Email Honey Pot - Trap and Classify

Adrian Baron-Hyppolite
adrianb@vt.edu
Virginia Tech
USA

Prateek Sethi
prateek20@vt.edu
Virginia Tech
USA

ABSTRACT

In this project, we aim to create an email honeypot to lure spammers. We use expired domains older than seven years to redirect and collect legit spam. We create multiple accounts and register them on platforms and websites belonging to various categories such as news, sports, social media, and shopping. We collected this data and made a real dataset that can be further expanded and used for classification studies. Additionally, using this data we train different classifiers and compare their performance for spam detection.

KEYWORDS

email honeypots, MX records, SPF, machine learning, classification

ACM Reference Format:
Adrian Baron-Hyppolite and Prateek Sethi. 2022. Email Honey Pot - Trap and Classify. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Email is the most common means of professional communication, but it is not limited to it. Emails are used to share information, documents, and files. They are also used by companies for promotion and marketing purposes as well as for surveys. This leads to a lot of unnecessary emails being generated known as spam. However, electronic mail service was not built with security as one of the pillars of design. This leads to opportunities for malicious attackers to gain access to private or confidential information.

As per Cisco, Spam email is unsolicited and unwanted junk email sent out in bulk to an indiscriminate recipient list[2]. Many times attackers embed links and scripts in these spams in order to fulfill their hostile intents. This makes it very important to identify and segregate spam. In this study, we aim to create an email honey pot using two techniques. The data that is collected is used to train three classifiers in order to accurately identify spam. The concept of honeypots has been used in different use cases including cybersecurity, cloud services, and on-premise infrastructure. It is mainly used to reduce the damage from malicious agents as well as for gaining information about attacks and vulnerabilities.

The first approach - MX Honeypot uses old expired domains, where the traffic of the expired domains is redirected to our honeypot. The second approach - Email Honeypot is based on interaction with the most popular websites belonging to various categories such as sports, news, social media, and shopping.

The MX honeypot builds on the concepts of email delivery. How the sender email service looks for the MX records in order to send the email to the receiver. The MX records consist of a list of mail servers that host the destination email address. In addition to this once this email reached the mail server. The mail server queries for the TXT records in order to get information about SPF and other security measures discussed in section-2

Spam detection can be passed as a simple classification problem where the algorithm has to identify based on the content of the email, if the message is spam or if it is not spam. Machine learning has been extensively used for solving classification problems. This study uses 3 different models in order to make intelligent decisions about whether the email is spam or not. These methods are discussed in detail in section-2. The code and the dataset can be accessed on github [8]

The rest of the paper is organized as follows; section-2 gives an overview of all the concepts used in the study. section-3 describes in detail the 2 approaches and the classification techniques. section-4 section describes our observations, section-5 section mentions some interesting finds. the section-6 section touches upon how this work can be expanded. Finally, section-7 section summarises the study.

2 BACKGROUND

Honeypots prove to be an effective way to deter malicious entities looking to deploy various phishing attacks using spam. In this project, an email Honeypot is created by directing the MX records of an expired domain to an email we created specifically to detect spam. Moreover, an additional four emails were created. The purpose of these emails was to subscribe to various websites stretching over the following categories: Social Networking and communities, News and Media, Sports, and Marketplace. This would, in turn, give an idea as to how often these websites sold user information.

Moreover, once these addresses began receiving emails, we collected them and put them through various spam classifiers that would inform us whether an email was spam or not. The email classifiers chosen were: The Multinomial naive Bayes model, the Bernoulli Naive Bayes model, the SVM model, and vectorization. The section will go as follows a breakdown of the records associated with email sending and receiving, a brief overview of critical email security concepts, Honeypots, and finally spam classifiers.

2.1 DNS Records

The Domain name system is a key element in internet communication. It helps identify the source and destination for communication.

Conference'17, July 2017, Washington, DC, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

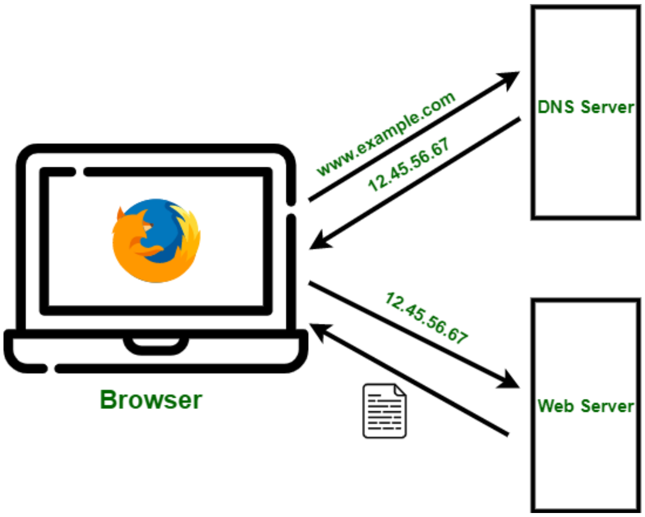


Figure 1: Accessing DNS records [4]

DNS is used as the base for many frameworks and mechanisms related to email. There are dozens of DNS records that have different purposes. Some of them are described below -

2.1.1 A records. A Records are responsible for saving IP addresses after they have been translated from domain names. Following the transition to IPv6, A records will become AAAA records; this is responsible for mapping IPv6 addresses with domain names as mentioned in [15].

2.1.2 MX records. The Mail Exchange records are responsible for directing emails to the proper server to which they belong to. In this project, we redirected the MX records of an expired domain to identify incoming spam easily.

2.1.3 TXT records. These records are responsible for holding all text-based information belonging to the outside domain for the configured domain. This helps identify ownership. Through the use of SPF, TXT records can publish on authorized mail servers. Unfortunately, due to the TXT record's ability to store all text-based information, it is susceptible to DNS tunneling. DNS tunneling occurs when a malicious user injects an unrelated stream of information during a DNS query.

2.2 Email Security

2.2.1 SPF. The Sender Policy Framework is a mechanism that can check incoming IP addresses to determine whether or not they can use a sender's domain. This mechanism can be used to stop phishing attacks. SPF has three major components, as stated in 2: specialized headers, an authentication technique, and the policy framework.

2.2.2 DKIM. Domain Keys Identified email is a key that helps authorize whether an email was sent by the approved owner of a particular domain. This is done through the use of a digital signature. This signature tips off the owner to the fact that this email is to be trusted and the message has not been altered in any way by a malicious user, providing guaranteed data integrity.

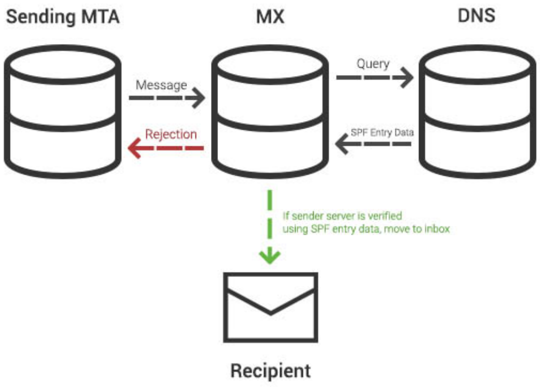

Sending MTA
MX
DNS
Message
Rejection
Query
SPF Entry Data
If sender server is verified
using SPF entry data, move to inbox
Recipient

Figure 2: Email sender authentication process [14]

2.2.3 DMARC. In an attempt to block domain and email spoofing, Domain-based Message Authentication Reporting and Conformance was created. This is an authentication standard meant to use the DNS and SPF framework to verify email senders.

2.3 Email Honeypot

2.3.1 Honeypot. A Honeypot is a cyber defense mechanism that functions by luring attackers into a designated space where one can identify and address malicious attacks. An attacker is lured into thinking they have infiltrated their desired system only to have just exposed sensitive information about their identity. Some papers have explored using Honeypot data to apply geospatial tools like the one mentioned in [5]. This would effectively expose an attacker by giving away their location. This tool would also be useful in determining where a majority of spam comes from.

2.3.2 Honeypot for emails. An email Honeypot functions by directing spam into a designated email folder to detect phishing attacks and expose the identities of these attackers. In this project, an expired domain's MX records were directed to an email we created that served as our honeypot for detecting spam.

2.4 Spam Classification

Nowadays, many email services such as Gmail, Yahoo, etc provide spam detection functionalities. It is reasonable to say that spam detection is a necessary part of the electronic mail ecosystem. [1] shows that 14.5 billion spam emails are sent every single day, earning their senders a daily average of 7,000 dollars. According to the same research, 46 % of emails are spam.

Many different machine-learning approaches have been used to tackle the problem of spam detection [3]. In this study, we have used the most commonly used classifiers - Multinomial naive Bayes, Bernoulli Naive Bayes, and Support Vector Machine (SVM). They are discussed in more detail below:

2.4.1 Naive Bayes. The Bayes classifier assigns the given example to the most closely related class based on the features of the given example. It makes an assumption that all the features are independent of each other. This is not true in most cases, but the classifier still performs reasonably well [9]. Naive Bayes is used extensively for text classification [11]. Hence is a good fit for spam detection. For our use case, we tested the Multinomial naive Bayes [12] and the Bernoulli Naive Bayes classifier [13]

2.4.2 SVM. The main objective of the support vector machine algorithm is finding a hyperplane in an N-dimensional space, where N is the number of features. This hyperplane should divide the space such that it distinctly classifies the data points.

3 METHODOLOGY

For this project, we used two techniques to create honeypots. For the first technique, we created email accounts and exposed them to a different categories of websites. For the second approach, we redirected the emails intended for all the usernames of an expired domain name. These emails were forwarded to the MX honeypot email address.

3.1 Project setup

Four email accounts were created fitting the four categories chosen for this project; for each category, five sites were selected as seen in table 1. The purpose of registering with these websites was to examine spam and determine whether or not these websites were selling information to third parties. Social sites were of particular interest due to the fact that many of these sites use social engineering tactics to attract users. We figured that companies that are heavy on collecting user data would be more inclined to sell this information to third parties.

3.2 Expired Domain Procurement

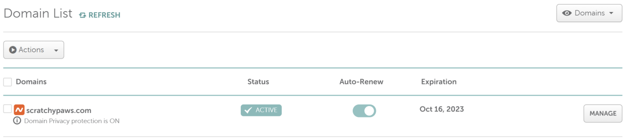
In order to create the MX honeypot we are required to acquire domains that were created a long time ago but are not being used currently. These can be expired domains or abandoned domains. As suggested in the study [10] many of the most popular TLDs in target embedding certificates are those that can be registered for free: .ga, .ml, .cf, .tk, and .gq. We tried to acquire these domains for free but the process was not straightforward and mostly led us to paid websites to acquire these domains.

In order to overcome this issue, we bought a domain directly from NameCheap [7]: which is an ICANN-accredited domain name registrar providing domain name registration and web hosting based in Phoenix, Arizona, US. The domain is called scratchpaws.com and was created 7 years ago. This domain has been idle for more than a year and was available for a reasonable price.

Once we purchased the domain the ownership was transferred to us Fig-3 and we were given access to a dashboard Fig-4 that had information on the domain and also gave us access to the DNS records of the domain. Through this dashboard, we can manage DNS options, security solutions, and other products and services.

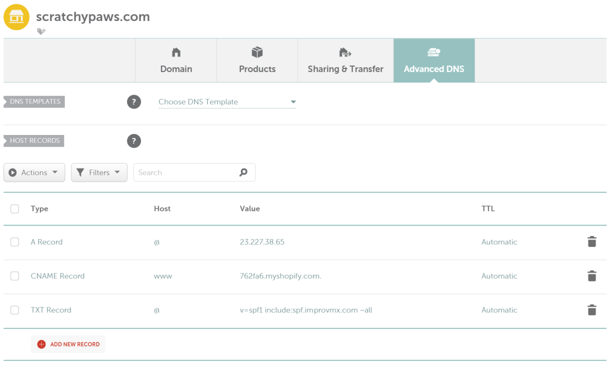
3.3 Redirecting Emails

The next challenge is to redirect all the emails that are coming to this domain to a central location. We created a new email address - 'emailhoneypottest@gmail.com' and directed all the emails to it. The process of enabling email forwarding is described below.



Domains	Status	Auto-Renew	Expiration
scratchypaws.com Domain Privacy protection is ON	✓ ACTIVE	●	Oct 16, 2023

Figure 3: Procuring domain scratchypaws.com



Type	Host	Value	TTL
A Record	@	23.227.38.65	Automatic
CNAME Record	www	762fa6.myshopify.com.	Automatic
TXT Record	@	v=spf1 include:spf.improvmx.com ~all	Automatic

Figure 4: namecheap Dashboard

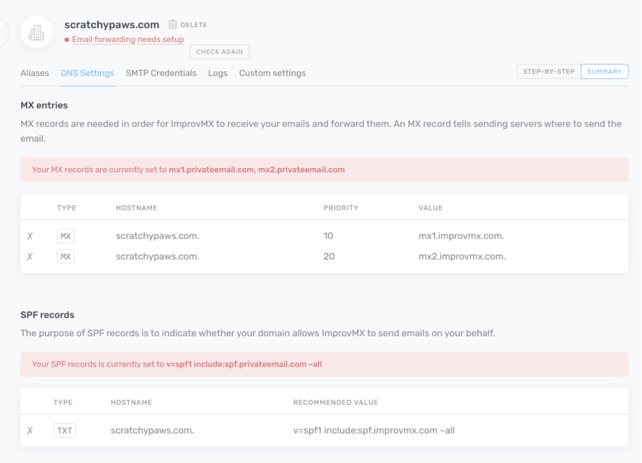
Due to the limitation of resources and the scope of the study, we did not host our own Domain name System and we relied on the DNS of a third-party platform called ImprovMX. This service provides us a Mail exchange server (MX) for redirecting the emails meant for a domain. This is done by updating the MX records and the TXT records on the DNS server for our domain. This information is used for establishing the SMTP connection between the sender and the receiver mail server and also for SPF.

After registering on ImprovMX we are prompted to update the MX records as well as the TXT records. These records are available through the domain name registrar. In our case it is NameCheap. We refer to the dashboard Fig-4 in order to navigate to these records. These records can be updated from the dashboard. Before the registration is completed we can see a warning on the ImprovMX portal as in Fig-5. The MX records and the TXT records need to be updated with the registrar which is Namecheap. This can be done through the dashboard Fig-4. Once the information is updated the MX records point to the email solver of ImprovMX. The email solver of ImprovMX can be configured to transfer all the emails to any email account. We used the email address - 'emailhoneypottest@gmail.com'. ImprovMX allows us to create Alias or usernames that are forwarded or special characters can be used to specify the alias. We use the $* special character which instructs ImprovMX to send all emails to any username on the domain scratchypaws.com to our email address. This can be seen in Fig-7.

3.3.1 Verification. For verification purposes, we sent an email from our personal account to random usernames on the scratchypaws domain - abc@scratchypaws.com. These usernames do not exist. Hence the delivery of these emails to our MX honeypot indicated that the emails have been successfully redirected. Fig-8 shows the received email in the MX honeypot which was entered into a random email abc@scratchypaws.com.

Table 1: Email Addresses used

Email Address	Website Category	Websites interacted with
adamap.social@gmail.com	Social Networking	Facebook, Twitter, LinkedIn, discord, ticktok
adamap.news@gmail.com	News and Media	CNN, HuffPost, NYT, BBC, Fox News
adamap.sports@gmail.com	Sports	ESPN, The Athletic, Foxsports, NBA, Sky Sports
adamap.shop@gmail.com*	Marketplace	Amazon, Target, Best Buy, Ebay, Macy's



● Email forwarding needs setup

MX entries

MX records are needed in order for ImprovMX to receive your emails and forward them. An MX record tells sending servers where to send the email.

Your MX records are currently set to **mx1.privateemail.com, mx2.privateemail.com**

	TYPE	HOSTNAME	PRIORITY	VALUE
✗	MX	scratchypaws.com.	10	mx1.improvmx.com.
✗	MX	scratchypaws.com.	20	mx2.improvmx.com.

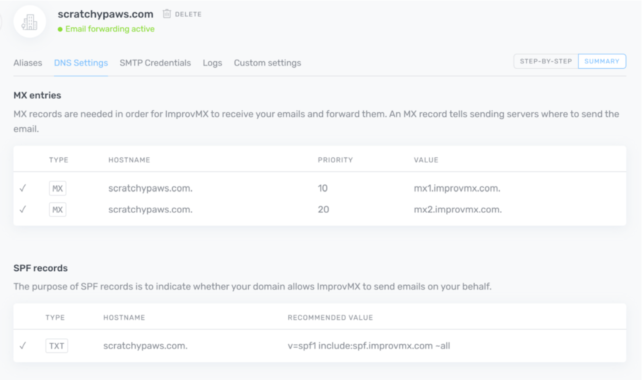
SPF records

The purpose of SPF records is to indicate whether your domain allows ImprovMX to send emails on your behalf.

Your SPF records is currently set to **v=spf1 include:spf.privateemail.com ~all**

	TYPE	HOSTNAME	RECOMMENDED VALUE
✗	TXT	scratchypaws.com.	v=spf1 include:spf.improvmx.com ~all

Figure 5: ImprovMX: Before updating MX and TXT records



● Email forwarding active

MX entries

MX records are needed in order for ImprovMX to receive your emails and forward them. An MX record tells sending servers where to send the email.

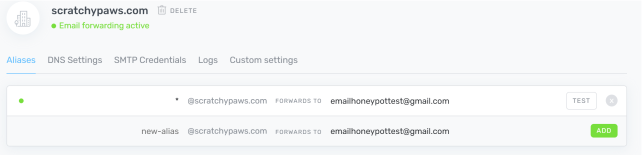
	TYPE	HOSTNAME	PRIORITY	VALUE
✓	MX	scratchypaws.com.	10	mx1.improvmx.com.
✓	MX	scratchypaws.com.	20	mx2.improvmx.com.

SPF records

The purpose of SPF records is to indicate whether your domain allows ImprovMX to send emails on your behalf.

	TYPE	HOSTNAME	RECOMMENDED VALUE
✓	TXT	scratchypaws.com.	v=spf1 include:spf.improvmx.com ~all

Figure 6: ImprovMX: After updating MX and TXT records



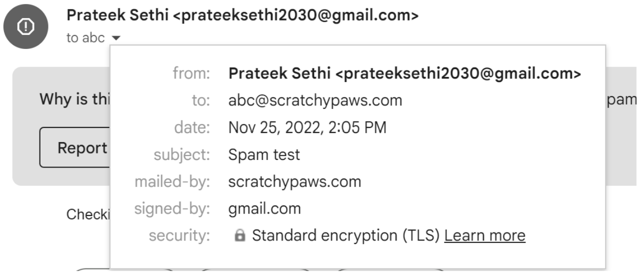
● Email forwarding active

●	*	@scratchypaws.com	FORWARDS TO	emailhoneypottest@gmail.com
	new-alias	@scratchypaws.com	FORWARDS TO	emailhoneypottest@gmail.com

Figure 7: ImprovMX: Setting Alias

3.4 Data Collection

Once the two setups were created we let the setup run from 16th October 2022 to 1st December 2022. The data from these two setups



Prateek Sethi <prateeksethi2030@gmail.com>
to abc

from:	Prateek Sethi <prateeksethi2030@gmail.com>
to:	abc@scratchypaws.com
date:	Nov 25, 2022, 2:05 PM
subject:	Spam test
mailed-by:	scratchypaws.com
signed-by:	gmail.com
security:	Standard encryption (TLS) Learn more

Figure 8: Verificarion of MX Honeypot

was then exported as CSV files. We were able to collect 237 emails in total from the 5 email accounts that we created. The breakup of the emails is given in the table-2 and table-9.

Due to the shortage of time, we were not able to collect sufficient data in the various honeypots. In order to train an efficient classifier for spam detection we require much more data. We used data from Kaggle [6] an online platform for data scientists and machine learning practitioners. We downloaded an email dataset consisting of 5730 rows and 2 columns. We used this data for training our Machine learning models.

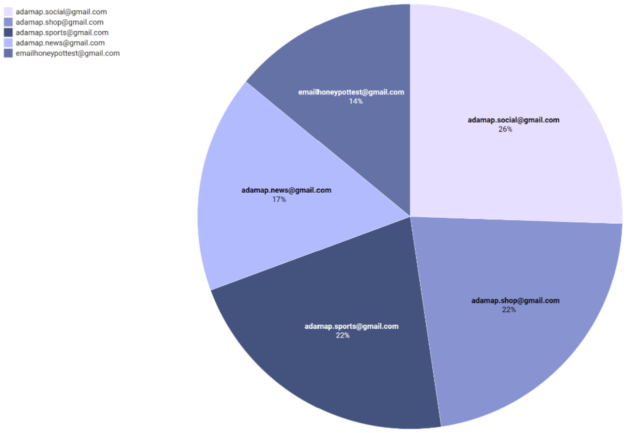


Figure 9: Proportion of mails per category

3.5 Model Training

Using the collected data and the dataset from Kaggle we trained 3 machine learning models; Multinomial naive Bayes, Bernoulli Naive

Table 2: Email Data Collected

Email address	Setup	Count
adamap.social@gmail.com	Email Honeypot	60
adamap.news@gmail.com	Email Honeypot	39
adamap.sports@gmail.com	Email Honeypot	51
adamap.shop@gmail.com	Email Honeypot	52
emailhoneypottest@gmail.com	MX Honeypot	33

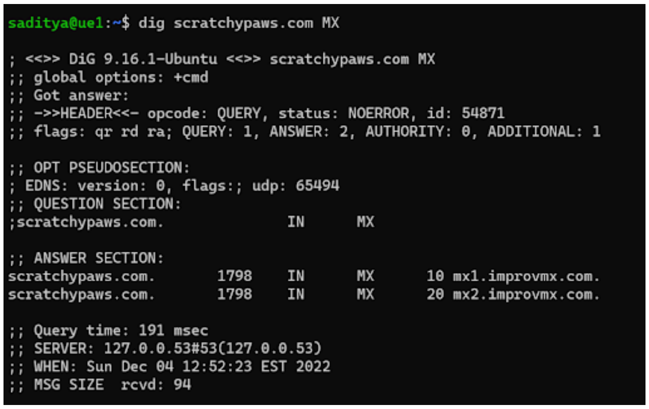
Bayes, and Support Vector Machines (SVM). The observations and results are described in section-4

4 RESULTS

We were able to successfully set up the honeypot using both techniques MX Honeypot - We acquired an expired domain and redirected all the email traffic from that domain to our email address. Email Honeypot - The accounts were successfully created and registered to websites belonging to popular categories.

For the classification of spam, all three classifiers performed really well. In order to compare the performance we noted the Precision, recall, and the f1-score for each model. Precision is the ratio of TP (True Positives) to the sum of TP and FP(False Positives). whereas, Recall is the ratio of TP to the sum of TP and FN(False Negatives). If a true spam email is wrongly identified as a real email, that is False Positive. On the other side, if a real email is identified as spam, that is False Negative.

The performance of Multinomial naive Bayes, Bernoulli Naive Bayes, and Support Vector Machines (SVM) can be seen in table-3, table-4 and table-5 respectively. The Multinomial naive Bayes classifier has the highest R^2−Score whereas SVM has the lowest R^2−score amongst the 3 as seen in table-6.

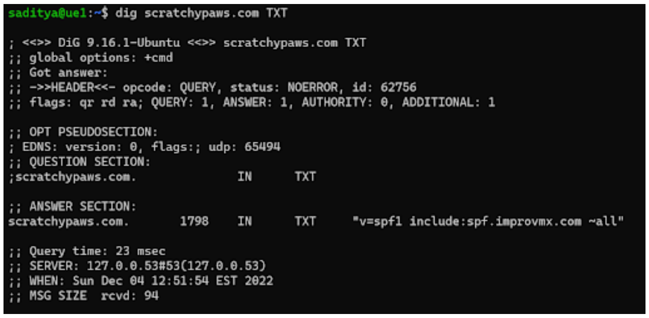


```
saditya@ue1:~$ dig scratchypaws.com MX

; <<>> DiG 9.16.1-Ubuntu <<>> scratchypaws.com MX
;; global options: +cmd
;; Got answer:
;; ->>HEADER<<- opcode: QUERY, status: NOERROR, id: 54871
;; flags: qr rd ra; QUERY: 1, ANSWER: 2, AUTHORITY: 0, ADDITIONAL: 1

;; OPT PSEUDOSECTION:
; EDNS: version: 0, flags:; udp: 65494
;; QUESTION SECTION:
;scratchypaws.com.              IN      MX

;; ANSWER SECTION:
scratchypaws.com.       1798    IN      MX      10 mx1.improvmx.com.
scratchypaws.com.       1798    IN      MX      20 mx2.improvmx.com.

;; Query time: 191 msec
;; SERVER: 127.0.0.53#53(127.0.0.53)
;; WHEN: Sun Dec 04 12:52:23 EST 2022
;; MSG SIZE  rcvd: 94
```

Figure 10: Using Dig command to verify MX record

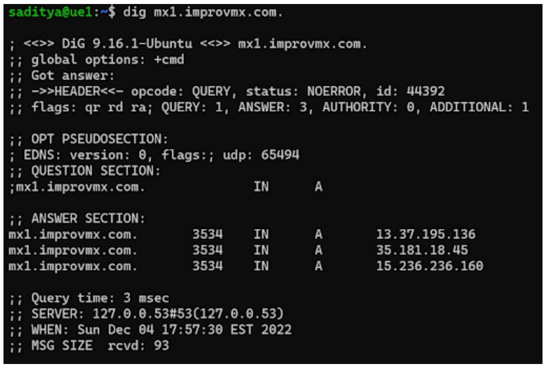
5 DISCUSSIONS

As mentioned previously, we could redirect the MX records of an expired domain to our email server. This effectively sent all incoming emails meant for the expired domain to the honeypot. We also had many other emails attached to various categories of websites in which we could test our spam classifiers and demonstrate their performance successfully. Moreover, ideally, with more time, far more spam would’ve been sent to our email honeypot, allowing us to test these classifiers on our data. As seen through our results,



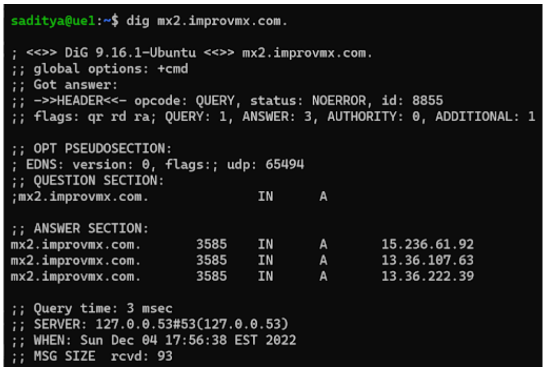
```
saditya@ue1:~$ dig scratchypaws.com TXT

; <<>> DiG 9.16.1-Ubuntu <<>> scratchypaws.com TXT
;; global options: +cmd
;; Got answer:
;; ->>HEADER<<- opcode: QUERY, status: NOERROR, id: 62756
;; flags: qr rd ra; QUERY: 1, ANSWER: 1, AUTHORITY: 0, ADDITIONAL: 1

;; OPT PSEUDOSECTION:
; EDNS: version: 0, flags:; udp: 65494
;; QUESTION SECTION:
;scratchypaws.com.              IN      TXT

;; ANSWER SECTION:
scratchypaws.com.       1798    IN      TXT     "v=spf1 include:spf.improvmx.com ~all"

;; Query time: 23 msec
;; SERVER: 127.0.0.53#53(127.0.0.53)
;; WHEN: Sun Dec 04 12:51:54 EST 2022
;; MSG SIZE  rcvd: 94
```

Figure 11: Using Dig command to verify TXT record



```
saditya@ue1:~$ dig mx1.improvmx.com.

; <<>> DiG 9.16.1-Ubuntu <<>> mx1.improvmx.com.
;; global options: +cmd
;; Got answer:
;; ->>HEADER<<- opcode: QUERY, status: NOERROR, id: 44392
;; flags: qr rd ra; QUERY: 1, ANSWER: 3, AUTHORITY: 0, ADDITIONAL: 1

;; OPT PSEUDOSECTION:
; EDNS: version: 0, flags:; udp: 65494
;; QUESTION SECTION:
;mx1.improvmx.com.              IN      A

;; ANSWER SECTION:
mx1.improvmx.com.       3534    IN      A       13.37.195.136
mx1.improvmx.com.       3534    IN      A       35.181.18.45
mx1.improvmx.com.       3534    IN      A       15.236.236.160

;; Query time: 3 msec
;; SERVER: 127.0.0.53#53(127.0.0.53)
;; WHEN: Sun Dec 04 17:57:30 EST 2022
;; MSG SIZE  rcvd: 93
```

Figure 12: A Record: Mail server 1



```
saditya@ue1:~$ dig mx2.improvmx.com.

; <<>> DiG 9.16.1-Ubuntu <<>> mx2.improvmx.com.
;; global options: +cmd
;; Got answer:
;; ->>HEADER<<- opcode: QUERY, status: NOERROR, id: 8855
;; flags: qr rd ra; QUERY: 1, ANSWER: 3, AUTHORITY: 0, ADDITIONAL: 1

;; OPT PSEUDOSECTION:
; EDNS: version: 0, flags:; udp: 65494
;; QUESTION SECTION:
;mx2.improvmx.com.              IN      A

;; ANSWER SECTION:
mx2.improvmx.com.       3585    IN      A       15.236.61.92
mx2.improvmx.com.       3585    IN      A       13.36.107.63
mx2.improvmx.com.       3585    IN      A       13.36.222.39

;; Query time: 3 msec
;; SERVER: 127.0.0.53#53(127.0.0.53)
;; WHEN: Sun Dec 04 17:56:38 EST 2022
;; MSG SIZE  rcvd: 93
```

Figure 13: A Record: Mail server 2

Table 3: Classifier performance: Multinomial naive Bayes

	Precision	Recall	f1-Score
0	1.00	0.98	0.99
1	0.95	0.99	0.97
Average	0.976	0.976	0.976

Table 4: Classifier performance: Bernoulli naive Bayes

	Precision	Recall	f1-Score
0	0.99	0.99	0.99
1	0.96	0.96	0.96
Average	0.976	0.976	0.976

Table 5: Classifier performance: SVM

	Precision	Recall	f1-Score
0	0.95	1.00	0.97
1	0.98	0.84	0.91
Average	0.967	0.917	0.939

Table 6: Classifier performance R2-Score

Classifier	R^2−Score
Multinomial Naive Bayes	0.985
Bernoulli Naive Bayes1	0.982
SVM	0.975

SVM produces the lowest performance results of each classifier, while Bernoulli and Multinomial produce identical precision, recall, and f1-score results on average.

Following honeypot implementation tests using a series of classifiers showed which emails were classified as spam. Each classifier boasted great results that demonstrated remarkable accuracy of over ninety-five percent.

Honeypot data is precious as it can alert a user to where a majority of spam could come from the data pulled from our classifiers; we would be able to determine whether the data received is, in fact, spam. Once we have determined whether it is spam, we can apply other tools to it to choose the attack being deployed.

6 FUTURE WORKS

Future works for Honeypot detection should include implementing and creating more classifiers. As well as application on a much larger set of spam emails. Several works have used honeypot data to expose attacks on various networks further. For instance, in [5], they use amazon honeypot data to feed into a geospatial tool that can track a spammer's location. This is a beneficial study because it can provide crucial information on which countries and regions produce the most spam emails. Furthermore, once an attacker's location and identity are exposed, you can blacklist said attacker using honeypot data.

Furthermore, using honeypots can be extended beyond just stopping phishing attacks and preventing DDOS attacks. In [16], They use honeypots to detect potential DDOS attacks. The critical component of this honeypot is the intrusion prevention system(IPS); this tool essentially functions as a protective layer that handles all malicious emails with DDOS potential.

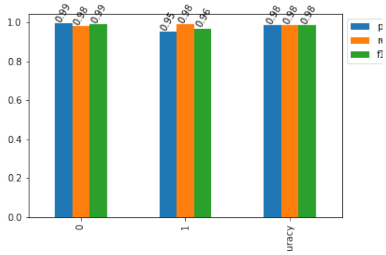


Figure 14: Performance of Multibilomian Naive Bayes

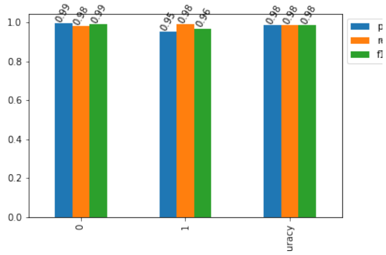


Figure 15: Performance of Bernoulli Naive Bayes

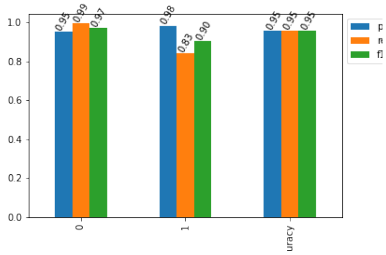


Figure 16: Performance of SVM

7 CONCLUSION

Attacks on email systems are common; malicious agents constantly attempt phishing attacks and even DDOS attacks. It is for this reason that honeypots are very valuable. Honeypots are a great way to mitigate and expose active threats that are present in a network by luring spammers and other malicious agents into attacking a system they think is legitimate. In our project, we created a honeypot, redirected an expired domain's MX records to the honeypot, and ran a series of classifiers on our honeypot data set.

Through this project, we gained a solid foundational understanding of how various components of email security work. For example, we not only fortified our knowledge of A records, MX records, and TXT records, but we also got hands-on experience manipulating them, specifically MX records. We were able to redirect the MX records of an expired domain to our honeypot test email. Furthermore, we then explored several classifiers: the Multinomial Naive Bayes, SVM, and Bernoulli. These classifiers use machine learning to discover characteristics of an email that mimic that of a spam email. We were able to successfully implement these classifiers and gauge their performance.

REFERENCES

[1] Emily Bauer. 2022. 15 Outrageous Email Spam Statistics that Still Ring True in 2018. https://www.propellercrm.com/blog/email-spam-statistics [Online; accessed 5. Dec. 2022].

[2] Cisco. 2022. What is Spam Email? https://www.cisco.com/c/en/us/products/security/email-security/what-is-spam.html [Online; accessed 7. Dec. 2022].

[3] Emmanuel Gbenga Dada, Joseph Stephen Bassi, Haruna Chiroma, Shafi'i Muhammad Abdulhamid, Adebayo Olusola Adetunmbi, and Opeyemi Emmanuel Ajibuwa. 2019. Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon* 5, 6 (June 2019), e01802. https://doi.org/10.1016/j.heliyon.2019.e01802

[4] GeeksforGeeks. 2022. Working of Domain Name System (DNS) Server - GeeksforGeeks. https://www.geeksforgeeks.org/working-of-domain-name-system-dns-server [Online; accessed 5. Dec. 2022].

[5] IEEE. 2022. IEEE Xplore Full-Text PDF:. https://ieeexplore-ieee-org.ezproxy.lib.vt.edu/stamp/stamp.jsp?tp=&arnumber=9377145&tag=1 [Online; accessed 5. Dec. 2022].

[6] Kaggle. 2022. spam email detection dataset. https://www.kaggle.com/datasets/studymart/spam-email-detection-dataset [Online; accessed 7. Dec. 2022].

[7] NamesCheap. 2022. Buy a domain name - Register cheap domain names from $0.99 - Namecheap. https://www.namecheap.com [Online; accessed 6. Dec. 2022].

[8] Adrian Baron-Hyppolite Prateek Sethi. 2022. Email Honeypot code and dataset. https://github.com/Pseth3/emailhoneypot [Online; accessed 7. Dec. 2022].

[9] Irina Rish. 2001. An Empirical Study of the Naïve Bayes Classifier. *IJCAI 2001 Work Empir Methods Artif Intell* 3 (01 2001).

[10] Richard Roberts, Yaelle Goldschlag, Rachel Walter, Taejoong Chung, Alan Mislove, and Dave Levin. 2019. You Are Who You Appear to Be: A Longitudinal Study of Domain Impersonation in TLS Certificates. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security* (London, United Kingdom) *(CCS '19)*. Association for Computing Machinery, New York, NY, USA, 2489–2504. https://doi.org/10.1145/3319535.3363188

[11] stanford. 2009. Naive Bayes text classification. https://nlp.stanford.edu/IR-book/html/htmledition/naive-bayes-text-classification-1.html [Online; accessed 6. Dec. 2022].

[12] stanford. 2009. Relation to multinomial unigram language model. https://nlp.stanford.edu/IR-book/html/htmledition/relation-to-multinomial-unigram-language-model-1.html [Online; accessed 6. Dec. 2022].

[13] stanford. 2009. The Bernoulli model. https://nlp.stanford.edu/IR-book/html/htmledition/the-bernoulli-model-1.html [Online; accessed 6. Dec. 2022].

[14] taguchi.com. 2022. What Are Spf, Dkim, Dmarc & Bimi? - Taguchi Support. https://support.taguchi.com.au/knowledge-base/tracking-and-reporting/deliverability/what-are-spf-dkim-dmarc-bimi.html [Online; accessed 5. Dec. 2022].

[15] Rebekah Taylor. 2021. Know the eight most common DNS records – BlueCat Networks. *BlueCat Networks* N/A (April 2021), N/A. https://bluecatnetworks.com/blog/know-the-eight-most-common-dns-records

[16] R. Venkatesan, G. Ashwin Kumar, and M. Ragu Nandhan. 2018. *A NOVEL APPROACH TO DETECT DDOS ATTACK THROUGH VIRTUAL HONEYPOT.* IEEE, N/A. 1–6 pages. https://doi.org/10.1109/ICSCAN.2018.8541209