figure 5 Conversion of Text Data to Contextual Embedding in BERT Model

The encoder processes the preprocessed data to generate contextualized word embeddings. The encoder is made up of multiple transformer layers,that capture the contextual relationships between the words in the input text. The contextualized embeddings are then passed to the next transformer layer until the final layer generates a fixedsize vector representation for the input data.

6.1.4 Feature Extraction and Selection

Feature extraction and selection are important to balance the overfitting and overgeneralization of ham and spam in SMS messages classification. The featuresextracted from the preprocessed data was doneusing both text-based and metadata-based approaches. The text-based features include word frequency and part-of-speech tags, which capture the linguistic characteristics of spam messages while the metadata-based features include sender information and time of message, which provide additional information about the likelihood of a message being spam.

In order to represent the SMS messages as numerical vectors for classification,the pre-trained BERT model was utilized to extract contextualized word embeddings for each token in the tokenized SMS messages. BERT generates these embeddings by considering the entire sentence's context, capturing the nuanced meaning of words based on their surroundings. These embeddings contain rich semantic and syntactic information, enabling better representation of words in the SMS messages. The dataset was split into: 4540 rows for training, 1396 rows for testing and 1050 rows for validation. The training dataset was used to train the BERT model for the spam detection task, involving the fine-tuning of model parameters and weights. Subword tokenization was performed by the pre-trained BERT model to handle vocabulary's variability and capture the meaning of rare or complex words in the messages, especially local languages such as Pidgin English and others. Test dataset was used for model evaluation.

6.1.5 Classification

To classify SMS messages as either ham or spam, both rule-based filtering and machine learning-based classification methodswere utilized. The rule-based filter was applied to the preprocessed data to quickly discard obvious spam messages. This filter was based on a set of rules defined using the characteristics of known spam messages while the remaining messages were classified using machine learning-based classification. During this phase, machine learning classifiers such as: Logistic Regression, Naive Bayes, Random Forest, Gradient Boosting, and SVMin conjunction with BERT model were utilized.

7 Results and Discussion

7.1 Vectorization

To run machine learning algorithms, text files must beconvertedinto numerical feature vectors, this process is known as Vectorization [6]. Machine learning models operate exclusively on numerical data, to achieve this, a matrix was created to contains words and its frequency of occurrence. The two techniquesused for the creation of the document term matrixin this study areBag of words (BoW) and Term frequency-inverse document frequency (TF-IDF).

1. Bag of words (BoW):Bag of Words is a way of extracting the features from the set of text messages[12]. Figure 7 shows a matrix called bagOfWordswhichdescribes the text based on the frequency of the word appearing in the documentand was implemented using python’s packageCountVectorizer.

2. Term frequency-inverse document frequency (TF-IDF): TF-IDF matrix helps in understanding the importance of the word in the corpus of documents [13]. To implement, the TF-IDFVectorizer() function was used with the n-gram as a unigram. Figure 7 shows the matrix of converted dataset to numerical data used for the model. The machine learning models were implemented after the creation of the matrix.

	Text_Length	1	2	3	4	5	6	...	3615	3616	3617	3618	3619	3620	3621
0	29	0	0	0	0	0	0	...	0	0	0	0	0	0	0
1	101	0	0	0	0	0	0	...	0	1	0	0	0	0	0
2	141	0	0	0	0	0	0	...	0	0	0	0	0	0	0
3	59	0	0	0	0	0	0	...	0	0	0	0	0	0	0
4	31	0	0	0	0	0	0	...	0	0	0	0	0	0	0

Figure6

Generated Bag of Words Matrix for Spam dataset

	Text_Length	1	2	3	4	5	...	3616	3617	3618	3619	3620	3621
0	29.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0
1	101.0	0.0	0.0	0.0	0.0	0.0	...	0.286945	0.0	0.0	0.0	0.0	0.0
2	141.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0
3	59.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0
4	31.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0

Figure 7 Generated TF-IDF Matrix for Spam dataset

7.2 Comparing Research Result

It is essential to strike a balance between minimizing false positives (to avoid inconveniencing users) and false negatives in SMS spam classification (to catch all spam) is essential. Table 1 shows the comparison of the result of machine learningmodelsused in this study.

Table 1 Comparison of utilized Models

	ML Classifiers	Training Accuracy	Training Execution Time	Testing Accuracy	Testing Execution Time	Recall	Precision	FP	FN
1	SVM+BERT	93.78%	8 seconds	94.35%	1 seconds	95%	98%	47	16
2	Logical Regression+BERT	95.29%	2 seconds	96.14%	0.8 seconds	97%	99%	29	14
3	Gradient Boosting+BERT	95.75%	3 minutes 9seconds	96.5%	1 minutes 28 seconds	98%	98%	22	17
4	**Naives Bayes+BERT**	**96.83%**	**1 seconds**	**97.31%**	**0.3 seconds**	**99%**	**98%**	**13**	**17**
5	Random Forest+BERT	92.31%	16minutes1 1 seconds	92.55%	14 minutes 4 seconds	93%	98%	63	20

From Table 1, the best performed model in term of accuracy, low latency, precision, and recall is Naives Bayes + BERT model. The model has an accuracy of 97.31%, precision of 98%, recall of 99%, epoch time of 1 seconds on the training dataset and inference time of 0.3 seconds on test dataset. The model also has the lowest less total count of False-Positive (FP) and False-Negative (FN): 30 compared to other models used in the study.

7.3 Confusion Matrix of the Models

Confusion matrix visualizes and summarizes the performance of a classification algorithm[14]. In classification problems, reviewing the confusion matrices of the models used is crucial for assessing their performance. Table 2 shows the confusion matrices for all the models in this study.

Table 2 Confusion matrix of utilized Models

	ML Classifiers	Confusion Matrix	Predicted Spam SMS Message	Predicted Non-Spam SMS Message
1	SVM + BERT	Truly Spam	255	16
		Truly Non-Spam	47	1078
2	Logical Regression + BERT	Truly Spam	257	14
		Truly Non-Spam	29	1096
3	Gradient Boosting + BERT	Truly Spam	254	17
		Truly Non-Spam	22	1103
4	**Naives Bayes + BERT**	**Truly Spam**	**261**	**13**
		Truly Non-Spam	**17**	**1105**
5	Random Forest + BERT	Truly Spam	251	20
		Truly Non-Spam	63	1062

7.4 Comparison between Developed and Existing Models

Table 3 shows the comparison of the existing models and the proposed model.

Table 3 Comparison of Existing models with Proposed Model

References	Method	Model	Accuracy	Comments
S. Gupta, Saha, and Das[14]	TF-IDF vectorization algorithm	Did not state the specific machine learning model utilized	96.5%	The authors employed and assessed using only one vectorization technique.
Abayomi- Alli, Olusola, Sanjay Misra, and Adebayo Abayomi- Alli[8].	Word Embedding (Word2vec Algorithm)	BiLSTM (Bidirectional Long Short-Term Memory)	98.6%	Very high preprocessing time

Baaqeel, Hind, and Rachid Zagrouba[4]	Word Tokenization	Kmeans-SVM	98.8%	Did not investigate alternative combinations of unsupervised and supervised algorithms for building the hybrid model.
Current Research	Contextual Sentence Embedding	Naive Bayes + BERT	97.3%	Class imbalance issuerectified, detect local languages (Pidgin), less execution time

8 Conclusion

This study aims to design aproficient machine learning model capable of identifying spam SMS, thus addressing the issue of telephone network abuse prevalent in Nigeria. In addition to this primary goal, a supplementary contribution of this research involves juxtaposing the outcomes of fully developed models against existing counterparts using evaluation metrics and visual representations. These dual contributions are poised to significantly benefit mobile phone users by expeditiously mitigating the challenge of spam SMS, consequently streamlining its detection process. Natural language processing approachwas used to achieve this objective, with a particular emphasis on understanding the semantics of the text. In this study, a hybrid model,Naïve Bayesand BERT model achieved an impressive accuracy rate of 97.3%, precision of 98% and recall of 99% with execution time of 0.3seconds.All these indicate that the model is highly effective in identifying ham messages correctly while also being able to distinguish spam messages in less time.Looking ahead, extending the scope to encompass non-English languages presents an intriguing prospect for advancing spam SMS detection in upcoming research.

Looking ahead, there is significant potential for this study to be expanded. In the future, we can focus on collecting new latest data and testing the proposed model.To enhance the efficacy of the results, an avenue for improvement involves training machine learning models using localized datasets sourced from diverse countries, while also considering datasets with substantial records. This approach holds the potential to bolster the dependability of the predictive models. Furthermore, extending the scope to encompass non-English languages presents an intriguing prospect for advancing spam SMS detection in upcoming research endeavors.

9 References

[1] N. Mageshkumar, A. Vijayaraj, N. Arunpriya, and A. Sangeetha, “Efficient spam filtering through intelligent text modification detection using machine learning,” *Mater. Today Proc.*, vol. 64, pp. 848–858, 2022, doi: https://doi.org/10.1016/j.matpr.2022.05.364.

[2] S. Agarwal, S. Kaur, and S. Garhwal, *SMS spam detection for Indian messages*. 2015. doi: 10.1109/NGCT.2015.7375198.

[3] N. Hussain, H. Mirza, and I. Hussain, “Detecting Spam Review through Spammer’s Behavior Analysis,” *ADCAIJ Adv. Distrib. Comput. Artif. Intell. J.*, vol. 8, p. 61, Mar. 2019, doi: 10.14201/ADCAIJ2019826171.

[4] D. Saraswathi and D. Sowmya, “SMS Spam Classification Using PSO-C4.5,” *Lect. Notes Electr. Eng.*, vol. 967, pp. 41–47, 2023, doi: 10.1007/978-981-19-7169-3_4.

[5] N. Jindal and B. Liu, “Analyzing and Detecting Review Spam,” *Seventh IEEE Int. Conf. Data Min. (ICDM 2007)*, pp. 547–552, 2007, [Online]. Available: https://api.semanticscholar.org/CorpusID:12335888

[6] M. Rubin Julis and S. Alagesan, “Spam detection in sms using machine learning through text mining,” *Int. J. Sci. Technol. Res.*, vol. 9, no. 2, pp. 498–503, 2020.

[7] S. Gupta, S. Saha, and S. K. Das, “SMS Spam Detection Using Machine Learning,” *J. Phys. Conf. Ser.*, vol. 1797, p. 12017, Feb. 2021, doi: 10.1088/1742-6596/1797/1/012017.

[8] O. Abayomi-Alli, S. Misra, and A. Abayomi-Alli, “A deep learning method for automatic SMS spam classification: Performance of learning algorithms on indigenous dataset,” *Concurr. Comput. Pract. Exp.*, vol. 34, no. 17, p. e6989, 2022, doi: https://doi.org/10.1002/cpe.6989.

[9] P. K. Roy, J. P. Singh, and S. Banerjee, “Deep learning to filter SMS Spam,” *Futur. Gener. Comput. Syst.*, vol. 102, pp. 524–533, 2020, doi: https://doi.org/10.1016/j.future.2019.09.001.

[10] J. Ma, Y. Zhang, J. Liu, K. Yu, and X. Wang, *Intelligent SMS Spam Filtering Using Topic Model*. 2016. doi: 10.1109/INCoS.2016.47.

[11] S. DasGupta, S. Saha, and S. K. Das, “SMS spam detection using machine learning,” *J. Phys. Conf. Ser.*, vol. 1797, no. 1, 2021, doi: 10.1088/1742-6596/1797/1/012017.

[12] K. Yadav, S. K. Saha, P. Kumaraguru, and R. Kumra, “Take Control of Your SMSes: Designing an Usable Spam SMS Filtering System,” in *Proceedings of the 2012 IEEE 13th International Conference on Mobile Data Management (Mdm 2012)*, in MDM ’12. USA: IEEE Computer Society, 2012, pp. 352–355. doi: 10.1109/MDM.2012.54.

[13] A. Basu, C. R. Watters, and M. A. Shepherd, “Support vector machines for text categorization,” *36th Annu. Hawaii Int. Conf. Syst. Sci. 2003. Proc.*, p. 7 pp.-, 2003, [Online]. Available: https://api.semanticscholar.org/CorpusID:14913198

[14] S. Ballı and O. Karasoy, “Development of content-based SMS classification application by using Word2Vec-based feature extraction,” *IET Softw.*, vol. 13, no. 4, pp. 295–304, 2019, doi: 10.1049/iet-sen.2018.5046.