N°d'ordre :.....
N°de série :.....

RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITE ECHAHID HAMMA LAKHDAR - EL OUED
FACULTY OF EXACT SCIENCES
Computer Science department

ACADEMIC MASTER

Domain: **Mathematics and Computer Science**
Industry: **Computer Science**
Specialty: **Distributed Systems and Artificial Intelligence**

Theme

Spam Detection Using Machine Learning

Presented by:

- **LABRECHE RIDHA**
- **LAOUAR MOHMMED BACHIR**

Sustained on 20-06 2021 from the jury:

BENALI ABDELKAML	MCA	President
NIDIOUI MOHAMMED ABDELHAMID	MAA	Reporter
LEJDEL IBRAHIM	MCA	**Supervisor**

University Year: 2020/2021

Dedicated to

“

In the name of God, I thank him for guiding me to the straight path. I dedicate this humble work to:

To my dear parents MOHAMMED LAKHDAR and FATMA who were always beside me, and who gave me a great example of hard work and perseverance,

To my small family, my wife, my children, Khadija, Arwa, Suhaib, Abd al-Rahman,"Y A K E E N " ,

To my brothers and sisters, each in his name; For all family members

All my friends

To all my dear teachers who taught me

T h a n k s.

”

- LABRECHE RIDA

Dedicated to

“

In the name of God, I thank him for guiding me to the straight path. I dedicate this humble work to:

To my dear parents who were always beside me, and who gave me a great example of hard work and perseverance,

To my small family, my wife, my children, AYMEN, ABDELLAH, YaKeen , ***”Z A Y E D”***

To my brothers and sisters, each in his name; For all family members

All my friends

To all my dear teachers who taught me

T h a n k s.

”

- LAOUAR MOHAMMED BACHIR

Acknowledgements

First and foremost, I thank God Almighty for giving me the courage and patience to complete this task successfully, and I especially thank my en-cadrante**DR LEJDEL IBRAHIM** / NAOUI MOHAMMED ANOUAR for her competent assistance, patience, and encouragement. His criticism has been invaluable in the organization of work and the improvement of the quality of various services.

Also, I'd like to express my heartfelt gratitude to Mohamed Fouad, our dear friend, for his invaluable assistance in providing us with valuable information as well as encouragement. I hope the members of the jury will find here an expression of heartfelt gratitude for the honor they have bestowed upon us by taking the time to read and evaluate this work.

I'd also like to thank the educational team for their efforts in providing us with an excellent education. Abbas, and Khabash Muhibdin.

Finally, I'd like to thank everyone who helped make this project a reality, no matter how big or small.

Abstract

Email has developed into one of the most powerful and cost-effective means of communication. The rise in the number of email subscribers has resulted in a large spike in spam emails in recent years. Since spammers are still trying to find a way past the existing filters, it has become important to create new spam filters.

As part of this work, emails were classified using six important algorithms: (SVM, NB, RandomForest, XGBoost, Decision Tree and KNN).

The effectiveness of these algorithms is tested with different representation methods using the dataset and the best is selected to filter spam

---

Keywords : spam Emails, machine learning, Classification, Nave Bayes, SVM ,K Closest Neighbor,RandomForest, XGBoost, Decision Tree .

Résumé

Le courrier électronique est devenu l'un des moyens de communication les plus puissants et les plus rentables. L'augmentation du nombre d'abonnés aux e-mails a entraîné une forte augmentation des spams au cours des dernières années. Étant donné que les spammeurs tentent toujours de trouver un moyen de contourner les filtres existants, il est devenu important de créer de nouveaux filtres anti-spam.

Dans le cadre de ce travail, les e-mails ont été classés à l'aide de six algorithmes importants : (SVM, NB, RandomForest, XGBoost, Arbre de décision et KNN).

L'efficacité de ces algorithmes est testée avec différentes méthodes de représentation utilisant l'ensemble de données et la meilleure est sélectionnée pour filtrer le spam.

Mots clés : Pourriel, Algorithme d'apprentissage automatique, Machine à Vecteurs de Support, Naïve Bayes, K voisins les plus proches, RandomForest, XGBoost, Arbre de décision .

ملخص

أصبح البريد الإلكتروني أحد أقوى وسائل الاتصال وأكثرها فعالية من حيث التكلفة أدت الزيادة في عدد المشتركين في البريد الإلكتروني إلى زيادة حادة في الرسائل العشوائية في السنوات الأخيرة. ونظرًا لأن مرسلي البريد العشوائي يحاولون دائمًا إيجاد طريقة لتجاوز عوامل التصفية الحالية، فقد أصبح من المهم إنشاء عوامل تصفية جديدة

بشكل عام يعد تصنيف النص الجانب الأكثر أهمية في تصفية البريد الإلكتروني. يعرف المصنف على انه نظام يستخدم طرق التصنيف لتحديد ما إذا كانت الرسائل الواردة بري عشوائي أو بريد أصلي. تُستخدم خوارزميات التعلم الآلي كواحدة من أكثر أنظمة التصنيف كفاءة

في إطار هذا العمل ، تم تصنيف رسائل البريد الإلكتروني باستخدام ستة خوارزميات مهمة: آلة متجه الدعم , المصنف بايز ساذج, غابة عشوائية, شجرة قرار XGBoost, KNN

يتم اختبار كفاءة هذه الخوارزميات بطرق تمثيل مختلفة باستخدام مجموعة البيانات واختيار الاحسن منها لتصفية البريد العشوائي

كلمات مفتاحية : رسائل البريد الإلكتروني العشوائية , التصنيف , ساذج, غابة عشوائية, شجرة قرار,الة متجه الدعم

Contents

List of Figures

List of Tables

Liste of algorithms

NB	*Naive Bayes*
RF	*Random Forest*
KNN	*K-nearest neighbour KNN*
SVM	*Support Vector Machine*
DT	*Decision Tree*
XGBoos	*XGBoos*

General Introduction

Contexte

Electronic mail is one of the most widely used Internet services, and it is without a doubt the technology that has influenced our behavior on a large scale. The value of electronic mail is directly related to the development of the Internet, as many Web sites are now dedicated to it, and nearly everyone who has access to the Internet has at least one electronic address that they check on a regular basis, which explains the billions of emails sent and received every day. Electronic mail is also an important tool for consumers, providing a convenient and cost-effective way to share information. When we compare e-mail to other forms of communication (written, phone), we see that the benefits outweigh the drawbacks. His power lies in the process by which communications are transported, the speed at which emails circulate, the economy, the availability at any time, regardless of time zone, and their ability to deliver to multiple recipients at the same time. The informatized conception of these e-mails provides incomparable benefits, such as the submission of selected documents.The computerized design of these e-mails offers unparalleled advantages, such as submitting electronic documents by attachment, archiving of messages is much easier to achieve than with written or telephone correspondence, and e-mail makes it possible to "Take care of messages in a fast, reliable and automated way, such as keyword analysis, automatic sorting by subject."

Problem Statement

Users are easily inundated with irrelevant or unsolicited e-mails, also known as spam. Spam is quickly becoming a major concern on the internet. Spam is a massive global phenomenon. Spam is defined as follows by the CNIL (Commission Nationale de l'Informatique et des Libertés): "Spamming" or "spam" is the mass distribution of unsolicited e-mails to people with whom the author has never had contact and whose email address he has inadvertently captured. There were various anti-spam tactics.[15] The first entails alternatives based on email address en-têtes, such as black and white lists. The second type is those that focus on the textual content of the message, such as machine learning-based autofilter.

Objectives of the Study

There are several projects that use automatic learning approaches to solve the problem of spam filtering. Anti-spam filtering based on message textual content is an example of text triage, which involves categorizing textual documents into a set of predefined groups. The goal of a classification system is to perform classification tasks with an appropriate level of precision. There is now a very long list of classifiers that are based on algorithms. In the context of this thesis, a comparative analysis will be conducted using six learning algorithms: Support Vector Machine (SVM), Nave Bayes et K Plus Proche Voisin.RandomForest, XGBoost,Decision Tree . The main objective is to choose the best algorithm to classify the received emails.

Thesis Organization

Our study is divided into four sections:

The first chapter attempts to describe spam by explaining its purposes, content and implications and describes the different methods used to identify this form of email. The second chapter presents several published works on spam filtering, the thrid chapter present the prop sed approach ,the fourth chapter explains the architecture of the system, the methods used for deployment and the experimental results, which are presented with illustrative statistics to facilitate interpretation and comparison. of these results in order to choose the best Classifier.

Chapter 1

DEFINITIONS AND CONCEPTS

1.1 Introduction

Spam is a major issue for Internet consumers. Recent spikes in spam rates have caused widespread anger among internet users. Several alternatives to the dilemma have been proposed.

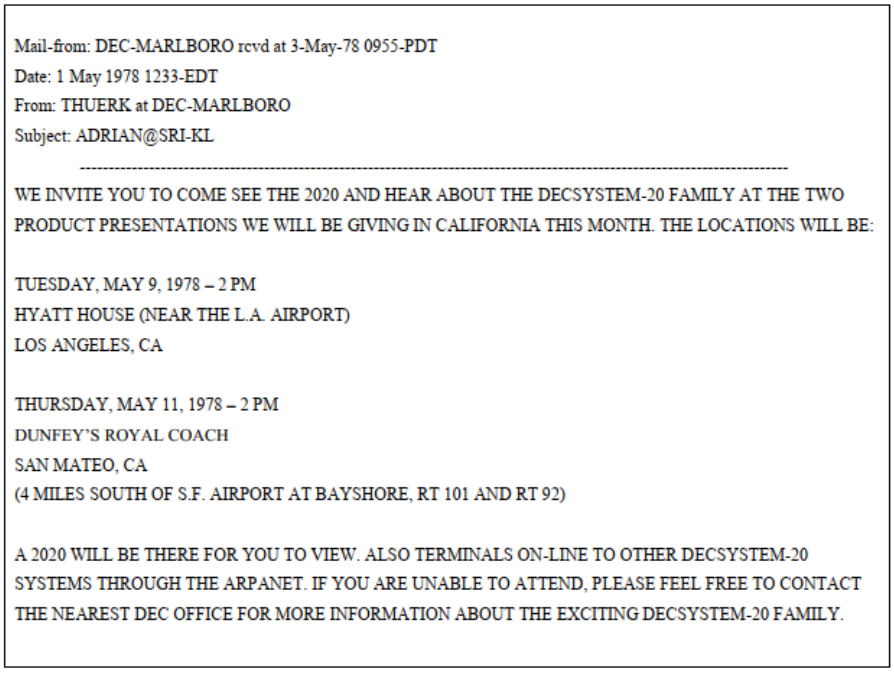
In this chapter, we will first discuss the origins, goals, content, and effects of spam, as well as the different strategies used to define this type of email.

1.2 What is Spam?

Spam is unsolicited electronic mail sent in bulk to a large number of recipients for advertising or malicious purposes.[14]

Spam can also refer to messages sent through other electronic means of communication, such as instant messages, blogs, forums, and more recently mobile networks, via SMS or MMS. Despite the differences in communication methods, the strategies for sending and detecting messages are strikingly similar. The first spam message was sent on May 3, 1978.(Figure 1.1) On that day, GaryThuerk, a DEC3 sales representative, sent an email to 393 people inviting them to see the new device.[20]

This advertisement was broadcast on Arpanet1. Email clients weren't as sophisticated as they are now. As a result, the spammers (spammers) had to enter all recipient addresses manually. Only 320 recipients received during the first broadcast due to lack of SNDMSG buffering space they were using.



Mail-from: DEC-MARLBORO rcvd at 3-May-78 0955-PDT
Date: 1 May 1978 1233-EDT
From: THUERK at DEC-MARLBORO
Subject: ADRIAN@SRI-KL

WE INVITE YOU TO COME SEE THE 2020 AND HEAR ABOUT THE DECSYSTEM-20 FAMILY AT THE TWO PRODUCT PRESENTATIONS WE WILL BE GIVING IN CALIFORNIA THIS MONTH. THE LOCATIONS WILL BE:

TUESDAY, MAY 9, 1978 – 2 PM
HYATT HOUSE (NEAR THE L.A. AIRPORT)
LOS ANGELES, CA

THURSDAY, MAY 11, 1978 – 2 PM
DUNFEY'S ROYAL COACH
SAN MATEO, CA
(4 MILES SOUTH OF S.F. AIRPORT AT BAYSHORE, RT 101 AND RT 92)

A 2020 WILL BE THERE FOR YOU TO VIEW. ALSO TERMINALS ON-LINE TO OTHER DECSYSTEM-20 SYSTEMS THROUGH THE ARPANET. IF YOU ARE UNABLE TO ATTEND, PLEASE FEEL FREE TO CONTACT THE NEAREST DEC OFFICE FOR MORE INFORMATION ABOUT THE EXCITING DECSYSTEM-20 FAMILY.

Figure 1.1: The first spam

1.3 Spam targets and statistics

Spam was initially used solely for promotional purposes. It is now highly developed, diverse, and complex, with increasingly malicious goals. Spam has evolved in a variety of ways. Not only in terms of quantity, but also of content (see figure 1.2). Spam currently serves many purposes; here is a non-exhaustive list:

- **phishing** : The goal is to impersonate an organization known to successfully steal confidential information. For example, we may receive an email that appears to be from our bank or another site where we have personal information. You are asked to click a link in this email (for various reasons, update, etc.), a web page is displayed when you click the link ... The user is asked to enter their bank account information or any other known personal information. Some of the best sites for fighting phishing are eBay, Paypal, and Bank of America.

- **Publicity** : "The goal is to promote the benefits of each product. Pharmaceuticals, consumer goods, a wide range of applications and gambling are just a few examples. They can also serve as political, educational or religious ideas and / or organizations.

- **Scam** : This is an attack designed to take advantage of the credulity of recipients to collect money from them. The most common example is the Nigerian scam: a dignitary from an African country asks you to act as an intermediary for a large financial transaction, promising you a good percentage of the sum. must donate money.

- **Canular** : the goal is to disseminate information that seems very sensitive, often with an emergency flavor: false virus warning, chain of solidarity ..... For example: "a new very dangerous virus is spreading, we must circulate information ";" underwear is infected with a dangerous bacteria ".

- **Malware** : Is software designed to infiltrate or damage a computer system. It is commonly believed to contain computer viruses, worms, Trojans, spyware, and adware. This type of program is sent as a non-suspect attachment. When a user opens the file, malware is installed. The interdependence of spam and malware has given rise to malware that delivers unwanted e-mail. These infected hosts are referred to as "zombie computers". Many people believe that most spam emails are sent by botnets, which make up a network of PCs that are malware that infiltrate or damage your computer. It is widely believed to be infected with computer viruses, worms, Trojans, spyware and adware..

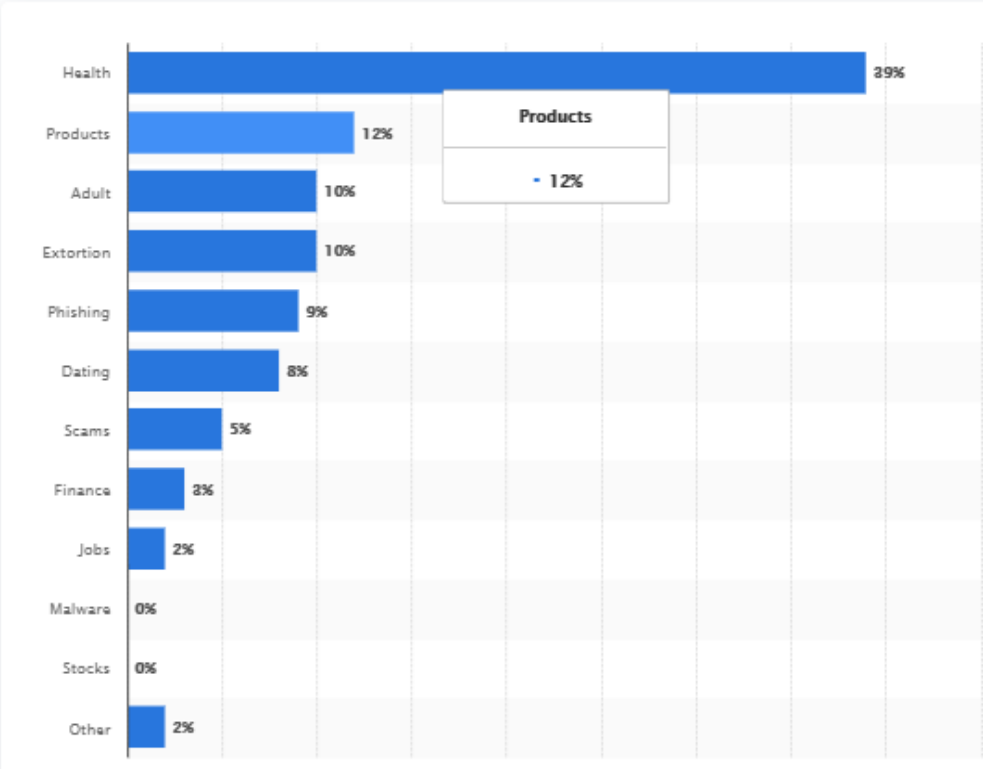


Figure 1.2: Distribution of spam by content
[10]

Figure 1.3 depicts the global spam rate between 2012 and 2018; spam accounted for 55% of all emails in 2018.

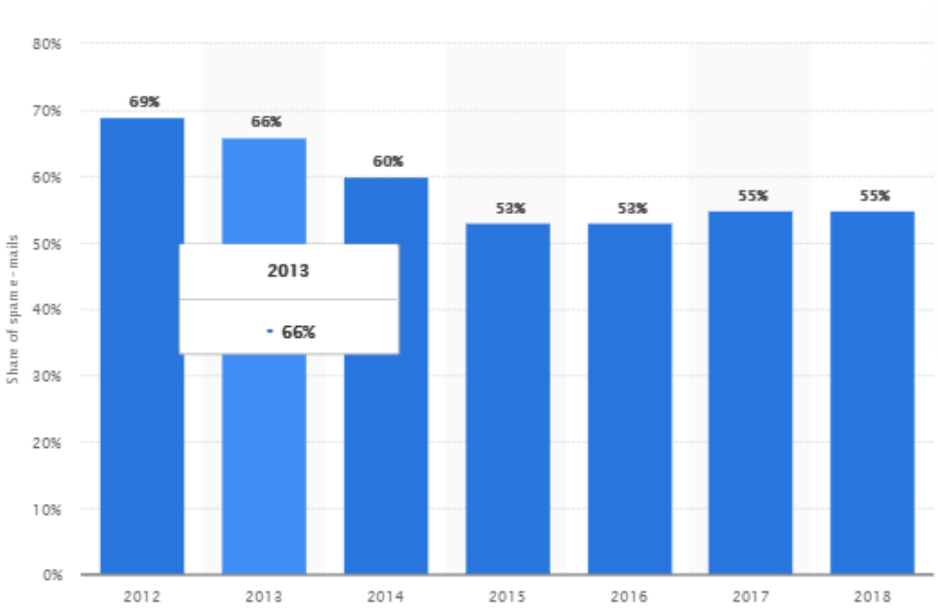


Figure 1.3: Global e-mail spam rate from 2012 to 2018
[9]

1.4 Spam impacts

In this part, we look at how spam affects users, businesses, and ISPs.

1.4.1 Delay of time

.

- Mailboxes are unusually full.
- Delete unwanted emails.
- Configure and maintain filters.
- Consult with rejected emails to uncover correct messages due to the risk of losing important poorly categorized emails by anti-spam tools.

1.4.2 Loss of bandwidth and space

- Specially for modem users.
- Sometimes spam attachments can be big.

1.4.3 Significant financial losses at the corporate and ISP levels

- High repair costs and help fight spam.
- Lack of profitability for the employee

1.5 Spam filtering techniques

There are several methods for countering spam that can be combined: predictive analysis (Bayesian filter), keyword scanning, white lists, and black lists. When different types of spam try to get around them, these control mechanisms must evolve on a regular basis.

There are two spam detection options: detection at the ISP mail server level and detection at the end user level. These tools are classified into two categories: envelope filtering and content filtering.

1.6 Envelope filtering

This method of filtering only extends to the message's header, which also provides enough detail to differentiate spam. This approach, when used at the ISP server level, has the advantage of being able to block emails right before their body is received, significantly reducing traffic on the SMTP gateway. The following methods are included in this category:

1.6.1 Blacklist filtering

The blacklist filtering technique offers network-level security by categorising forwarded emails based on the sender's address, IP address, or DNS address. These data are extracted from the email's header and compared to a pre-defined list, and if any of them match, the email is marked as spam.

When these details are compared to the list, the email is denied. As a result, this strategy filters SPAM emails to provide network protection. The company in charge of delivering and applying this filter is Internet Server Providers (ISP).[12]

1.6.2 Whitelist filtering

These lists are made up of a pre-declared list of (email addresses, domain names, and IP addresses) from which the recipient agrees to receive emails. By default, all hosts are considered protected, and spammers may use their addresses to send spam. The whitelist, like the blacklist, needs constant updating and refreshing.[18]

1.6.3 Domain verification filtering

The grey list is a combination of the white and black lists. When an anonymous person sends an email to a specific mailbox, the email is suspended and replaced by an auto-reply address with a connection to complete the sending. This is done to detect bots; spammers are unaware that they must issue a clarification in order for the message to be remembered. If an authentic email is expected and the source is not blacklisted or whitelisted, the email would be graylisted. If the sender replies to the confirmation request (typically by clicking on a web link), their email address will be added to the whitelist and their messages will be sent to you. In fact, it is the whitelist's complex opening.

1.6.4 Filtration based on a gray list

Recipients are set up to only accept communications from specific locations. Emails with non-listed domains would not be received. This effectively eliminates a substantial volume of email.[3]

1.7 Content filtering

This process of filtering happens at the consumer level, where their content is checked for spam that has gone through the envelope filter. The following methods are included in this category.

1.7.1 Keyword filtering

To decide the emails are spam, the administrator should consult the list of keywords to be discovered. For instance, all emails containing the words ”cash” or ”drugs” would be flagged as spam. The keywords in the emails are used to build this filter. The scan is fast, but it can also be effective. Since manual monitoring is required, spammers purposely put codes between keywords to prevent this filter. For instance, we can find M.O.N.E.Y. Alternatively, m * o * n * e * y

1.7.2 Filtering by characters

This enables you to block emails that contain specific characters or fonts. Or any of the languages that are used.

1.7.3 URL filtering

This involves comparing hypertext links in messages to a pre-saved archive of ”bad URLs”, or consulting web blacklists in real time. Spammers will also try to hide a hyperlink by filtering URLs to avoid crawling.

1.7.4 Bayesian filters

The most popular machine learning method for filtering spam is NaiveBayes, Naive Bayes is a probability classifier.[13] It calculates and uses the probability of occurrence of certain words / phrases in the most popular examples (messages) to classify a new example Naive Bayes has been found to be effective in classifying text documents. Bayesian filters (statistical method) Filters work by analyzing the words in the letters inside an email to calculate the likelihood that the message is spam or not.

1.7.5 Support Vector Machine

Text document classification has been accomplished using Support Vector Machines (SVM). SVM has sparked considerable interest in their application to spam filtering. SVMs are kernel methods that integrate data representing text documents into a vector space where linear algebra and geometry can be performed. In this vector space, SVM attempts to build a linear separation between two classes.[1]

A support vector machine is a classifier with a linear binary maximum margin. It can be interpreted as locating a hyperplane in a feature space that is linearly separable and separates the two classes with the greatest possible margin. Because the instances closest to the hyperplane are closest to the hyperplane, they are referred to as ”support vectors.” Because they support the hyperplane on both sides of the margin, the instances closest to it are referred to as ”support vectors.”

1.8 Machine learning

Machine learning refers to any process of building a model of reality from data, whether refining an existing partial or less general model or developing a new model entirely. There are two trends in learning, that resulting from artificial intelligence qualified as ”symbolic7” and that resulting from statistics and qualified as digital. We distinguish between different types of learning: unsupervised learning and supervised learning:

1.8.1 Supervised learning

It is an inductive mechanism that involves the automatic construction of a classifier that learns the characteristics and properties of target groups from previously categorized (or labeled) instances. This learning method is framed by the training of the classification function on the divisions (or classes) as well as on their characteristics. Supervised learning solves two types of problems: supervised classification and regression. These two categories of problems are distinguished by the names assigned by the specialist.

Definition :supervised classification (also called classification or inductive classification) Supervised classification attempts to predict whether records belong to groups that are known a priori. It is therefore a compilation of methods aimed at determining the membership of an individual in a community.

In this thesis we are only interested in supervised classification. Thus, supervised learning is a method (see Figure 2.1) that allows you to build models from a series of learning examples which are then used as a guide by the categorization algorithm.

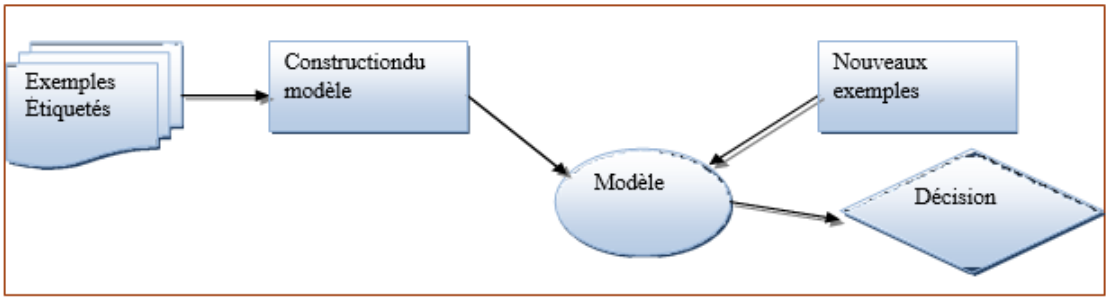


Figure 1.4: Principle of supervised learning

1.8.2 Unsupervised learning

Unsupervised learning In this type of learning, there are no predefined classes, the goal is to make the strongest possible groupings of objects in which opinions vary too little according to their values. This type of learning is also called ”observation” or ”discovery” learning.

The best known of the unsupervised problems is unsupervised classification or clustering. The classes, which will be called clusters, are formed by grouping together data

that has certain characteristics in common. To build a grouping of these data, we have three choices to make:

- Choose a measure of resemblance (or similarity) between the data

- Choose the type of structures we want to obtain: partition, hierarchy, tree ...

- Choose the method to obtain the desired structure.

1.9 Conclusion

In this chapter, we present concepts of spam, its goals and consequences, as well as the various approaches to combating spam, which are divided into two categories: The first includes email header-based options including blacklists, whitelists, and graylists. The second category of solutions includes those that are content-based message text-based, such as machine learning-based filtering.

Chapter 2

RELATED WORK

2.1 Introduction

This chapter provides an overview of some of the most important studies for detecting spam:

2.2 Drucker et al.

Drucker et al. [5] compared the efficiency of the SVM linear classifier with those of RIPPER, Rocchio and decision trees. He was the first to try out a large set of experimental setups on term selection and different learning algorithms. They arrive at the following conclusions:

- SVM (with a binary representation) and decision trees (with a TF representation) are the two best classifiers, but SVMs make it possible to reach positive defaults lower and more easily. between the use of a list of stopwords or not, it is preferable that a list of stopwords is not used.

- Dans un choix entre l'utilisation d'une liste de stopwords ou non, il est préférable qu'une liste de stopwords ne soit pas utilisée.

- The learning by using the decision trees is enormously long.

- The RIPPER and Rocchio methods are not efficient for spam filtering.

2.3 Saumya Goyal and al.

in 2016, Saumya Goyal end al. [7] propose the use of a despam detection mechanism based on the KNN algorithm and decision tree, they apply these algorithms on real twitter datasets. To analyze the proposed mechanism, the WEKA7 tool is used. Performance measures such as TP Rate, FP Rate, Precision, Recall and F-Measure are used to evaluate the proposed mechanism. They obtained the following results

Table 2.1: Performance measures with the decision tree [7]

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
KNN	1	0.508	0.904	1	0.949	SPAM
	0.492	0	1	0.492	0.659	NORMAL
	0.912	0.42	0.92	0.912	0.899	Weighted Avg
Decision Tree	1	1	0.827	1	0.905	spam
	0	0	0	0	0	normal
	0.827	0.827	0.683	0.827	0.748	Weighted Avg

The obtained results show that the KNN algorithm is more efficient than the decision tree.

2.4 Nurul Fitriah Rusland and al

In 2017, Nurul Fitriah Rusland end al [13] tested a naive bayes algorithm for filtering spam on two corpora (Spam Data which contains 9324 e-mails and 500 attributes and SPAM-BASE which contains 4601 emails and 58 attributes) and test its performance. The system architecture is as follows

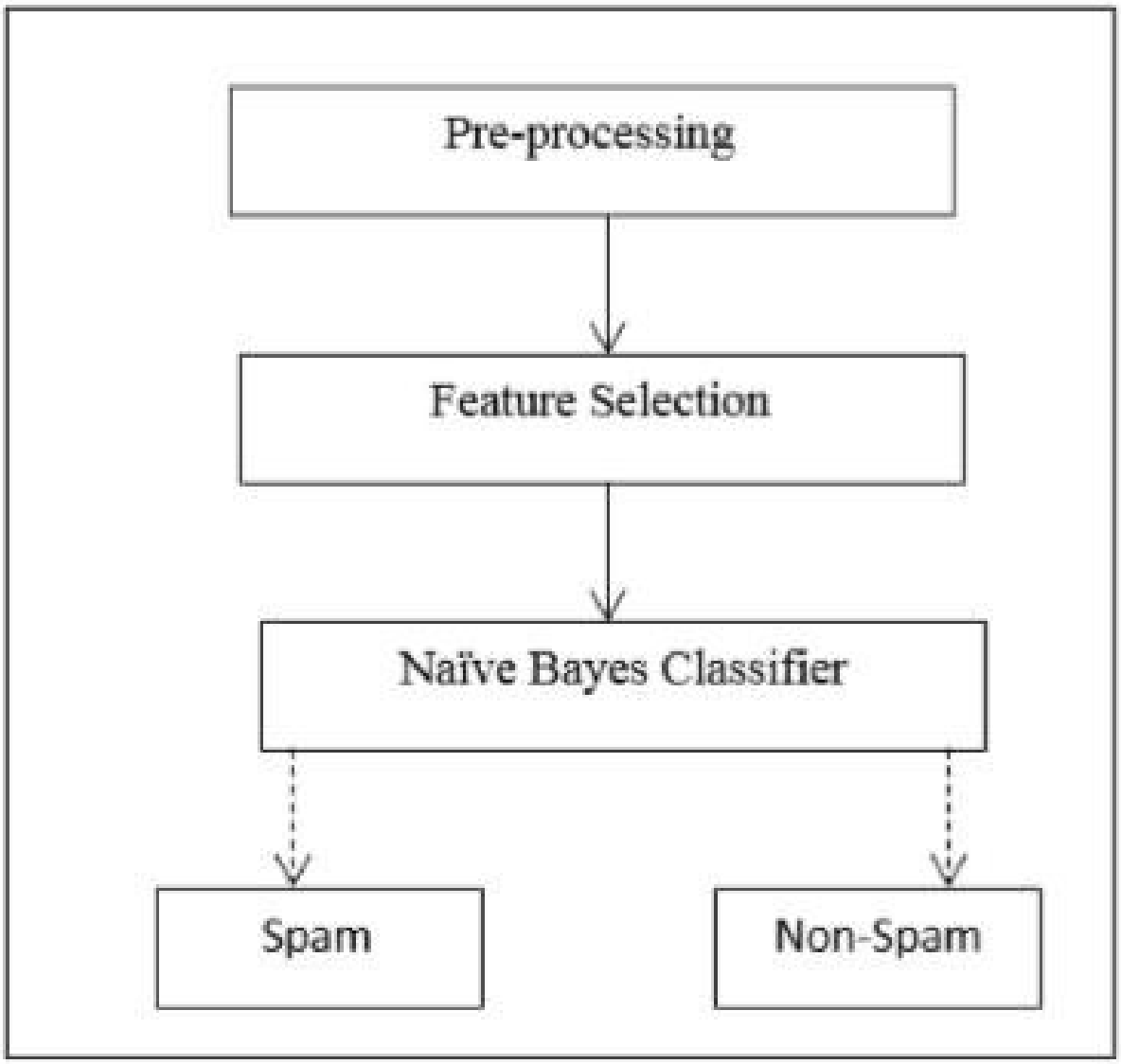
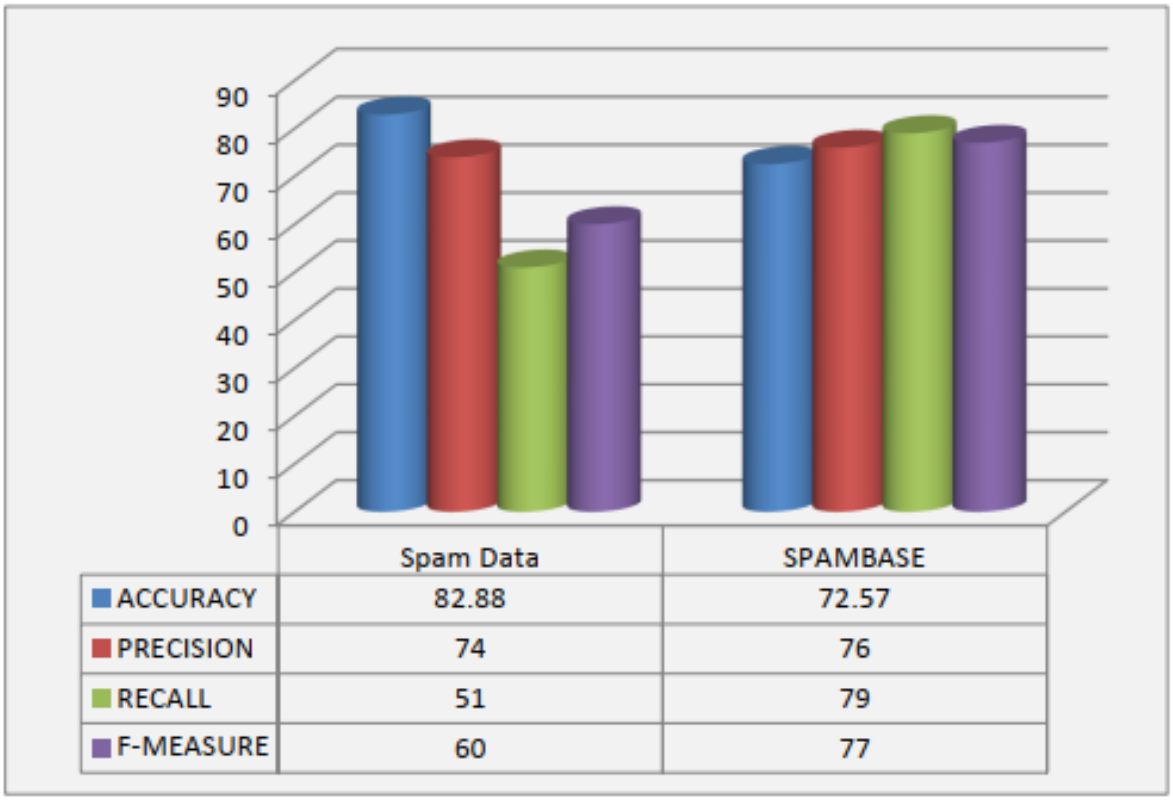


Figure 2.1: Spam filter using the naive bayes algorithm
[13]

The WEKA method is used to test the output of this filter based on its accuracy, recall, and F-measure. They discovered the following results:



	Spam Data	SPAMBASE
ACCURACY	82.88	72.57
PRECISION	74	76
RECALL	51	79
F-MEASURE	60	77

Figure 2.2: Evaluation results with the two corpora

They found that the performance of this filter is also based on the corpora used. As we can see, corpora that have fewer instances and attributes (here SPAMBASE) can give good results

2.5 Anju Radhakrishnan and Vaidhehi V

Anju Radhakrishnan and Vaidhehi V. [17] use two essential algorithms, Nave Bayes and J48 Decision Tree, to measure their utility in email classification. The corpus used is Enron, and the frequency value is TF-IDF. Classifiers are often checked for various attribute scales. The test results are as follows:

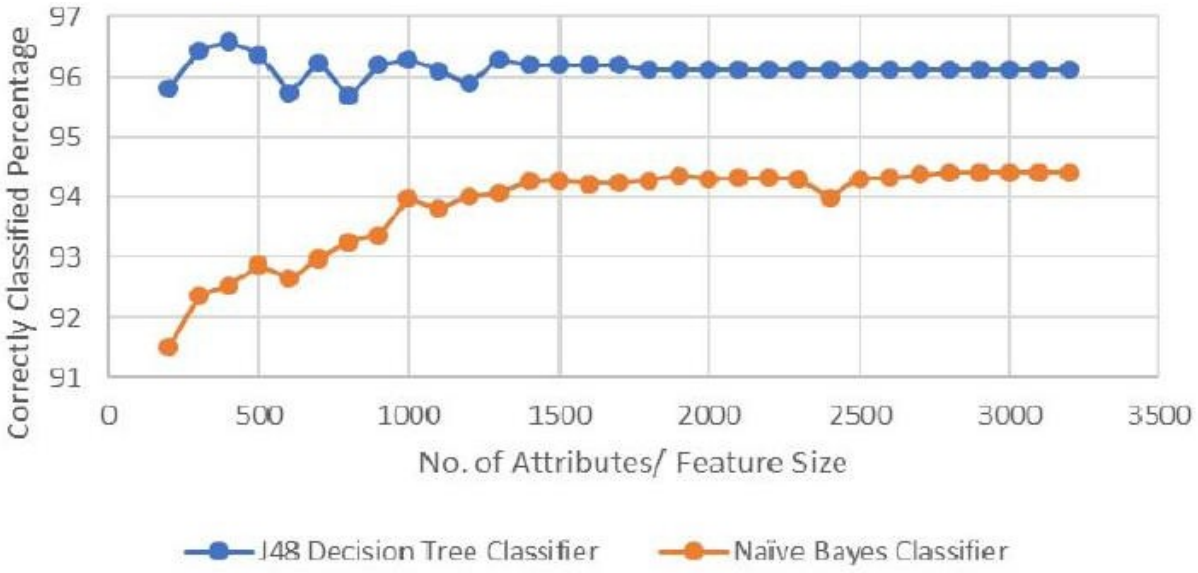


Figure 2.3: The test results for the two classifiers NB and J48. [17]

Email classification experiments have shown that the J48 Decision Tree classifier is more efficient than the Naïve Bayes classifier for the Enron corpus. It gives an accuracy of 96.5971 % in the classification of emails with a size of 400 attributes.

2.6 Shradhanjali and Verma Toran

Shradhanjali and Verma Toran [19] propose the use of a new method for spam detection using SVM and attribute extraction which achieves 98% accuracy. The architecture of the proposed system is shown in the following figure:

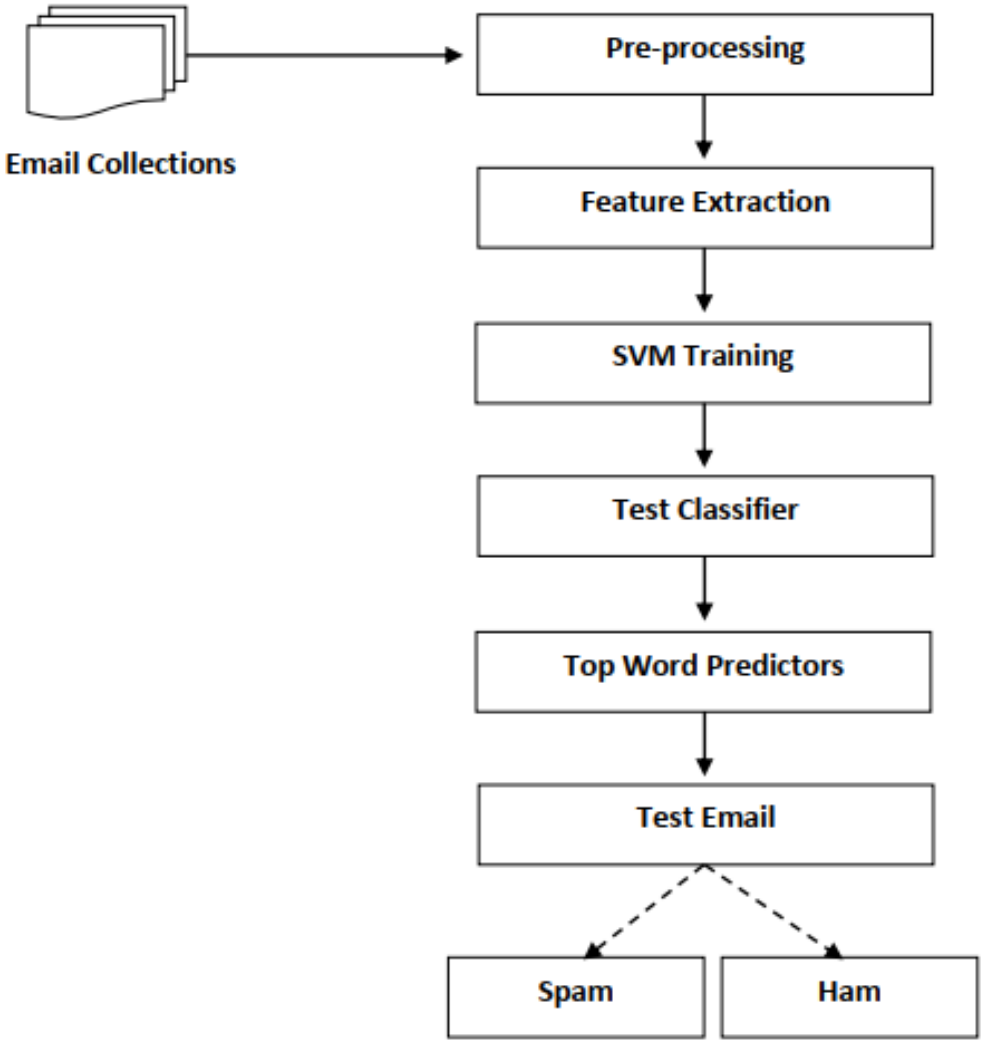


Figure 2.4: Spam filter using SVM and attribute extraction [19]

Pretreatment: all numbers, special symbols, URLs, and HTML tags are omitted during the preprocessing phase. The root is used to remove the superfluous alphabet from the vocabulary.

Extracting attributes: The attributes are derived throughout the pretraitement. Following the extraction of characteristics, education is carried out. Emails are fed into the SVM classifier as data during practise.

Classifier evaluation: Following training, test e-mails are distributed for review. the system's precision The precision would be up to 98 percent. Finally, the classifier is checked for an e-mail (la classe spam ou la classe authentique).

2.7 Jawale Diksha .S and.al

In 2018, Jawale Diksha.S [8] suggest the use of a spam classifier hybrideNB-SVM that combines the benefits of Nave Bayes (NB) and Support Vector Machine (SVM). NB is a quick classification algorithm, and SVM has a high recall and precision rate. The learning data are processed by the NB algorithm, which computes the probability for and word and message and compares it to a threshold that classifies the data. The data treated by NB are fed into SVM to improve precision. This classifier's design is as follows:

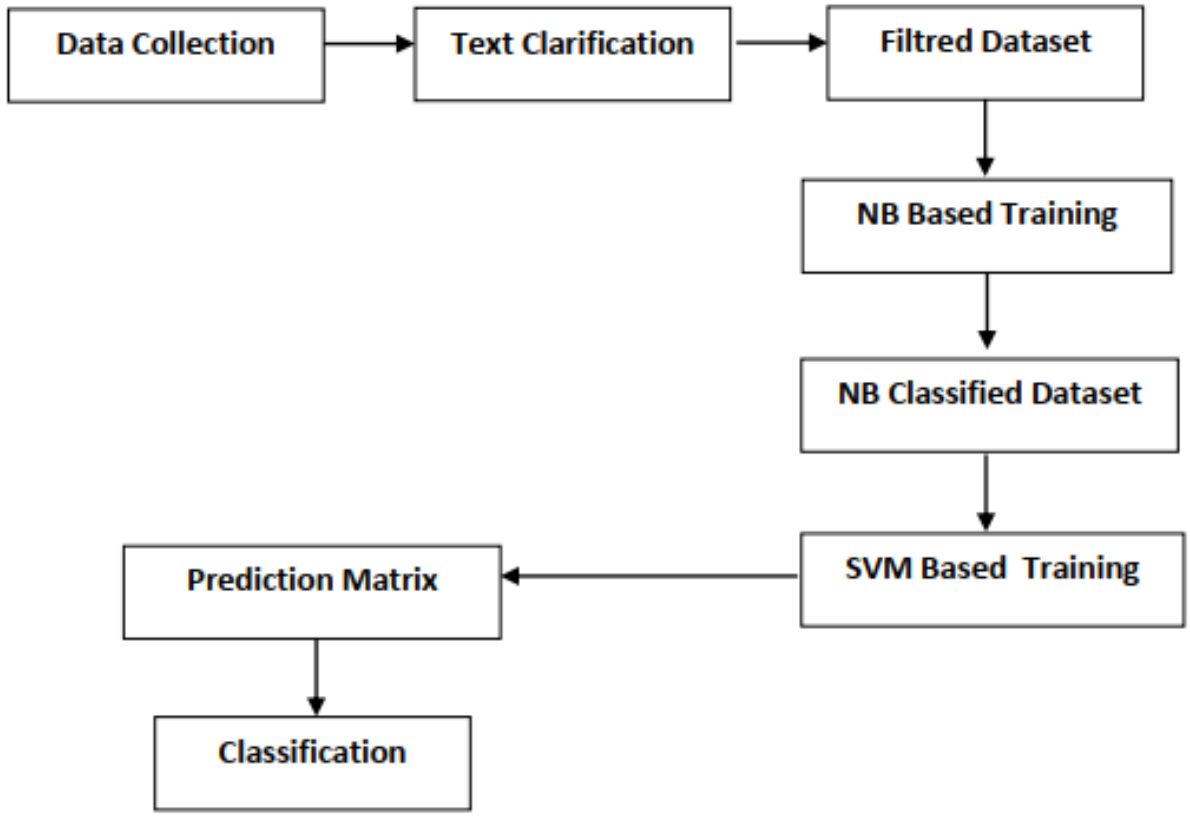


Figure 2.5: NB-SVM architecture
[8]

We use two algorithms in a spam detection system: NB and SVM. It has the highest accuracy SVM and a high rated pace. The researchers achieved 96.65 percent accuracy in the training process and 95.78 percent accuracy in the evaluation phase using NB. It achieved 99.43 percent accuracy in the training phase and 97.13 percent accuracy in the evaluation phase using SVM. These two algorithms were combined to achieve the highest precision, quickest classification time, and smallest data set possible. Aside from this NB-SVM algorithm mix, it achieved 99.44 percent accuracy in preparation and 97.57 percent accuracy in testing. This suggests that introducing them as a group is more accurate than implementing them individually.

2.8 W.A. Awad and S.M. ELseuofi

W.A. Awad1 and S.M. ELseuofi2 in 2011 [2]provides some of the most popular machine learning methods and their applicability to the email spam classification problem. Descriptions of the algorithms A comparison of their performance was presented on the SpamAssassin SPAM group, where the experiment showed very promising results especially in the algorithms that are not common in commercial email filtering packages, and the spam recall in the six methods was the lowest between the values of accuracy and accuracy, while in terms of accuracy we could We find that Naïve bayes and coarse groups have very satisfactory performance among other methods. More research should be done to escalate the performance of Naïve bayes and the artificial immune system either by the hybrid system or by solving the feature-dependence problem in a nave bayes classifier, or by a hybrid system. Immunity by rough combinations. Finally, hybrid systems seem to be the most effective way to create a successful anti-spam filter nowadays.

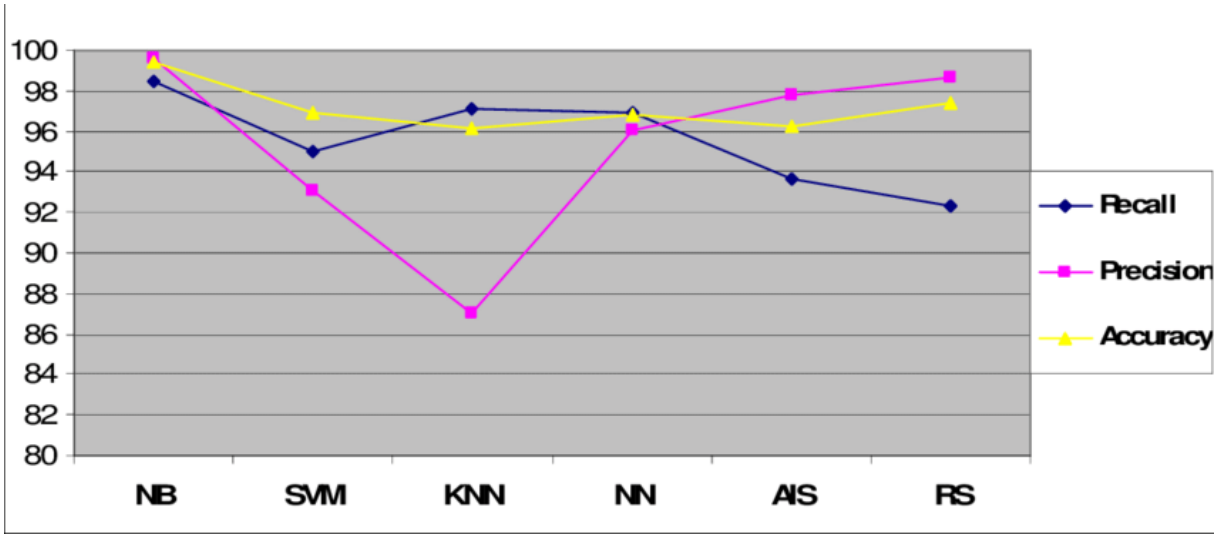


Figure 2.6: Performance of six machine learning algorithms by selecting top 100 features [2]

2.9 Sumant Sharma , Amit Arora

Sumant Sharma and Amit Arora[16] introduced different algorithms for machine learning in 2013. To classify spam. In the spam database containing 55 key spam features, we compared 24 algorithms (Table 1). We prioritize accuracy and efficiency. Algorithm to distinguish between spam and unwanted email messages As a general rule, emails associated with a single account An experiment to assess accuracy and performance As shown in (Table 2), ten classifiers are used. Experiment results are most evident in Spambase datasets. Implement the proposed method. It reads 94.28. The random panel accuracy and score value is 95.32, which is high in all other respects.

Table 2.2: compared 24 algorithm

[16]

ALGORITHIM	CCI	TP	FP	TN	FN
NB	88.54	2596	192	1478	335
BN	88.56	2596	192	1479	334
NBU	88.54	2596	192	1478	335
LOGISTIC	92.95	2654	134	1623	190
MLP	93.28	2630	158	1662	151
SGD	93.21	2655	133	1637	176
SMO	91.71	2659	129	1630	183
VP	92.32	2615	173	1644	169
KSTAR	9.56	2665	123	1640	173
DT	91.71	2666	122	1554	259
RIP	92.32	2634	1544	1614	199
PATR	93.06	2631	157	1651	162
DS	77.2	2041	747	1511	302
J48	92.34	2618	170	1631	182
RF	93.89	2673	115	1647	166
RT	91.54	2586	202	1626	187
BAGGING	92.93	2650	138	1626	187
LOGICBOOST	89.76	2589	199	1541	272
MCC	92.95	2654	134	1623	190
RS	92.37	2667	121	1583	230
CVR	92.15	2637	151	1603	210
FC	92.34	2618	170	1631	182
RC	94.28	2680	108	1658	155

Table 2.3: accuracy and performance OF ALGORITHMS

[16]

ALGOS	CCI	PPV	TRP	ACC	F1
BAYSNET	88.56	93.113	88.6	88.56	90.8
LOGICBOOST	89.76	92.862	90.4	89.7	91.66
RANDOMTREE	91.71	92.754	93.2	91.54	93
JRIP	92.32	94.476	92.34	92.32	93.71
J48	92.34	93.902	93.5	92.34	93.7
MULTILAYER PRESAPTRON	93.28	94.332	94.5	93.28	94.45
KSTAR	93.56	95.588	93.9	93.56	94.73

2.10 Conclusion

In this chapter, we presented some previously published works on spam filtering. We go through their guiding principles or architectures, as well as their outcomes.

Chapter 3

COMPARATIVE ANALYSIS

3.1 Introduction

In this chapter, we compare spam filter accuracy results based on six algorithms (SVM, NB, RandomForest, XGBoost, Decision Tree and KNN algorithms). Based on these results, we will discuss and select the best algorithm among those investigated to classify received emails.

3.2 Classifiers

This section will give a detailed description of the classifiers used

3.2.1 Naive Bayes

The naive bayes classifier belongs to the family of probabilistic algorithms and used bayes theorem to categorize sample data. Bayes theorem : Given a hypothesis H and evidence E, Bayes' theorem states that the relationship between the probability of the hypothesis P(H) before getting the evidence and the probability P(H|E) of the hypothesis after getting the evidence is :

$$P(H/E) = \frac{P(H) * P(E/H)]}{P(E)} \tag{3.1}$$

The probability of each category is calculated and outputs the one with highest probability

3.2.2 K-nearest neighbour KNN

The k-nearest neighbour (K-NN) classifier is considered an example-based classifier, that means that the training documents are used for comparison rather than an explicit category representation, such as the category profiles used by other classifiers. As such, there is no real training phase. When a new document needs to be categorized, the k most similar documents (neighbours) are found and if a large enough proportion of them have been assigned to a certain category, the new document is also assigned to this category, otherwise not . Additionally, finding the nearest neighbours can be quickened using traditional indexing methods. To decide whether a message is spam or ham, we look at the class of the messages that are closest to it. The comparison between the vectors is a real time process.[2]

3.2.3 Support Vector Machines SVM

SVM is a supervised algorithm which is popular for text classification algorithm due to high speed and good performance. Based on the training set provided, it outputs a hyperplane which is a line in two dimension that best separates the categories. This hyperplane is called the decision boundary. In phishing detection, input is represented by a set of features for instance, presence or absence of certain word and output which is 1

or -1 indicates whether the email is phished or not. This algorithm plots each node from a dataset within a dimensional plane and through classification technique the cluster of data is separated by a hyperplane into their respective groups . The hyperplane can be described as equation:H=VX+c where c is a constant and V is the vector.

3.2.4 Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over-fitting to their training set.

3.2.5 Decision Tree

DT is a classification and prediction approach that is commonly used in data mining. C4.5 is a common and effective decision tree approach that has been used to filter e-mail messages in several studies. A decision tree in antispam email is a tree whose internal nodes contain words from the email message and the leaf nodes are either spam or legitimate.

3.2.6 XGBoos

XGBoost is a novel gradient tree boosting method introduced by Chen and Guestrin [4]. It first applies a set of Classification and Regression Trees (also known as CART) as weak learners and then boosts the performance of the trees by creating an ensemble of trees that minimize a regularized objective function. The algorithm introduced concepts such as sparsity-aware split finding in each tree, cache-friendly approximate algorithms to determine splitting points, and efficient out-of-core calculation to the methods of gradient tree boosting to create an algorithm with very fast computational speed while maintaining good prediction ability.

3.3 General architecture

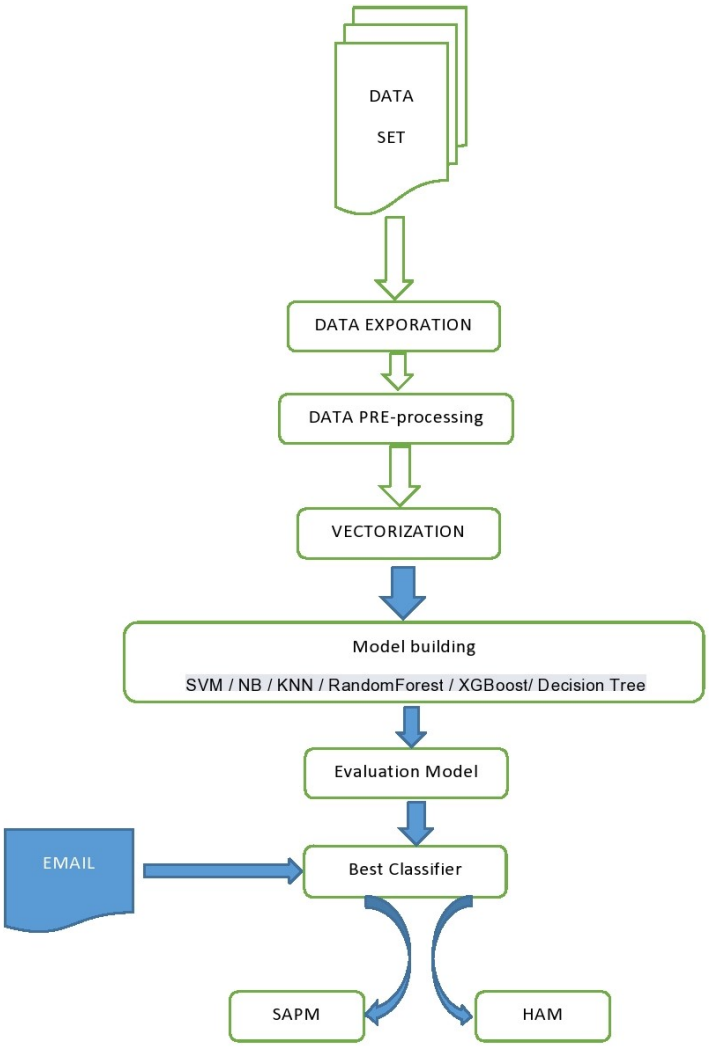


Figure 3.1: System architecture

3.4 Detailed architecture

In this section , we detail each of the phases of our system

3.4.1 Datat Set

The dataset consists of 5572 messages in English. The data is designated as being ham or spam. Dataframe has two columns. The first column is “Target” indicating the class of message as ham or spam and the second “Text” column is the string of text. We will call all the libraries we use in the application Especially those that relate to word processing and machine learning.

3.4.2 Data Exploration

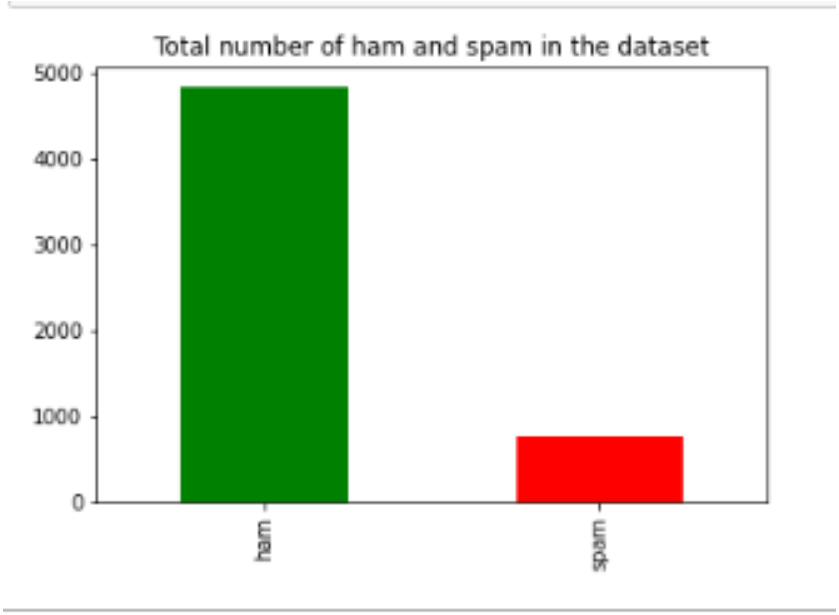


Figure 3.2: Total number of ham and spam in the dataset

we will Noting From the above countplot the data imbalance is quite evident

Feature Engineering

For the purpose of data exploration, we will creating new features

- No of Characters: Number of characters in the text message
- No of Words: Number of words in the text message
- No of sentence: Number of sentences in the text message

Table 3.1: Feature engineering

	count	mean	std	min	25%	50%	75%	max
No of Character	5572.0	80.118808	59.690841	2.0	36.0	61.0	121.0	910.0
No of Words	5572.0	18.698492	13.737477	1.0	9.0	15.0	27.0	220.0
No of sentence	5572.0	1.991565	1.501427	1.0	1.0	1.5	2.0	38.0

We will noting from the pair plot, we can see a few outliers all in the class ham. This is interesting as we could put a cap over one of these. As they essentially indicate the same thing ie the length of Email.

Next, we shall be dropping the outliers

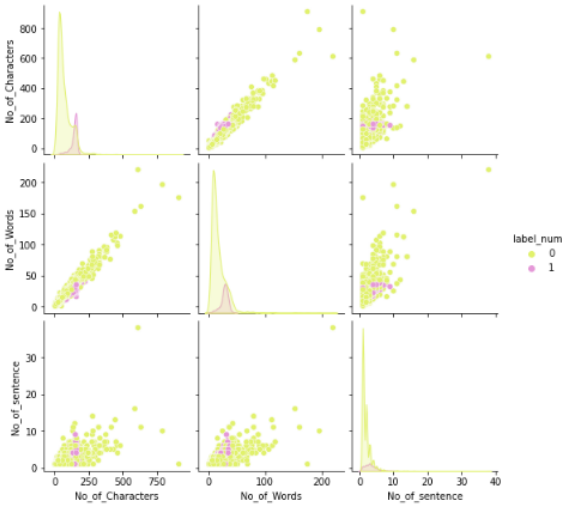


Figure 3.3: eature engineering 1

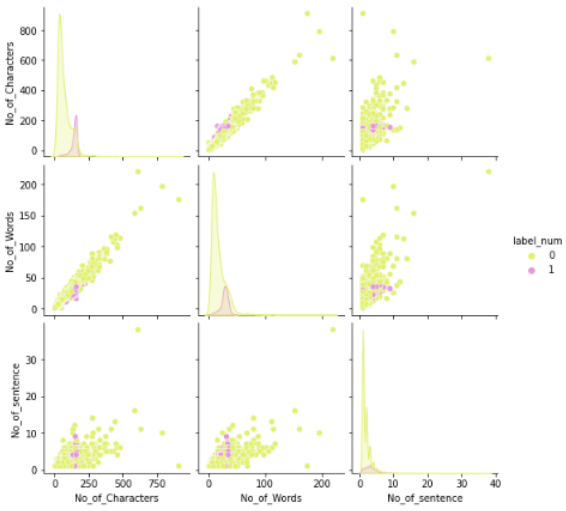


Figure 3.4: eature engineering 2

Figure 3.5: WordCloud: spam messages

Figure 3.6: WordCloud: ham messages

3.4.3 Data Pre preprocessing

Cleaning Text

Lets have a look at a sample of texts before cleaning

The First 5 Texts:

```
The First 5 Texts:
Go until jurong point, crazy.. Available only in bugis n great world la e buffet...
Cine there got amore wat...
Ok lar... Joking wif u oni...
Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to
87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's
U dun say so early hor... U c already then say...
Nah I don't think he goes to usf, he lives around here though
```

Figure 3.7: The First 5 Texts

The data cleaning process NLP is crucial. The computer doesn't understand the text. for the computer, it is just a cluster of symbols. To further process the data we need to make the data cleaner.

- In the first step we extract only the alphabetic characters by this we are removing punctuation and numbers.
- In the next step, we are converting all the characters into lowercase.

```
The First 5 Texts after cleaning:
go until jurong point crazy available only in bugis n great world la e buffet cine
there got amore wat
ok lar joking wif u oni
free entry in a wkly comp to win fa cup final tkts st may text fa to to receive
entry question std txt rate t c s apply over s
u dun say so early hor u c already then say
nah i don t think he goes to usf he lives around here though
```

Figure 3.8: The First 5 Texts after cleaning

This text will be then used in further processing

Tokenization

Tokenization is the method where the sentences within an email are broken into individual words (tokens). These tokens are saved into an array and used towards the testing data to identify the occurrence of every word in an email. This will help the algorithms in predicting whether the email should be considered as spam or ham [6].

Removing Stopwords

Stopwords are frequently occurring words(such as few, is, an, etc). These words hold meaning in sentence structure, but do not contribute much to language processing in NLP. For the purpose of removing redundancy in our processing, I am removing those. NLTK library has a set of default stopwords that we will be removing

3.4.4 Model Building

Steps involved in the Model Building

- Setting up features and target as X and y
- Splitting the testing and training sets
- Build of model for six different classifiers.
  - 1. Naïve Bayes
  - 2. RandomForestClassifier
  - 3. KNeighborsClassifier
  - 4. Support Vector Machines
  - 5. XGBoost
  - 6. Decision Tree
- Fit all the models on training data

Table 3.2: Accuracy of classifier

Classifier	Accuracy
Naïve Bayes	97.2197%
RandomForestClassifier	97.4888%
KNeighborsClassifier	92.7354%
Support Vector Machines	97.9372%
XGBoos	97.6682%
Decision Tree	96.5022%

3.4.5 Model evaluation

Testing the models on Testset

- Accuracy Report

The research was aimed at finding the highest accuracy for detecting the emails correctly as ham and spam. The module from the Scikit-learn library called 'Accuracy' helped analyse the correct number of emails classified as 'Spam' and 'Ham'. This can be measured by equation- :

$$\frac{(TN + TP)}{(TP + FN + FP + TN)} \tag{3.2}$$

Evaluating the Dataset for training and testing data to provide better accuracy and showed improvement. This could vary on the dataset size and the information separated during the split. It should be noted that the higher the rate of training data than testing data, better the performance achieved. This is a good sign, since when considered as a real-world example, the models will have bigger weight for training data than testing. According to the experiments, have improved the accuracy of all five models. The SVM is the algorithm that has performed better than all the other algorithms

3.5 Conclusion

Several different methods have been used in machine learning. In this chapter, we have tried to introduce a little, giving the principle of each method.

The question now arises: which of these methods is better for classifying spam? Comparing these methods requires applying them to the same data, using the same performance measures, or adopting a controlled evaluation method.

Chapter 4

IMPLEMENTATION AND RESULTS

4.1 Introduction

After a series of stages in the development process, the main goal of software implementation is to achieve an efficient system of solving problems using tools and algorithms.

The previous chapter provided a global and detailed design; this chapter provides a view of the system and the tools used to create a reliable system. In this chapter, we will discuss the development environment, including the various important libraries that were used, as well as the data structures that were chosen to implement this type of system. This chapter concludes with a conclusion.

4.2 The choice of programming language:

To choose a programming language that specializes in machine learning, and image processing, it must consider the skills listed in current job postings as well as the libraries available in different languages that can make the learning process. deep. Python is the language of most affected programming in machine learning and deep learning. Python is followed by Java, then R, then C ++.

4.2.1 Python

Python is a high level, general purpose interpretation programming language . Created by Guido van Rossum and first published in 1991, the philosophy of Python design emphasizes code readability with its notable use of large white space.[11]

His Language constructs and its object-oriented approach aim to help programmers write clear and logical code for small and large-scale projects, it provides constructs that enable clear programming on both small and large scale. Python has a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional, and procedural, and has a large and comprehensive standard library. Python interpreters are available for many operating systems. the official python website link http:///www.python.org

4.2.2 The Anaconda Environment:

Anaconda brings together a set of tools revolving around programming languages Python and R: it provides including the two runtime environments. This distribution of Python is oriented ”Data Science” (science data) and ”Machine Learning” automatic): in these areas, it is certainly the most popular. Anaconda installs equally well on Windows, MacOs or Linux. You can find the Anaconda software from its official website https: HTTP://www.anaconda.com/. Anaconda offers its own package manager called ”Conda”.

4.2.3 Jupyter Notebook

A Jupyter notebook is a document that includes live interactive code, output from the code, and additional descriptive rich text elements. In SAS University Edition, you can create SAS notebooks, in which you write SAS code, as well as Python 3 notebooks, in which you write Python code. Both types of notebooks use SAS code to generate results. Jupyter notebooks can consist of three types of cells:

Code - contains code that can be sent to the SAS kernel for execution. Results are displayed in the cell footer. Markdown - contains descriptive text that can be used as explanatory notes in the notebook. Markdown cells can include some limited formatting. Raw - enables you to write output directly. Raw cells are not evaluated by the notebook. Jupyter notebooks can contain multiple cells that you can collapse, expand, and rearrange. Each notebook has one interactive session connected to the SAS kernel. When you open a notebook, the associated SAS kernel is automatically launched. The following notebook contains a markdown cell and a code cell. The markdown cell includes the title, and the code cell includes both the code input and results output.

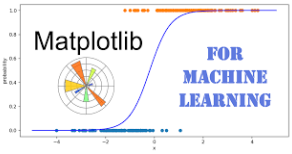
4.3 Libraries used:

4.3.1 Pandas:

Pandas is an open source library licensed under BSD providing high performance data structures and easy to use, as well as data analysis tools for the Python programming language. Pandas is a project sponsored by NumFOCUS. This will participate in successful development of pandas as a world-class open source project and will make it possible to donate to the project. Pandas is a python library that allows you to easily handle data to analyze: • Manipulate data tables with variable labels (columns) and of individuals (lines). • These tables are called DataFrames, similar to DataFrames under R. • One can easily read and write these DataFrames from or to a tabulated file. • We can draw graphs from these DataFrames thanks to Matplotlib

4.3.2 Matplotlib

Matplotlib is a library of the Python programming language for plotting and visualizing data in the form of graphs5. It can be combined with the scientific computation python libraries NumPy and SciPy6. Matplotlib is distributed freely and free under a BSD4-style license. Its current stable version (2.0.1 in 2017) is compatible with Python version 3.



Several points make this library interesting:

- Export possible in many raster (PNG, JPEG ...) and vector (PDF, SVG ...) formats
- Extensive online documentation, many examples available on the internet Strong, very active community
- Pylab interface: faithfully reproduces MATLAB syntax
- High level library: ideal for interactive computing

4.3.3 Nltk

Natural Language Toolkit (NLTK) is a Python software library for automatic language processing developed by Steven Bird and Edward Loper of the Department of Computer Science at the University of Pennsylvania. In addition to the library, NLTK provides graphical demonstrations, sample data, tutorials, as well as programming interface (API) documentation.

4.3.4 SCIKIT-LEARN

Scikit-Learn (SKLearn) is an environment that is incorporated with Python programming language. The library offers a wide range of supervised algorithms that will be suitable for this project [13]. The library offers high-level implementation to train with the 'Fit' methods and 'predict' from an estimator (Classifier). It also offers to perform the cross validation, feature selection, feature extraction and parameter tuning [14].

4.4 Application interface:

In this section, we are interested in presenting our system via displaying it. some screenshots of the many interfaces created We attempted to design a graphical interface that displays as much info as feasible.

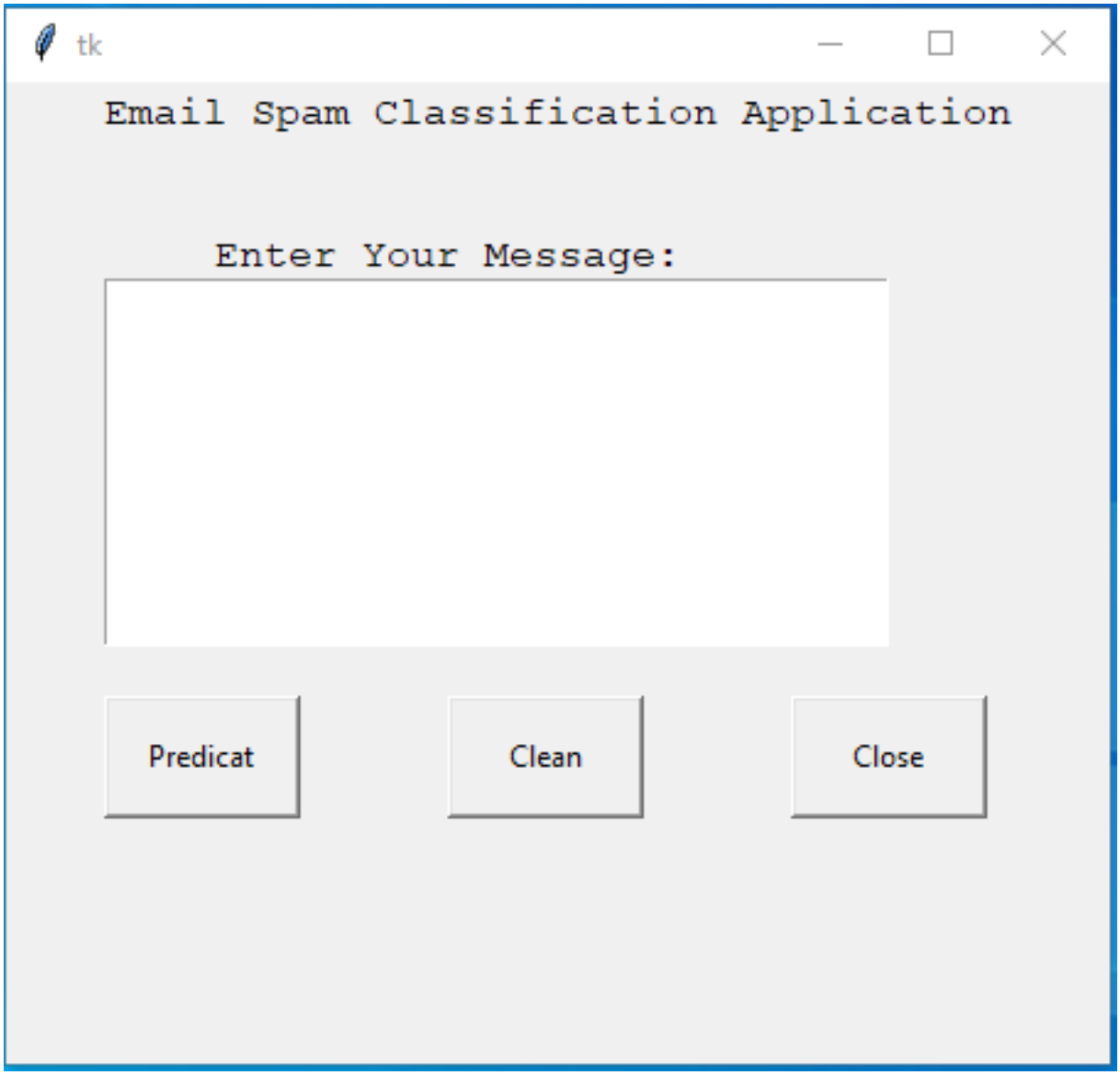


Figure 4.1: the application interface:

HAM email interface

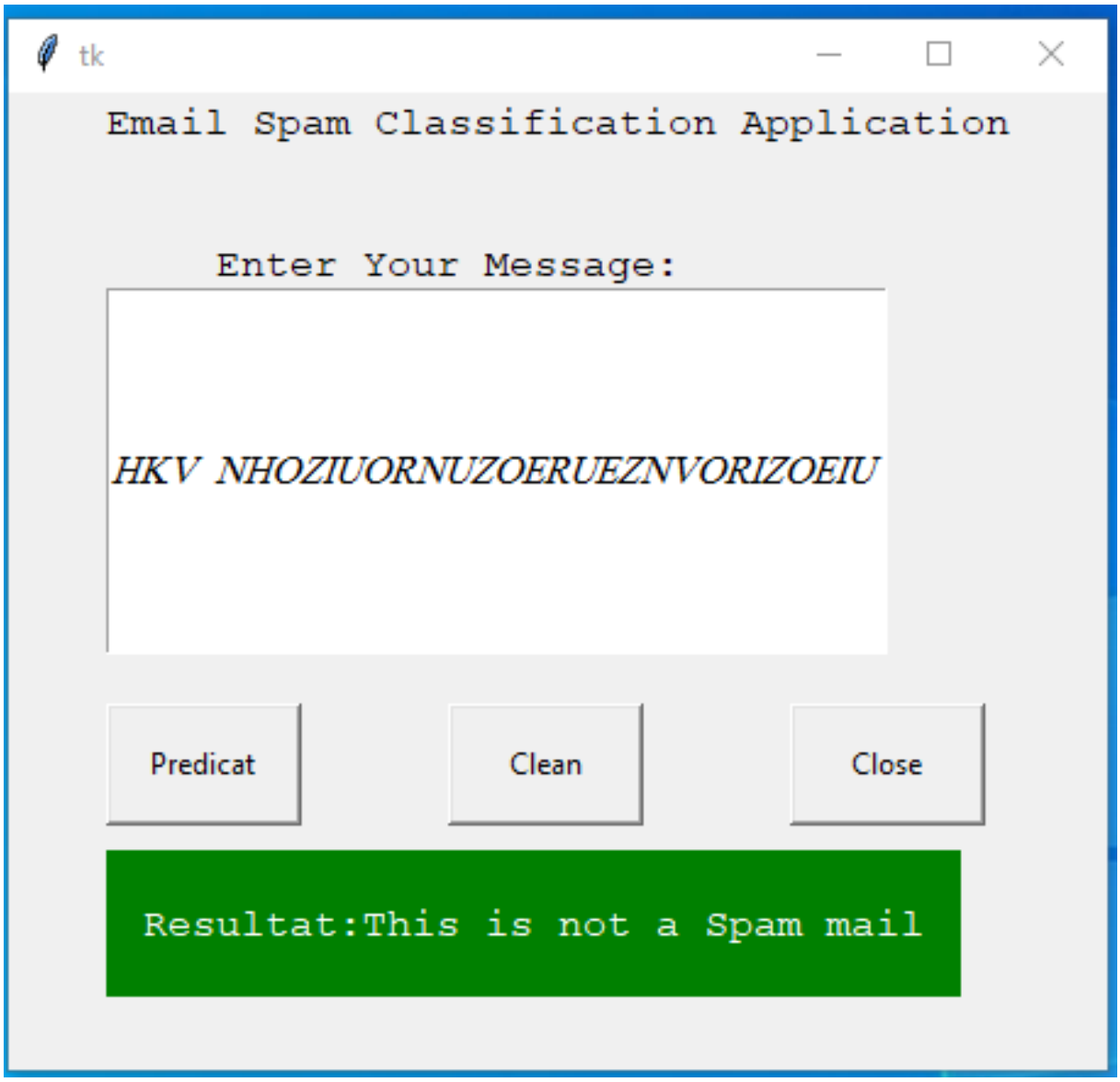


Figure 4.2: interface ham email

SPAM email interface

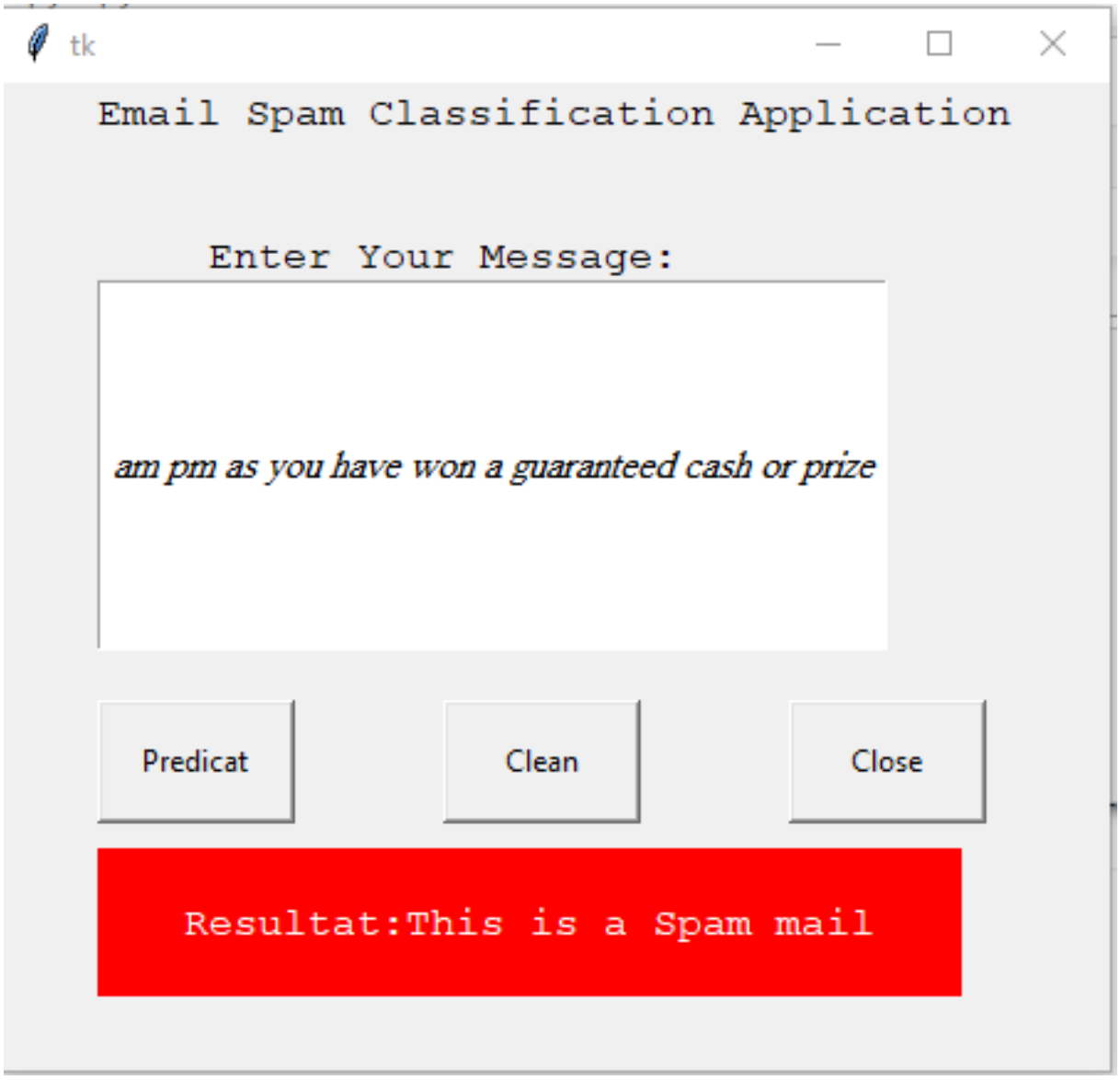


Figure 4.3: interface spam email

4.5 Conclusion

In this chapter, we demonstrated the architecture of our system spam filtering , as well as a comprehensive view of it. We analysed the outcomes of several SVM, NB,Random-Forest,XGBoost,Decision Tree and KNN algorithms with different representations and decided that the SVM technique was the best classifier due to these good findings.

Conclusion

General conclusion

Spam has evolved into one of the most vexing and damaging aspects of Internet technology. Traditional spam filters are incapable of dealing with the massive volumes of spam that evade anti-spam defences. Spam problems are becoming more severe, necessitating the use of more effective and efficient technologies to combat them. Researchers now have a new technique to tackle spam thanks to machine learning technologies. Text classification has been effectively implemented using machine learning. Because the email has a body, the ML technique may readily be utilised to classify spam email.

Email can now be categorized with minimal human involvement, making control easier and more precise. The efficiency of the spam filter can be enhanced by applying preprocessing procedures to the dataset. Based on a comparison of algorithm accuracy results explored in our research,SVM demonstrates potential and superior anti-spam methods.

Future Work

Feature selection techniques need to be further improved in order to continuously manage the new technologies that spammers have developed over time. Therefore, we recommend Developed a tool to extract new functionality from raw email messages in order to improve the accuracy of spam detection and handling of spammer techniques

Bibliography

[1] Ola Amayri and Nizar Bouguila. "A study of spam filtering using support vector machines". In: *Artificial Intelligence Review* 34.1 (2010), pp. 73–108.

[2] WA Awad and SM ELseuofi. "Machine learning methods for spam e-mail classification". In: *International Journal of Computer Science & Information Technology (IJCSIT)* 3.1 (2011), pp. 173–184.

[3] Nouman Azam. "Comparative Study of Features Space Reduction Techniques for Spam Detection". In: *Department of Computer Engineering College of Electrical and Mechanical Engineering National University of Sciences and Technology* (2002).

[4] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.

[5] Harris Drucker, Donghui Wu, and Vladimir N Vapnik. "Support vector machines for spam categorization". In: *IEEE Transactions on Neural networks* 10.5 (1999), pp. 1048–1054.

[6] Simran Gibson et al. "Detecting Spam Email". In: (2020).

[7] Saumya Goyal, RK Chauhan, and Shabnam Parveen. "Spam detection using KNN and decision tree mechanism in social network". In: *2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC)*. IEEE. 2016, pp. 522–526.

[8] Diksha S Jawale et al. "Hybrid spam detection using machine learning". In: *International Journal of Advance Research, Ideas and Innovations in Technology* 4.2 (2018), pp. 2828–2832.

[9] Joseph Johnson. *Global e-mail spam rate from 2012 to 2018 Number Of Clusters: 3 Must Know Methods*. en-US. URL: https://www.statista.com/statistics/263452/most-common-content-of-spam-messages-worldwide-by-category/ (visited on 05/03/2021).

[10] Joseph Johnson. *Most prevalent spam content categories worldwide in 2019 Number Of Clusters: 3 Must Know Methods*. en-US. URL: https://www.statista.com/statistics/263452/most-common-content-of-spam-messages-worldwide-by-category/ (visited on 05/03/2021).

[11] Joseph Johnson. "Most prevalent spam content categories worldwide in 2019 Number Of Clusters: 3 Must Know Methods". en-US. In: *Datanovia* (). URL: https://www.python.org//.

[12] Gerhard Paaß and André Bergholz. "„AntiPhish-Machine Learning for Phishing Detection "." In: ().

[13] Nurul Fitriah Rusland et al. "Analysis of Naïve Bayes algorithm for email spam filtering across multiple datasets". In: *IOP conference series: materials science and engineering*. Vol. 226. 1. IOP Publishing. 2017, p. 012091.

[14] Omar Saad, Ashraf Darwish, and Ramadan Faraj. "A survey of machine learning techniques for Spam filtering". In: *International Journal of Computer Science and Network Security (IJCSNS)* 12.2 (2012), p. 66.

[15] Enrique Puertas Sanz, José María Gómez Hidalgo, and José Carlos Cortizo Pérez. "Email spam filtering". In: *Advances in computers* 74 (2008), pp. 45–114.

[16] Sumant Sharma and Amit Arora. "Adaptive approach for spam detection". In: *International Journal of Computer Science Issues (IJCSI)* 10.4 (2013), p. 23.

[17] Aditya Shrivastava and Rachana Dubey. "Classification of Spam Mail using different machine learning algorithms". In: *2018 International Conference on Advanced Computation and Telecommunication (ICACAT)*. IEEE. 2018, pp. 1–10.

[18] Thamarai Subramaniam, Hamid A Jalab, and Alaa Y Taqa. "Overview of textual anti-spam filtering techniques". In: *International Journal of Physical Sciences* 5.12 (2010), pp. 1869–1882.

[19] Toran Verma et al. "E-Mail spam detection and classification using SVM and feature Extraction". In: *International Journal of Advance Research, Ideas and Innovations in Technology* (2017).

[20] Jonathan A Zdziarski. *Ending spam: Bayesian content filtering and the art of statistical language classification*. No starch press, 2005.