International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 November 2023*

Tamil Spam E-Mail Filtering for Text Message Using Bayesian Filtering Algorithm

Dr.T.Ebanesar [1 *], Dr.A.Alphonse John Kenneth [2], Dr. Sam Abraham[3], Dr. K. Joylin Bala[4] ,G. Borgia Crusu Venthan[3]

[1]Assistant Professor, Department of Computer Science, Malankara Catholic College, Mariagiri, Tamilnadu, India
Affiliated to Manonmaniam Sundaranar University, Tirunelveli – 627 012
[2]Senior Lecturer, Department of Computer Science, DMI - St. Eugene University, Lusaka, Zambia
[3]Assistant Professor, Department of Computer Application, Malankara Catholic College, Mariagiri, Tamilnadu, India
Affiliated to Manonmaniam Sundaranar University, Tirunelveli – 627 012
[4]Assistant Professor, Department of Computer Science, Infant Jesus College of Arts and Science for Women
Mulagumoodu, Tamilnadu, India
Affiliated to Manonmaniam Sundaranar University, Tirunelveli – 627 012
******Corresponding Author*** **E-mail:** ebanesarcs@gmail.com

The **T**ext **Re**trieval **C**onference's (**TREC**) **Spam Track** [1] states email spam as **"**Unsolicited, unwanted email that was sent indiscriminately, directly or indirectly by a sender having no current relationship with the recipient.**"**

ABSTRACT

In the contemporary digital era, email has become an essential component of daily modern living. The growth of email usage and cloud storage has been increasing drastically. Email is one of the most important technology innovations and a great medium of communication in the 21st century. But, every email account is receiving a large number of spam emails every day. Email spam is a rampant activity in cyberspace. The growth of Email spam has been increasing drastically. Spam email is a kind of Email-Based Advertisement. Mostly, spam emails are come from the advertisers to promote their growing business. The prime aim of this advertisement is to increase sales and income. The emails of such kind are called Unsolicited Bulk Email (UBE).

Spam emails pose a significant issue, steadily affecting users of email. Spam emails are the unwanted emails sent by unwanted users (Spammers) from unknown email addresses. Spam has become a nightmare for email users because it clutters mailboxes, occupies unnecessary disk spaces and makes us spend a considerable amount of time deleting spam messages. In order to filter spam emails, there is a need to design a powerful and intelligent spam filtering method using machine learning techniques. At present, we use the English language to send emails. Numerous techniques and algorithms have been suggested and put into practice to filter out spam in the English language. But, there is no perfect spam filtering method in the Tamil language so far. In order to filter the spam emails in Tamil, the researcher has designed a new spam filtering method in the Tamil language called Tamil Spam Filter (TSF).

The primary objective of this paper is to help Tamil speaking people who are spread all over the world in sending emails in Tamil language and help them categorize the send and received emails as good emails (HAM) or SPAM emails using Bayesian Machine Learning algorithm. This paper focuses on fast email sending and cloud storage with 100% spam protection in the Tamil language using the Bayesian Spam Filtering algorithm.

***KEYWORDS* –**Email Spam, Cloud Storage, Spam Filtering, Machine Learning, Naive Bayesian, Unsolicited Bulk Email (UBE), Email-Based Advertisement, Tamil Spam Filter (TSF), Bayesian Spam Filtering Algorithm

1. INTRODUCTION

Email stands out as a highly efficient and rapid means of communication, serving both personal and business needs. Unfortunately, the escalation of email spam has become a significant issue. As the usage of email continues to rise, the influx of spam emails is also on the rise, inundating the email accounts of users. Statistics indicate that the Internet contends with a staggering 103,447,520 [2] spam emails every minute. Each day, regular email users find themselves inundated with a substantial volume of unwanted promotional emails, overwhelming their Inboxes.

Spam email is a visible spy and a thief. We are deceived by its attractive content. Spam emails are tricky and can harm users. They try to get important information like credit card or login details and may even hack into

***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 November 2023*

social media accounts. Table 1 shows the number of people using email and how many emails are sent and received every day worldwide.

Table 1: Global Email Users and Daily Email Sent and Received

Sl.No.	**Year**	**Email users in Million***	**Growth**	Daily Email Sent and Received**(Billion)**	**Growth**
1	2018	3,823	-	281.1	
2	2019	3,930	3%	293.6	4.4%
3	2020	4,037	3%	306.4	4.4%
4	2021	4,147	3%	319.6	4.3%
5	2022	4,258	3%	333.2	4.3%

***1 Million= 10 Lakhs**

In March 2019, spam messages constituted 56 percent of email traffic. China emerged as the leading contributor to unsolicited spam emails this year, accounting for 15 percent of the global spam volume [4]. Table 2 provides detailed statistics on the worldwide sending and receiving of spam emails.

Table 2: Statistical Data on the Sending and Receiving of Spam Emails

Sl.No.	**SPAM Emails Sent and Received**	**Total Numbers**
1	Spam emails sent every day	124 Billion*
2	Spam emails received by an individual each day.	6
3	Received spam per person annually	2,200
4	Individuals replying to spam emails	20%

***1 Billion =100 Crores**

2. LITERATURE REVIEW

This section reviews various techniques employed in previous research efforts within the field of spam email classification.

Paul Graham. P, [3][2002] published an article “A Plan for Spam” outlining the spam-filtering techniques employed in the Spam Proof Web-Based Mail Reader. The article introduces the incorporation of Content-Based Filters along with the Bayesian Spam Filtering Method. Graham stated: “I think it’s possible to stop spam, and that content-based filters are the way to do it. Merely looking for the word ‘click’ will catch 79.7% of the emails in his spam corpus, with only 1.2% false positives”.

Paul Graham. P, [7][2003] discussed efforts to enhance the effectiveness of his earlier work, "A Plan for Spam." During the period from December 10, 2002, to January 10, 2003, he received approximately 1750 spam emails. Among these, only 4 successfully by passed the filters, resulting in a filtering rate of around 99.75%.

Awad W.A et al.[5][2011] introduced machine learning methods for email classification. The article explores diverse classification algorithms, including Naïve Bayes, K-NN, ANN, Rough Sets (RS), SVM, and Artificial Immune System. The experiment employed the SpamAssassin corpus, comprising 6000 emails with a spam rate of 37.04%. Among the six classifiers, Naïve Bayes (NB) demonstrated the most outstanding performance, achieving an accuracy of 99.46%.

Prof. Toran Verma et al. [6][2017] proposed a spam filtering method utilizing the Support Vector Machine (SVM) algorithm. The dataset was sourced from the Apache Public Corpus. Pre-processing involved the removal of numbers, special characters, URLs, and HTML tags. Additionally, a word stemming process was implemented, followed by mapping all words from the dictionary using a Vocab file. The SVM algorithm was then applied to the

***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 November 2023*

training set, achieving accuracy of 99.85%, and on the testing set, an accuracy of 98.9%.

3. FRAMEWORK OF THE PROPOSED TAMIL SPAM FILTERING METHOD

Tamil language is vocabulary-resourceful. For Tamil speaking people, when comparing with English language, things are easily understood when we speak or read in Tamil language. In the realm of spam filtering, a corpus is formed by pre-classified examples and the acquired rules are employed to categorize emails as either spam or ham. To combat spam, the information technology community initiated the development of spam filters to automatically identify and filter out undesirable incoming messages. Various strategies have been suggested and put into practice in the ongoing effort to counteract spam. Figure 1 illustrates the block diagram of the suggested method for Tamil spam filtering.

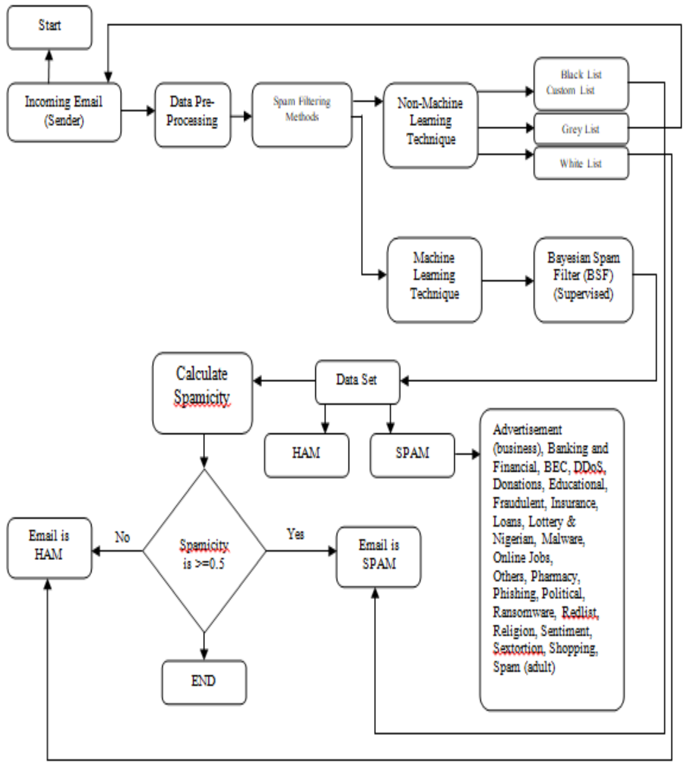

Start
Incoming Email (Sender)
Data Pre-Processing
Spam Filtering Methods
Non-Machine Learning Technique
Black List Custom List
Grey List
White List
Machine Learning Technique
Bayesian Spam Filter (BSF) (Supervised)
Calculate Spamicity
Data Set
HAM
SPAM
Advertisement (business), Banking and Financial, BEC, DDoS, Donations, Educational, Fraudulent, Insurance, Loans, Lottery & Nigerian, Malware, Online Jobs, Others, Pharmacy, Phishing, Political, Ransomware, Redlist, Religion, Sentiment, Sextortion, Shopping, Spam (adult)
Email is HAM
No
Spamicity is >=0.5
Yes
Email is SPAM
END

Figure 1: Block Diagram for the Proposed Method of Filtering Tamil Spam

3.1 DATA PRE-PROCESSING

In SPAM filtering process, pre-processing stands out as a crucial phase in email classification. In the present scenario, a substantial portion of real-world data is characterized by incompleteness, meaninglessness, and noise. The primary objective of data pre-processing is to eliminate superfluous noisy data that lacks valuable information, including the removal of anomalous data from the dataset..In this research work, there are seven steps in the preprocessing task for spam email classification. The data pre-processing involves actions such as HTML stripping, elimination of English letters, removal of numbers, exclusion of special characters, stopping words removal, elimination of duplicate words and tokenization. The effectiveness of the machine learning algorithm is significantly influenced by the quality of the data. Figure 2 depicts the block diagram of the data pre-processing procedure.

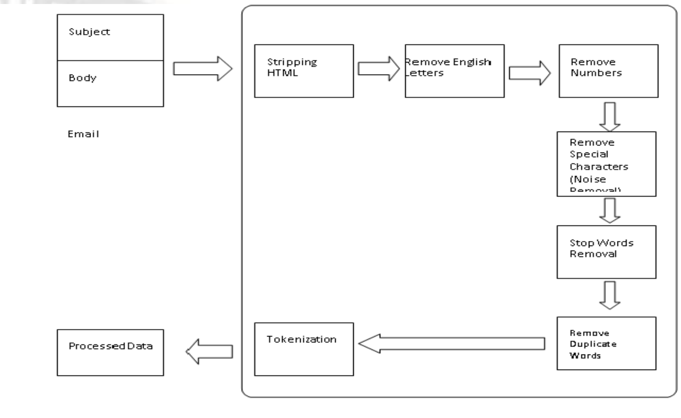


Figure 2: Block Diagram illustrating the Data Pre-Processing Process

Table 3 presents the operational principles of the data pre-processing steps.

.Table 3: Principles Underlying the Data Pre-Processing Steps

Sl. No.	Incoming Email Data Pre-Processing Steps	ஈ1, இலவச கடன் பலறனை தொழில் தொடங்க ரூ. 1 கோடி வரைகூ தொழில் பல
1	Stripping HTML Tags	இலவச கடன் பலறனை தொழில் தொடங்க ரூ. 1 கோடி வரைகூ தொழில் பல
2	Remove English Letters	இலவச கடன் பல தொழில் தொடங்க ரூ. 1 கோடி வரைகூ தொழில் பல
3	Remove Numbers	இலவச கடன் பல தொழில் தொடங்க ரூ. கோடி வரைகூ தொழில் பல
4	Remove Special Characters	இலவச கடன் பல தொழில் தொடங்க ரூ கோடி வரை தொழில் பல
5	Stop Words Removal	இலவச கடன் தொழில் தொடங்க ரூ கோடி வரை தொழில்
6	Remove Duplicate Words	இலவச கடன் தொழில் தொடங்க ரூ கோடி வரை
7	Tokenization	“இலவச”, “கடன்”, “தொழில்”, “தொடங்க”, “ரூ”, “கோடி”, “வரை”

4. SPAM FILTERING METHOD

A spam filtering method is software designed to identify and block undesired and unsolicited emails, preventing them from reaching a user's inbox. This method can be implemented using Machine Learning technique.

4.1 MACHINE LEARNING TECHNIQUE

Machine Learning involves the exploration of computer algorithms that enhance their performance autonomously through experience and data utilization. It encompasses machines acquiring information and devising processes based on the received information, without the original program designer explicitly guiding the machine on the creation and application of processes.

4.1.1 BAYES THEOREM

Rev. Thomas Bayes introduced Bayes Theorem in 1763 through his publication titled "An Essay towards solving a problem in the Doctrine of Chances." Within this work, Bayes elucidates how conditional probability can be employed to estimate the likelihood of specific events occurring based on the occurrence of certain external events.

Bayes theorem is used to learn from prior email messages. Then the filter can calculate a spam probability score against each new email message entering into the inbox.

4.1.2 BAYESIAN SPAM CLASSIFICATION

The foundation of Bayesian spam classification theory lies in the Bayes theorem. This theorem posits that the future probability of an event can be computed by considering its prior occurrence to estimate its frequency. The Naïve Bayes classification model employs prior probability word categories and the distribution of conditional probability types to determine the probability that an unknown text belongs to a specific class.

4.3 WHY BAYESIAN FILTERING IS BEST?

1) The Bayesian spam filtering approach serves as a globally recognized and multilingual anti-spam filter applicable to any language.
2) Bayesian filters, recognized as the most sophisticated type of content-based filtering, utilize the principles of mathematical probability to discern whether messages should be classified as ham or spam..
3) The main advantage of Naïve Bayesian classifiers is that it is trained based on one’s emails.
4) The Bayesian filter computes the probability associated with a new email message. If the calculated probability surpasses or equals the predefined threshold score (>=0.5), the email is categorized as SPAM.

4.4 BAYESIAN FILTERING OF PAUL GRAHAM

In August 2002, Paul Graham authored an article titled "A Plan for Spam," delving into the shortcomings of rule-based filtering in effectively handling spam emails. He introduced a novel approach to spam filtering based on the Bayes theorem.

Bayesian filtering algorithms employ probabilistic reasoning to categorize whether a mail message is spam or ham. In the context of emails, the algorithm determines the probability of an email being spam based on the presence of specific words. What sets Bayesian filters apart from other filters is their learning capability.

Paul Graham recommends the use of a modified Bayes theorem for probability calculations. Bayes theorem, when combining multiple probabilities is represented in the Equation (1).

$$P(B \wedge C) = \frac{P(C)P(B|A \wedge C)}{P(B|C)}$$

……………..Equation (1)

In Paul Graham's Bayes spam filtering rule, it is presumed that the probability of an email being spam for each individual word is already known. This probability is computed using the following equation (2)

$$P(\mathrm{Spam}|\mathrm{word}) = \frac{\frac{b}{nbad}}{\frac{g}{n\,good} + \frac{b}{n\,bad}}$$

…Equation (2)

Here, b and g represent the occurrences of a word in spam and ham emails, respectively. Additionally, nbad and ngood denote the total numbers of spam and ham emails received, respectively.

4.5 SPAM FILTER SENSITIVITY

To detect spam emails, spam filters conduct various tests on both the content and subject line of each message. Consequently, each message is assigned a set of spam scores. Typically, the default sensitivity of the spam filter is configured to classify any email messages with a score equal to or exceeding 0.5 as spam.

. 4.6 THRESHOLD SCORE

In this research work, we employ the threshold value for learning. The spam threshold score is utilized to assess an email and gauge the probability of it being classified as spam.When an email is received, a spam filter scans it and runs tests against the pre-configured rule set and custom rules. Spam filter classifies the email as spam only if it exceeds the configured threshold score.

The threshold score plays a crucial role in distinguishing emails as either ham or spam messages. In this research, a spam threshold score of >=0.5 is employed. If the threshold score surpasses this value, the incoming email is categorized as spam. The block diagram of the

***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 November 2023*

proposed spam filter with the threshold score is illustrated in figure 3.

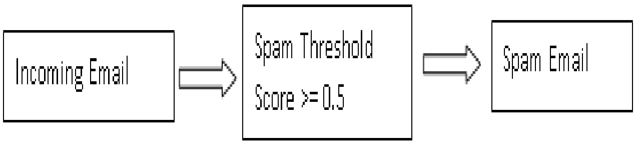


Figure 3: Depiction of the Block Diagram Showcasing the Threshold Score of the Proposed Spam Filter

5. DATA SET

Datasets are an integral part of the field of Machine Learning. The modern world's most valuable resource is data. In this data-driven era, businesses can make instant decisions, and every user experience is improved through the use of data.. Data is the hardest part of Machine Learning and the most important piece to get right. Data scientists spend 80% of their working time on preparing, managing and handling data for analysis.

"Data is more valuable than Gold" is the mantra of modern data computing. Enormous amounts of data are being generated every minute across the globe. The entry of AI and ML has facilitated the processing of this data. A dataset is a collection of data in which data is arranged in some order. The block diagram of the dataset is shown in figure 4.

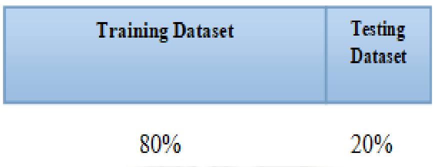
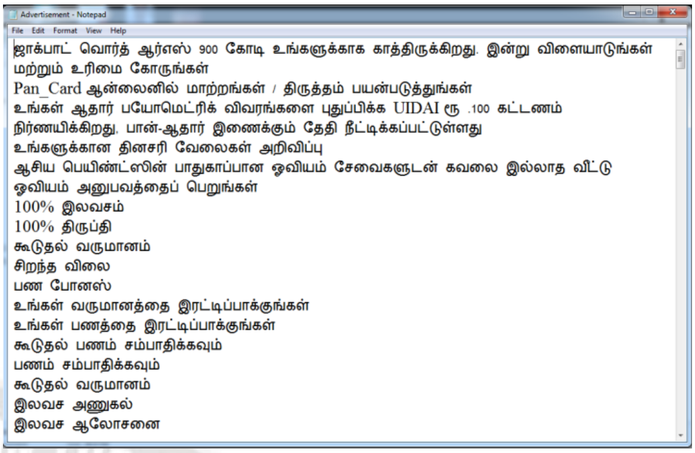


Figure 4: Block Diagram of the Dataset

5.1 TAMIL DATASETS FOR SPAM AND HAM

The dataset used for this research work was taken from two email IDs of rediffmail.com and gmail.com. In this study, we used 23 distinct datasets for training and testing, as indicated in figure 1. These datasets are stored in plain text format (txt file). The description of the text format is given in figure 5.



ஜாக்பாட் வொர்த் ஆர்எஸ் 900 கோடி உங்களுக்காக காத்திருக்கிறது. இன்று விளையாடுங்கள்
மற்றும் உரிமை கோருங்கள்
Pan_Card ஆன்லைனில் மாற்றங்கள் / திருத்தம் பயன்படுத்துங்கள்
உங்கள் ஆதார் பயோமெட்ரிக் விவரங்களை புதுப்பிக்க UIDAI ரூ .100 கட்டணம்
நிர்ணயிக்கிறது. பான்-ஆதார் இணைக்கும் தேதி நீட்டிக்கப்பட்டுள்ளது
உங்களுக்கான தினசரி வேலைகள் அறிவிப்பு
ஆசிய பெயிண்ட்ஸின் பாதுகாப்பான ஓவியம் சேவைகளுடன் கவலை இல்லாத வீட்டு
ஓவியம் அனுபவத்தைப் பெறுங்கள்
100% இலவசம்
100% திருப்தி
கூடுதல் வருமானம்
சிறந்த விலை
பண போனஸ்
உங்கள் வருமானத்தை இரட்டிப்பாக்குங்கள்
உங்கள் பணத்தை இரட்டிப்பாக்குங்கள்
கூடுதல் பணம் சம்பாதிக்கவும்
பணம் சம்பாதிக்கவும்
கூடுதல் வருமானம்
இலவச அணுகல்
இலவச ஆலோசனை

Figure 5: The Description of the Text Format of the Dataset

5.2 TMS DATASET

TMS stands for Tamil Mail Spam. In the English language, a dataset for spam email is available in huge volumes. But in the Tamil language, there is no more dataset for spam email. So, for this research, researcher created a new dataset named TMS. It contains a collection of spam emails collected from different sources from internet in the Tamil language. TMS dataset contains a total of 1667 emails. Among them, 1462 emails are spam and the remaining emails are ham emails. The summary of the TMS dataset is presented in table 4.

Table 4: **The TMS Dataset Summary**

Sl.No.	Name of the Dataset	Number of Emails	Spam Emails	Ham Emails
1	TMS	1667	1462	205

Figure 6 shows the summary of TMS dataset using a bar chart.

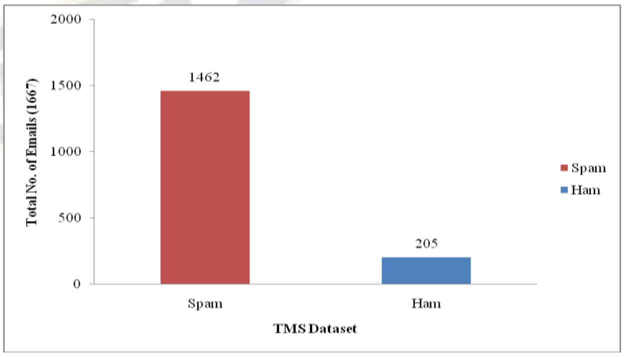


Figure 6: Summary of TMS Dataset using a Bar Chart

Table 5 shows the percentage summary of the TMS dataset.

Table 5: The Percentage Summary of the TMS Dataset

Sl.No.	Name of the Dataset	Number of Emails	Percentage of Spam Emails	Percentage of Ham Emails
1	TMS	1667	87.70%	12.30%

Figure 7 shows the percentage summary of TMS dataset using bar chart.

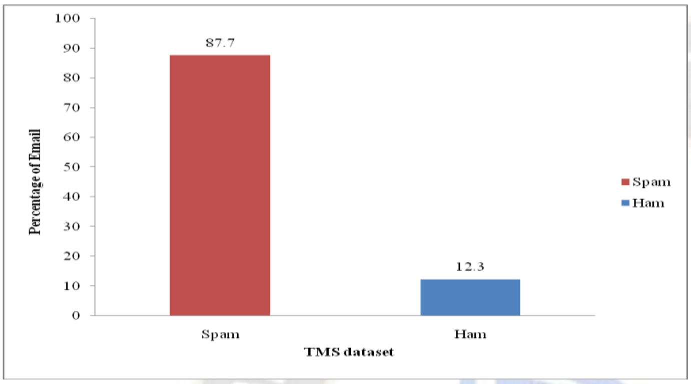


Figure 7: Percentage Summary of TMS Dataset using a Bar Chart

6. IMPLEMENTATION

This research work has been executed using ASP.NET with C#. The setup requires Visual Studio 2012 and SQL Server 2012. The successful implementation of these results has been integrated into the email system.

7. RESULTS AND DISCUSSIONS

This section provides an in-depth explanation of the proposed spam filtering in the Tamil language, along with a performance analysis and the corresponding experimental outcomes.

7.1 THE CONFUSION MATRIX

A confusion matrix is a tabular representation commonly employed to illustrate the performance of a classification model or classifier on a set of test data with known true values. It provides a visual representation of algorithm performance. Table 6 displays the confusion matrix for a binary classifier.

		Actual Class	
		Spam (Positive)	**Ham (Negative)**
Predicted Class	**Spam (Positive)**	TP	FP
	Ham (Negative)	FN	TN

Table 6: Confusion Matrix of Binary Classifier

- True Positive (**TP**): the number of spam emails correctly classified as spam
- True Negative (**TN**): the number of ham emails correctly classified as ham
- False Positive (**FP**): the number of spam emails classified as ham
- False Negative (**FN**): the number of ham emails classified as spam
- Precision (**P**): is the percentage of the messages classified as spam will actually be spam.

P= (TP/ (TP+FP))

- Recall(**R**): is the percentage of all spam emails that are correctly classified as spam.

R= (TP/ (TP+FN))

We can compute the accuracy test from the confusion matrix. The mathematical formula of accuracy is shown given below.

Accuracy= (TP+TN)/ (TP+TN+FP+FN)

7.2 PERFORMANCE METRICS

Performance Metrics are used to calculate the accuracy of a spam filtering classification model. All possible performance measures are listed in table 7.

Table 7: Performance Metrics of a Classification Model

Sl. No.	Performance Metrics	Formula
1	False Positive Rate (Specificity)	FPR = (FP/(FP+TN))
2	True Positive Rate (Sensitivity)	TPR = (TP/(TP+FN))
3	Precision (P)	P = (TP/(TP+FP))
4	Recall (R)	R = (TP/(TP+FN))
5	F1-Score	F1 Score = ((2*P*R)/(P+R))
6	Error Rate (ERR)	ERR = ((FP+FN)/(TP+TN+FP+FN))
7	Accuracy (ACC)	ACC = ((TP+TN)/(TP+TN+FP+FN))

8. EXPERIMENTS

In this research, two experiments are done. In the first experiment, a straightforward method is employed to assess the model's performance. In the second experiment, the Holdout Method is utilized for evaluating the model's performance.

8.1 EXPERIMENT 1

The collected data is partitioned into two sets: the training dataset and the testing dataset. The training dataset comprises 1239 emails, while the testing dataset includes 428 emails. Table 8 illustrates the summary of results for the testing dataset at various threshold values.

Table 8: Results Summary for Testing Dataset across Various Threshold Values in Experiment 1

Sl. No.	Threshold Values	Testing Dataset = 428 (374 Spam + 54 Ham)							
		TP	TN	FP	FN	FPR	TPR	Filtering Time (MS)	Accuracy (%)
1	> = 0.1	372	54	0	2	0	0.9946	619	99.06
2	> = 0.2	370	50	4	4	0.07407	0.9893	613	98.13
3	> = 0.3	371	47	7	3	0.1296	0.9919	614	97.66
4	> = 0.4	373	52	2	1	0.3709	0.9973	615	99.29
5	> = 0.5	374	54	0	0	0	1	605	100
6	> = 0.6	364	25	29	10	0.5370	0.9732	606	90.88
7	> = 0.7	357	22	32	17	0.5925	0.9545	608	88.55
8	> = 0.8	363	21	33	11	0.6111	0.9705	610	89.71
9	> = 0.9	352	5	49	22	0.9074	0.9411	612	83.41
10	> = 1	350	0	54	24	1	0.9358	609	81.77

Figure 8 shows the graphical representation of different threshold (T) values and accuracy of the experiment 1.

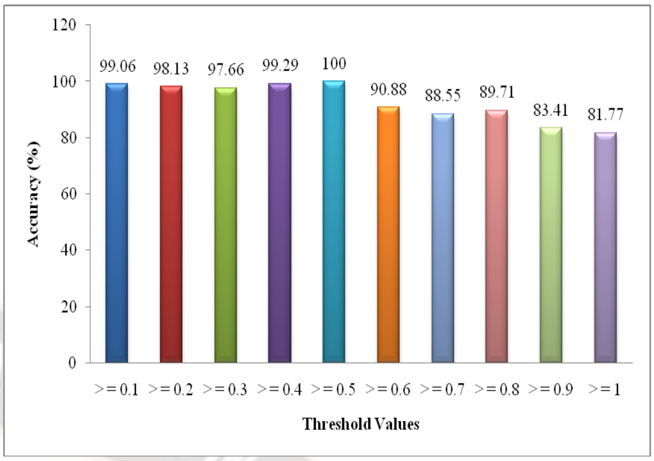


Figure 8: Graphical Representation of Different Threshold Values and Accuracy of Experiment 1

8.2 EXPERIMENT 2

Experiment 2 was done by using the Holdout method of Cross Validation technique. In the Holdout method, the dataset could be divided in a ratio of 60:40 or 70:30 or 80:20. The good rule of thumb is to use 30% of the dataset for testing. For testing purposes, we divide the total dataset in the ratio of 70:30. The total dataset contains 1667 email messages. Table 9 shows the summary of the training and testing dataset ratio for the holdout method.

A total of 500 emails are used to test the performance of the classifier. Among them, 350 emails are spam and 150 emails are ham. In experiment 2, the total percentage of spam and ham emails utilized was 30%.

Table 9: The Summary of Training and Testing Dataset Ratio for Holdout Method

Sl.No.	Total Dataset	The Ratio of Training Dataset (70%)	The Ratio of Testing Dataset (30%)
1	1667	1167	500

Table 10 shows all the evaluation metrics of the classifier.

Table 10: Results and Evaluation Metrics of Experiment 2

Sl.No.	Evaluation Metrics	Threshold Values									
		>=0.1	>=0.2	>=0.3	>=0.4	>=0.5	>=0.6	>=0.7	>=0.8	>=0.9	>=1
1	False Positive Rate (FPR)	0.1	0.08	0.06	0.333	0	0.433	0.5	0.533	0.566	1
2	True Positive Rate (TPR)	0.9	0.91	0.945	0.977	1	0.885	0.871	0.877	0.862	0.857
3	Precision (P)	0.9545	0.963	0.973	0.985	1	0.826	0.802	0.793	0.780	0.666

4	Recall (R)	0.9	0.91	0.945	0.977	1	0.885	0.871	0.877	0.862	0.857
5	F1-Score	0.926	0.93	0.958	0.980	1	0.854	0.835	0.832	0.818	0.749
6	Matthews Correlation Coefficient (MCC)	0.77	0.80	0.87	0.938	1	0.478	0.398	0.376	0.324	-0.218
7	Error Rate (ERR)	0.1	0.084	0.056	0.026	0	0.21	0.24	0.246	0.266	0.4
8	Accuracy (ACC)	90%	91.6%	94.4%	97.4%	100%	79%	76%	75.4%	73.4%	60%

Examining table 10, reveals that the spam classification for the testing dataset in iteration 1 of fold 1, at the threshold value of >=0.5, achieves 100% accuracy.In other threshold values, accuracy may be increased or decreased. But the threshold value is >=0.5, the classifier gives 100% accuracy. Like this, in other iterations, it was achieved 100% accuracy when the threshold value >=0.5.

9. CONCLUSION

The proliferation of spam presents a significant security concern, causing disruptions for end-users and consuming computing power and organizational resources. To address these challenges, various spam filtering methods and solutions have been examined in the literature review section. As technology advances, spammers are employing sophisticated techniques to send deceptive spam emails to users.The main challenge which we faced during this research was the dataset collection in Tamil language.

However, the Bayesian algorithm has proved as a powerful and successful classifier to send spam emails to the spam folder after filtering the emails. The experimental results indicate that this model achieved overall classification accuracy. The objective of this research has been successfully achieved.

First, for spam email classification in Tamil language using machine learning approach, the Bayesian algorithm is the best classifier. It shows the proposed Tamil Spam Filter (TSF) using Bayesian classifier shows the best accuracy of 100 % at the threshold value >=0.5 when compared to other threshold values.

REFERENCES

[1]. Cormack, G. Trec 2005 spam track overview. In Proceedings of TREC 2005 (Gaithersburg, MD, 2005)

[2]. https://www.iflscience.com/technology/how-much-data-does-the-world-generate-every-minute/

[3]. https://www.paulgraham.com/spam.html

[4]. https://www.statista.com/statistics/420391/spam-email-traffic-share/

[5]. Awad, W. A. and Elseuofi, S.M (2011), “Machine Learning Methods for E-mail Classification”, International Journal of Computer Applications (0975-8887),16(1),February 2011

[6]. Prof Toran Verma T and Shardhanjali (2017) “Email Spam Detection and Classification using SVM and Feature Extraction”, International Journal of Advance Research Ideas and Innovations in Technology, Volume: 3, Issue: 3 Page 1491-1495.

[7]. https://paulgraham.com/better.html