Development of Short Message Service Spam Filtering Using Naive Bayes Algorithm

Mmuo Grace C[1]., Akpado Kenneth. A[2]., Okafor Chukwunenye S[3]., Obiora-Dimson Ifeyinwa[4], Onuzulike Vincent C[5].

[1,2,3,4,5]*(Department of Electronics & Computer Engineering / Nnamdi Azikiwe University, Awka, Nigeria)*

Abstract

Short Message Service (SMS) is a text message service available in smart-phones as well as basic phones. The daily increase in the number of mobile device users drastically increases SMS traffic which in turns increases the number of spam messages. Spam is a very serious universal problem that causes a lot of damages. Several studies have been presented, including implementations of spam filters to prevent spam from reaching their destination. Mobile spam is a growing problem that keeps increasing day by day. The complexity of the messages that spammers impose has made it harder to classify spam. The difficulty in sifting through SMS spam is that messages are typically short, with phrases made up of acronyms and abbreviations, which makes them more ambiguous. In this paper, a supervised machine learning algorithm called Naïve Bayes Algorithm is used as an effective technique in SMS spam filtering. The dataset used for training and testing models was downloaded from UCI machine learning repository that contains 5574 English raw text messages; where 4827 are ham messages and 747 are spam messages. The results obtained show accuracy score of 98.7%, precision 96.1%, recall 94.5% and F1-score 95.3%. The experimental results have shown that, Naïve Bayes Machine learning model performs better than other supervised machine learning models during the training and testing of SMS spam detection and filtering.

Keywords: Spam, Short Message Service (SMS) Machine Learning, Naive Bayes Algorithm

Introduction

Short Message Service (SMS) is one of the most ubiquitous communication services presently. It is standardized in the GSM mobile communication systems, that is to say, it is a popular means of mobile communication. SMS can be sent and received simultaneously with GSM voice, text and image. Smart phones have become common devices during the past few years, incorporated with multiple wireless networking technologies to support additional functionality and services. Although it was created as a component of the Global System for Mobile Communications (GSM), it is currently accessible over many different networks, including Code Division Multiple Access (CDMA) (Le Bodic, 2005). Mobile phone users with SMS enabled devices had reached 6.1 billion users by the year 2015 (Ezpeleta, 2017). As the popularity of smart phones surged, frequent users of text messaging began to see an increase in the number of commercial advertisements, spams being sent to their telephones through text messaging. Recently, there has been a dramatic increment in the volume of SMS spam (Mobile SMS Marketing, 2010).

Any unsolicited, bulk-sent digital communication that is not requested is referred to as spam. When a communication is sent without the recipient's express consent, it is said to be unsolicited. When a message is sent in bulk, it indicates that it is a part of a larger collection of communications with nearly identical content (Forbes, 2002). Spam is an issue about consent of the recipient and not necessarily message content. A message is therefore said to be spam only if it is both unsolicited and bulk. Spam is commonly used as a medium for advertising products and services. According to Spamaus (2021), ninety-eight percent of all spam messages sent are a form of products or services advertisement which would apparently seem harmless while the remaining two percent pose the threat to the internet community and results in great losses. Spam includes, phishing mails that steal users' login information and identities, messages from internet fraudsters that mislead internet users to pay money in exchange for great riches or reward which are non-existent, mal-spam that intend to make users download malicious web content attached to the messages.

The complexity of the messages that spammers impose has made it harder to classify spam. As a result, many techniques have been created to filter spam. The spam generally has an important economic impact to end users and service providers. The importance of increasing of this problem has motivated the development of a set of techniques to fight it (Mobile SMS Marketing, 2010). Because consumers view every SMS they get, SMS spam directly affects users, which is why it has a greater negative impact on users than email spam. There are several strategies that can be employed to prevent spam. One significant and well-liked method of preventing spam is filtering. It is the automatic categorization of messages into SMS that are not spam and those that are. The problem with filtering SMS spam is that brief messages frequently contain only a few words, and occasionally these words are made up of idioms and acronyms (Guzella et al., 2009). In this paper, an anti-spam filtering technique called Naive Bayes Algorithm; based on Artificial Intelligent System (AIS) is used to reduce the effect of spam in text messages.

Literature Review

Because of its great scalability, spam filtering has long been a problem of interest to scholars worldwide. This section evaluated previous research that has been done using various methodologies and technology. This procedure brings to light concerns that have not yet been addressed in the examined works as well as the comparative differences between the reviewed works and the generated research study.

Table 1: Summary of the Literature Review

AUTHORS	TITLE	METHODS	RESULTS
Anchal Ora (2019)	Detection in Short Message Service using Natural Language Processing and Machine Learning Technique.	Natural Language Processing (NLP) and Machine Learning (SVM, Random forest and Bernoulli Naïve Bayes) techniques.	The result focused on a sample dataset was used to find an effective solution on Spam Detection in Short Message Service Using Natural Language Processing and Machine Learning Techniques.
Dilip Singh et.al (2020).	Automated SMS Classification and Spam Analysis using Topic Modeling.	Latent Dirichlet Analysis (LDA) and Non-negative Matrix Factorization (NMF).	The research developed an automated framework for filtering SMS spam with the use of various classifiers using modeling approaching like LDA and NMF
Gomathan et.al (2020)	Mobile SMS Spam Filter Techniques using Machine Learning Techniques.	K-Neighbor classifier, Decision tree classifier, Random forest classifier, SVM classifier and Logistic regression.	The result emphasized more on the comparison between the machine learning algorithms to predict messages and calculate the accuracy criterion.
Kanza and Hamid (2021)	Detection of SMS Spam and Filtering using Data Mining Methods	Deep learning technique Data-mining.	The result concentrated on presenting the literature review of the machine and deep learning techniques used in the detection, classification, and spam filtering for SMS spam.
Arulprakash and Jansi (2022)	Eshort Message Service Spam Detection and	Decision tree classifier, K-Nearest Neighbor classifier, Support	The result discussed the design of a universal SMS spam filter that is efficient and

	Filtering using Machine Learning Approach.	Vector Machine classifier, Naïve Bayes classifier.	user-friendly, taking into account user preferences and distinguishing useful messages operators.
Aju Omojokun and Adedeji Joy (2022)	An Email Spam Filtering Model using Ensemble of Machine Learning Techniques.	Decision tree, Support Vector Machine and Multilayer Perception Technique.	The result dwelled more on the development of spam email filtering model using Ensemble of Decision Tree, Support Vector Machine and Multilayer Perception Technique as a solution approach rather than using Naïve Bayes Machine algorithm approach.
Sankar et.al (2023)	SMS Spam Detection using Machine Learning.	Machine Learning models (SVM, NBA) and Deep Learning models (LSTM, DNN, RNN).	The work developed a method for improving SMS spam filtering performance by combining two of data-mining task association and bracket.
Vaman Saeed (2023)	A Method for SMS Spam Message Detection using Machine Learning.	Machine Learning: J48, KNN DT.	The work developed a technique using supervised machine learning algorithms in identifying spams.

Research Gap

Based on the several academics' previously mentioned contributions, it is clear that this thesis will advance knowledge by introducing the best supervised machine learning algorithm for SMS spam filtering - the Naïve Bayes Algorithm. Anchal Ora (2019) found an efficient solution for spam detection in short message services using the Bernoulli Naïve Bayes algorithm. Multinomial naïve bayes algorithm was used by Arulprakash and Jansi (2022) to address the construction of an efficient and user-friendly universal SMS spam filter that takes into consideration user preferences and distinguishes useful message operators; however, the dataset was limited to European languages. The development of this work was done using a multinomial naïve bayes method. Multinomial Naïve Bayes Algorithm is the real type of naïve bayes algorithm utilized in this research work; it is an optimized strategy used to filter out more accurate SMS. The dataset is an internationally recognized English raw set that was obtained from the UCI machine learning library. When Naïve Bayes Algorithm is compared to other supervised machine learning reviewed, it is used to train and predict messages more quickly.

Materials and Methods

The materials used in this research work are grouped into hardware and software requirements. The hardware requirement includes: standard Desk top computer or Personal Computer (PC) having at least the following properties: high speed processor**,** 1.30 GHz operating frequency**,** 64-Bit system type, 4GB RAM and40 GB hard disk. The software requirements include online installation of Google Collaborator, python IDE, Microsoft document, and other platform for the execution of programs. Google Collaborator is useful in performing tasks and tracking the progress throughout the development activity.

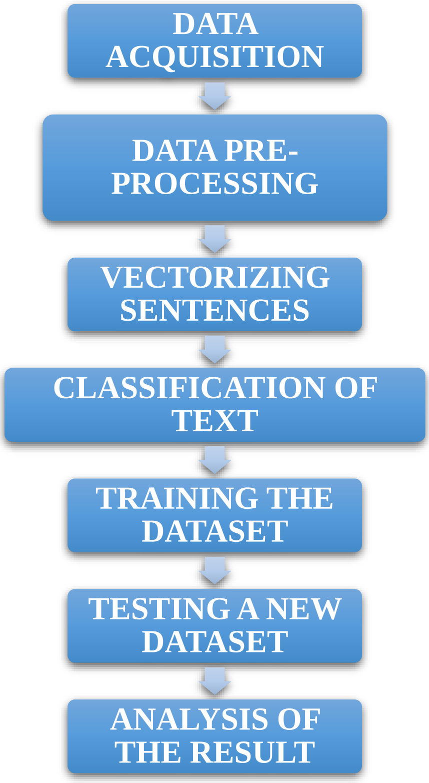

Figure 1: Block Diagram of Development of SMS Spam Filtering using Naive Bayes Algorithm

Data Acquisition

The SMS Spam collection data set; one of the largest and publicly used SMS spam datasets is used in this project to train and detect spam messages. The first step is to download the dataset from the UCI Machine Learning Repository. The dataset considered in the current research is available on Kaggle, a machine learning repository. This study finds that there are only 5,574 labelled messages in the dataset; with 4827 of messages belonging to ham messages while the other 747 messages belong to spam messages as presented in the table below.

Table 2: A Depiction of a Taken Dataset

Messages	Amount	Percentage (%)
Hams	4827	86.6
Spams	747	13.4
Total	5574	100

Data Pre-processing

Data pre-processing is mainly cleaning of data by removing unwanted rows, columns, missing values, etc. The purpose of pre-processing is to convert the downloaded raw data into a form that fits machine learning because, structured and clean data allows us to get more precise results from an applied machine learning model. This involves data formatting and data cleaning.

Vectorizing Sentences

This step has to do with feature extraction and selection of ham and spam in SMS text messages. The process of removing stop words such as “the”, “to”, “or”, “a” “an”, “in”, “with”, etc from the document during training the model is known as vectorization of sentence. Those words removed are said to be ineffective for communicating important information and little or nothing to the comprehension of the text which is the process that convert the input data (collection of text documents) from its raw format (text) into vector of real numbers which is the format that machine learning (ML) model support.

Features Extraction: This is the process of extracting features from messages and it is very important on which the accuracy of the machine learning algorithms depends. The accurate analysis of SMS dataset and the extracted features leads to increase the probability detection of spam message. There are two types of SMS messages namely the "ham" message and the "spam" message. There are numerous ways to distinguish between spam and ham. Extracting and well features from each message help in filtering SMS messages efficiently.

Featurization**:** Featurization is a process by which some form of data such as text data, graph data, time-series data etc. is changed or converted into numerical vectors. In Featurization; data do not need to be in the form of numerical vector. Direct processing of row text data is not possible for the machine learning model. But in the end, machine learning models work with numerical features. So, it is important to change some type of data into numerical vector so that we can leverage the whole power of linear algebra (making the decision boundary between data points) and statistics tools with other types of data.

Text Classification

Text classification is all about predicting and discovering the category of given data points. Categories are sometimes called labels, goals or classes. The predictive modelling of a classification is the task of approximating a mapping function (f) from the input variables (X) to separated output variables (y). As example that can be defined as a classification problem is detecting of spam message and this classification is binary because there are only two classes which are spam (1) and ham (0).

Training Data stages

Data from datasets is what machine learning algorithms use to learn. They examine the data for trends, get an understanding of it, use it to inform their judgments, and assess how accurate those decisions were. As a result, the machine learning model is trained using training data to learn how to make predictions or carry out a specified activity. It is usually marked, indicating that each data point's output from the model is known. Prior to producing predictions, the model needs to become adept at identifying patterns within the data.

Testing Data stages

In this stage, once the machine learning model has been trained on the dataset, there is need to test the model using unseen data in order to evaluate its performance. This unseen data is called the testing data. Thus, testing data is used to evaluate the machine learning model’s performance. It is typically different from training data and not labelled. This means the model’s output is unknown for each data point. On the testing data, the model's predictive accuracy is assessed. In machine learning, testing data is used to ensure that the model works for the given testing data. The testing data should meet two criteria:

i. It should represent the actual dataset that the model will be used on. This implies that the feature distribution in the testing data should match that of the original dataset.
ii. Testing data should be large enough to generate meaningful predictions. This means the testing data should be large enough to provide a statistically test of the model’s performance.

Therefore, the testing data should be new, “unseen” data model that has not been seen before. This is because the model must have already learned the patterns in the training data, so testing it is the ability to generalize to a new data in essential.

Result Analysis stage

Result analysis displays information about the quality of the model in SMS spam filtering system. Different processes and measures were taken during the training and they are confusion matrix and performance metrics. They are used to measure the performances of the developed model.

a. Confusion Matrix

The confusion matrix is a metric that often used to measure the performance of a classification algorithm. True to its name, the terminology related to the confusion matrix can be rather confusing, but the matrix itself is simple to understand. In other words, confusion matrix is the most commonly used evaluation metrics in predictive analysis mainly because it is very easy to understand and it can be used to compute other essential metrics such as accuracy, recall, precision, etc. It is an “[N x N]” matrix that describes the overall performance of a model when used on dataset, where N is the number of class labels in the classification problem.

b. Performance Metrics

In order to evaluate or determine the accuracy of the SMS spam filtering techniques, there is need to apply performance metrics to the model. These are referred to as assessment metrics or performance metrics. The performance metrics help us to know how the model has performed for the given data. The performance metrics are called precision, Recall, and Accuracy as shown below:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{Total}}$$

Where;

True Positive (TP): is defined as amount of samples that are properly classified.

True Negative (TN): is defined as amount of samples that are properly rejected from the class.

False Positive (FP): is defined as amount of samples that are wrongly rejected from the class.

False Negative (FN): is defined as amount of samples wrongly classified to the correct class.

Total (T): represents "True Positive + False Positive + False Negative + True Negative" (i.e. **TP + FP + FN + TN**).

RESULT

The performance of the text classification model; Naive Bayes Algorithm is evaluated on the downloaded test dataset. The key metrics considered were accuracy, precision, recall, and F1-score respectively.

Table 3: Model Performances for Naive Bayes (NB) Supervised Machine Learning Algorithm

S/N	Metrics	Predicted Score (training model output) in Percentage	Model's Performances
1.0	Scikit learning NB	98.7%	Accuracy score
2.0	Confusion-metric	94.5%	Sensitivity
3.0	Confusion-metric	99.4%	Specificity
4.0	Confusion-metric	96.1%	Precision
5.0	Confusion-metric	94.5%	Recall
6.0	Confusion-metric	95.3%	F1-score

From the table above, the results of SMS spam detection using Naive Bayes algorithm was evaluated using metrics such as accuracy, precision, recall, and F1 score. The Naive Bayes algorithm correctly classified 98.7% of the SMS messages in the test set. It achieved a precision of 96.1%, which means that, of all the messages that were classified as spam, 96.1% were actually spam. The recall of 94.5% indicates that, of all the spam messages in the test set, 94.5% were correctly identified by the Naive Bayes algorithm. The F1 score of 95.3% is the harmonic mean of precision and recall and provides a balanced measure of the algorithm's performance. Overall, these results demonstrate that the Naive Bayes algorithm is the most effective in detecting SMS spam compare to other supervised machine learning algorithms discussed in this work. However, it's important to note that the performance of the algorithm can vary depending on the quality and representativeness of the dataset, the choice of pre-processing and feature extraction techniques. Naïve Bayes (NB) is versatile, supporting both binary and multinomial classification tasks. It shows impressive performance, especially in applications involving short texts, such as tweets. In certain scenarios, particularly with the aid of feature selection, NB can outperform more complex classifiers.

Receiver Operating Characteristic (ROC)

An ROC curve, or receiver operating characteristic curve, is like a graph that shows how well a classification model performs. It helps us see how the model makes decisions at different levels of certainty. The curve has two lines:

1. It shows how often the model correctly identifies positive cases (true positives) and
2. How often it mistakenly identifies negative cases as positive (false positives).

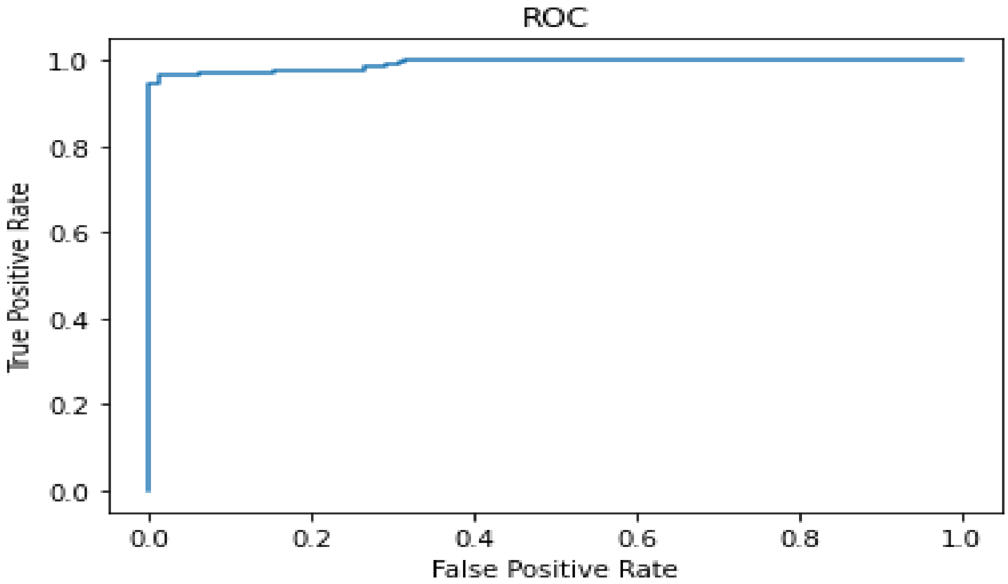

Figure 2: ROC Graph of Naïve Bayes Algorithm

From the figure above, the ROC curve shows the trade-off between TPR (sensitivity) and FPR (specificity). Classifiers that give curves closer to the top-left corner indicate a better performance. On other words, the closer the ROC is to the upper left corner of the graph (as indicated with arrow), the higher the accuracy model's performances, this is because, in the upper left corner, the sensitivity = 1 while false positive rate (FPR) = 0 (specificity = 1). Hence, the ideal ROC curve has its area under curve (AUC = 1).

Comparison of Developed Model with Existing Models

In this section, the developed model (NBA) is compared with the existing models that are discussed in the literature review of this work. The researcher Anchal Ora (2019) achieved an accuracy of 96.5% using Bernoulli Naïve Bayes algorithm. Arulprakash and Jansi (2022) obtained an accuracy of 95%. The table below shows the comparison of developed model with existing models.

Table 4: Comparison of Developed Model with Existing Models

References	Algorithms	Accuracy	Comments
Anchal Ora (2019)	Natural Language Processing (NLP) and Bernoulli Naïve Bayes (BNB)	96.5%	The researcher used another type of Naïve Bayes algorithm which is different from the one used in the design of this work. The Bernoulli Naïve Bayes algorithm is used in combination with Light GBM and TF-IDF which are NLP algorithm to achieve an overall accuracy of 96.5%.
Arulprakash and Jansi (2022)	Multinomial Naïve Bayes (MNB)	95%	The researchers used on data-sets sourced from European countries and European language only for its dataset. This makes it more difficult to be adopted as a robust and

			universal method of sms spam filtering technique.
Current Thesis	Multinomial Naïve Bayes (MNB)	98.7%	A larger dataset downloaded from UCI machine learning repository was used in the design of this work. The dataset is universally accepted and contains 5574 messages. This model is used to train and predict messages at a faster rate and a higher accurate result was obtained.

Conclusion

In conclusion, the SMS spam filter project has shown that machine learning algorithms can effectively solve the problem of spam filtering and the implementation of such solutions can be made simple with the use of advance python and its libraries in Google collaborator platform. The Naive Bayes algorithm correctly classified 98.7% of the SMS messages (accuracy) in the test set. It achieved a precision of 96.1%, which means that, of all the messages that were classified as spam, 96.1% were actually spam. The recall of 94.5% indicates that, of all the spam messages in the test set, 94.5% were correctly identified by the Naive Bayes algorithm. The F1 score of 95.3% is the harmonic mean of precision and recall and provides a balanced measure of the algorithm's performance.

The Naive Bayes algorithm is a popular and effective machine learning algorithm for SMS spam detection and filtering. It works well for text classification tasks because it can handle high-dimensional feature spaces and noisy data. The algorithm is relatively simple and fast to train, making it suitable for real-time SMS spam detection applications. To achieve optimal performance with the Naive Bayes algorithm, it's important to carefully pre-process and extract features from the text data, tune the hyper-parameters of the algorithm, and evaluate its performance using appropriate metrics. Additionally, it's essential to use a representative and diverse dataset that reflects the target population of SMS messages. Overall, Naive Bayes algorithm provides a practical and reliable approach for SMS spam detection and filtering that can be deployed in various real-world scenarios.

References

Aju Gabriel O. and Adedeji Joy (2022). *An Email Spam Filtering Model using Ensemble of Machine Learning Techniques.* International Journal of Computer Applications Technology and Research. Vol. 11, Issue No.: 03, pp. 66-71.

Anchal Ora 201). *Spam Detection in Short Message Service using Natural Language Processing and Machine Learning Techniques.* National College of Ireland Project Submission Sheet, School of Computing.

Arulprakash M. and Jansi K. R. (2021). *Eshort Message Service Spam Detection and Filtering using Machine Learning Approach.* Turkish Journal of Computer and Mathematics Education. Vol. 12, No. , pp. 721-727.

Dilip Singh, Shreya, Mahapatra and Aripita Sharma (2020). *Automated SMS Classification and Spam Analysis using Topic Modeling.* Retrieved from https://ieeexplore.ieee.org/document/9170710.

Ezpeleta E. (2017). *Short Messages Spam Filtering Combining Personality Recognition and Sentiment Analysis.* International Journal Uncertain Fuzziness Knowledge-Based System, pp. 175-189.

Forbes S. (2002). *Web of Deception: Mis-information on the Internet Information Today Inc. C.S. (2018) Security report.*

Gomathan Sai, Pradeepini G., Vadderswaram and Guntur (2020). *Mobile SMS Spam Filter Techniques using Machine Learning Techniques.* International Journal of Scientific and Technology Research. Volume 9, ISS: 3, pp. 384-389.

Guzella T. S. and Camihas W. M. (2009). *A Review of Machine Learning Approaches to Spam Filtering.* Elsevier, Expert Systems with Applications: 10206-10222.

Kanza, Hamif and Hamid Ghous (2021). *Detection of SMS Spam and Filtering using Data Mining Methods.* International Research Journal of Modernization in Engineering, Technology and Science. Vol. 3, e-ISSN: 2582-5208.

Lebodic G. (2005). *Mobile Messaging Technologies and Services SMS, EMS and MMS.* 2nd ed. John Wiley and Sons Ltd.

Mobile SMS Marketing (2010). Available at: https//www.mobilesmsmarketing.com/live-examples.php

Sankar E., Shekhar Babu and Tridev M. (2023). *SMS Spam Detection using Machine Learning.* Indian Scientific Journal of Research in Engineering and Management. Vol. 07, Issue 04. Doi: 10.55041/IJSREM 18832.

Spamhaus (2021). The Spamhaus Project- the Definition of Spam. Available at: https://www.spamhaus.org

Vaman Ashqi Saeed (2023). A Method for SMS Spam Message Detection using Machine Learning. Artificial Intelligence and Robotics Development, Journal. Vol. 3, pp. 214-228. Issue: Doi: https://doi.org/10.52098/airdj.202366.