National College of Ireland

The Mechanism to detect spam emails in Marathi language using NLP

MSc Research Project
Cyber Security

Onkar Vilas Bhanarkar
Student ID: X19114761

School of Computing
National College of Ireland

Supervisor: Dr. Imran Khan

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Onkar Vilas Bhanarkar
Student ID:	X19114761
Programme:	Cyber Security
Year:	2020
Module:	MSc Research Project
Supervisor:	Dr. Imran Khan
Submission Due Date:	17/08/2020
Project Title:	The Mechanism to detect spam emails in Marathi language using NLP
Word Count:	4975
Page Count:	16

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	August 15, 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

The Mechanism to detect spam emails in Marathi language using NLP

Onkar Vilas Bhanarkar
X19114761

Abstract

Communication has increased extremely nowadays. Today's generation considers email as the fastest medium of communication within a shorter duration and for longer distance. Spam is junk email or email which users do not want in their inbox. The English and Marathi languages are completely different. Hence, detecting spam emails in the Marathi language is difficult as general research in spam filtering in other languages will not apply to that in the Marathi language. Several methods exist for finding spam mails. These methods are broadly classified as context-based or non-context-based. Most of the algorithms and techniques that are used for Spam classification in English and other languages are discussed and evaluated from different researches in this paper. Moreover, we have developed a tool by machine learning techniques that are appropriate in the Marathi Language. In our work, we have performed spam detection for emails in the Marathi Language. Experimental results were compared with respect to different machine learning models for classification to suggest an optimal solution for this problem.

1 Introduction

In today's era, electronic mails play a vital role in every field as they are an integral part of corporate life. With digitalization, cyber crime increases proportionally. Spamming email is the most commonly used method for initiating a cyber attack. Email spam has been increasing exponentially over the last 10 years leading to various scams and the national security can also be compromised due to these kinds of cyber attacks leading to huge problems caused for governments around the world. Even though extensive research and counter measures have been taken, we have yet to eradicate spamming completely. As of now, many organizations have successfully created software and hardware-based spam filters. These filters work on the concept of white listing and black listing email addresses. Although they give fairly good results, hardware spam filters are not a feasible solution for every application due to high costs. The software filters are less costly as compared to hardware filters. The two main approaches for the software filters are machine learning and non-machine learning.

There is a lot of research being done in software-based spam filters. Apart from machine learning, research shows impressive works towards non machine learning approaches. English is a universal language. Therefore, researchers have done work related spamming email detection using sentimental analysis. i.e. the content- based spam filters. These filters are not able to detect the emails from other languages like Marathi, Hindi,

or Urdu as every language has a different structure. That's why it is a very tedious task to differentiate spam emails from a genuine one.

Almost 60 million people are using Marathi language on social media as well as in emails all over the world. Often attackers use language manipulation techniques to trick the victim and avoid the general English- based filters. Therefore, our system is in accordance with these manipulation techniques in Marathi language. Some researchers have already done similar work in other languages like Urdu, English, Spanish, Chinese, Persian but not in Marathi. On social media platforms, algorithms are doing a good job at detecting pornographic and abusive content based on English language. However, similar content in Marathi is not being regulated due to language barrier. This issue is creating problems for users all over the world. Due to this predicament, we have been motivated for research on email spamming in Marathi language.

1.1 Research Question

Will the use of machine learning along with NLP techniques efficiently detect spam emails in the Marathi language ?

2 Related Work

The researchers have already drawn some disputes on spamming emails. Analysts have studied the characteristics of fake emails, email servers, and concluded the outputs regarding the detection of spam emails. On one hand, a lot of research has been done on other languages leading to multiple approaches and algorithms but its application for the Marathi language is questionable. Detecting language manipulation for any language is tedious and a never ending problem since the language changes constantly due to slangs and external influences like trends, events, etc. So in terms of Marathi language, the structure is quite difficult to understand but as the researchers have done work in other languages, it is possible on the basis of those approaches to detect manipulated spam emails in the Marathi language. Previous approaches have been done using techniques such as NLP and deep learning.

In the following section, the previous work has been explained in detail. The review is divided into the following subsections. 2.1 Spam detection in English language and 2.2 spam detection in other languages.

2.1 Spam detection in English language

Nowadays social media is being used widely in every corner of the world for different purposes such as news articles, blogs, videos, and chatting. Some of the social media platforms such as Twitter, Facebook, and eCommerce websites which are flooded with fake reviews, fake profiles, fake tweets, etc. As twitter is now quite popular for sharing ongoing events, it is regularly flooded by fake news, spams, frauds. The authors named Chao Chen, Yu Wang, Jun Zhang, Yang Xiang, Wanlei Zhou [1] have developed the Detection of spam tweets using machine learning approaches [1]. The authors collected the real time tweet data of 2 million tweets. They have used random forest classifier, moreover, they proposed some models to identify the relation between their followers, using content and graph based features. This paper includes pre-labeled data but the drawback is, this technique is limited for the English language.

Facebook is the most used social media platform in this world, due to the huge number of profiles some people use this platform to fulfill their personal agenda using different types of spams. In order to counter this problem, author Mai-Vu Tran1 [2] and his colleagues use Facebook profiles for their research. This research paper is based on the detection of fake profiles on Facebook, using content classification and user behavior. They proposed the hybrid approach with the maximum entropy model to classify whether the comments are spam or not. As per their research F1 measure in spam is 92.59 and non spam is 92.76. A study has shown that an average user comments up to 5 times in a minute if the number of comments increases then it is most probably done by a bot. This approach is quite simple for extracting user behavior features, based on this feature their accuracy is increased by 2%.

As of now, the above papers show that even though they have used less data and it has a lot of parameters on the basis of which they used the machine learning model as per their data specification. In addition, their text data is not enough for text classification that's why the author mentions this drawback in their respective paper. Also, we get to know that human behavior can be considered for feature extraction.

All over the world rate online shopping is highly elevated, it is common practice for people to check the reviews of products before buying. The reviews are generally about the overall performance, efficiency, and handling experience, therefore the sales of products are directly dependent on the user reviews. Users post their reviews directly on the eCommerce platforms which in turn influences the buying populace. These reviews are beneficial for both buyers and manufacturers. Due to the nature of these reviews, it is easy to mislead people on such platforms using profile spam and bots. These reviews could destroy the product's reputation and damage the platform's authenticity. This serious problem was discovered by Nitin and Bing [3]. They consider the reviews which are used by both manufacturers as well as the consumer. Moreover, these reviews are classified into three categories such as false opinions, true opinions, and opinions which are related to respective brands. They have used supervised machine learning algorithms like Naïve Bayes, Support vector machines, and Logistic regression for detecting the reviews. Among these three algorithms, Logistic regression is most efficient. In another research [4] about online paid reviews of restaurants, it proposes supervised learning approaches such as Logistic regression and Support vector machines. They explore a lot of features that made their model quite efficient but the only drawback is insufficient data.

As of now, we have seen review spamming and its detection. Many researchers have applied only machine learning algorithms on classified data. The supervised algorithms like Naive Bayes, Support vector machines, Logistic regression are being used as classified in spam detection models.

Detection of spamming SMS can be seen in the study done by Sethi, Bhandari, and Kohli[5]. In this paper, they have used the stop words that is NLP operation to pre-process the database. Authors used multiple machine learning algorithms such as Naive Bayes, Random forest, and Logistic regression for which they got accuracy 98.4%, 97.09%, 94.3% respectively. Adi Wijaya and Achmad bisri [6] proposed the solution on email spam, they have used the logistic regression and decision tree algorithms to detect the spam emails. The drawback of this research is the absence of analysis of Decision tree training data set, due to the presence of multiple parameters their accuracy is dependent on it. Moreover, due to unwanted information, the accuracy of the model and its performance is affected. Similarly, the studies by Masurah Mohamad and Ali Selamat [7] have used the TF-IDF(Term frequency inverse document frequency) for the feature selection. Before

the feature selection process they classified emails into two parts that is TEXT and Image. As for image recognition they have used Optical character recognition(OCR). They have worked on only 169 emails.

As of now, our literature shows different studies in different fields, we get to know different researchers have used different machine learning algorithms to check the accuracy. That's why we will consider a number of machine learning algorithms to compare its accuracy, moreover, if the dataset is insufficient the accuracy varies. We decided to go with a large amount of data.

2.2 Spam detection in other Languages

In this particular section, we are going to focus on research being done on spam detection in other languages. Reading other papers can be beneficial to our research because they can provide us with different approaches. While studying these papers we are mostly focused on statistics like accuracy, datasets, algorithms, and shortcomings. Unlike spam detection in English, resources like NLP libraries are not readily available for other languages compelling many researchers to create custom libraries and synthetic datasets.

Mohit Agrawal and R. Leela Velusamy [8] proposed an unsupervised approach to detect spam messages. They implemented the Reliability-based Stochastic Approach for Link Structure Analysis (RSALSA) technique on a dataset that has 13,188 messages and 28,988 spam reports. They achieved accuracy of HITS: 81.75%, SALSA: 87.08% and R-SALSA: 89.25%. Kashif Mehmood and Hammad Afzal [9], have completed research in bilingual tweets using DMNBText and Naïve Bayes techniques. They had 1463 tweets, in that, almost 450 tweets(30%) were spammed, the rest of 70% were original tweets. They were successfully able to distinguish the Roman Urdu language and English Language with accuracy as Naive Bayes: 95.42%, DMNB Test: 95.12%, and LibLinear: 94.60%. furthermore, Kashif Mehmood, Hammad Afzal, Awais Majeed, Hassan Latif [10], have successfully able to detect the Urdu Spam from SMS using Naïve Bayes Multinomial, DMNBText, LibSVM, Liblinear, Sequential Minimal Optimization (SMO). In both papers, the objective is similar but the dataset is different. In this paper, [10] accuracy decreased by 2% but they got 93.33% accuracy with Sequential Minimal Optimization (SMO).

On the other hand Tanvirul Islam, Subhenur Latif and Nadim Ahmed [11], proposed research under the title of Using Social Networks to Detect Malicious Bangla Text Content, in that they used different social networking platforms for collecting data. The research is about detecting inappropriate texts in Bangla from other languages. They got 82.44% accuracy with the Multinomial Naïve Bayes (MNB) classifier. Moreover, A similar kind of research has been done by Chunyu He and Yijie Shi [12]. They used WeChat accounts, Tou Tiao comments, and Netease news. Using the Support vector machines, machine learning technique where they were able to detect Chinese comments based on Chinese characteristics which give results as Precision:90.36% Recall: 94.93% F1: 91.04%. But they are unable to detect the malicious Chinese comments.

In the above research Urdu, Bangla, and Chinese languages have a different type of language structure so it will not help directly in my research but It has different types of algorithms so that it helps me to choose a suitable algorithm.

The research has been done on Text and image classification from spam emails by Anirudh Harisinghaney, Aman Dixit, Saurabh Gupta, and Anuja Arora [13]. They have used the K-nearest neighbors algorithm, Naive Bayes algorithm and reverse DBS CAN

algorithm on 200399 messages which belonged to 158 users with an average of 757 messages per user which gives accuracy 87%. In this paper, we understand the dataset is larger so the accuracy depends on the dataset and type of data. Feature selection and similarity coefficient based method for email spam filtering by Ali Ahmed A.Abdelrahim, Ammar Ahmed E. Elhadi, Hamza Ibrahim and Naser Elmisbah [14] gives the concept of feature selection in Urdu language.

Michal Prilepok and Milos Kudelka proposed spam detection using the nearest community classifier which gives 93.78% accuracy [15]. Simranjit Tuteja and Nagaraju Bogiri [16] suggested a spam filter that uses BPNN classification and K-Means clustering for 200 emails. The data set is less so they got 98.4 accuracies. similar kind of research with another ensemble classification technique which gives 98.5% accuracy [17]. Alaa Mustafa El-Halees [18] has researched the Arabic language. The research is to detect Arabic spam emails using a different type of machine learning algorithm. In that, he first classified languages. We need to know that first, we have to classify the data. After classification, it gives almost 89.77% accuracy but it is for the Arabic language. Mohammad Ehsan Basiri, Neshat Safarian and Hadi Khosravi Farsani [19] have recommended a system that detects spam reviews In the Persian language. They have used a Decision tree algorithm with F-1 measure 0.78 which is quite good for the Persian language.

So far, we have seen numerous studies that try to provide a solution for content-based spam filtering with the help of classification. This is an approach which categorizes the mails and then the user declares a certain category as spam. Disregarding the languages used, the accuracy of all these proposals seemed like an important rating. However, it is surprising how there hasn't been sufficient research in the supervised learning department for spam filtering. At first, supervised learning might lead to an overestimated solution to the problem at hand. However, when we consider the attackers, they seem to have proceeded from bulky bodied spams or simple phishing attempts to an ever-evolving area of manipulation. Thus, for the security of vulnerable masses, we must use the advancements made in the field of data science to produce an aware system to detect manipulation.

This takes us to the next challenge of looking at studies done in the Marathi language relating to language processing. Detection of a paraphrase using machine learning by Darshana S. Bhole and Sandip S. Patil [20] they have used the Support vector machine technique. Again, this is an unsupervised method but for language detection. Snehal V. Pawar and Swati Mali [21] proposed Sentiment Analysis in Marathi Language using machine learning. This method was the most inspiring as it included an aware system. In our system, natural language processing will be used to tackle manipulative speech in emails. It is believed that this will solve the question of detection.

3 Methodology

In this research, we follow the knowledge discovery database methodology [22] to collect useful information from data. It consists of various steps such as data selection, data pre-processing, data transformation, data mining, and evaluation of data. In every step, respective operations are there. The following figure explains the KDD approach. The remaining steps discussed in architecture.

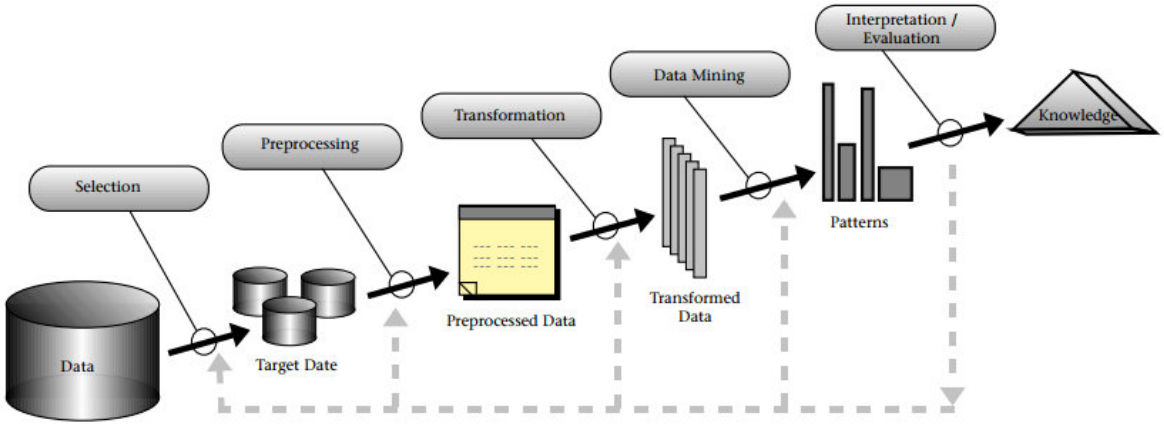

Figure 1: Knowledge Discovery in Databases

3.1 Data Description

In this chapter, the email data is confidential data, the Marathi language is expanded widely, but for the research, open source data is not available, that's why we decided to create a synthetic dataset for the research. The data consist of overall 5569 entries with respective labels (ham, spam). The ham data consists of 4822 records and spam consists of 747 email records.

Table 1: statistics of dataset

Total Count	Spam	Ham
5569	747	4822

The above table shows the statistics of the dataset. Pre-processing has been done on same dataset. In the next phase the pre-processing part has been broadly described.

3.2 Data Pre-processing

Before using data in machine learning models the pre-processing of data is essential, generally, in pre-processing, unwanted data or noise is eliminated. This helps the model to perform efficiently and gives a good result. As in the initial phase of pre-processing firstly checked for the blank or null values. The blank or null values need to be removed. After handling the null or missing values, data transferred in another file.

In the core part of pre-processing that is text analysis, in-text analysis there is a need to perform some activities so that it is understandable to the model. All activities are listed below.

- Removing the numeric values from the data.
- Removing English words from data.
- Removing Email address from data.

- Removing web addresses including HTTP and HTTPS from data.
- Removing Currency symbol from data.

In the next phase of the pre-processing of data is tokenization. There are mainly 2 libraries available for tokenization in the Marathi language. Neural language toolkit for Indic Languages (INLTK) and The classical Language toolkit (CLTK), both are python libraries. Both libraries provide similar kinds of functionalities. INLTK libraries are used to tokenize the text.

The words such as आहे, या, आणि, व, नाही, आहेत are called stop words. This type of words do not contain any useful data.

3.3 Feature extraction and transformation

The feature extraction is an important part of the machine learning model, which has an extensive amount of data. Choosing correct features, which gives the correct accuracy and reduces the burden on the system. In this system, we used the words as our features. In this research, we used the top 1500 frequent words as features in the whole dataset. Basically, the feature extraction process comes after the tokenization. The reasoning behind considering only one feature is, the words come repeatedly in both emails. As the structure of spam emails is different from the ham emails in Marathi language.

3.4 Data mining models

In order to classify the spam and ham emails, in our research we consider the different types of machine learning models. Such as support vector machines, K-Nearest Neighbors, Decision tree, random forest, Logistic regression, Stochastic Gradient Descent classifier, Naive Bayes, and Support vector machine linear. Among them, Logistic regression provided more accuracy than the other data mining algorithms.

Logistic Regression: When considering classification methods in unsupervised learning, our strategy depends on the type of data we need to analyze. In our case, we are working with categorical data. This attracts us towards logistic regression. When it comes to probabilistic prediction, model fitting on categorical data is facilitated by Logistic regression. Here, the dependent variable is of the binary form where the data is coded to have two levels, 0 or 1. In relation to our report, we will use Logistic regression to predict if an email is a ham (0) or spam (1).

To understand Logistic regression, we must take a look at its parent class of algorithms, Generalized Linear Models (GLM). They have a fundamental equation of the form:

$$g(E(y)) = \alpha + \beta x1 + \gamma x2 \tag{1}$$

Where g() is considered the link function with $E(y)$ as it's target variable. On the other side α, β, and γ are to be predicted by the model. To convert this generalized equation into a Logistic regression equation, we must transform the link function. Since we know the response has to be binary, we can expect g() to have targets around (p/(1-p)). Furthermore, we need to model a non-linear association in a linear way. This can be done easily by a logarithmic transformation on (p/(1-p)). On the predictor side, we will consider our email data. Therefore, our Logistic regression model becomes:

$$log(p/(1-p)) = \beta o + \beta(emails) \tag{2}$$

$$p = 1/(1 + e(-(\beta o + (emails)))) \tag{3}$$

Transforming this equation to find the probability:

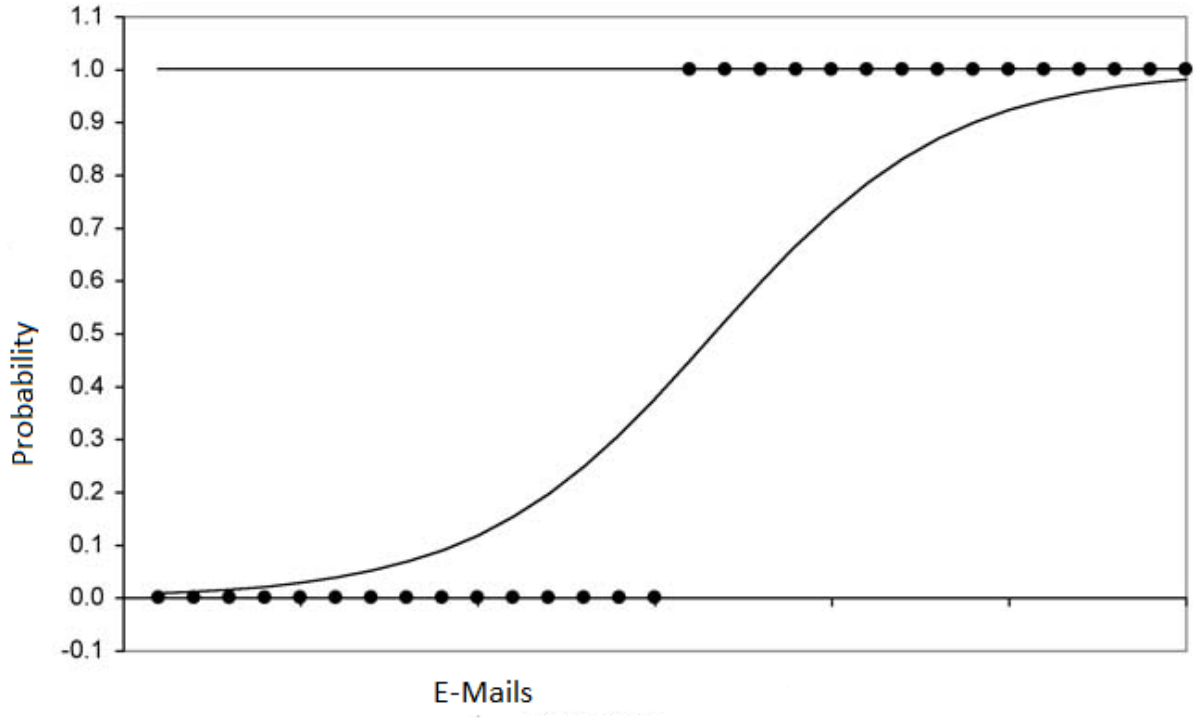

Figure 2: Logistic Regression Graph

3.5 Evaluation matrix

For investigating the result we performed or calculated the related parameters as these parameters used in previous related work. We analyze the result of all the data mining algorithms as mentioned above. Mainly we focused more on accuracy because, in this problem of spamming emails, even a single spam email can lead to serious problems incurring huge damages. Moreover, we used the various evaluation matrices such as accuracy, precision, recall, F1 score.

4 Design Specification

The figure shows the architecture of research. In the first stage data gathered by the sources. In the next phase, data pre-processing is carried out, including various tasks. Such as label encoding, elimination of null values, removing English words, numeric values, websites, and currency symbols. Later pre-processed data used to extract the features. Then we used our feature selection algorithm that is the top 1500 words. Next,

we split the data into two categories training and testing with the ratio 80:20 respectively. The next step is about training the machine learning model using a Logistic regression algorithm. Which train the model and that we can use on the testing data. After the prediction, we calculated the results in terms of the confusion matrix.

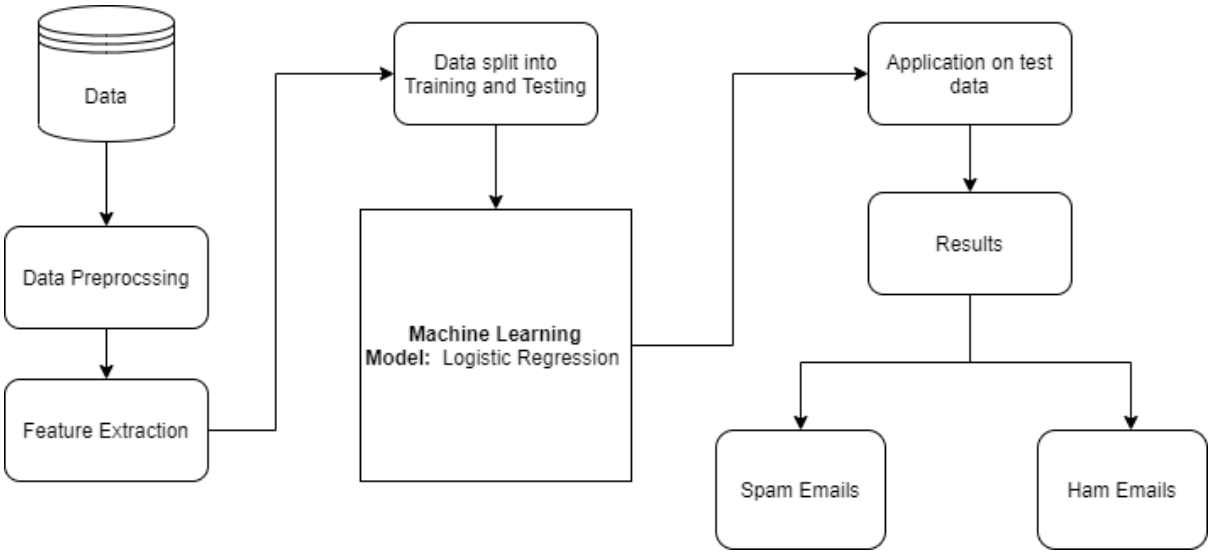

Figure 3: Architecture - Detection of Marathi Spam Emails

5 Implementation

This section provides detailed implementation of research. For the implementation, we have used the python 3.6.9 version. Python is the default choice of programming because of their features related to machine learning and it is easy to use as Python provides many machine learning and NLP related libraries. We have used mainly nltk, pandas, etc as it is basic required libraries for every machine learning programming. As in every machine learning project, the process is similar to the one we followed in our program.

Firstly we uploaded the dataset file in google collab. Next step using label encoding we encoded the data. As 0 is ham and 1 spam. Later in pre-processing, we have used regular expression. The reason behind using this is, these regular expressions are designed as per our data.

Table 2: Regular Expressions

Regular Expressions	**Category**
r'^.+@[^\.].*\.[a-z]{2,}$	Email
r'^http\://[a-zA-Z0-9\-\.]+\.[a-zA-Z]{2,3}(/\S*)?$	Web Address
r'£|\$'	Currency
r'\d+(\.\d+)?'	Phone Number
r'[a-zA-Z]'	English Words
r'[!@#$%^&*()_+,.;?]'	Common Symbols

In the next step, we have used the cltk libraries for removing stop words. The next step, tokenization, is very important. As python libraries provide the various tokenization algorithms for example "text.split()" is a basic tokenization algorithm. In our case, data is in the Marathi language so these techniques are not that useful. As CLTK and INLTK both libraries are efficient enough to provide the tokenization. In our case, we have used the INLTK tokenization algorithm, for better accuracy

The next step is feature extraction, on this step our whole model is dependent. The features are extracted from the tokenize words. As we implemented a lot of features but they didn't work. So at last we decided to take the top 1500 words in the database and along with train the different models for checking accuracy.

```
[ ]  word_features = list(all_words.keys())[:1500]
```

```
[ ]  def find_features(message):
         words=tokenize(message ,'mr')
         features = {}
         for word in word_features:
             features[word] = (word in words)

         return features
     features = find_features(processed[0])
     for key, value in features.items():
         if value == True:
             print (key)
```

```
_जु
र
ाँग
_बिंदू
_वेडा
_पर्यंत
_जा
_फक्त
_महान
```

Figure 4: Feature Extraction

The accuracy is a very important evaluation of our research. Because our research is in the field of cyber security. Single spam email may give much more damage. For example, if any system can give an accuracy of 90%. That means in 100 emails that system can classes the 90 emails correctly. Therefore, this is a huge thing. That's why accuracy plays a crucial role in cyber security based spam detection applications.

The next step is choosing the appropriate model for machine learning. Multiple classifiers were used, and after analyzing their accuracy we decided to implement the logistic regression. The table shows the accuracy of different machine learning classifiers.

In this model, we decided to go with an 80:20 pattern for selecting the database. That is 80% is for training data set, and 20 % testing dataset. In the last step, we used this model to calculate evaluation matrix parameters, such as accuracy, F1 score, recall, precision.

Table 3: Accuracy Table

Model	**Accuracy**
K-Nearest Neighbors	91.95
Decision Tree	95.04
Random Forest	97.7
Logistic Regression	98.27
SGD Classifier	97.7
Naive Bayes	97.63
SVM Linear	97.84

6 Evaluation

In this section, we have considered the performance of contrasting machine learning algorithms on content based data. This evaluation is calculated in terms of accuracy, precision, recall, and F1 score. Considering the importance of accuracy, The lowest accuracy is 91.95% and the highest is 98.27%, with respective algorithm K-nearest neighbors and Logistic regression. In the classification, we have seen many algorithms give accuracy in the range 95% to 98%. As the researcher got the accuracy in other languages in the range between 90% to 98%, as the structure of language and available resources also more as compared to the Marathi language. The researcher[15] got 98.4% Accuracy, although the size of the dataset is only 200 emails. The comparison to other researchers such as we got decent accuracy.

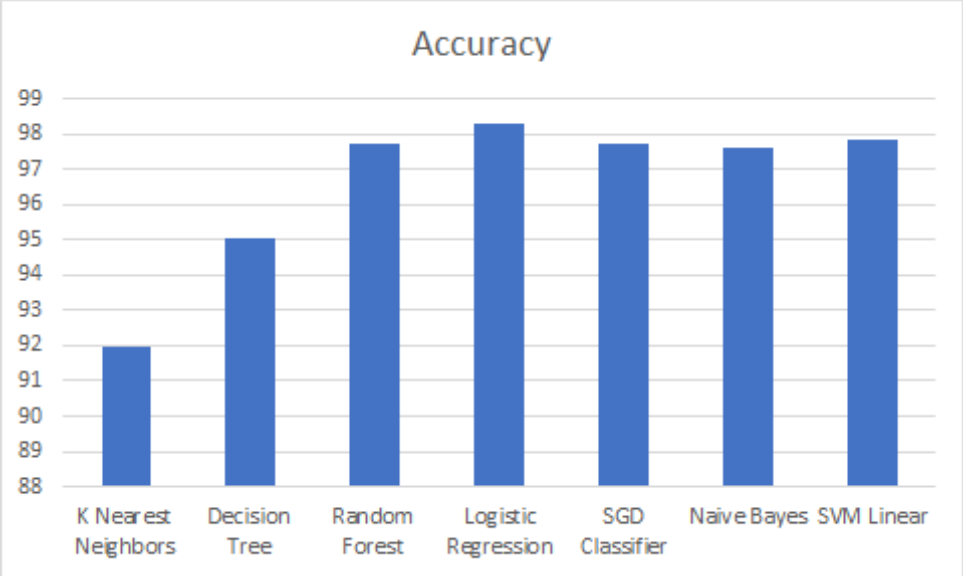

Figure 5: Accuracy

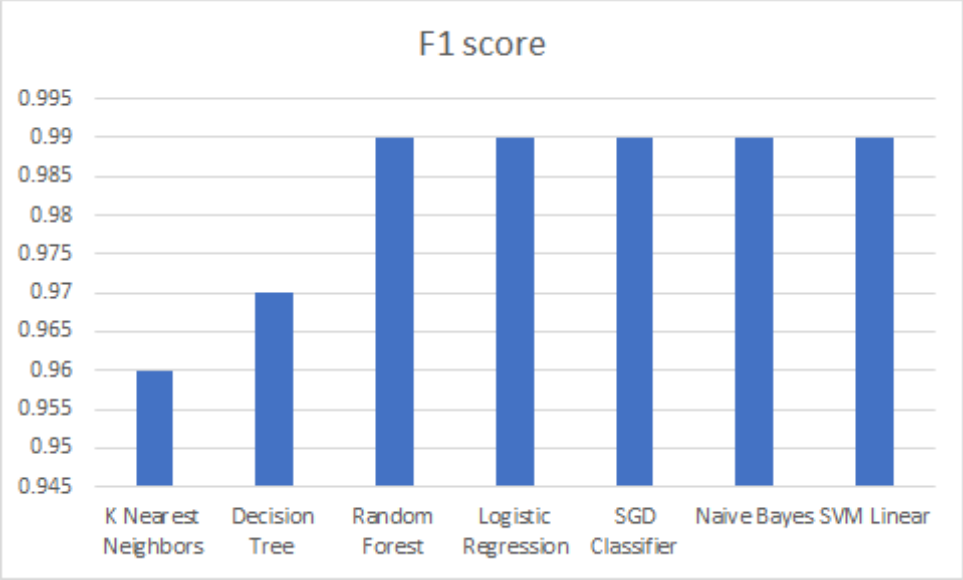

Figure 6: F1 Score

The F-1 score in the Persian language[19] has 0.78, compared to this we got 0.99 which is very close to 1. The above figure shows the graphical representation of the F-1 score. The following figure shows the precision and recall values. Precision evaluated on the basis of predicted positive values and Total positive values. The recall is interpreted as the number of correct samples are classified by the algorithm.

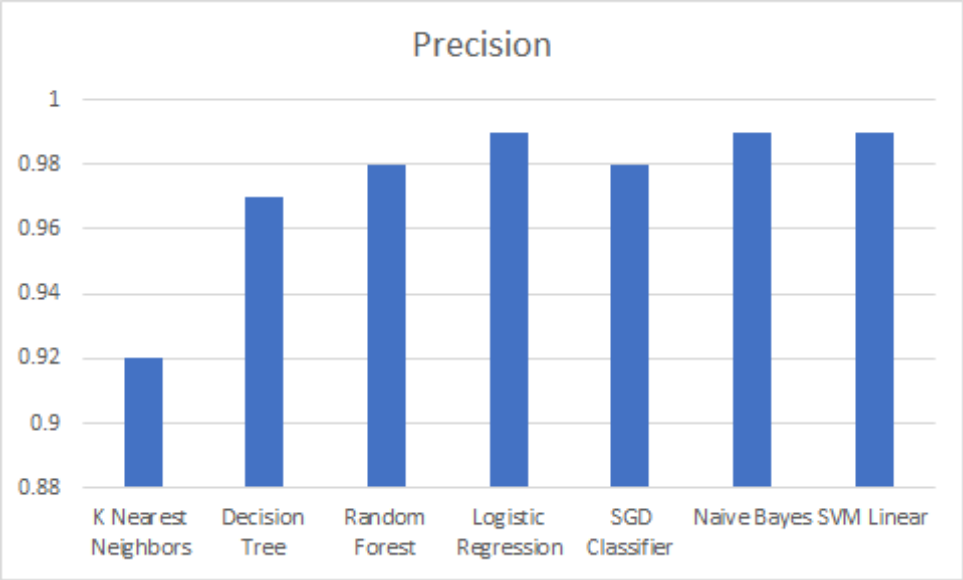

Figure 7: Precision

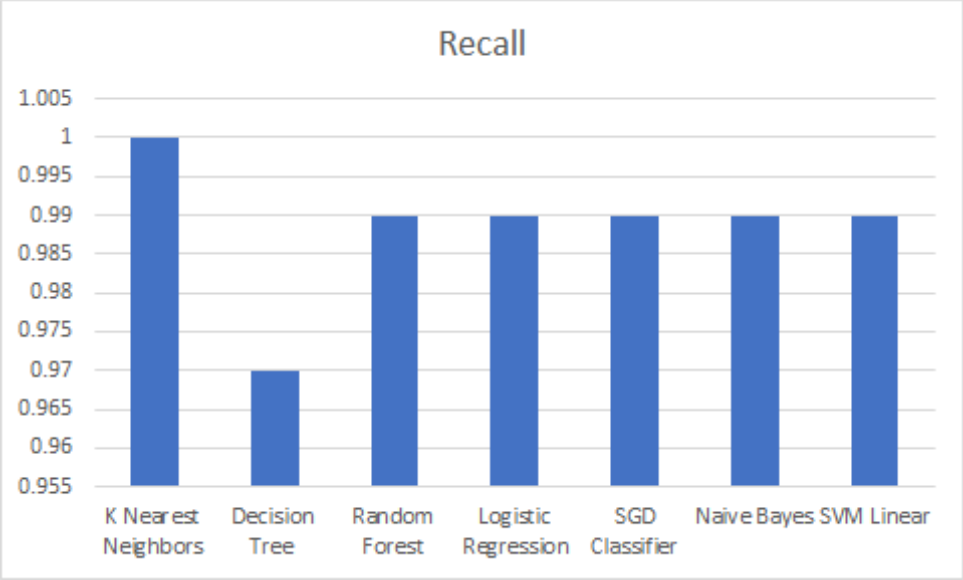

Figure 8: Recall

The following tables show the confusion matrix of Logistic regression. Out of the 1393 samples our algorithm successfully classified the 1369 samples. There are only 24 samples that are misclassified.

Table 4: Confusion Matrix

Actual	Predicted	
	Ham	Spam
Ham	1191	8
Spam	16	178

6.1 Discussion

Several challenges presented themselves during the classification of spam emails. capturing the behavior of spammers was a substantial challenge, as they are highly dynamic. Considering the accuracy of various models, many of them have performed well, but Logistic regression has outperformed with the lowest misclassification of emails. The current system is only about detection. Our objective is to correctly classify the spamming email. Our model has successfully classified the spam email and the model is less prone to error. The implemented approach would be an satisfactory fit for detection of spam emails in Marathi language. With the help of this system, we can build a prevention system.

The following are the few limitation of the systems.

- This research is only limited to Marathi language. It will not work with other languages. In our research, we only eliminated the English language from the data. If data in other languages is present, we could not detect these languages.
- Our research solely focuses on content-based spam detection, text feature extraction

was done using NLP, email header information has not been used for detection of spam.

- As our system is not able to prevent spammers directly but this system could be implemented on an email server. In such a case, we can implement the system on an email server which will act as an cost saving mechanism for consumer as well as organization perspective. But an email server is expensive for deployment and security.
- The intelligence achieved by this model would be limited and would have to be combined with another system for this intelligence to be truly used.

While using detection techniques for spammers it was found that machine learning models have the capability to easily classify spam vs ham. The above research displays the ability of several machine learning algorithms, out of which Logistic regression has an edge over the other models. These models have been evaluated using machine learning metrics. The following metrics are used for the evaluation of our machine learning model.

1. accuracy : we have selected Logistic regression as a model for predicting the Marathi email spam. the accuracy of model can be seen as 98.27% which peaks every other algorithm. from this we can evaluate that Logistic regression will perform better even when the size of the data-set changes.
2. F1 score : while computing values for F1 scores of various algorithms like Random forest, Logistic regression, Stochastic Gradient Descent, Naive Bayes and Support vector machine linear are all 0.99. this shows us that the balance of the precision and recall for these algorithms is on an exceptional level.
3. precision : precision value for Logistic regression is calculated as 0.99 where as when we compare the values for Naive Bayes and Support vector machine, we can see that they are both settled at 0.99 as well. Whereas when we compute the precision values of K-nearest neighbours, Decision tree and Stochastic Gradient Descent, there is decreased in the precision rate
4. recall : While performing the experimentation it was found that the recall value for Logistic regression is 0.99, while comparing to the other recall values we can see in the above graph the K-nearest neighbours algorithm has the value of 1. .

7 Conclusion and Future Work

It is decisive to indicate that spamming and phishing emails are destructive, and its consequences can be faced for a longer duration while potentially leading to system breakdowns. Our system provides a software filtering solution on detecting spamming emails which includes the three important steps like Tokenization, Feature extraction, and Machine Learning model Training. After tokenization using INLTK library, feature extraction was done on these distinct tokens. This feature extraction process uses 1500 common words which turned out to be the most appropriate technique for Marathi spam filtering. Furthermore, for training the machine, the learning model used Logistic regression for prediction to provide us with the most optimal spam filtering pipeline. We believe that

this system outperforms other machine learning algorithms when it comes to solving the problem of spamming in the Marathi language.

In the Future, The current system can be extended by combining multiple feature sets and datasets. Moreover, using the references of this system, a prevention module in email servers can also be made.

References

[1] C. Chen, Y. Wang, J. Zhang, Y. Xiang, W. Zhou, and G. Min, "Statistical features-based real-time detection of drifted twitter spam," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 4, pp. 914–925, 2017.

[2] T. Vuong, V. Tran, M. Nguyen, C. Thi Nguyen, T. Pham, and M. Tran, "Social-spam profile detection based on content classification and user behavior," in *2016 Eighth International Conference on Knowledge and Systems Engineering (KSE)*, 2016, pp. 264–267.

[3] N. Jindal and B. Liu, "Opinion spam and analysis," in *Proceedings of the 2008 International Conference on Web Search and Data Mining*, ser. WSDM '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 219–230. [Online]. Available: https://doi.org/10.1145/1341531.1341560

[4] M.-C. Ko, H.-H. Huang, and H.-H. Chen, "Paid review and paid writer detection," in *Proceedings of the International Conference on Web Intelligence*, ser. WI '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 637–645. [Online]. Available: https://doi.org/10.1145/3106426.3106433

[5] P. Sethi, V. Bhandari, and B. Kohli, "Sms spam detection and comparison of various machine learning algorithms," in *2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN)*, 2017, pp. 28–31.

[6] A. Wijaya and A. Bisri, "Hybrid decision tree and logistic regression classifier for email spam detection," in *2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE)*, 2016, pp. 1–4.

[7] M. Mohamad and A. Selamat, "An evaluation on the efficiency of hybrid feature selection in spam email classification," in *2015 International Conference on Computer, Communications, and Control Technology (I4CT)*, 2015, pp. 227–231.

[8] M. Agrawal and R. Velusamy, *PRISMO: Priority Based Spam Detection Using Multi Optimization: 6th International Conference, BDA 2018, Warangal, India, December 18–21, 2018, Proceedings*, 01 2018, pp. 392–401.

[9] H. Afzal and K. Mehmood, "Spam filtering of bi-lingual tweets using machine learning," in *2016 18th International Conference on Advanced Communication Technology (ICACT)*, 2016, pp. 710–714.

[10] K. Mehmood, H. Afzal, A. Majeed, and H. Latif, "Contributions to the study of bi-lingual roman urdu sms spam filtering," in *2015 National Software Engineering Conference (NSEC)*, 2015, pp. 42–47.

[11] T. Islam, S. Latif, and N. Ahmed, "Using social networks to detect malicious bangla text content," in *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, 2019, pp. 1–4.

[12] C. He and Y. Shi, "Research on chinese spam comments detection based on chinese characteristics," in *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*, 2018, pp. 2608–2612.

[13] A. Harisinghaney, A. Dixit, S. Gupta, and A. Arora, "Text and image based spam email classification using knn, naïve bayes and reverse dbscan algorithm," in *2014 International Conference on Reliability Optimization and Information Technology (ICROIT)*, 2014, pp. 153–155.

[14] A. A. A. Abdelrahim, A. A. E. Elhadi, H. Ibrahim, and N. Elmisbah, "Feature selection and similarity coefficient based method for email spam filtering," in *2013 INTERNATIONAL CONFERENCE ON COMPUTING, ELECTRICAL AND ELECTRONIC ENGINEERING (ICCEEE)*, 2013, pp. 630–633.

[15] M. Prilepok and M. Kudelka, "Spam detection based on nearest community classifier," in *2015 International Conference on Intelligent Networking and Collaborative Systems*, 2015, pp. 354–359.

[16] S. K. Tuteja and N. Bogiri, "Email spam filtering using bpnn classification algorithm," in *2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*, 2016, pp. 915–919.

[17] S. Suryawanshi, A. Goswami, and P. Patil, "Email spam detection : An empirical comparative study of different ml and ensemble classifiers," in *2019 IEEE 9th International Conference on Advanced Computing (IACC)*, 2019, pp. 69–74.

[18] A. El-Halees and A. Hammad, "An approach for detecting spam in arabic opinion reviews," *International Arab Journal of Information Technology*, vol. 12, 01 2015.

[19] M. E. Basiri, N. Safarian, and H. K. Farsani, "A supervised framework for review spam detection in the persian language," in *2019 5th International Conference on Web Research (ICWR)*, 2019, pp. 203–207.

[20] D. S. Bhole and S. S. Patil, "Detection of paraphrases for devanagari languages using support vector machine," in *2018 International Conference on Communication information and Computing Technology (ICCICT)*, 2018, pp. 1–5.

[21] I. J. Ijritcc, "Sentiment analysis in marathi language." [Online]. Available: https://www.academia.edu/36867981/Sentiment_Analysis_in_Marathi_Language

[22] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Magazine*, vol. 17, pp. 37–54, 1996.