Notes:

Yellow means I want to reword/rework it

There on comments on this page as well highlighting some of my main thoughts

Title

Abstract

[Todo: Abstract]

# Introduction

## What is Machine Learning

In the modern era of data, individuals are generating and collecting vast amounts of information which can be difficult to manage, comprehend and utilize effectively. To make sense of this data, the field of computer science and artificial intelligence has introduced a sophisticated tool: Machine Learning. Machine Learning (ML) was created to help process these large amounts of data, many of which are too vast for human [processing ?]. This is made possible by the ability of machine learning algorithms to automatically learn from data, identify patterns, and make predictions “without being explicitly programmed [[MIT](#_https://mitsloan.mit.edu/ideas-made)].” In other words, instead of being given specific instructions for what to do with the data, machine learning algorithms are designed to learn from patterns in the data itself. When these algorithms are given large datasets to analyze, they are able to identify underlying patterns and make predictions based on new, unseen data.

## Machine Learning in Cyber Security

This ability to learn and adapt to new data makes ML particularly versatile and ideal for a wide array of situations, especially pertaining to the constantly evolving nature of Cyber Security. [include more history here?]. While, historically, rule-based systems were the primary method of detecting threats, these methods often struggle to identify new and emerging threats, as they rely on pre-defined rule sets that can be easily exploited. However, the adaptability of ML algorithms enables them to identify anomalies that rule-based methods may miss; instead of relying on a set of pre-defined rule, they can create new rules as necessary. [todo: add more context]

An example of ML's superiority over rule-based systems can be seen in the context of spam detection. Spam emails and text messages are a major threat to cyber security and have many malicious intentions, including illegally accessing confidential data such as passwords and other important identifiers, spreading malicious software with harmful links or attachments, and more [[Detection](#_Detection)]. [todo: add more context]

This adaptability makes ML an effective tool for identifying new and emerging spamming techniques that may be missed by traditional rule-based methods.

[segue to spam detection here – In outline draft 3 under Objective [segue] /Topic Introduction]

## Introduction to our Study

[todo] – include dataset name, programming language (python), Kaggle, 3 types of models (KNN, naïve bayes, decision tree/random forest), NLP/ NLTK(define stop words)

# 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’.

## Data Configuration (Transformation?)

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, 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.

## Algorithms

Introduce datasets

[todo]

# 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/