Exploring Machine Learning Classifier Models for Spam Detection

# Abstract

This report presents a comparative analysis of three machine learning classifier models for spam detection in Short Messaging Service (SMS) messages. We trained and tested K-Nearest Neighbors (KNN), Decision Trees (DT)/ Random Forests (RF), and Naïve Bayes (NB) on a publicly available Kaggle dataset. Our methods discuss dataset preparation using natural language processing (NLP) preprocessing and tokenization techniques, feature engineering with regex, feature extraction using CountVectorizer, and our model selection process. Each model was tested and compared using various performance metrics including accuracy, confusion matrices, precision, recall, and F1-scores. These results are assessed using visualizations such as ROC, Precision-Recall Curve, and Bar Plots. The results show that all three models performed well, with Naive Bayes having the highest accuracy of 98.5%, followed closely by Decision Tree/Random Forest with 98.2%, and K-Nearest Neighbors with 97.3%. Our results display the effectiveness of various machine learning algorithms for spam detection and provide insights into the strengths and weaknesses of each approach, with hopes to help guide future research in this area.

Keywords: Spam, Ham, Naïve Bayes, Decision Trees, Random Forest, K Nearest Neighbors

# Introduction

## Background

The rapid advancement of technology has transformed global communication, providing numerous benefits such as instant messaging across the globe through email or SMS. However, this evolution of messaging technology has also resulted in an increase in spam and malware, which is often abused by marketers or threat actors to send unsolicited digital communication or to harm computer systems. More specifically, spam is unwanted electronic correspondence that is sent out in bulk, while malware refers to any software that intends to harm computer systems or gain unauthorized access to sensitive information. Examples of malware include viruses, worms, and other code-based entities that infect a host with malicious intent. Threat actors leverage spam methods to send phishing attempts via mass emails or Short Messaging Service (SMS) messages, aiming to acquire personally identifiable information (PII), credit card details, and other unauthorized access to computer systems. To combat these threats on a global scale, the field of computer science and artificial intelligence has leveraged a sophisticated tool: Machine Learning. Machine Learning (ML) was created to help process large amounts of data generated by individuals, many of which are too vast for human processing.

## Understanding Machine Learning

The ability of ML algorithms to automatically learn from data, identify patterns, and make predictions “without being explicitly programmed” [[MIT](#_https://mitsloan.mit.edu/ideas-made)], is what makes it an effective solution for managing and utilizing large amounts of data. In other words, instead of being given specific instructions for what to do with the data, ML algorithms are designed to learn from patterns in the data itself. Thus, by giving these algorithms large datasets to analyze, they can identify underlying patterns and make predictions based on new, unseen data. This adaptability is particularly important in the context of spam and malware, where ML algorithms can be trained to recognize patterns and identify potential threats, such as the aforementioned phishing attempts via mass emails or SMS messages.

## Machine Learning in Cyber Security

In addition to its effectiveness in detecting and preventing spam and malware, the ability to learn and adapt to new data has made ML an invaluable tool for cyber security as a whole. The constantly evolving nature of cyber threats requires a dynamic and flexible solution, which traditional rule-based systems struggle to provide. Historically, rule-based systems have been the primary method of detecting threats; however, these methods often struggle to identify new and emerging threats as they rely on pre-defined rule sets that can be easily exploited. On the other hand, ML algorithms can create new rules as necessary and adapt to new and emerging threats. As new threats are constantly emerging, ML enables cyber security experts to stay ahead of the curve and better protect individuals and organizations from the growing threat of cyber attacks.

Spam messages are a great example of these cyber attacks as they pose a significant threat to cyber security, with many malicious intentions such as illegally accessing confidential data such as passwords and other important identifiers, or spreading malicious software with harmful links or attachments [[Detection](#_Detection)]. SMS messages are no exception to the threat of spam, and detecting such messages in this form of communication is crucial for protecting individuals and organizations. Due to the limitations of rule-based systems, they have traditionally struggled in the past to detect spam and other cyber threats, as they rely on suspicious patterns for detection, such as misspelled words, malicious links, and other anomalies. This is where machine learning algorithms come into play, as they can adapt and create new rules as necessary to stay ahead of the constantly evolving nature of cyber threats. Thus, it becomes trivial for an attacker to modify message content in hopes of circumventing rules, as the rules are constantly updating.

## Research Objective and Scope

Our study aims to make a valuable contribution to the discussion on the most effective machine learning algorithm for identifying and filtering out spam messages in SMS communication. To achieve this, we will explore different methods of detecting spam using an open-source dataset from Kaggle and the Python programming language for preprocessing, training, testing, and analysis. Specifically, we will focus on three primary models, including Naïve Bayes, Decision Trees/Random Forests, and K-Nearest Neighbors (KNN), and provide a comprehensive comparison of their performance metrics to determine the best approach for accurate classification. Through our rigorous testing and analysis, we aim to contribute to the ongoing efforts to improve cyber security and help prevent spam messages from causing harm.

# Related Work

[todo]

# Methods

Before we begin training our models, we much first preprocess the data to ensure consistency and compatibility. We will procedurally analyze and modify our dataset to increase quality and resulting analytic scores, while avoiding over-generalization and over-fitting. This transformed dataset will then be trained and tested across each model individually.

## Introduce Dataset

We have selected the publicly available Kaggle dataset, “Spam (or) Ham,” to train and test our spam classification models [[SpamVHam](#_[SpamVHam])]. It is important to note that this dataset is a condensed version of The University of California, Irvine’s (UCI) ‘SMS Spam Collection Dataset’ [[OGSMS](#_[OGSMS])]. The original message collection was consolidated from various public sources, including 425 spam messages from Grumbletext, 3,375 ham messages from NUS SMS Corpus (NSC), 450 ham messages from Caroline Tag's PhD Thesis, and 1,002 ham and 322 spam messages from the SMS Spam Corpus v.0.1 Big [[OGSMS](#_[OGSMS])]. The version we are using was selected due to accessibility and minor pre-processing, as two messages were formatted incorrectly and resulted in missing values. As such, our version contains 5572 Short Message Service (SMS) messages, 5169 of which are unique. Each message is split between two columns: ‘Class’ which identifies whether it is spam or ham, and ‘sms’ which contains the plain-text version of the message.

## Data Analysis

To prepare our data for use in our ML algorithms, we first needed to review and analyze a series of properties and decide which, if any, transformations were required. We began by importing the comma-separated values (CSV) data file using the python library, Pandas. Pandas is an open-source python library which is commonly used in machine learning as it helps organize, manipulate, and analyze complex tabular data. Our process begins through a series of steps, including observing general information, checking for extraneous null values, viewing various observations, and analyzing statistical information.

We started by viewing the general and statistical information to confirm that each column, ‘Class’ and ‘sms,’ contained 5572 non-null entries. As shown in Figure 1, each column not only contained this expected value, but the column values for the ‘count’ row were equivalent. If this were not true, that would mean we had incomplete or missing data which would need to be trimmed or modeled accordingly. Similarly, we needed to check if there were any unnecessary null values anywhere in our dataset which could further negatively impact the ML algorithms. These are important steps as they are examples of noise which can cause overfitting and potentially “result in more complex models that miss the true pattern” [[MLR](#_[MLR])].

The statistical information in Figure 1 also shows the number of possible unique observations, the most common value, and the frequency of the most common value. We use this information to determine the completeness of our data, and verify the values contained are expected and acceptable. As we can see in Figure 1, the amount of unique observations for Class are the expected value of 2 for ‘spam’ and ‘ham.’ However, we see sms indicates a unique value of 5169, which is less than the expected 5572; as such, we must conclude that some messages are identical. This hypothesis is verified with the remaining top and frequency values which show that the most frequent message, "Sorry, I'll call later," occurs 30 times. Since , we must further conclude that there are additional repeated messages. Within the Class column, these two rows show us that most messages are categorized as `ham`, with 4825 occurrences, which leave the remaining 747 messages to be categorized as `spam`.

A screenshot of a computer screen

Description automatically generated with medium confidence

Figure 1: Statistical Analysis of Data

The last step before transforming our data involved observing a few examples and their corresponding features. This provides additional context which enables us to understand how the information is formatted and deduce if transformation is necessary. During this step, we concluded that the messages were unstructured, containing a mix of lower and upper case, punctuation, and stop words. Before we could continue with testing and training, the messages would require natural language preprocessing. Furthermore, our observations showed that the identifying classes were also in categorical form, e.g. ‘ham’ and ‘spam’. Since many ML algorithms execute complex mathematical computations, categorical data is not ideal. As such, both columns should be converted to an equivalent numerical form, e.g. ‘0’ for ‘ham’ and ‘1’ for ‘spam’.

## Dataset Preparation

The conversion to numerical form for Class is straightforward as there are only two categories, ‘spam’ and ‘ham.’ Since class is our target value and is contained within a one-dimensional array, the function from the Python library is our best option. To do this, we created a new column ‘is\_spam’ where we mapped each classification to its corresponding encoded values; each ‘spam’ classification was encoded as a 1 for true, and each ‘ham’ classification was encoded as a 0 for false.

However, the conversion for the sms column is a bit more complicated. First, it is important to consider that there are many nuanced variations between a legitimate text and spam - most notably excessive punctuation, web addresses, phone numbers, or promotional content which often results in longer, more complex messages. For example, Figure 2 shows the relationship between message lengths in both spam (red) and ham (green) messages. As shown by the dotted yellow line in each box plot, spam messages average length of is significantly higher than ham messages average length of . While ham messages come in a larger variation of sizes, Figure 2 shows that the entire interquartile range (IQR) of all ham messages in our dataset have message lengths that fall below the minimum value for spam messages, excluding outliers. This suggests that message length can be a useful feature for distinguishing between spam and ham messages, with longer message lengths being a potential indicator of spam. As such, we created a new column in our data, sms\_len, to keep track of the message lengths.

Chart, box and whisker chart

Description automatically generated

Figure 3: Comparing Message Lengths (Outliers not Shown)

Although message length can provide valuable insights into distinguishing between spam and ham messages, it is not the only feature that can help accurately classify these messages. In addition to length, the words used in a message are also critical for effective classification. However, not all words carry significance within the context for a message. For example, consider the sentence "I went to the store to buy milk." We could remove the words "I," "to," and "the," and still convey they original message meaning. These removed words are called "stop words" as they occur “very frequently and their presence doesn't have much impact on the sense of the sentence (NLPF).” The remaining sentence still contains the necessary useful information while being more efficient for the machine learning algorithms to analyze.

To begin the process of filtering out these common words from out dataset, we imported the stopword corpus provided by the Natural Language Toolkit (NLTK). This corpus includes a vast collection of the most frequently occurring words, allowing us to remove them efficiently without manually creating a list. As we iterated through the messages to remove these words, we also replaced all punctuation with a space, and extracted each word from the text into a list. Simply removing punctuation versus replacing it with a space is an important distinction in our case as our spam messages often include links for victims to follow. By replacing the punctuation with a space, 'https://www.here.com' becomes ['https’, ‘www’, 'here','com'] instead of 'httpswwwherecom'. When these words are tokenized later, this will enable the identification of the top-level domains; if we were to simply remove the punctuation, there would be no commonality between ‘herecom’ and ‘therecom’, despite both originally containg ‘.com’.

During this text iteration would be a good point to convert the messages to lowercase, however, we noticed that many spam messags user uppercase to attract user attention. As such, we chose to leave it in. Doing so could help the algorithms differentiate between the ham message “call me now” and the spam message “CALL NOW!”, both of which would look the same (‘call’,’now’) after removing stopwords if converted to lowercase. The resulting messages are then saved to a new column named ‘sms\_clean.’

## Pre-Training Setup

While our sms data is still in categorical form and requires conversion, this will be conducted after we split our training and testing data to avoid overfitting and ensure our models are able accurately predict new, unseen data. Furthermore, if we do not conduct a train/test split prior to the conversion, we could inadvertadly introduce bias to our model as it would have access to information from the testing set during the training phase, which it should not have access to. As such, we will begin by defining our variables, splitting the training and testing data, and then finish executing our data tranformation.

Since the goal of our algorithms is to classify whether a message is spam or ham, and this classification is dependent on each message, the discrete value from ‘is\_spam’ becomes our dependent variable, , and the string value from ‘sms\_clean’ becomes our independent variable, . After identifying and assigning our variables, we split the data into training and testing sets using sklearn's train\_test\_split function, allocating 75% for training and 25% for testing. To verify our split was successful, we compared the shapes of our new sets to confirm that our observation and feature values matched. To confirm that our split was successful, we checked the number of observations and features in each set. Our training set (X\_train and y\_train) contained 4179 observations and 1 feature, while our testing set (X\_test and y\_test) contained 1393 observations and 1 feature. We can verify that these values are correct by noting that 75% of 5572 is 4179 (5572 \* 0.75 = 4179) and 25% of 5572 is 1393 (5572 \* 0.25 = 1393). This split ensures that our model is able to accurately predict new, unseen data without being biased by information from the testing set during the training phase. We can now proceed with data transformation after the split to avoid overfitting.

Our method of categorical conversion utilizes sklearn’s CountVectorizer, where a message is separated into smaller units, called tokens. In our case, these tokens identify specific words in the message. After each word in a message tokenized, referred to as feature extraction, it is mapped to a distinct numerical identification and stored in a sparse matrix. It is important to note that word order is disregarded in this approach and the main focus is on frequency. This technique is commonly known as tokenization or the bag-of-words (BOW) approach where text data is represented as a "bag" of these tokens. For example, consider our data contained the message “I’m going to the store, are you going to join?” If we left in the stopwords, this approach would produce the matrix shown in Table 1.

Table 1: BOW Example

|  |  |
| --- | --- |
| Word | Count |
| I’m | 1 |
| going | 2 |
| to | 2 |
| the | 1 |
| store | 1 |
| are | 1 |
| you | 1 |
| join | 1 |

Since our dataset is quite large, this method produced a sparse matrix that is 4179 rows × 7373 columns for the training set, and 1393 rows × 7373 columns for the testing set. Since the row values still match our dependent variables,y, and maintain a 75:25 split, we can confirm the transformation was success. As our data has now been split and transformed successfully, we convert it to a numpy array, a ML preferred numerical optimized data structure, and begin training our models.

## Models

After analyzing and preprocessing our dataset, we began training and testing machine learning algorithms to analyze and classify the messages. For our analysis, we chose three popular classification algorithms: Naive Bayes, Decision Tree/Random Forest, and K-Nearest Neighbors (KNN). Naive Bayes is a probabilistic algorithm which ‘naively assumes independency between features,’ Decision Trees create a hierarchical “a tree-like structure” structure of rules to classify data points, while Random Forests extend Decision Trees by combining multiple trees to improve performance, and KNN classify a data point based on the most common label within a predefined k-nearest neighbors proximity [[MLF](#_[MLF])]. Each algorithm has unique strengths and weaknesses, and we analyzed their corresponding prediction metrics to explore how they performed on our dataset. By comparing these results, we were able to determine the most accurate and efficient algorithm for spam/ham classification.

1. *Naive Bayes:* Naive Bayes uses Bayes' theorem, a mathematical probability formula based on prior evidence, to calculate the probability of a new message being spam, based on the presence or absence of specific words in the message. This classifier is often considered naïve because it must assume all features, or words in the message, are equally important and independent of each other. While this may not always be the case, Naïve Bayes remains a popular method for many text classification tasks, including spam filtering. It can also be a beneficial simplification method when dealing with large, high-dimensional datasets like ours. When combined with our BOW approach, Naive Bayes can effectively capture the frequency of each word in the message and use it to make predictions.

To implement Naive Bayes, we used sklearn's MultinomialNB, which is specifically designed for text classification tasks like spam filtering. After fitting our model to the training data, we made predictions on the test data and used these to create a confusion matrix and calculate the accuracy score. Our confusion matrix shows that we correctly classified 1197 ham messages and 176 spam messages, with only 10 false positives and negatives. The accuracy score of 0.9856 indicates that our Naive Bayes model performed very well on this task.

1. *Decision Tree/ Random Forest:* Decision Trees create a hierarchical structure of fixed rules to classify data points. The algorithm begins at the root of the structure and recursively splits the data based on the decisions made at the internal feature nodes, where each node corresponds to a decision or question about a feature, and each edge represents a possible answer. The resulting values of these decisions are indicated by edges, which lead to other internal nodes or leaves, which represent a final classification [[AIML](#_[AIML])]. This node-edge traversal repeats until the algorithm reaches a leaf or other stopping criterion, such as a maximum depth or a minimum number of samples per leaf. At each leaf, the algorithm outputs the predicted class label for the data point that was passed down through the decision tree. Furthermore, Random Forests are an extension of Decision Trees that combine multiple trees to improve performance and reduce overfitting. Instead of using a single tree, the algorithm creates a collection of trees by randomly selecting a subset of features and data points for each tree. Each tree performs an its own classification and the final prediction is then determined by combining all of the predictions, such as by taking the majority vote [[AIML](#_[AIML])].

For our code, we used the sklearn functions DecisionTreeClassifier and RandomForestClassifier to create and fit multiple models, including a standard tree model, a model with entropy, a pre-pruned model with entropy, and a Random Forest model. After fitting and training each model, we evaluated their performance through a series of predictions on the test set, which were then used to calculate a confusion matrix and accuracy score. The accuracy score is the proportion of correct predictions out of all predictions made. Table 2 shows the layout of a confusion matrix, which provides the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for each model.

Table : Confusion Matrix Layout

|  |  |  |
| --- | --- | --- |
|  | Predicted Value  [0] | Predicted Value  [1] |
| True Value  [0] | True Negative (TN) | False Positive (FP) |
| *Predicted HAM Correctly* | *Predicted SPAM incorrectly,  was actually HAM* |
| True Value  [1] | False Negative (FN) | True Positive (TP) |
| *Predicted HAM incorrectly, was actually SPAM* | *Predicted SPAM Correctly* |

Our initial Decision Tree Classifier, which utilized function defaults, achieved an accuracy score of 97.84% with 162 TP’s and 6 FP’s, indicating that the model incorrectly classified 24 FN. While this is within an acceptable range, it was lower than the accuracy score obtained by the Naive Bayes model (98.56%). As such, we experimented with different hyperparameters, including the criterion, specified as entropy or Gini impurity, and the maximum depth for pre-pruning, to increase model metrics.

Since our default model utilizes Gini impurity, we selected the entropy criterion to improve the accuracy. The entropy criterion differs from the Gini impurity in that it measures the level of information gained by each split in the decision tree, whereas the Gini impurity measures the probability of a data point being misclassified. While this model performed equally at identifying TP (162 correct), it was much less accurate at identifying TN (1197 correct vs 1201 correct). As such, the accuracy score for this model decreased to 97.55% and we continued experimenting in an attempt to recreate a model which increased accuracy without a loss in TN. Next, pre-pruning was implemented to limit excess tree growth and prevent overfitting, resulting in a higher accuracy score of 97.55%. However, as the aim was to increase accuracy without sacrificing correct spam detection, a random forest variation was also tried. This method improved the accuracy score to 97.91% with no false positives, the highest among the models tried so far.

After comparing the metrics of each model, as shown in Table 3, we deduced that the Random Forest model seems to be the best decision tree model for our spam/ham classification. It achieved the highest accuracy score of 97.91%, has the highest TN values, and has no FP. Furthermore, the Random Forest model was able to maintain the same level of correct spam detection as the Pre-Pruned Model while identifying Ham messages more accurately, making it the best option overall.

Table 3: Comparison of DT Metrics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model Name | Accuracy Score | TP | TN | FP | FN |
| Normal Model | 0.972003 | 160 | 1194 | 13 | 26 |
| Entropy Model | 0.974156 | 158 | 1199 | 8 | 28 |
| Pre-Pruned Model | 0.975592 | 157 | 1202 | 5 | 29 |
| Random Forest | 0.979182 | 157 | 1207 | 0 | 29 |

1. *K-Nearest Neighbors:* K-Nearest Neighbors (KNN) is a classification algorithm that determines the class of an observation based on its proximity, defined as the Euclidean distance or straight line between two points, to its nearest neighbors within a region [[AIML](#_[AIML])]. The value of K indicates the number of neighbors to consider and can be calculated multiple ways. A common method to choose K is to calculate the square root of the number of observations in the training set, as used in our initial KNN model. However, as shown in Table 4, when we observe the testing and training metrics of our dependent variable, y, we can see that the majority of observations are ham and contain the value '0'. When there is this imbalance of data, using the square root of the number of observations to calculate K may result in a bias towards the majority class. This is because the majority class will have more neighbors and therefore influence the classification more than the minority class.

Table 4: Testing and Training Metrics for y

|  |  |  |
| --- | --- | --- |
|  | Testing Set  () | Training Set  () |
| count | 1393.000000 | 4179.000000 |
| mean | 0.133525 | 0.134243 |
| std | 0.340263 | 0.340954 |
| min | 0.000000 | 0.000000 |
| 25% | 0.000000 | 0.000000 |
| 50% | 0.000000 | 0.000000 |
| 75% | 0.000000 | 0.000000 |
| max | 1.000000 | 1.000000 |

To address this bias, we employed the Grid Search optimization method to calculate K for our KNN model by using GridSearchCV from the sklearn library. Grid Search is an “automated hyperparameter-tuning method” which implements a nested for loop approach to “brute-force” its way through a range of parameter values to find the optimal combination for the given model [[HPT](#_[HPT])]. Since grid search is a brute force method, it can be extremely computationally expensive. As such, we specified an upper limit of 15 for the search space. GridSearchCV then fits the KNeighborsClassifier model using different values of k in the search space and performs cross-validation to evaluate their performance, this is referred to as a k-fold cross-validation technique. The best k value is chosen based on the highest mean score across all cross-validation folds and used to fit and train our new model. While this method increased the accuracy score by 8.26% to 94.90%, successfully identifying 115 instances of spam, it remains limited in its Spam v Ham classification abilities.

Our models show that due to the high computational cost and sensitivity to bias, KNN is unsuitable for large datasets with many features, such as those typically encountered in spam filtering. For example, since KNN uses the Euclidean algorithm to calculate the distance between observations, As the number of features increases, the distance calculation becomes more “computationally complex,” reaching “exponential time [and] could lead to computational explosions [[MLA](#_[MLA])].” This can result in a curse of dimensionality, where the increase of features disproportionately and dramatically reduces the model’s performance.

# Compare Results

[todo]

# Conclusion

[todo]

# Resources

## [MIT]

* + <https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained#:~:text=Machine%20learning%20is%20a%20subfield,learn%20without%20explicitly%20being%20programmed>
  + Note: first quote was actually quoted in 1950s by AI pioneer [Arthur Samuel](https://en.wikipedia.org/wiki/Arthur_Samuel) – see if a og source is available

## [CS]

* https://www.cs.unc.edu/~jeffay/courses/nidsS05/ai/00816048.pdf

## [Classifiers]

* https://www.researchgate.net/publication/328907962\_A\_Comparative\_Study\_of\_Spam\_SMS\_Detection\_Using\_Machine\_Learning\_Classifiers

## [Detection]

* https://www.sciencedirect.com/science/article/pii/S1742287615000079?ref=pdf\_download&fr=RR-2&rr=7b77774e197530dd#page=10&zoom=100,0,0

## [SpamVHam]

* <https://www.kaggle.com/datasets/arunasivapragasam/spam-or-ham>

## [OGSMS]

* + <https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection>

## [MLR]

* + Machine Learning with R - Third Edition, Brett Lantz
  + <https://learning.oreilly.com/library/view/machine-learning-with/9781788295864/>

## [MLF]

* + Machine Learning Fundamentals - Hyatt Saleh
  + <https://learning.oreilly.com/library/view/machine-learning-fundamentals/9781789803556/>

## [AIML]

* + Artificial Intelligence and Machine Learning Fundamentals
  + <https://learning.oreilly.com/library/view/artificial-intelligence-and/9781789801651/>

## [HPT]

* + Hyperparameter Tuning with Python
  + <https://learning.oreilly.com/library/view/hyperparameter-tuning-with/9781803235875/>

## [MLA]

* + Machine Learning Algorithms - Second Edition
  + <https://learning.oreilly.com/library/view/machine-learning-algorithms/9781789347999/>