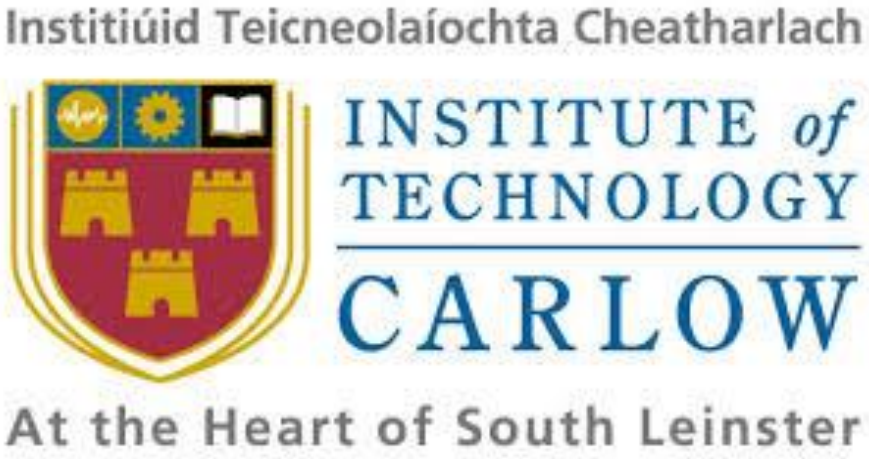

Email Spam Filter using Machine Learning

Research Manual

Author: Hazel Murphy

Student ID: C00230058

Project Supervisor: James Egan

Date: Friday 13th November 2020

Abstract

Emails are used daily worldwide. It is a method of exchanging messages between people using electronic devices. E-mail is one of the most secure methods for online communication and transferring data or messages through the internet. Around, 293.6 billion emails were sent per day in 2019. (How Many Emails Are Sent per Day: The Startling Truth [2020], 2020). However, as emails become more popular, spam has become a major problem on the internet. To try and stop the spread of spam, numerous techniques can be performed. These include spam filtering techniques. According to Statista, spam messages accounted for 53.95 percent of e-mail traffic in March 2020 (Spam statistics: spam e-mail traffic share 2019 | Statista, 2020). Discussed in the paper will be the different types of spam for example phishing, malspam, and advance-fee scams. Also, we will investigate the effectiveness of several spam filtering techniques and technologies. Our analysis was performed by simulating email traffic under different conditions. We show that algorithm-based spam filters perform best at server level and aim to justify that naïve Bayesian filters are the most appropriate for filtering at user level compared to support vector, random forest, and logistic regression.

Table of Contents

Introduction

In 1971, Ray Tomlinson sent the very first email. Tomlinson is to thank for creating this new era of communication. Billions of people send emails every day all over the world. Email began as an experiment to work out if two computers could exchange a message. By 1996, more electronic mail was being sent than postal mail. There are now over 2.6 billion active users and over 4.6 billion email accounts in operation. Email is the most vital and widely used communications medium on the web. (A brief history of email: dedicated to Ray Tomlinson - Phrasee, 2020)

Unfortunately, on May 3rd, 1978 around 400 of the 2,600 people who had email accounts on ARPANET received the first spam email. The email was sent by a Digital Equipment Corp marketing representative. These emails were labelled as spam until the early 1990s. The term spam originates from a now-legendary 1970 Monty Python's Flying Circus sketch. (Figario, Godiksen, Labs and Injaian, 2020)

Spam email is an unwanted email or also known as junk mail. This type of email is sent in bulk to your inbox. Spam emails affect all ages and all companies big or small they affect everyone with an email address (Services and Security, 2020). According to Statista, in 2019, 55% of all emails were spam. In 2019, 293.6 billion emails were sent per day, which means around 107 billion spam emails are sent per day. By 2024, the figure is expected to increase to over 361.6 billion daily mails. (Daily number of e-mails worldwide 2024 | Statista, 2020)

Spam emails can be performed in multiple ways. The most common is the bulk mailers technique. These bulk mailers can send huge volumes of email without going through a specific mail server or a particular ISP. Some bulk mailers can send approximately 250,000 messages an hour over a 28.8kb/s modem line. (Garcia, Hoepman and Nieuwenhuizen, 2004).

One of the most common types of spam email is phishing emails. The aim for the spammers is to gain sensitive information for example by clicking the link in the phishing email or signing/agreeing to a specific form.

At the end of 2006, the purpose of spam emails began to change. Spam emails were mainly based on words or phrases, but graphic images started becoming more common in spam emails. (Email spam, 2020)

The research aims to fight against spam and lower the number of spam emails being sent and received each day. This will be beneficial for both individual home users and organizations small or large. The less spam received reduces the likelihood of a computer running malware through malspam, reduces the downtime for an organization, reduces a spammer harvesting sensitive information via bank frauds and specific requests, reduces memory space, and much more.

Topics Researched

Email

An email is short for the term ‘electronic mail’. A man known as Ray Tomlinson sent the very first email in 1971. He created and brought an unforgettable new form of communication to light. The main duty of an email is to allow a user of an electronic device to send and receive emails that contain messages from another user which also obtains an email address. Emails can be sent and received anywhere in the world. When you are sending an email it can contain text, files, images, or attachments. (Peter, 2004)
This is an example of how an email address would look: test@test.com

- The text before the @ symbol will display a user’s name or a department within a company. In this example, ‘test’ is the users’ account name in the ByeHackers company.
- The @ symbol is required for all SMTP email addresses. The symbol is sometimes referenced as a divider in the address.
- The domain name is test.com

Emails are now used daily by billions of people across the world. They are used for both work and personal reasons. (Definitions and Hope, 2020)

Below is a high-level example of how email delivery works:

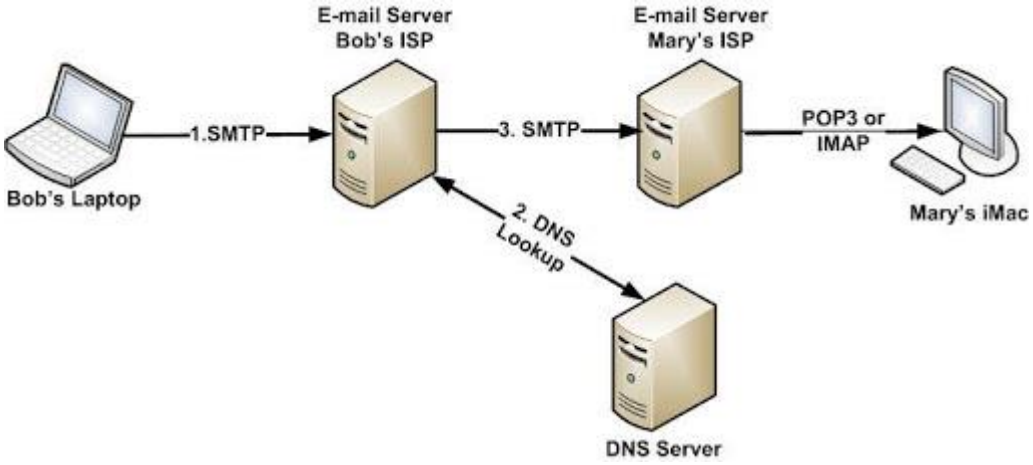

Figure 1: How an email operates (/chapter: How-Email-Works / THUNDERBIRD, 2020)

1. Firstly, you will send your email to your email service provider. This email will be sent from your PC. In this example, it will be sent from Bob's Laptop.
2. The email service provider will get the destination address. The destination address is Mary's address.
3. Your email service provider will send the message to Mary's email service provider.
4. Mary will then retrieve the email Bob sent from her email service provider.

(/chapter: How-Email-Works / THUNDERBIRD, 2020)

Structure of an email

An email contains two sections: the body and the header section.
Body: (Unstructured data) Includes data such as text, HTML mark-up, and attachments.
Header: (Structured) Includes tracing information. This information is included in the mail content.
SMTP states what should be included in the email header section. These fields are the subject, sender's name, e-mail ID, date, routing information, timestamp, etc. Every field in the header section has a specific name and meaning.

Received: Contains information like email servers, IP addresses, dates, etc.
From: The name of the sender for the mail (test@test.com)

To: The e-mail address of the person who is receiving the email (test123@testing.com)
Return-Path: An optional address specification to be used if an error is encountered.
Message-ID: one unique message identifier designated by the mail system.
X-mailer: The mail software used to create/send the message.
Subject: text to summarize the body of the email.
Content-type: Informs the email application how to interpret the text characters in the body of the email. (Bhowmick and Hazarika, 2017)
Body: Contains the main message of the email

An email address also contains two optional fields:
CC (Carbon copy): This field allows you to specify recipients who are not direct addressees (listed in the "To" field). E.g. you can address an email to John and CC Mary and Bob. This means Mary and Bob also receive a copy of the email. Also, everyone can see who received the email.
BCC (Blind carbon copy): This field is like CC. BCC is a way of sending copies of an email to other people. However, the difference between the two is that, while you can see a list of recipients when CC is used, you don't see this list with BCC. (Einstein and Einstein, 2020)

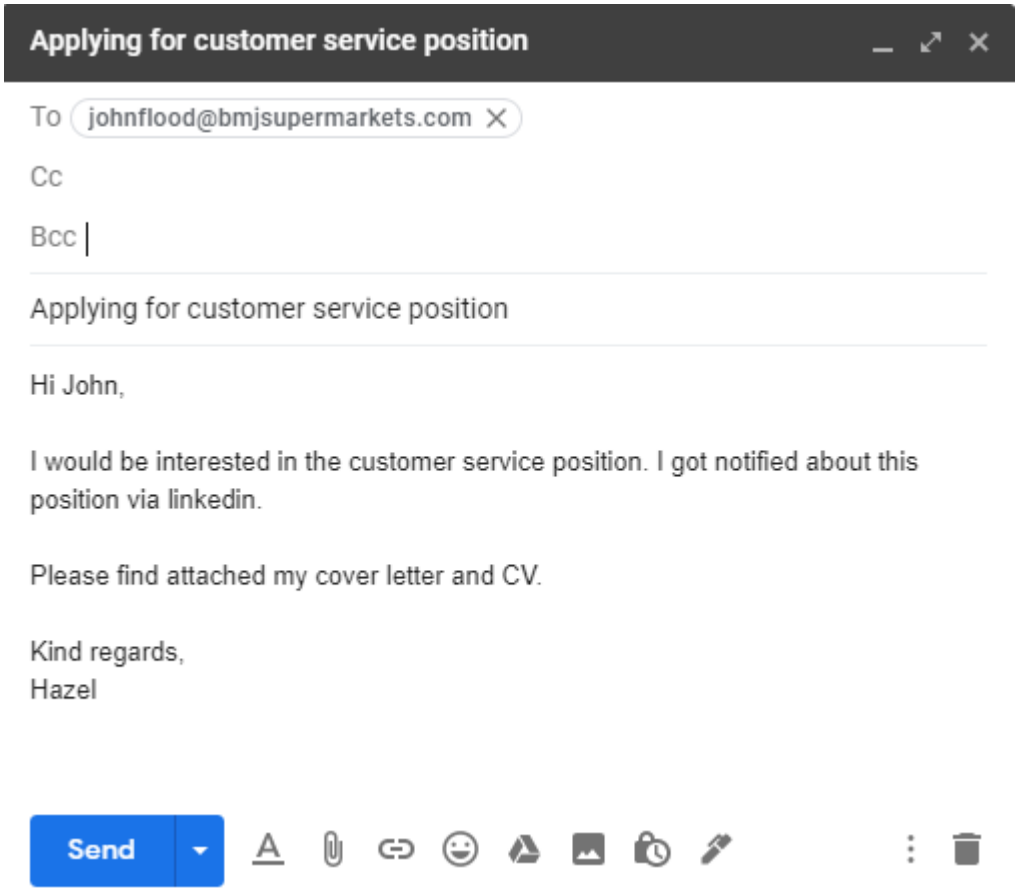

Figure 2: Sample email

Spam Email

Spam email is an unwanted email or also known as junk mail. This type of email is sent in bulk to your inbox. Spam emails affect all ages and all companies big or small they affect everyone with an email address (Services and Security, 2020). According to Statista, 55% of all emails were spam in 2019 alone. 293.6 billion emails were sent per day in 2019, which means around 107 billion spam emails are sent per day. By 2024, the figure is expected to increase to over 361.6 billion daily mails. (Daily number of e-mails worldwide 2024 | Statista, 2020)

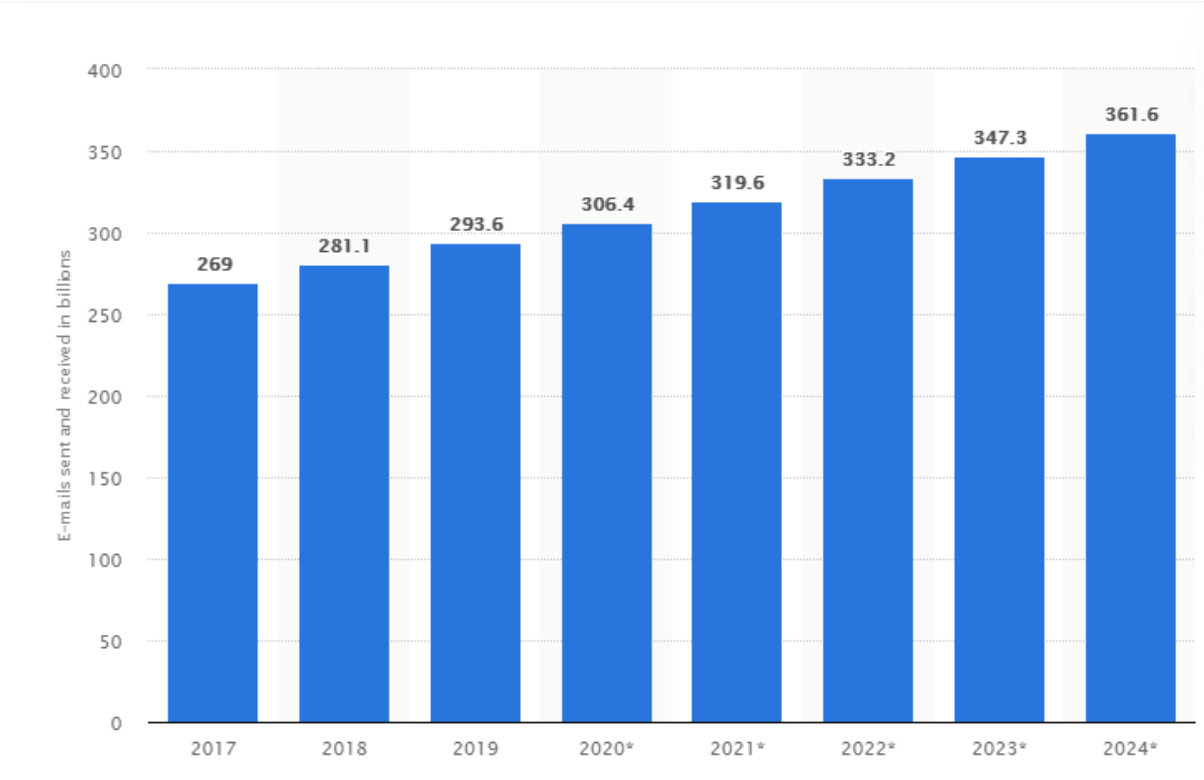

Figure 3: The number of sent and received emails per day worldwide from 2017 - 2024 in billions (Daily number of e-mails worldwide 2024 | Statista, 2020).

Spam emails can be quite dangerous. They may contain malicious links, if you click on these links, they can infect your computer with malware. A typical spam email often sounds urgent, which tries to make the user feel the need to act on it straight away without giving the user a chance to view the email properly and think before taking any action. (Controlling spam Emails at the routers - IEEE Conference Publication, 2020)

History of spam email

Spam emails were not known as spam until the early 1990s. The term spam originated from a 1970 Monty Python's Flying Circus sketch.

1978	First email spam was sent out to users of ARPANET
1993	First use of the term spam
1994	First large-scale spam distributed across USENET
1994 – 2003	Spam grows at an almost exponential rate
2003	First spamming "botnets" appeared
2004 – 2016	Spam volume has dropped to 66% of all email traffic

Figure 4: Timeline for the history of email spam from 1978-2016

1864

In 1864 the very first unsolicited electronic messages are believed to have been transmitted via telegraph. These messages were sent to wealthy Americans which included investment offers.

1978

The first recorded email spam message took place 112 years later. On May 3rd, 1978 around 400 of the 2,600 people who had email accounts on ARPANET received the first spam email. The email was sent by a Digital Equipment Corp marketing representative. Every ARPANET address on the west coast of the United States received this email. The content of the email contained information about the availability of a new model of a computer. Gary Thuerk was responsible for sending the email. Thuerk claimed that his email generated about $12 million in new sales. However, many people who received his email got highly irritated and

complained to the US Defence Department which ran ARPANET. (Figario, Godiksen, Labs and Injaian, 2020)
However, at the time this message which was sent in bulk was not classified as spam. The message, however, contained all the attributes of a spam email. Thuerk received numerous complaints but also an ARPANET representative made Thuerk promise not to do it again.

1993

In 1993, the term "spam" came into effect. The term applied to a USENET posting.
At the time, a man known as Richard Depew oversaw implementing a moderation system that allowed posts to be deleted after they had been posted, but as a result Depew accidentally slipped up. Depew software contained a bug. This bug caused the software to send roughly 200+ messages to the news.admin.policy discussion group. As time passed members of the news.admin.policy group began making jokes about the incident. A member of the group referred to the incident as "spamming." As a result, the term was used from then on. (Spam, 2020)

1994

The first large-scale spam hit the USENET, on January 18, 1994. The subject of this message was: *"Global Alert for All: Jesus is Coming Soon"*. It was created by a student and was available to every newsgroup. Many debates occurred across the USENET due to this message. In April of the same year, the Immigration-Law Services of Laurence A. Canter and Martha S. Siegel published an advertisement and USENET was overwhelmed by this. This caused a major outcry. The lawyers defended their practice and ignored the outcry. As a result, they wrote a book about it. The book was titled: "How to make a Fortune on the Information Superhighway". This meant spam was now going mainstream.

1994 - 2003

Throughout the years between 1994 - 2003 roughly 182.9 billion emails are sent and received per day worldwide. Spam grows at an almost exponential rate. It makes up for around 80 to 85 percent of email messages sent worldwide.

2003

The CAN-SPAM Act (Controlling the Assault of Non-Solicited Pornography and Marketing Act of 2003) was passed by the U.S. on December 16, 2003. It gave the Federal Trade Commission some limited regulatory powers to curb “spammers”. (CAN-SPAM Act of 2003, 2020)

2004 - 2016

It is a major battle between spammers and the rest of the world every day. It started fully in 2002 when the ePrivacy Directive was introduced by the EU. In article 13 of this directive, it stated to send unsolicited marketing communications without first getting prior agreement from the recipient was illegal. The directive established an opt-in rule. (A history of email spam - ThinkAutomation, 2020)

Numerous technologies and techniques are being created and tested which will be used to send spam and to block spam. Some examples of these include zombie networks, email harvesting botnets, neural network-based spam blockers, DMARC. According to “the 2014 Internet Security Threat Report, Volume 19” published by Symantec Corporation, spam volume has dropped to 66% of all email traffic (down from ~85%). (Figario, Godiksen, Labs and Injaian, 2020)

However, this certainly does not mean that the battle between everyday email users and spammers is over yet.

Different types of spam emails

Many different types of spam emails exist. These vary from marketing spam example: get rich quick schemes and you also have serious threats example where cyber criminals attempt to hack into your accounts and they aim to gain your sensitive data and spread malware. Marketing spam may not be a major threat as these types of emails are usually filtered by your email software. If one of these emails are not filtered out is it very easy to identify as spam and then you can flag it for removal.

Phishing emails

Scammers try to gain your personal details for example your bank account details such as debit/credit card numbers, usernames, passwords, or other sensitive information this is known as phishing.
Common phishing includes:

- A request for payment of an outstanding invoice.
- A request to reset your password or verify your account.
- Verification of purchases you never made.

Adam Kujawa, Director of Malwarebytes Labs, says of phishing emails: “Phishing is the simplest kind of cyberattack and, at the same time, the most dangerous and effective. That is because it attacks the most vulnerable and powerful computer on the planet: the human mind” (What is Spam? Definition & Types of Spam | Malwarebytes, 2020).
Phishing can also take place over the phone. For example, scammers may phone you and state they work for PayPal. They tell you there is a problem with your account, and they can fix this issue remotely, but you must pay first before they fix the issue. They would then request you to call out your card details to them. (Phishing - CCPC Consumers, 2020)

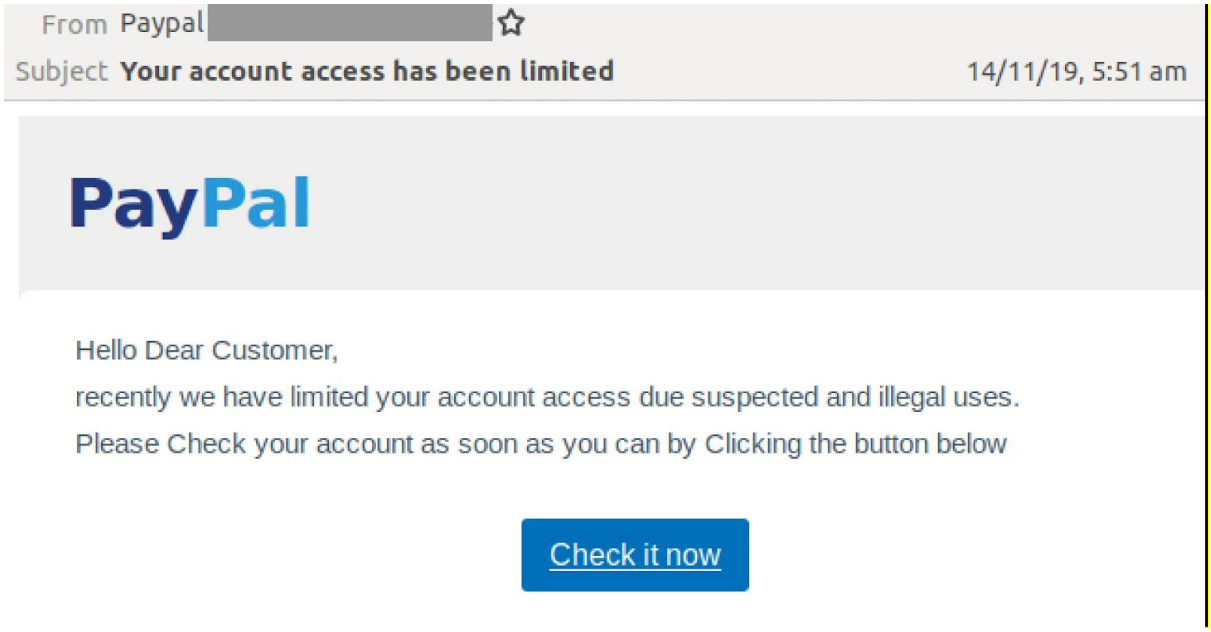

Figure 5: Sample of a phishing email (Dewan, 2020)

Advance-fee scams

This type of scam began in Nigeria and is also known as the "Nigerian scam". Not much spam begins in Nigeria. According to Cisco Talos, Nigeria ranks 68 in the top spam senders. This scam involves a mysterious sender offering you a vast reward in exchange for a cash advance, usually as some sort of processing fee, required to unlock the larger sum. Once you wire the cash to the cybercriminal, the sender disappears with your money.

Malspam

Malware that is spread via spam is known as malspam. It depends on social engineering as the aim is to get the user to perform one of the following: click on a download link or open an attachment contained in the email that infects your computer with malware. The download would be one of the following formats: Word, PowerPoint, or PDF files. However, the documents would include malicious code which would be hidden in the scripts/macros. Once the document is opened by the user the scripts will then run/execute, retrieving the malware payload from the command and control (C&C) servers run by the cybercriminals.
The payloads would be all different; it may transform your device into a botnet. It would do this to send out more spam emails. Payload may often be a Trojan. The Cybercrime Tactics and Techniques Report states the majority of malware attacks in 2018 for both businesses and consumers were identified as Trojans of some kind. (Dewan, 2020)

Sweepstakes winners

Sweepstakes winners are a common spam that takes place. This involves the spammer creating an email which states for example that you own a prize or sweepstakes, and you have to respond fast which maybe by clicking on a link or submit your personal information to claim your prize. If you have not entered a competition or don't recognize the email address you definitely shouldn't respond or click on any link within the email.

Commercial advertisements

In the United States, it's subject to the guidelines in the CAN-SPAM act that when an organization retrieves your email address, by default they add your email to their newsletter which could be sent to you regularly. This is a way for them to sell or advertise their product at a very low cost. To avoid this, you should always make sure to unlock the opt-in box when you are filling out an online form.
However, most of these emails are harmless. By law, they must have a visible opt-out or unsubscribe option. (Dewan, 2020)

How to stop spam

There are numerous ways to prevent spam:

1. Install a spam filtering tool

For my project, I will be creating a spam filter tool using four different machine learning algorithms. These algorithms are Naïve Byes, Supper Vector, Random Forest, and Logistic Regression.

2. Strengthen the SMTP protocol
3. Do not respond to spam

In a spam survey conducted by the Messaging, Malware, and Mobile Anti-Abuse Working Group, 46% of respondents said they clicked or replied to spam out of curiosity. (Dewan, 2020)

4. Turn macros off

In some cases, you may be requested to "enable macros," to open an attachment. If you know the sender and you are expecting the email you still should confirm with them before enabling macros.

5. Become familiar with how a phishing email is usually displayed

All employers should provide training to their employees.

6. Use multi-factor authentication

A user will only gain access to the application if they have presented two or more requirements successfully.

7. Install cybersecurity

If you have good cybersecurity software on your device and if you click on a link that contains malware the software should be able to recognize the malware and shut down the malware. Once the malware is shut down it cannot do damage to your device or network. (Dewan, 2020)

The effects of spam emails

In March 2020, spam messages made up 53.95% of email traffic. The average email user will receive 16 malicious spam emails every month. This means over a year an individual would receive around 192 spam emails a year. This causes major effects for business and especially small to medium-sized business as it will create a significant about of downtime. (business, 2020)

Comparison between Q1 - Q2 2019 and 2020

- **2019**

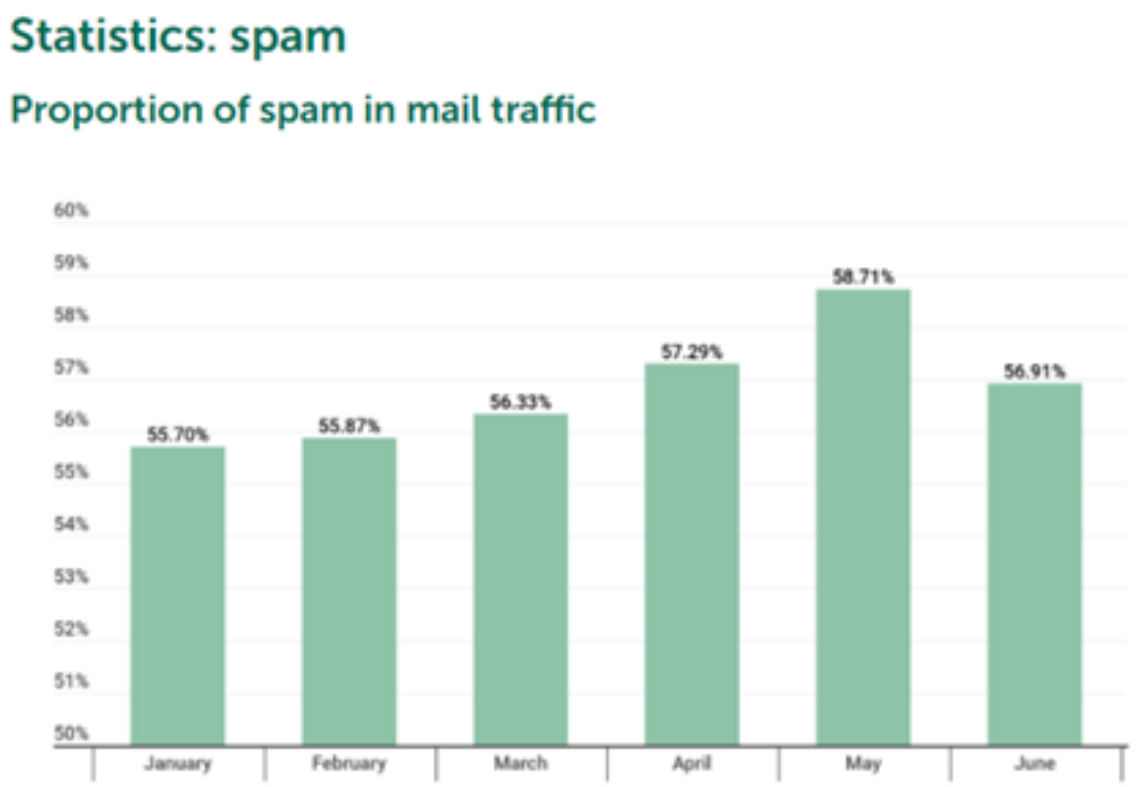

Figure 6: Proportion of spam in global mail traffic, Q1 2019 – Q2 2019 (Spam and phishing in Q2 2019, 2020)

In Q2 2019, the largest share of spam was recorded in May (58.71%). The average percentage of spam in global mail traffic was 57.64%, up 1.67 p.p. against the previous reporting period. (Spam and phishing in Q2 2019, 2020)

- **2020**

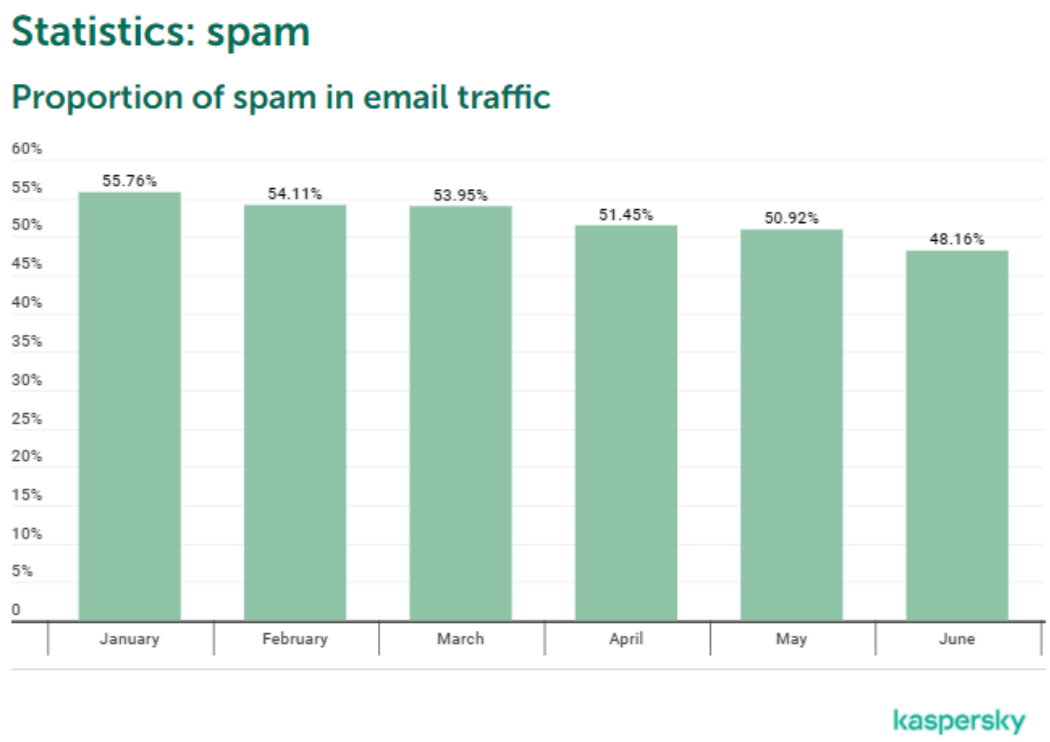

Figure 7: Proportion of spam in global email traffic, Q1 2020 – Q2 2020 (Kulikova, Sidorina and Shcherbakova, 2020)

In Q2 2020, the largest share of spam (51.45 percent) was recorded in April. The average percentage of spam in global email traffic was 50.18%, down by 4.43 percentage points from the previous reporting period. (Kulikova, Sidorina and Shcherbakova, 2020)

According to the reports generated by Securelist, we can see the results are becoming more positive from Q2 2019 to Q2 2020 spam in global email traffic has fallen by 7.46%. (Vergelis, 2020)

Techniques to perform spam

One of the most popular and most used techniques to perform spam is known as the bulk-mailing technique. Bulk mailers are used by the spammer to send the spam mail. These have the capability of sending huge volumes of email without going through a specific mail server or a particular ISP. Some bulk mailers can send approximately 250,000 messages an hour over a 28.8kb/s modem line (Garcia, Hoepman and Nieuwenhuizen, 2004). There is a large number of features within the bulk-mailer that allows the spammer to be able to hide their tracks and makes it near impossible to trace also the bulk-mailing techniques try to bypass spam filtering.

The mail server of the spammer's ISP is most of the time not used by the bulk-mailer. This means they use an open relay or connect directly to the destination mail server.

Spam filter

A spam filter is a type of program developed to spam messages from reaching your inbox. A spam filter works by searching for certain elements within the message. A classifier is using the information within the email for example, the sender, receiver, subject of the message, and the body (which is the main content of the message).

However, before the classifier uses this information the following steps are performed:

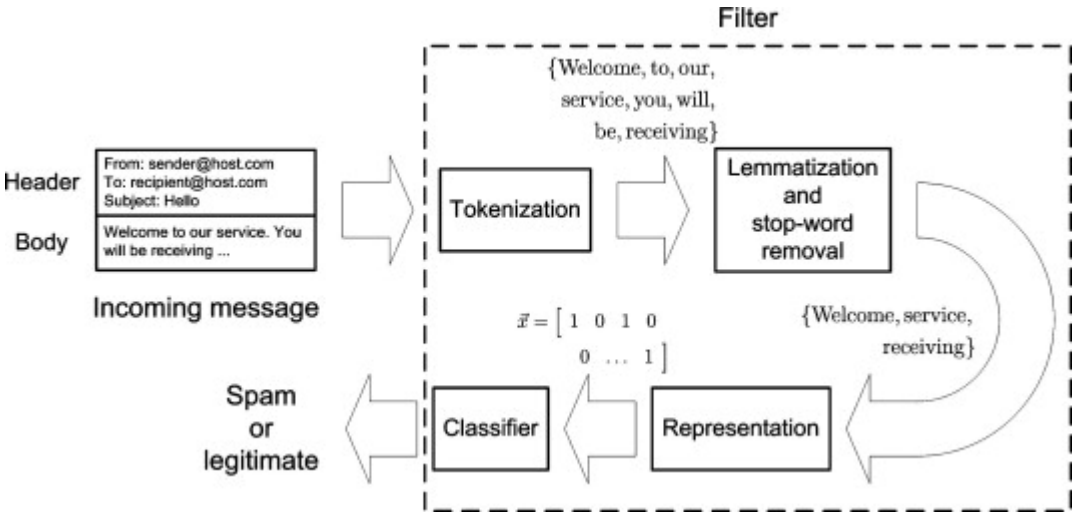

Figure 8: An illustration of some of the main steps involved in a spam filter (Thakur, Shinde, Mane and Thakare, 2020)

1. **Tokenization:** extracts words in the context of the email.
2. **Lemmatization:** decreases words to their root form e.g. “taking” to “take”.
3. **Removing stop-words:** the search engine has been programmed to ignore e.g. “the”. Also removing words that occur quite often in many messages e.g. “to”, “a”, “for”.
4. **Representation**: translates the set of words in the message to the specific format needed by the algorithm used for machine learning.

(Thakur, Shinde, Mane and Thakare, 2020)

Email servers

Relevance to the project

To create a spam filtering tool, we need to set up an email server to allow us to send and receive emails. The server will obtain various protocols such as SMTP, POP3, and IMAP.

Discovery

An email server is a server that handles and delivers emails over a network (internet). As seen in the example below the mail server can receive emails from the client computer and then delivers the email across the internet to another mail server. A mail server can also operate in reverse, the mail server can deliver emails to the client computer. The client computer would usually be the location where you would read your emails. An example would be your computer at home or in your office at work. (rubecula et al., 2020)

A client computer may also be a smartphone where it contains email capabilities. A function of the mail server will give the system administrator the required access to create and manage email accounts for any domains hosted on the server.

For example, if the server hosts the domain name "test.com," it can provide email accounts ending in "@test.com." (How Does An Email Server Work? [Technology Explained], 2020)

Figure 9: How an email server operates (What is a Mail Server and How Does it Work? (Article), 2020)

When the user clicks "send" in their email message this email will then connect to a server on the network. This server is known as an SMTP server.
To download your email message to your email program this program will connect to the POP3 server. (What is a Mail Server and How Does it Work? (Article), 2020)

Email server protocol's

SMTP

SMTP is the abbreviation for Simple Mail Transfer Protocol. The protocol's responsibility is to handle the process of email exchange and delivery across IPs. The DNS server provides the hostname to find a record. The record contains the IP address for the host. An IP (Internet Protocol) address is made up of numbers. Each device that is connected to a computer network and is using the Internet protocol for communication will have an IP address assigned to them. An IP address operates two functions: host/network interface identification and location addressing.
Example: When you send an email to your friend/work colleague using a mail client (Gmail, Outlook, Thunderbird), this email is collected by the outgoing server which is the SMTP server, this will begin the conversation with your friend/work colleague's incoming email.

The machines "talk" SMTP. The protocol performs multiple processes like the following:

- provides a standard
- provides a reliable set of guidelines
    - SMTP makes all the servers identify themselves and communicate to know who is that the sender, who the recipient is, where the content must go etc.

SMTP's main role is to properly deliver the email and take care of possible issues. (What is SMTP - smtp mail server - professional SMTP service provider, 2020)

POP3

POP is the abbreviation for Post Office Protocol. The protocol's responsibility is to receive messages and are used to process incoming mail. The POP3 protocol is an application-layer Internet standard protocol and this version is the most common in use today. (hjp: doc: RFC 1939: Post Office Protocol - Version 3, 2020)

IMAP

IMAP is the abbreviation for Internet Mail Access Protocol. IMAP is more advanced than POP3. The protocol's responsibility is to allow the user to keep messages on a server that you can access from multiple computers.

For example: With this protocol, the user can use an email client such as Thunderbird to gain access to their email accounts from home and Outlook to do this from work and still keep all your messages on the server.

The email client retrieves emails from the mail server over a TCP/IP connection.

IMAP vs POP

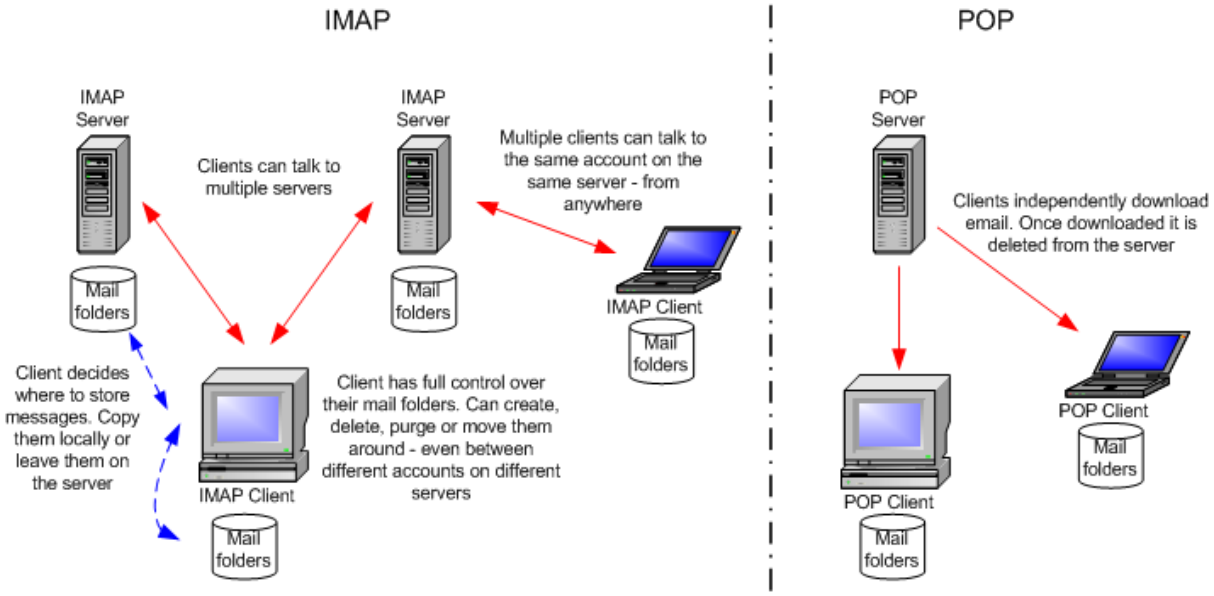

Figure 10: The difference between two protocols IMAP and POP (IMAP vs. POP3, 2020)

How did this help?

The research demonstrates not just one server is involved in sending and receiving an email message. Multiple steps take place that a user cannot see. Even though multiple steps occur in the process of sending and receiving emails, it is incredible how fast this operation can be completed. This shaped my understanding of email servers and fed into the research of the technical side of this project.

Spam filtering techniques

Relevance to the project?

Spam is growing daily and can be extremely dangerous. Spam may contain malicious content which can result in cyber-attacks or even spread viruses. This makes spam filtering very important. Everyone needs the best security possible as one single email could make a huge impact on any size organization or an individual.

Discovery

During my research, I have discovered there are many different options to filter spam emails. There are non-machine learning filtering techniques and machine learning filtering techniques.

Non-Machine learning spam filtering techniques

- **Blacklist**

This is a common spam-filtering method. System administrators used this method to create a pre-set list of senders. This allows them to then block any messages received from these senders in the future. Any senders on this list mean all their messages will be delivered straight into the user's spam folder.

- **Real-Time Blackhole List**

This list is dynamic. It consists of IP address owners that are active spammers or spam sources. It may include ISP's with customers that are known spammers or ISP servers that are hijacked for spamming purposes. This list prevents systems from being a victim of spam by halting spam.

- **Whitelist**

This list is the opposite of a blacklist. It allows you to decide which users can send you mail. All these email addresses would be stored on a trusted-users list.

- **Greylist**

This filtering technique is relatively new. The greylist works by rejecting any messages from any unknown users and will then send an error/fail message to the mail server. However, if the mail server attempts to send the message a second time then the greylist will assume the

mail is not spam and lets it proceed to the recipient's inbox. This will mean the sender's email address and IP address will be added to the list which contains all of the allowed senders.

Machine learning spam filtering techniques

- **Content-Based Filtering Technique**

Content-based filtering takes place after a mail is fully received. This means the filtering can be based on known keywords which may be contained in the subject and body of the message, common features of spam and the use of signatures from databases on the internet. The filtering focuses on words/phrases which are found in each message and then decide whether the email is spam or not spam.

Naive Bayesian filtering is a content-based mechanism for spam filtering. Thus, this is the type of spam filtering that will be used in my project.
Naive Bayes works on Bayes theorem of probability to predict the class of unknown data sets. All word orders are treated the same. Bayes' theorem will classify each object by looking at all its features individually. The theorem computes by using the below equation. Bayes' theorem is used several times in the context of spam:

1. Compute the probability that the message is spam, knowing that a given word appears in this message.
2. Compute the probability that the message is spam, consider all words.
3. Rare words

$$\Pr(S|W) = \frac{\Pr(W|S) \cdot \Pr(S)}{\Pr(W|S) \cdot \Pr(S) + \Pr(W|H) \cdot \Pr(H)}$$

Figure 11: Bayes' theorem (Naive Bayes spam filtering, 2020)

$\Pr(S|W)$ is the probability that a message is a spam, knowing that the word "replica" is in it;
$\Pr(S)$ is the overall probability that any given message is spam;
$\Pr(W|S)$ is the probability that the word "replica" appears in spam messages;
$\Pr(H)$ is the overall probability that any given message is not spam (is "ham");
$\Pr(W|H)$ is the probability that the word "replica" appears in ham messages.

Figure 12: Detail of each component (Naive Bayes spam filtering, 2020)

For my project:
object = an email
features = unique words in the email
In figure 8 above, the feature would be processed through the following steps: tokenization - lemmatization – representation, and finally the classifier.

Stats have proven at the very minimum the current probability of any message being spam is 80%.
Pr(S) = 0.8
Pr(H) = 0.2

Most Bayesian spam detection software states there should be no prior reason for any incoming message to be spam instead of non-spam. It considers both cases to own equal probabilities of fifty percent.

Pr(S) = 0.5
Pr(H) = 0.5

No prejudging is performed with these on any incoming mail. However, this assumption permits simplifying the general formula.

$$\Pr(S|W) = \frac{\Pr(W|S)}{\Pr(W|S) + \Pr(W|H)}$$

Figure 13: The spamminess of a word formula (Naive Bayes spam filtering, 2020)

This function is the same as saying: “what percentage of occurrences of the word "replica" appear in spam messages?". This is known as the spamminess of the word "replica".

Pr(W|S) - This is the number approximated to the frequency of messages containing "replica" in the messages identified as spam during the learning phase.

Pr(W|H) – Is approximated to the frequency of messages containing "replica" in the messages identified as not spam during the learning phase. The sets of the learned message should be large for these approximations to make sense.

However, Bayesian spam software tries to consider several words and combine their spamicities to determine a message's overall probability of being spam.

The majority of Bayesian spam filtering algorithms are based on formulas that are strictly valid only if the words present in the message are independent events. The below formula is used in the situation like the probability of finding an adjective in the English language is affected by the probability of having a noun. However, this is a useful idealization.

$$p = \frac{p_1 p_2 \cdots p_N}{p_1 p_2 \cdots p_N + (1 - p_1)(1 - p_2) \cdots (1 - p_N)}$$

Figure 14: Combining individual probabilities (Naive Bayes spam filtering, 2020)

- p is the probability that the suspect message is spam;
- p_1 is the probability $p(S|W_1)$ that it is a spam knowing it contains a first word (for example "replica");
- p_2 is the probability $p(S|W_2)$ that it is a spam knowing it contains a second word (for example "watches");
- etc...
- p_N is the probability $p(S|W_N)$ that it is a spam knowing it contains an Nth word (for example "home").

Figure 15: Detail of each component in the formula

The value of p is compared to a given threshold. This will determine if the message is spam or not spam.

However, if the value of p is below the threshold, the message is considered as likely not spam. (Naive Bayes spam filtering, 2020)

- **Case Base Spam Filtering Method**

Case base filtering may also be known as sample base filtering. This type of filtering method is a very popular method and preferred by many. This filtering method takes all emails both

span and non-spam emails from each user's email. This process is done by using a collection model. To transform the email pre-processing steps are performed. These steps are completed using client interface, feature extraction, and selection, grouping of email data, and evaluating the method. Once the data is collected it is classified into two different vector sets. Finally, the machine learning algorithm is used to train datasets and test them to decide whether the incoming emails are spam or non-spam.

- **Heuristic or Rule-Based Spam Filtering Technique**

This filtering technique examines various patterns which are mainly common expressions against the mail chosen by using already created rules. If similar patterns are detected this will result in the score of that specific mail to increase. On the other hand, if no patterns correspond this will deduct from the score for this mail. If the mail score reaches a specific number, it is then classified as spam otherwise it is not spam. Spam Assassin is an example.

- **Previous Likeness Based Spam Filtering Technique**

Incoming emails are classified as spam or non-spam by comparing the mails against stored examples such as training emails, the technique uses instance-based or memory-based machine learning methods to complete the process. The emails features/attributes are then used to create a multi-dimensional space vector. The multi-dimensional space vector is used to plot new instances as points. The new instances are afterward allocated to the most popular class of its K-closest training instances. This technique uses the k-nearest neighbor (kNN) for filtering spam emails.

- **Adaptive Spam Filtering Technique**

This technique groups the messages in different classes to detect if the message is spam or not. It divides an email corpus into various groups, each group has an emblematic text. Each incoming mail is compared to each group. The similarity is graded as a certain percentage and this result then decides which group the mail belongs to. (Dada et al., 2019)

How did this help?

These studies show the different methods which can be implemented by a user and/or organizations to help in the fight against spam. I found content-based filtering is the best

approach to use and to implement Naive Bayes which works on Bayes theorem, SVM, Random Forest, and Logistic Regression. Bayes theorem is a useful and easy-to-use algorithm to determine if mail is spam or not spam. Naive Bayes has many advantages such as:

1. It is for a specific user mail inbox
2. You can train the algorithm on a per-user basis
3. Avoid false positives regularly
4. Handles both continuous and discrete data

Machine Learning

Relevance to the project?

Studies have shown to classify emails as spam or not spam, using machine learning algorithms makes this project achievable.

Discovery

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence. Using machine learning your system can complete the following with very minimum interaction from humans:

- learn from data
- identify patterns
- make decisions

The iterative aspect of machine learning is important because as models are exposed to new data, they can independently adapt. They learn from previous computations to produce reliable, repeatable decisions and results.

The following is required to create a high-performance machine learning system:

- Data preparation
- Algorithms
- Automation and iterative processes
- Scalability
- Ensemble modeling

There are various users and/or organizations that use machine learning.
Some examples are:
Financial services: (to identify important insights in data, and prevent fraud)
Health care: (Medical experts would use this technology to aim them in analysing data to help them find a red flag or any tends to improve treatments)
Government: (used to detect fraud and minimize identity theft)
Oil and gas: (This technology is used to analyse minerals in the ground to find new energy sources. The use of machine learning for oil and gas is still expanding today)
Retail: (Machine learning is used to provide you with recommended items to buy online which is based on purchases you have made in the past. Machine learning analyses your buying history) (Insights, 2020)

Methods of Machine learning

The most popular machine learning methods are as follows:

Supervised machine learning algorithm

Algorithms are trained using labelled examples. Supervised machine learning is the method that will be used to implement this project. The learning algorithm receives a set of inputs along with the corresponding correct outputs, and the algorithm learns by comparing its actual output with correct outputs to find errors. It then modifies the model accordingly. Examples of supervised machine learning include: Naive bayes, Regression, Tree Decision, Random Forest, KNN, Logistic Regression.

Semi-supervised machine learning algorithm

Compared to supervised machine learning semi-supervised uses labelled (small amount) and unlabelled data (large amount). Unlabelled data requires less effort to acquire and isn't as expensive as labelled data.
Examples of semi-supervised machine learning include classification, regression, and prediction.

Unsupervised machine learning algorithm

It is used against data that have no pre-existing labels. This algorithm looks for previously undetected patterns in a data set. The main goal is to explore the data and find some structure within it. It works well on transactional data.

Examples of unsupervised machine learning include nearest-neighbor mapping and k-means clustering.

Reinforcement machine learning algorithm

This algorithm operates on a trial-and-error basis. There are three main features:

- the agent (the learner)
- the environment (everything the agent interacts with)
- actions (what the agent can do)

The objective is for the agent to choose actions that maximize the expected reward over a given amount of time. The agent will reach the goal much faster by following a good policy. This type of algorithm can be used for high-dimensional control problems as well as various industrial applications. The main goal is to learn the best policy. (Insights, 2020)

Algorithms of Machine Learning

1. Linear Regression

This algorithm is a type of supervised machine learning. Unlike other algorithms which classify values into categories linear regression predicts values within a continuous range. The algorithm then needs to create a link between independent and dependent variables by fitting a best line.

This best fit line is known as the regression line and is represented by a linear equation Y= a *X + b.

- Y – Dependent Variable
- a – Slope
- X – Independent variable
- b – Intercept

Linear Regression is mainly of two types:

Simple Linear Regression: characterized by one independent variable.

Multiple Linear Regression: characterized by multiple independent variables.

While finding the best fit line, you can fit a polynomial or curvilinear regression. And these are known as polynomial or curvilinear regression.

2. Logistic Regression

Logistic regression is also known as logit regression. This type of model is linear and is implemented for classification. It is used to estimate discrete values based on a given set of independent variable(s). A logit function is used as the algorithm predicts the probability of an event occurrence and fits this data into the logit function. The output values will lie between 0 and 1 as expected. For this algorithm 1 means, the email is spam where 0 means the email is ham. Compared to naïve bayes, logistic regression would be a conditional model. Logistic Regression is categorized as a "state of the art" machine learning algorithm, and this algorithm is used by Google.

An example of logistic regression is as follows:

Model

Output = 0 or 1

Hypothesis => Z = WX + B

hΘ(x) = sigmoid (Z)

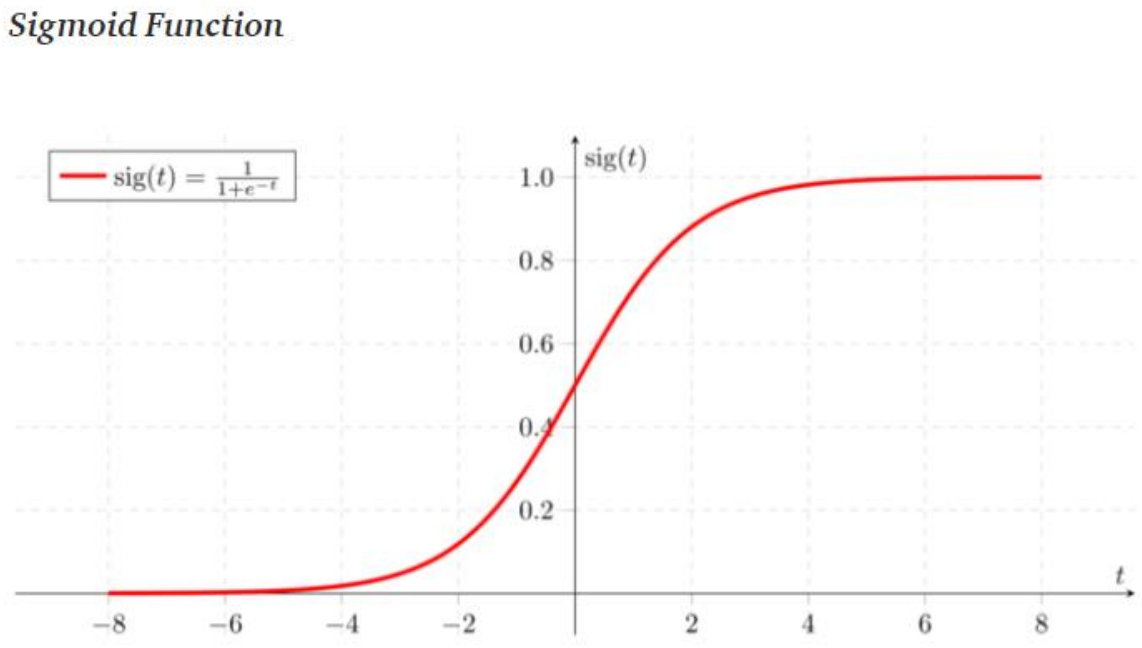

Figure 16: Sigmoid Function

For this example, if 'Z' in the equation goes to infinity, this means Y = 1. However, if 'Z' in the equation goes to negative infinity, this means Y = 0.

There are various types of Logistic regression such as:

1. Binary Logistic Regression

This type has two outcomes, which are spam or not spam, 1 or 0. This is the type I will be using in my implementation.

2. Multinomial Logistic Regression

This type has three or greater unordered categories. Example: Predicting what sport is everyone's favorite (rugby, tennis, soccer)

3. Ordinal Logistic Regression

This type has three or greater ordered categories. Example: Hotel star rating from 1 to 5.

(Logistic Regression — Detailed Overview, 2021)

3. Decision Tree

This algorithm is a type of supervised machine learning. It is mainly used for classification problems. Decision tree works for both categorical and continuous dependent variables. The population is split into at least two sets. These sets are homogeneous sets. Distinct groups are created which are dependent on a single variable or significant attributes. The pattern is displayed like a tree structure. The main aim of this algorithm is to produce a Decision tree model and train the model for it to forecast the value of a goal variable centered on several input variables.

Some types of Decision tree include: NBTree classifier, C4.5/J48 and Logistic model tree induction.

4. Naïve Bayes

Naive Bayes is a supervised machine learning algorithm. It is a classification technique based on Bayes' theorem with an assumption of independence between predictors. Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

The Naive Bayesian model is easy to build and particularly useful for very large data sets. Naive Bayes is known to outperform even highly sophisticated classification methods.

Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c).

Advantages of Naive Bayes Classifier

- Simple and easy to implement
- Small amount of training data is required
- Works for both continuous and discrete data
- High scalability
- High speed
- Work for real-time predictions

(Dada et al., 2019)

5. K - Nearest Neighbor

KNN is the abbreviation for K -Nearest Neighbor. It can be used for both classification and regression problems. KNN is mainly used in the industry for classification problems. This algorithm is quite simple. It operates by storing all the available cases and classifies new cases by a majority vote of its k neighbor. The case being assigned to the class is most common amongst its K nearest neighbor measured by a distance function.
There are some aspects that should be considered before choosing KNN:

- Expensive
- Variables should be normalized else higher range variables can bias it
- Works on pre-processing stage more before going for KNN

6. K-Means

This is a type of unsupervised algorithm. The purpose of K-Means to solve clustering problems. In the below example assume k is the cluster.
How K-means forms cluster:

1. Picks k number of points for each cluster known as centroids.
2. Each data point forms a cluster with the closest centroids i.e. k clusters.
3. Using already created cluster members this is how you determine the centroid of each cluster. Here we have new centroids.
4. As we have new centroids, repeat steps 2 and 3. Find the closest distance for each data point from new centroids and get associated with new k-clusters. Repeat this process until convergence occurs i.e. centroids do not change.

To retrieve the value of k, a centroid belongs to each cluster, so the sum of the square of the difference between the centroid and the data points within a cluster constitutes within the sum of the square value for that cluster. When all the clusters sum of square values is calculated this result determines the total within the sum of square values for the cluster solution.

7. Random Forest

RF is the abbreviation for Random forest. It is an example of ensemble learning approach and regression technique, various amount of decision trees are formed for the same issue and the results are joint together to retrieve a superior classification. RF solves problems that classify

data into groups. Random Forest is known to be a very successful machine learning algorithm to detect spam emails.

Some decision trees may be formed by the coder at the training step and used for predicting the group. This task is achieved by acknowledging the groups selected for every decision tree and the highest value number of votes group is noted as the result. This type of algorithm is known to have far less error and very pleasing f-scores when comparing to decision trees. Compared to SVM, RF performs the same or even better at times. Execution speed is fast with this algorithm. It has been proved that using RF to process a dataset containing 50,000 cases and 100 variables, outputs 100 trees in 660 seconds on a computer with a processor speed of 800Mhz.

In figure 17 outlines the steps involved to use Random Forests Algorithm for Email Classification:

1: **Input** X: number of nodes
2: **Input** N: number of features in the Email Message
3: **Input** Y: number of trees to be grown
4: **while** termination conditions is not true **do**
5: Select a self-staring Email Message S indiscriminately from the training corpus Y
6: Create tree $R_{??}$ from the selected self-starting Email Message S
7: Choose n features arbitrarily from N; where n≪N
8: Compute the optimal dividing point for node d among the n features
9: Divide the parent node to two offspring nodes through the optimal divide
10: Execute steps 1–3 till the maximum number of nodes (x) is created
11: Create your forest by iterating steps 1–4 for Y number of times
12: **end while**
13: generate result of every created trees $\{R_t\}_1^Y$
14: use a new Email Message for every created trees beginning at the root node
15: designate the Email Message to the group compatible with the leaf node.
16: merge the votes or results of every tree
17: **return** Final Email Message Classification (Spam/Non-spam email) group having the highest vote (G).
18: **end**

Figure 17: Random Forests Algorithm for Email Classification

(Dada et al., 2019)

8. SVM

SVM is the abbreviation for Support Vector Machines. SVMs are supervised learning algorithms. Classification and regression problems are resolved using SVM. This algorithm classifies data into groups. In some cases, SVM isn't as efficient as other classification algorithms due to its pace. The algorithm has a very high accuracy rate which is why it is preferred. It gains high accuracy due to its capacity to model multidimensional borderlines that are not sequential or straightforward. SVM is known to be easy to train compared to other algorithms. In the training process, SVM engages with data from email corpus. Applications such as email classification, digital handwriting recognition, text categorization, speaker recognition, etc should use SVM.

In figure 18 outlines the steps involved to use the SVM algorithm:

1: **Input** Sample Email Message x to classify
2: A training set *S*, a kernel function, $\{c_1, c_2, \ldots c_{num}\}$ and $\{\gamma_1, \gamma_2, \ldots \gamma_{num}\}$.
3: Number of nearest neighbours k.
4: **for** $i = 1$ to num
5: set $C=C_i$;
6: **for** $j = 1$ to q
7: set $\gamma=\gamma$;
8: produce a trained SVM classifier $f(x)$ through the current merger parameter (C, γ);
9: **if** ($f(x)$ is the first produced discriminant function) **then**
10: keep $f(x)$ as the most ideal SVM classifier $f^*(x)$;
11: **else**
12: compare classifier $f(x)$ and the current best SVM classifier $f^*(x)$ using k-fold cross-validation
13: keep classifier with a better accuracy.
14: **end if**
15: **end for**
16: **end for**
17: **return** Final Email Message Classification (Spam/Non-spam email)
18: **end**

Figure 18: SVM Algorithm

(Dada et al., 2019)

9. Dimensionality Reduction Algorithms

It is the method of reducing the number of random variables to be regarded by obtaining a set of major variables. It can be divided into two features which are feature selection and feature extraction. This algorithm helps with many other algorithms such as: Decision Tree, Random Forest, etc.

10. Gradient Boost

This algorithm is used for regression and classification problems. A prediction model is produced with this algorithm. The model is in the form of an ensemble of weak prediction models, which are mainly decision trees.

11. Adaboost

AdaBoost may also be known as meta-learning. It is an ensemble learning method. It was developed to increase the efficiency of binary classifiers. It uses an iterative approach to learn from the mistakes of weak classifiers and turn them into strong ones.
(Subramaniam, Thamarai & Jalab, Hamid & Taqa, Alaa. 2010) (Bhowmick, Hazarika, 2016)

How did this help?

This research demonstrates machine learning should be used to produce a spam filter technique. Machine learning provides a large variety of algorithms. After completing this research, I will be using the following algorithms within my implementation: Naive Bayes algorithm which relies on Bayes theorem, SVM, Random Forest, and Logistic Regression. Google uses algorithms such as logistic regression and neural networks to classify emails. I will perform testing on these algorithms and choose the best fit. The testing will score each model on their accuracy, recall, f1 score, and prediction.

(Dada et al., 2019)

Testing Machine Learning algorithms

To test and measure the performance of the classifiers we can use a tool known as “confusion matrix”. This tool can be used to calculate evaluation scores such as accuracy, f1-score, precision and recall and sensitivity and specificity. It is a table that counts the number of times every combination of known outcomes occurred in combination with each prediction type. The table consists of rows and columns. The rows relate to the labels (spam or non-spam) and the columns relate to the prediction which the specific model you have chosen makes. In figure 19 below we can see the first cell of the table is prediction = non-spam and truth = non-spam. This result relates to the 264 non-spam emails which are in the test dataset, and the chosen machine learning model correctly predicts are non-spam. These are known as true negatives which as correctly predictions (negative).

```
confmat_spam <- table(truth = spamTest$spam,
                      prediction = ifelse(spamTest$pred >
0.5,
                                          "spam", "non-spam"))
print(confmat_spam)
##          prediction
## truth    non-spam spam
##   non-spam   264   14
##   spam        22  158
```

Figure 19: Sample Confusion matrix

In the below table we can see each cell within the table of a two-by-two confusion matrix, is identified with its own name.

	Prediction=NEGATIVE (predicted as non-spam)	Prediction=POSITIVE (predicted as spam)
Truth mark=NEGATIVE (non-spam)	True negatives (TN) `confmat_spam[1,1]=264`	False positives (FP) `confmat_spam[1,2]=14`
Truth mark=POSITIVE (spam)	False negatives (FN) `confmat_spam[2,1]=22`	True positives (TP) `confmat_spam[2,2]=158`

Figure 20: Two-by-two confusion matrix

As mentioned above the confusion matrix can be used to calculate evaluation scores such as accuracy, f1-score, precision and recall.

1. Accuracy

Accuracy is scored on the number of objects correctly categorised divided by the total number of objects. The results means the fractions of the time that the classifier is correct. If your model has good accuracy your result will be around 0.9987.

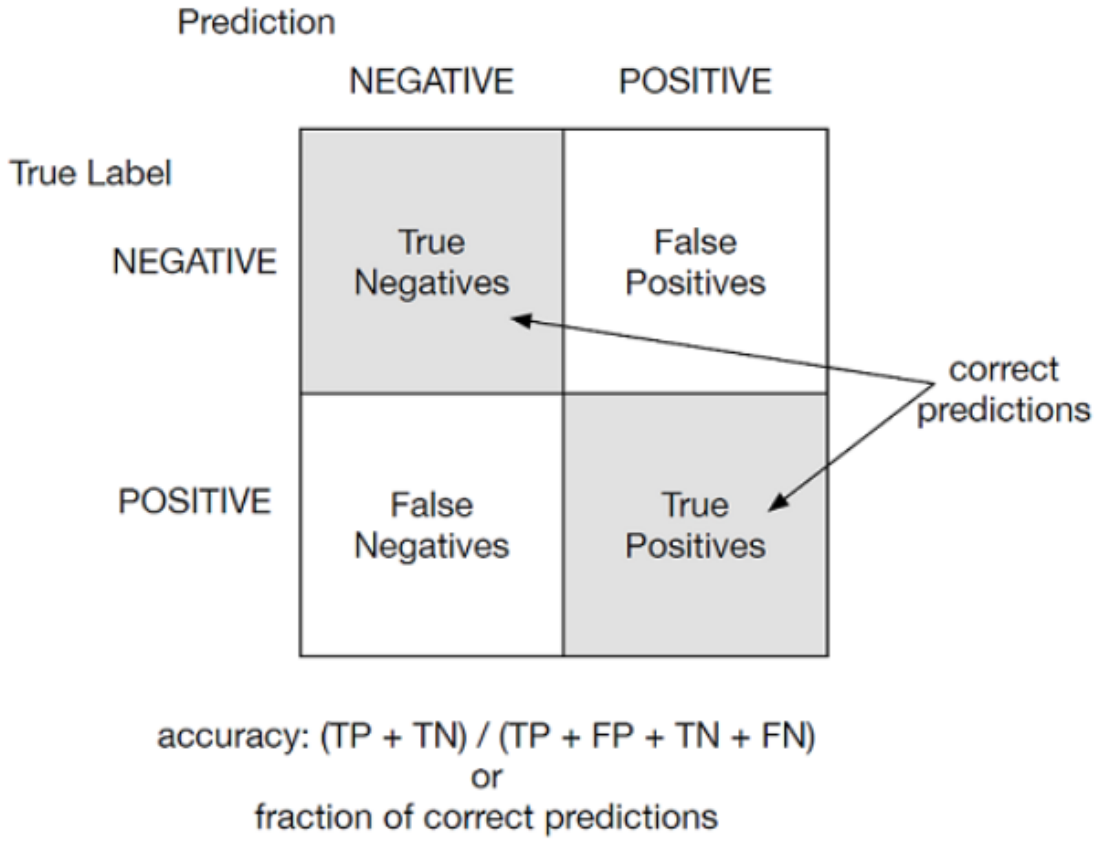

Figure 21: Accuracy formula

2. Precision and Recall

Precision and Recall are a pair of numbers. The result for precision is based on the ratio of true positives to predicted positives.

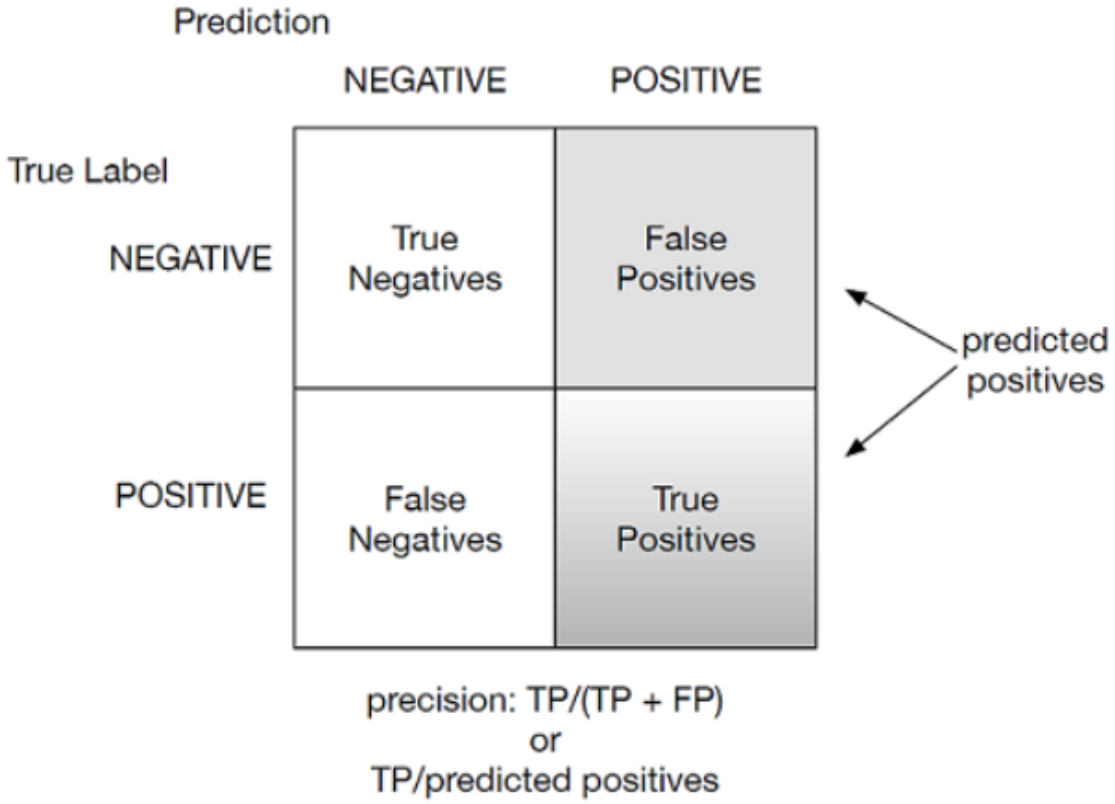

Figure 22: Precision formula

The precision result might or might not be close to the result you receive from accuracy. If your result for precision is 93% this means 9% of emails detected as spam were non-spam emails. This rate is not acceptable as you could miss out on important emails.

The acquaintance score to precision is known as recall. Recall is the ratio of true positives overall positives.

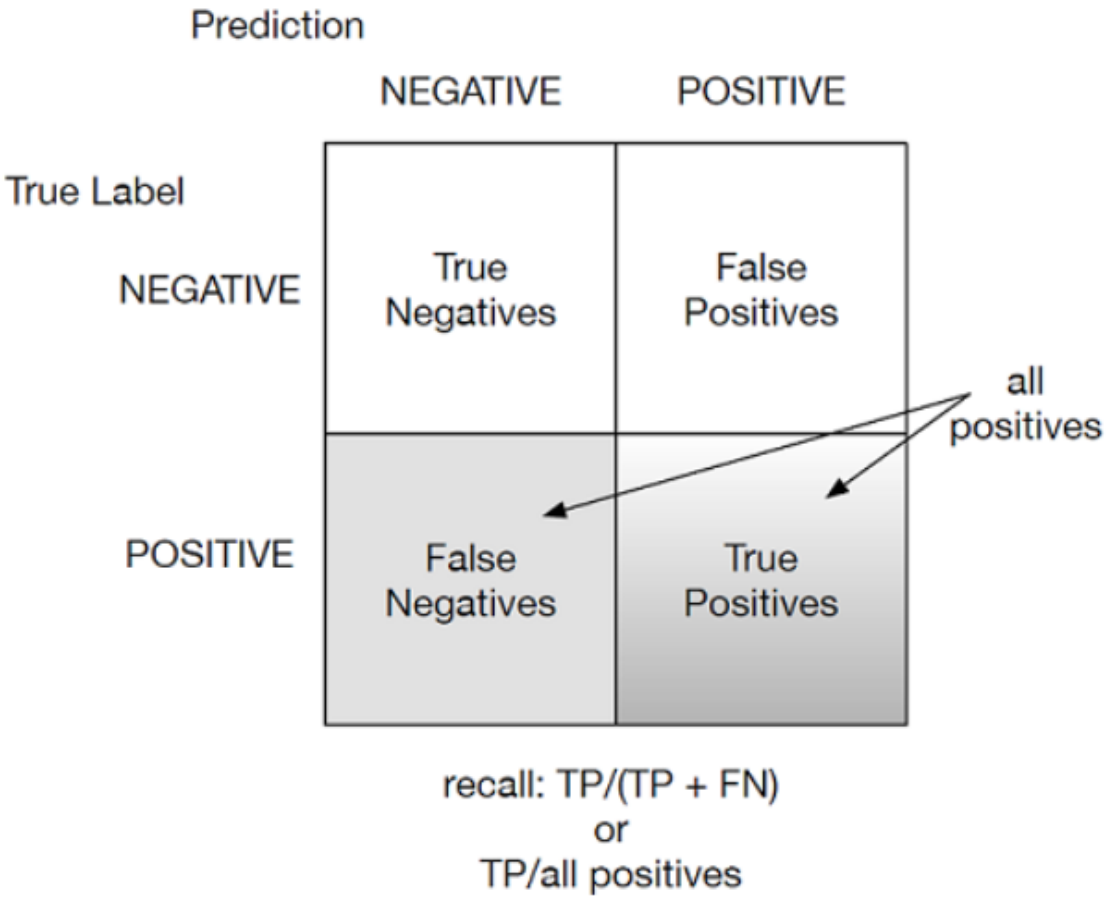

Figure 23: Recall formula

If your recall result is 87% this means roughly 11% of spam mail sent to us still makes its way to our inbox. This means we are likely to open these mails thinking they are not spam and this could lead to data loss, downtown and much more. These results mean there is high recall and precision is highlighted over recall. This scenario is acceptable for a spam filtering tool, as it is less important to filter every spam mail from our inbox than to lose important non-spam mail from our inbox.

3. F1

F1 is mainly used when you are trying to decide on just one spam filter to implement, where for example four of your filters have different precision and recall values. F1 score is usually the number that helps people to make their final decision. This score calculates a trade-off between precision and recall. It's defined as the harmonic mean of the precision and recall.

4. Sensitivity and Specificity

When testing your models, you should include performance metrics. sensitivity and specificity are a performance metric who are independent of the class prevalence. sensitivity and specificity are very popular in medical research.

Sensitivity is the same as recall and is also called the true positive rate. Specificity is called the very same as sensitivity which is the “true negative rate”. This is the ratio of true negatives to all negatives.

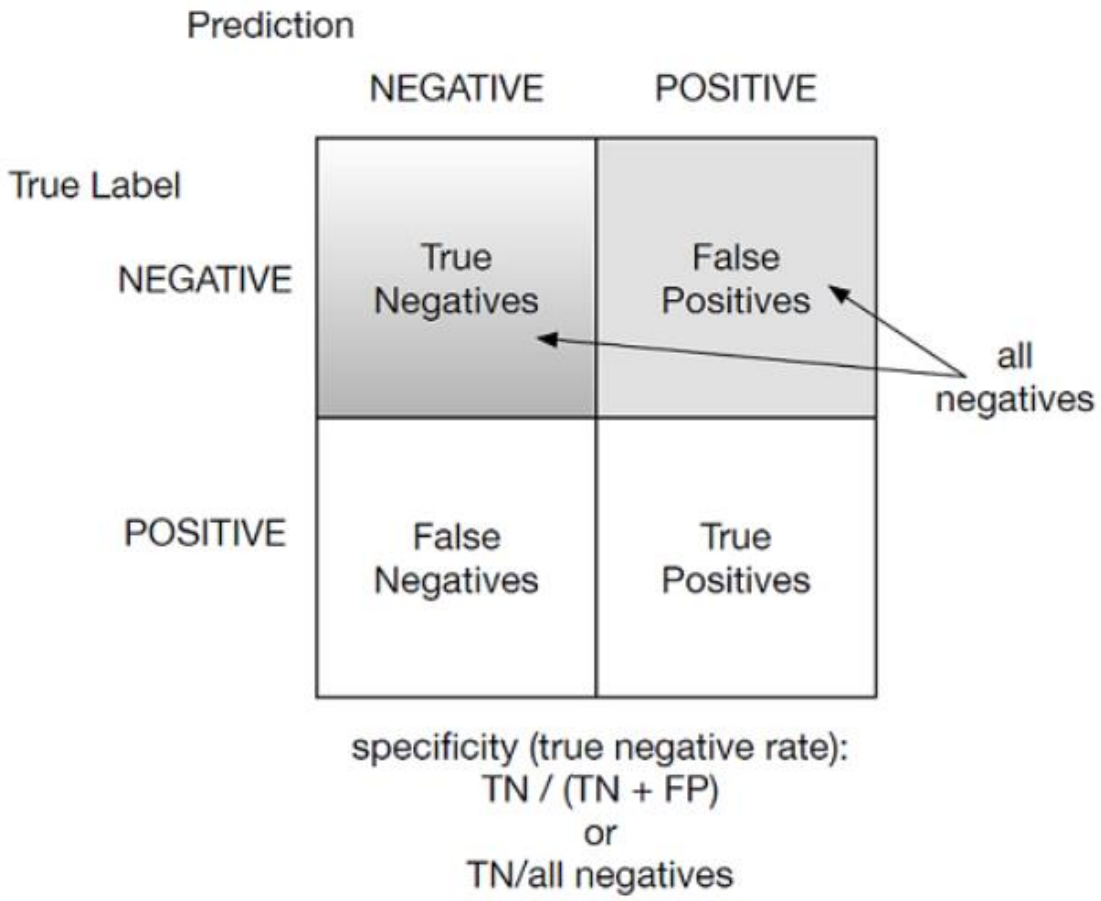

Figure 24: Specificity

(Evaluating a Classification Model with a Spam Filter - Manning, 2021)

Researched technology stack

After extensive research on the functioning of email servers, I started to piece together the technology stack components required to achieve the spam classification tool. I have made the decision to make a local email server, storage and build a web application that a user can log into. The web applications main function will be to show the emails for the logged-in user as normal mail or classified as spam. The admin user for the web application will be able to view spam reports such as “the amount of spam within the last seven days” and “the account who has received the most spam”.

Email server software

To make a local email server, an email server software must be used to deal with the SMTP, IMAP and POP3 protocols, sending, receiving and storing emails. During the course of research, I considered many different email server software’s, with the main requirements being that you can set up a local server and connect to an external database.

hMailServer

hMailServer is a free open-source email server software for Windows. It is written in C++ and C# (Knafve, 2020). It seems to be an attractive option for usage as it takes only a few minutes to install, deals with SMTP, POP3 and IMAP protocols. The software allows you to store the emails in an external database of your choice (Knafve, 2020). The software allows you to set up a local server and resolve issues on local domains that don't exist in a DNS server (Knafve, 2020).

Apache James

Apache James is an open-source email server software that runs on a java virtual machine. It also seems to be an attractive option and covers more protocols than hMailServer. The documentation for the email server software is extensive, however the perceived effort in using this software is higher than using hMailServer due to not having worked with the java virtual machine environment more. (Apache James Server 3.0 - Apache James Server 3 - Installation, 2020)

Storage

A database will be required to store the emails the email server is sending and receiving, it will also store the user login information for the web application and the spam classification information. The database will also need to be used by the web application so the web application can display the emails, the content of the emails, email spam score and display them in the web application as either normal mail or spam.

MySQL

MySQL is an open-source relational database management system. It is possible to set up a MySQL database server on a windows machine. (MySQL :: MySQL Community Edition, 2020)It is compatible with hMailServer.

Web application

When researching the technology to build the web application and spam classification module of this program, there were a couple of requirements that needed to be met. The language I use needs to be a good choice for machine learning, a good choice for use as a web application and needs to be thoroughly documented to assist in creating the web application.

Python

Python is a high-level, easy to use, programming language. It is commonly used in data science projects due to its fast-prototyping attributes and the endless amounts of packages it has to assist you in what you need to do fast. For example, scikit-learn is a machine learning library for use with classification, regression and clustering (scikit-learn: machine learning in Python — scikit-learn 0.23.2 documentation, 2020). Python also has libraries such as pandas and numpy for working with big data sets. (Top Python Libraries: Numpy & Pandas, 2020). These kits seem to be used in spam classification, one method I found showed the use of NLTK, Pandas and Numpy as the three packages required to classify spam using python.(Spam Classifier in Python from scratch, 2020). It also has web framework libraries that you can use to create web applications easily, such as Flask and Django.

Flask

Flask is a micro web framework for python. It is used for creating web applications and APIs using python. A ‘micro’ framework means that the whole web application does not need to fit into one file, it just means flask keeps the core simple but extensible. It has a website of extensive documentation; it is deemed good for use as a web application technology. (Foreword — Flask Documentation (1.1.x), 2020)

SQLAlchemy

SQLAlchemy is an object-relational mapper for use with Python. It gives developers full power and flexibility with SQL databases right from their code. It is commonly used in conjunction with Flask. (Object-relational Mappers (ORMs), 2020). Considering the web application will need to pull email data from a MySQL database this type of tool will be required to work alongside flask.

Java

Java was considered an option due to previous experience with working with the language. It does have several machine learning libraries, possibly more due to its maturity in comparison to Python. Java runs quite a bit more efficiently than python, however the difficulty level in programming with Java is much higher. (Bhatia, 2020)

Summary and Conclusion

Spam is a continuous issue around the world every day. It is highly distracting and gains the attention of user's very easily. To fight against spam effective technologies and methods need to be used, such as machine learning techniques. These types of techniques have shown a way on how to counter spam mails.

Due to having a windows machine, I have decided to use hMailServer as the email server software. I will run that locally and create a local MySQL database to store the emails. For the web application I have decided to use Python and Flask as my language and web framework. For machine learning and spam classification I am going to use NLTK for processing the messages, Pandas for loading data, NumPy for generating random possibilities for train-test splits.

The following is a high level three tier diagram to show how this is going to be implemented. It shows the sender's client sending an email which reaches hMailServer which acts as an SMTP, MTA and deals with IMAP and POP3 protocols. This information is stored in a message repository in .eml format. Our web application will be hosted on a local web server, which has access to a local MySQL database. The flask application will host code to periodically check for new emails, parse the stored email messages into storable string format and store it in the MySQL database. Spam scoring will take place in the flask application and this score will also be stored in the MySQL database. The front end of the flask application will provide the user a way to log in, when logged in there will be pages separating the inbox sent mail and spam messages. The admin account for the web application will have additional features available such as; view all registered users of the application and view spam reports.

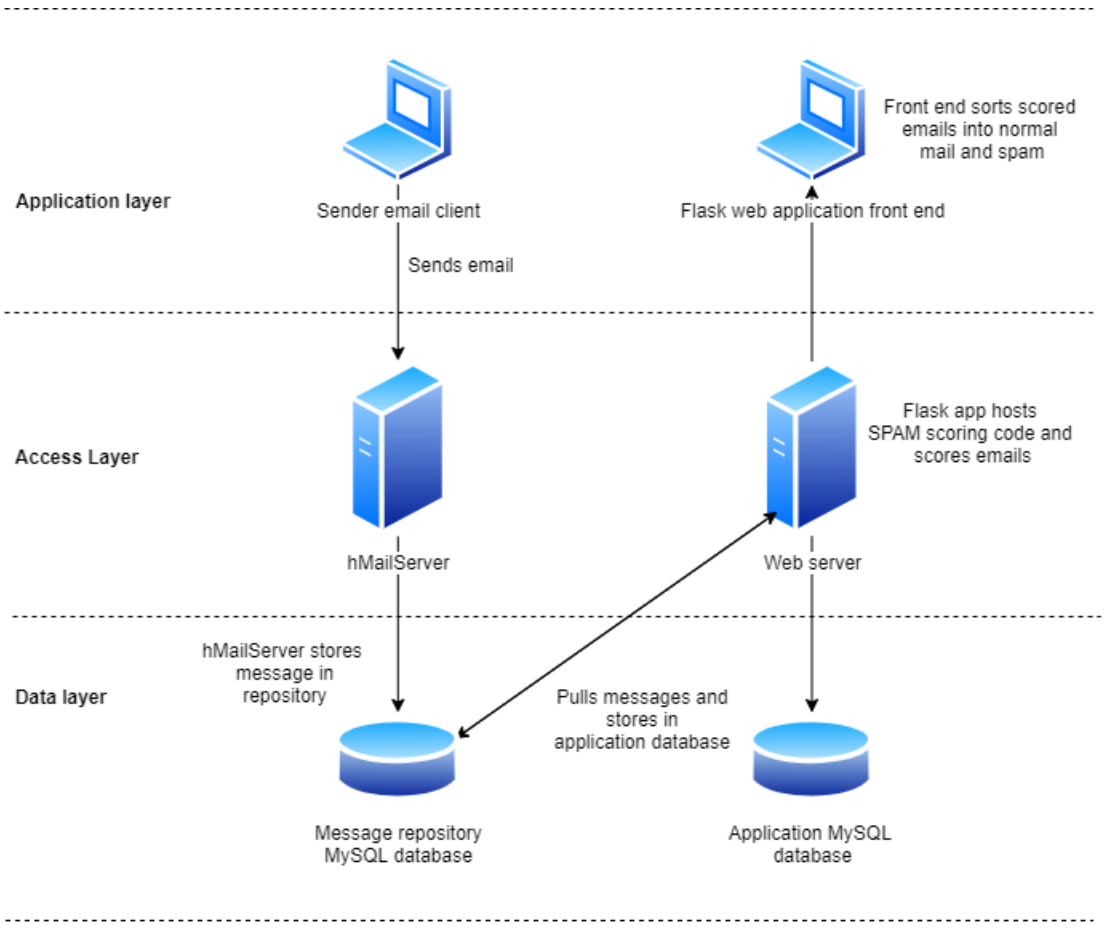

Figure 25: How each technology will be implemented

Glossary

SMTP

Abbreviation for Simple Mail Transfer Protocol. It is a communication protocol for electronic mail transmission.

Email server

It is a computer system which sends and receives emails.

DNS server

Abbreviation for Domain Name System. This server translates domain names into IP addresses. This makes it possible for DNS clients to reach the origin server.

DNS lookup

Gather all of the DNS Records of a domain and shows as received.

POP3

Abbreviation for Post Office Protocol. This protocol allows email clients to retrieve email from a mail server. Version 3 is the version in common use now.

IMAP

Abbreviation for Internet Message Access Protocol. This protocol allows you to access your email wherever you are, from any device.

ARPANET

Abbreviation for Advanced Research Projects Agency Network. It was the first wide-area packet-switching network with distributed control It was also one of the first networks to implement the TCP/IP protocol suite.

USENET

It is a newsgroup precursor to the Internet and is a worldwide distributed discussion system that can basically be regarded as a hybrid between email and web forums.

CAN-SPAM Act

The Controlling the Assault of Non-Solicited Pornography And Marketing Act of 2003, signed into law by President George W. Bush on December 16, 2003, established the United States' first national standards for the sending of commercial e-mail and requires the Federal Trade Commission to enforce its provisions.

DMARC

Domain-based Message Authentication, Reporting & Conformance, is an email authentication, policy, and reporting protocol.

ISP

An Internet service provider is an organization that provides services for accessing, using, or participating on the Internet.

Social Engineering

Is a manipulation technique that exploits human error to gain private information, access, or valuables.

Botnet

A number of Internet-connected devices, each of which is running one or more bots. Botnets can be used to perform Distributed Denial-of-Service (DDoS) attacks, steal data, send spam, and allow the attacker to access the device and its connection.

Trojan

It is a type of malware that is often disguised as legitimate software.
Multi-Factor Authentication

is used to ensure that digital users are who they say they are by requiring that they provide at least two pieces of evidence to prove their identity.

Open relay

It is a Simple Mail Transfer Protocol server configured in such a way that it allows anyone on the Internet to send email through it.

Discrete Values

This is information that can only take certain values. E.g. Binary 0 or 1.

Homogeneous

Used to describe multiple things that are all essentially alike or of the same kind.

Bibliography

TechJury. 2020. *How Many Emails Are Sent Per Day: The Startling Truth [2020]*. [online] Available at: <https://techjury.net/blog/how-many-emails-are-sent-per-day/#gref> [Accessed 5 November 2020].

Statista. 2020. *Spam Statistics: Spam E-Mail Traffic Share 2019 | Statista*. [online] Available at: <https://www.statista.com/statistics/420391/spam-email-traffic-share/> [Accessed 8 November 2020].

Phrasee. 2020. *A Brief History Of Email: Dedicated To Ray Tomlinson - Phrasee*. [online] Available at: <https://phrasee.co/a-brief-history-of-email/> [Accessed 10 November 2020].

Figario, S., Godiksen, B., Labs, S. and Injaian, J., 2020. *Do You Know The History Of Spam? - Socketlabs Email Delivery Solutions*. [online] SocketLabs. Available at: <https://www.socketlabs.com/blog/know-history-spam/#:~:text=Though%20it%20wasn't%20called,Monty%20Python's%20Flying%20Circus%20sketch.&text=The%20history%20of%20spam%20can,early%20as%201864%2C%20via%20telegraph.> [Accessed 26 October 2020].

Services, P. and Security, E., 2020. *What Is Spam Email?*. [online] Cisco. Available at: <https://www.cisco.com/c/en/us/products/security/email-security/what-is-spam.html> [Accessed 22 October 2020].

Statista. 2020. *Daily Number Of E-Mails Worldwide 2024 | Statista*. [online] Available at: <https://www.statista.com/statistics/456500/daily-number-of-e-mails-worldwide/> [Accessed 25 October 2020].

Garcia, F., Hoepman, J. and Nieuwenhuizen, J., 2004. Spam Filter Analysis. Security and Protection in Information Processing Systems, pp.395-410.

En.wikipedia.org. 2020. *Email Spam*. [online] Available at: <https://en.wikipedia.org/wiki/Email_spam> [Accessed 10 November 2020].

Peter, I., 2004. The History Of Email. [online] Nethistory.info. Available at: <http://www.nethistory.info/History%20of%20the%20Internet/email.html> [Accessed 10 November 2020].

Definitions, E. and Hope, C., 2020. *What Is E-Mail?*. [online] Computerhope.com. Available at: <https://www.computerhope.com/jargon/e/email.htm> [Accessed 20 October 2020].

FLOSS Manuals. 2020. */Chapter: How-Email-Works / THUNDERBIRD*. [online] Available at: <http://write.flossmanuals.net/thunderbird/how-email-works/> [Accessed 20 October 2020].

Bhowmick, A. and Hazarika, S., 2017. E-Mail Spam Filtering: A Review of Techniques and Trends. Lecture Notes in Electrical Engineering, pp.583-590.

Manning. 2021. Evaluating a Classification Model with a Spam Filter - Manning. [online] Available at: <https://freecontent.manning.com/evaluating-a-classification-model-with-a-spam-filter/> [Accessed 8 November 2020].

Einstein, M. and Einstein, M., 2020. *The Six Key Components To Properly Structure Business Email Messages. — Email Overload Solutions*. [online] Email Overload Solutions. Available at: <https://www.emailoverloadsolutions.com/blog/structure-business-email> [Accessed 22 October 2020].

Ieeexplore.ieee.org. 2020. *Controlling Spam Emails At The Routers - IEEE Conference Publication*. [online] Available at: <https://ieeexplore.ieee.org/abstract/document/1494611> [Accessed 26 October 2020].

Google Books. 2020. *Spam*. [online] Available at: <https://books.google.ie/books?hl=en&lr=&id=QF7EjCRg5CIC&oi=fnd&pg=PR4&dq=history+of+a+spam+email&ots=gawRE5EoNc&sig=BfFfGJj6ReNw3kE-mtqU5sEIxgU&redir_esc=y#v=onepage&q=history%20of%20a%20spam%20email&f=false> [Accessed 26 October 2020].

En.wikipedia.org. 2020. *CAN-SPAM Act Of 2003*. [online] Available at: <https://en.wikipedia.org/wiki/CAN-SPAM_Act_of_2003> [Accessed 12 November 2020].

ThinkAutomation. 2020. *A History Of Email Spam - Thinkautomation*. [online] Available at: <https://www.thinkautomation.com/histories/the-history-of-email-spam/> [Accessed 29 October 2020].

Malwarebytes. 2020. *What Is Spam? Definition & Types Of Spam | Malwarebytes*. [online] Available at: <https://www.malwarebytes.com/spam/> [Accessed 29 October 2020].

CCPC Consumers. 2020. *Phishing - CCPC Consumers*. [online] Available at: <https://www.ccpc.ie/consumers/money/scams/phishing/?gclid=Cj0KCQjwit_8BRCoARIsAIx3Rj4PZJUDTdL_QzTSRW28_LA-tAiQ0nhl5Z0BRvEyqixSofjCfyCKbToaAjYtEALw_wcB&gclsrc=aw.ds> [Accessed 30 October 2020].

Dewan, A., 2020. *Email Scam Spoofs Paypal Once Again; Informs Users Their Account Access Is 'Limited'*. [online] Mailguard.com.au. Available at: <https://www.mailguard.com.au/blog/email-scam-spoofs-paypal-once-again-informs-users-their-account-access-is-limited> [Accessed 30 October 2020].

business, T., 2020. *The Impact Of Spam And Spoofed Emails On Your Business - Graphus*. [online] Graphus. Available at: <https://www.graphus.ai/the-impact-of-spam-and-spoofed-emails-on-your-business/> [Accessed 30 October 2020].

Securelist.com. 2020. *Spam And Phishing In Q2 2019*. [online] Available at: <https://securelist.com/spam-and-phishing-in-q2-2019/92379/> [Accessed 1 November 2020].

Kulikova, T., Sidorina, T. and Shcherbakova, T., 2020. *Spam And Phishing In Q2 2020*. [online] Securelist.com. Available at: <https://securelist.com/spam-and-phishing-in-q2-2020/97987/> [Accessed 2 November 2020].

Vergelis, M., 2020. *Spam And Phishing In Q2 2019*. [online] Securelist.com. Available at: <https://securelist.com/spam-and-phishing-in-q2-2019/92379/> [Accessed 11 November 2020].

Thakur, D., Shinde, P., Mane, D. and Thakare, S., 2020. *Classification For Spam Filtering Using Naive Bayes*. [online] Semanticscholar.org. Available at: <https://www.semanticscholar.org/paper/Classification-for-Spam-Filtering-using-Naive-

Bayes-Thakur-Shinde/d01bac7cf3f95b14745cbe7e376e98f6097dd090> [Accessed 3 November 2020].

rubecula, E., failure, D., Networks, F., Intel, C., transmission, b., processes, e. and reactions, m., 2020. *US7617305B2 - Email Server System And Method - Google Patents*. [online] Patents.google.com. Available at: <https://patents.google.com/patent/US7617305B2/en> [Accessed 3 November 2020].

MakeUseOf. 2020. *How Does An Email Server Work? [Technology Explained]*. [online] Available at: <https://www.makeuseof.com/tag/technology-explained-how-does-an-email-server-work/#:~:text=When%20we%20send%20an%20email,Simple%20Mail%20Transfer%20Protocol)%20server.> [Accessed 4 November 2020].

Samlogic.net. 2020. *What Is A Mail Server And How Does It Work? (Article)*. [online] Available at: <https://www.samlogic.net/articles/mail-server.htm> [Accessed 5 November 2020].

smtp mail server - professional SMTP service provider. 2020. *What Is SMTP - Smtp Mail Server - Professional SMTP Service Provider*. [online] Available at: <https://www.serversmtp.com/what-is-smtp/> [Accessed 5 November 2020].

Hjp.at. 2020. *Hjp: Doc: RFC 1939: Post Office Protocol - Version 3*. [online] Available at: <https://www.hjp.at/doc/rfc/rfc1939.html> [Accessed 5 November 2020].

BevSites Help Center. 2020. *IMAP Vs. POP3*. [online] Available at: <https://bevmedia.freshdesk.com/support/solutions/articles/3000083258-imap-vs-pop3> [Accessed 6 November 2020].

En.wikipedia.org. 2020. *Naive Bayes Spam Filtering*. [online] Available at: <https://en.wikipedia.org/wiki/Naive_Bayes_spam_filtering> [Accessed 10 November 2020].

Leung, K., 2007. Naive Bayesian Classifier. [online] Cis.poly.edu. Available at: <http://cis.poly.edu/~mleung/FRE7851/f07/naiveBayesianClassifier.pdf> [Accessed 10 November 2020].

Dada, E.G., Bassi, J.S., Chiroma, H., Abdulhamid, S.M., Adetunmbi, A.O. and Ajibuwa, O.E. (2019). Machine learning for email spam filtering: review, approaches and open research problems. Heliyon, 5(6), p.e01802. Available at: < https://www.sciencedirect.com/science/article/pii/S2405844018353404> [Accessed 12 November 2020].

Insights, S., 2020. *Machine Learning: What It Is And Why It Matters*. [online] Sas.com. Available at: <https://www.sas.com/en_ie/insights/analytics/machine-learning.html> [Accessed 7 November 2020].

Medium. 2021. Logistic Regression — Detailed Overview. [online] Available at: <https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc> [Accessed 7 November 2020].

Subramaniam, Thamarai & Jalab, Hamid & Taqa, Alaa. (2010). Overview of textual anti-spam filtering techniques. International Journal of the Physical Sciences. 5. 1869-1882. [Accessed 9 November 2020].

Bhowmick, A., & Hazarika, S.M. (2016). Machine Learning for E-mail Spam Filtering: Review, Techniques and Trends. ArXiv, abs/1606.01042. [Accessed 9 November 2020].

Knafve, M., 2020. *Hmailserver - Free Open Source Email Server For Microsoft Windows*. [online] Hmailserver.com. Available at: <https://www.hmailserver.com/> [Accessed 10 November 2020].

Knafve, M., 2020. *Connect To Mysql - Hmailserver - Free Open Source Email Server For Microsoft Windows*. [online] Hmailserver.com. Available at: <https://www.hmailserver.com/documentation/v5.4/?page=howto_connect_to_mysql> [Accessed 10 November 2020].

Knafve, M., 2020. *Set Up Local / Stand-Alone Server - Hmailserver - Free Open Source Email Server For Microsoft Windows*. [online] Hmailserver.com. Available at: <https://www.hmailserver.com/documentation/v5.6/?page=howto_set_up_local> [Accessed 10 November 2020].

James.apache.org. 2020. *Apache James Server 3.0 - Apache James Server 3 - Installation*. [online] Available at: <https://james.apache.org/server/3/install.html> [Accessed 10 November 2020].

Mysql.com. 2020. *Mysql :: Mysql Community Edition*. [online] Available at: <https://www.mysql.com/products/community/> [Accessed 10 November 2020].

Scikit-learn.org. 2020. *Scikit-Learn: Machine Learning In Python — Scikit-Learn 0.23.2 Documentation*. [online] Available at: <https://scikit-learn.org/stable/> [Accessed 10 November 2020].

Medium. 2020. *Top Python Libraries: Numpy & Pandas*. [online] Available at: <https://towardsdatascience.com/top-python-libraries-numpy-pandas-8299b567d955> [Accessed 10 November 2020].

Medium. 2020. *Spam Classifier In Python From Scratch*. [online] Available at: <https://towardsdatascience.com/spam-classifier-in-python-from-scratch-27a98ddd8e73> [Accessed 10 November 2020].

Flask.palletsprojects.com. 2020. *Foreword — Flask Documentation (1.1.X)*. [online] Available at: <https://flask.palletsprojects.com/en/1.1.x/foreword/> [Accessed 10 November 2020].

Fullstackpython.com. 2020. *Object-Relational Mappers (Orms)*. [online] Available at: <https://www.fullstackpython.com/object-relational-mappers-orms.html> [Accessed 10 November 2020].

Bhatia, R., 2020. *Why Do Data Scientists Prefer Python Over Java?*. [online] Analytics India Magazine. Available at: <https://analyticsindiamag.com/why-do-data-scientists-prefer-python-over-java/#:~:text=In%20terms%20of%20toolset%2C%20Java,of%20concurrency%2C%20Java%20beats%20Python.> [Accessed 10 November 2020].