iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VII **Month of publication:** July 2021

DOI: https://doi.org/10.22214/ijraset.2021.36383

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VII July 2021- Available at www.ijraset.com

Email Spam Detection using SVM

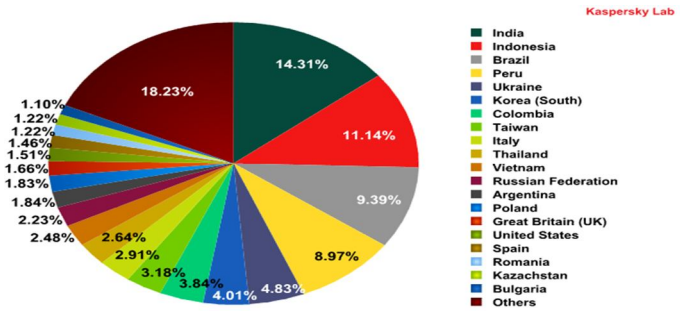
Azhar Baig[1], Rexina D[2], Neelagal Gnaneswari[3], N Anjana[4], P Aishwarya[5]
[1, 2, 3, 4, 5]*Ballari Institute of Technology and Management*

Abstract: ***E-mail contributes to internet messaging as a necessary component. Spam mails are unwanted messages that appear in large numbers and are exploited by spammers to divulge personal information of the user. These e-mails are frequently company/control announcements or malware that the user receives suddenly. Email spamming is one of the Internet's unsolved challenges, causing inconvenience to users and loss to businesses. Filtering is one of the foremost widely used and important methods for preventing spam emails. Email filters are commonly wont to organize incoming emails, protect computers from viruses, and eliminate spam. We present this method to classifying spam emails using support vector machines during this study, the SVM outperformed other classifiers.***

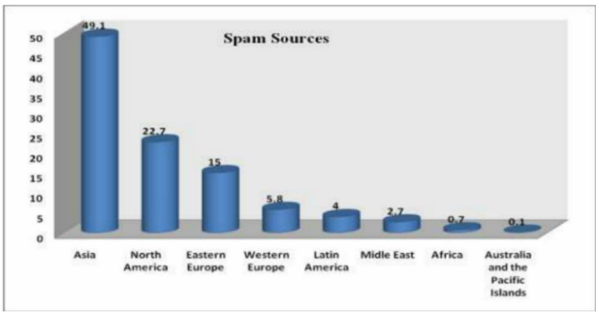
I. INTRODUCTION

E-mails have recently become a convenient and vital mode of communication for all Internet users. Spam, also called unsolicited commercial e-mail or bulk e-mail, could be a scourge of e-mail communication. Spam is typically compared to junk on paper. In truth, spam filtering is an application that categorizes emails and encompasses a high likelihood of detecting spam. Spam may be a continuing problem with no perfect solution, and there's no full spam treatment approach.

Separating real emails from spam has recently been greatly improved and enhanced. in step with Kaspersky Laboratory's recent analysis, over 65.7 percent of all emails are classified as spam. During this case, a big amount of bandwidth is squandered, and an overflow occurs during the e-mail transmission. Inline with data, the US, China, and Asian countries are the highest three suppliers of spam, with 21.9 percent, 16.0 percent, and 12.5 percent, respectively. FIG[1] shows the survey graph by Kaspersky where the spam generated by each country has represented a chart. Fig[2] shows spam generated, represented as per regions. In reality, distinguishing spam from legal emails could also be thought of as a kind of text classification, because the shape of all emails is usually textual, and therefore the type must be established by receiving the spam. SVM are supervised learning models that outperform other models in terms of generalization . They examine data and detect patterns, and are used for classification and multivariate analysis. SVM could be a depiction of the cases as points in space, smapped stated that trials of the various categories are separated by as far as feasible



FIG[1] Spam generated by countries.



FIG[2] Spam generated by countries represented in regions.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VII July 2021- Available at www.ijraset.com

II. PRELIMINARY DISCUSSIONS

In order to come to a decision upon the simplest algorithm for getting the simplest results, we checked the accuracy of various algorithms for our dataset. The algorithms that we checked with were Naïve Bayes, BernoulliNB, Decision Tree Classifier, and SVM. The accuracy provided by SVM is that the best as compared to others. The SVM yielded an accuracy of 97.29%, whereas Naïve Bayes yielded 85.60%, Decision Tree Classifier yielded 86.24% and BernoulliNB yielded 80.20%. This accuracy test helped when making a decision on the most effective algorithm for the Project.

A large type of classification algorithms is linked to the spam identification area, with support vector machine classification being particularly well recognized for its good generalization performance impact. SVM could be a sophisticated data categorization algorithm. Despite the very fact that it's thought to be easier to utilize than Neural Networks. The support vector machines were extensively used as a section of a content-based spam detection system in the exhibition. SVM is an excellent key for the little sample size delinquent as it uses a separating hyperplane to varnish the sorting. This paper suggests that the use of support vector machine to spot spam emails as it yields better results for spam detection

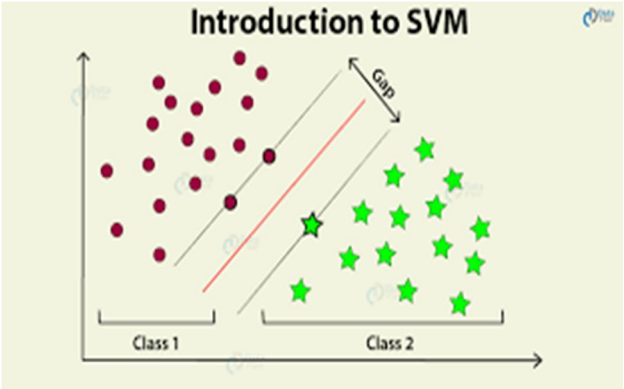


FIG [3] SVM Classification

The margin is defined by the biggest distance between the samples of the 2 classes and is calculated using the space between the margin's nearest instances, called support vectors. Many implementations of those methods utilize Kernel functions rather than linear hyperplanes. Support Vector Machines [3] may be a supervised learning method for automatic pattern recognition. SVM learns from a training data set to a classifier which separates a collection of positive examples from a collection of negative samples of introducing the most margin between the 2 data classes.

III. PROPOSED SYSTEM

In our proposed system, we use the SVM algorithm, for the classification of spam emails from the ham mails. We are going to be training a classifier to classify whether a given email, x, is spam (y=0) or non-spam(y=1). Specifically, we want to convert each email into a feature vector.

A. Proposed Algorithm flow

Algorithm 1 Pseudo-code of SVM algorithm

Inputs:Determine the various training and test data.
Outputs:Determine the calculated accuracy.
Select the optimal value of cost and gamma for SVM.
while (stopping condition is not met) **do**
    Implement SVM train step for each data point.
    Implement SVM classify for testing data points.
end while
Return *accuracy*

FIG [4] Pseudo-code Of SVM Algorithm

The flow of the System

B. Algorithm

S1: Choose the dataset

S2: Using the tokenization and word count algorithms, extract features.

S3: Use the SVM Classifier for training the dataset.

S4: Determine the likelihood of spam and ham messages.

Probablityspam = (sum (trainmat (spamindices,)) + 1) ./(spamwc+numbertokens) Probabilityham=(sum(trainmat(nonspamindices,)) + 1) / (nonspamwc + numbertokens)

S5: Dataset Testing

log_x=testmat*(log(probabititytokensspam))' + log(probspam)

log_y=testmat*(log(probabilitytokensham))'+log(1 - probabilityspam)

if output = log_x > log_y then document are spam else the document are non-spam

S6: Classification of the spam mails from the ham mails.

The flow of proposed System using the SVM algorithm is shown below in FIG [5]

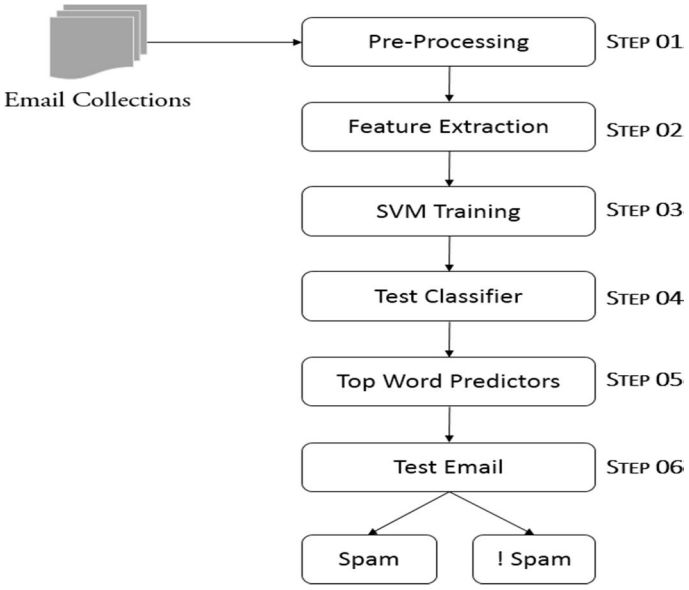


FIG [5] Flow of Proposed System.

IV. METHODOLOGY

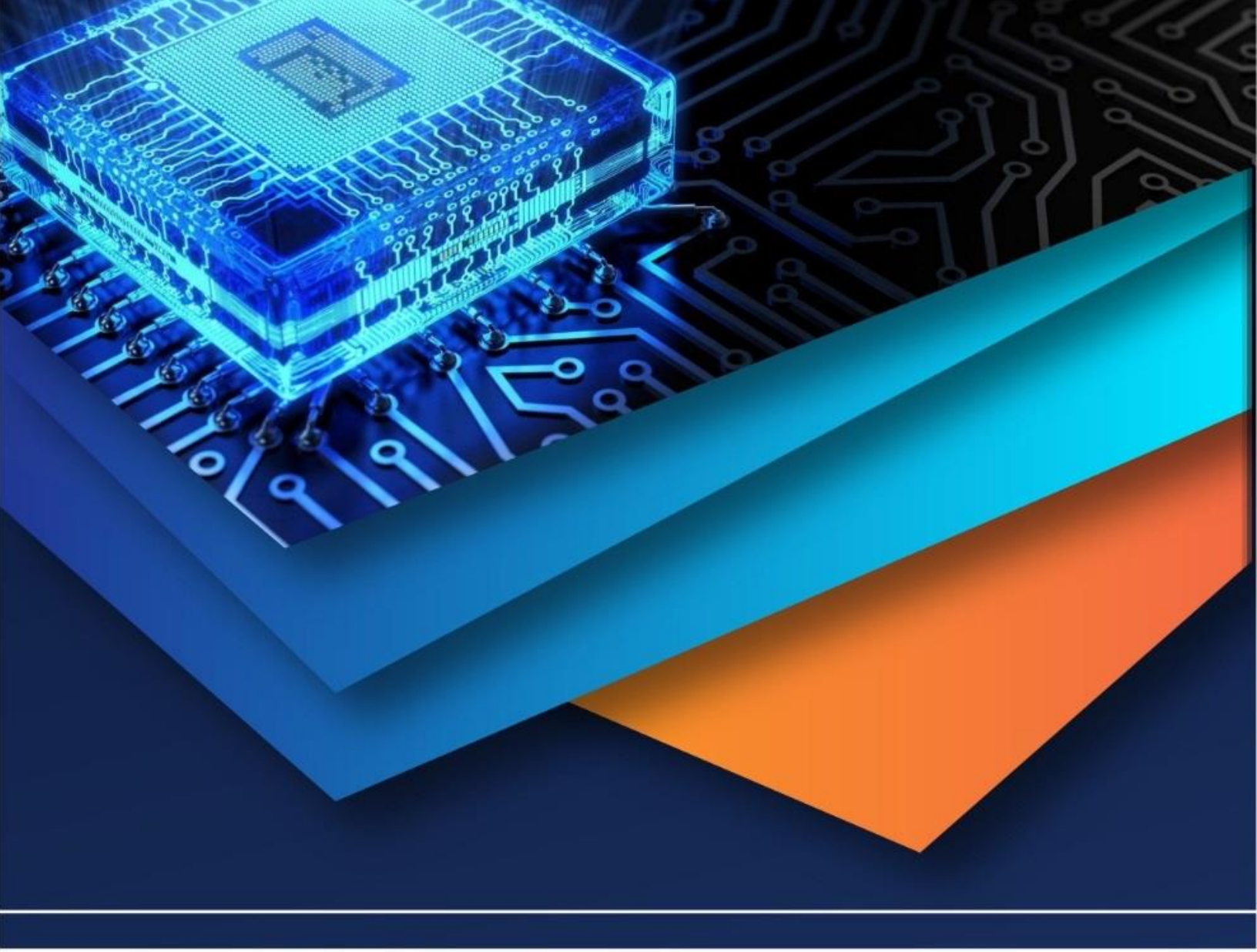
1) *Pre-processing:* All of the numbers and special symbols are deleted during the pre-processing stage, and a few texts are added in their place. Also, Site address and tags are detached. Word stemming could be used as a technique for removing letters from words that are not needed.
2) *Feature Extraction:* The feature is retrieved after pre-processing. The basic idea underlying feature extraction is that the words present in the email set obtained after the first step are mapped to their corresponding index. A feature vector is then created, with all mapped indexes stuffed with 1, therefore the remainder with 0.
3) *SVM Training:* The dataset is then split into train and test sets; the training set is going to be trained under the Support Vector Machine Algorithm (FIG [4]).
4) *Test Classifier:* Now that our Training set is trained under SVM, the subsequent step is to check our built model by using the testing dataset, to test the accuracy of the model built.
5) *Top Word Predictors:* Post training and testing, our model will begin to acknowledge the best words, that fall under the spam and ham classes. Then the foremost frequent words within the category of either spam or ham will be taken as predictors of that individual class.
6) *Test Mail:* Now that our model is trained, tested and top predictors of sophistication are made known to the model, the actual test of our model begins, that's using it on data except for those tested already. And see, how accurately it classifies the spam mails from Ham.

V. CONCLUSION

The support vector machine's findings outperformed the other three approaches by a decent margin. The SVM was discovered to outperform the alternative three, within the case of spam emails, classifiers are used. It's reasonable to conclude that SVMs are capable of more precise spam detection compared to other approaches. So, SVM was chosen as our algorithm for spam detection. The tactic of Spam classification using SVM is covered during this paper.

REFERENCES

[1] "An Efficient Email Spam Detection using Support Vector Machine", IJITEE ISSN: 2278-3075, Volume-9 Issue-2, December 2019 by K Sai Prasanthi, T Deepika, S Anudeep, M Sai Koushik.
[2] "Email Spam Classification by Support Vector Machine" 2018 International Conference on Computing, Power and Communication Technologies (GUCON) Sep 28-29, 2018 by Manmohan Singh, Rajendra Pamula, Sudhanshu Kumar Shekhar
[3] "Efficient Support Vector Machines for Spam Detection: A Survey", IJCSIS Vol. 13, No. 1, January 2015 by Zahra S. Torabi, Mohammad H. Nadimi-Shahraki, Akbar Nabiollahi
[4] "SVM-Based Feature Selection and Classification for Email Filtering" by Sebast´ıan Maldonado1, Gaston L'Huillie
[5] "A Novel Technique of Email Classification for Spam Detection", IJAIS Volume 5, No 10,2013 by Vinod Patidar, Divakar Singh, Anju Singh
[6] "Overview of textual anti-spam filtering techniques". International Journal of Physical Sciences, 5(12):1869–1882, 2010 by Thamarai Subramaniam, Hamid A Jalab, and Alaa Y Taqa.
[7] "A study of spam filtering using support vector machines. AI Review", 34(1):73–108, 2010 by Thamarai Subramaniam, Hamid A Jalab, and Alaa Y Taqa.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY