WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Literature Survey on Spam Email Detection

***Pritesh A. Patil*[1], *Prayag P. Bhosale*[2]**

[1]*Assistant Professor, Department of Information Technology, AISSMS's Institute of Information Technology, Pune-411001, INDIA*
[2]*TE. BE (Information Technology), AISSMS's Institute of Information Technology, Pune-411001, INDIA*

ABSTRACT

Nowadays, emails are used in almost every field, from business to education. Ham and spam are the two subcategories of emails. Email spam, often known as junk email or unwelcome email, is a kind of email that can be used to hurt any user by sapping their time and computing resources and stealing important data. Spam email volume is rising quickly day by day. Today's email and IoT service providers face significant and tremendous challenges with spam identification and filtration. Email filtering is one of the most important and well-known methods among all the methods created for identifying and preventing spam.

This has been accomplished using a number of machine learning and deep learning techniques, including Naive Bayes, decision trees, neural networks, and random forests. By categorizing them into useful groups, this study surveys the machine learning methods used for spam filtering in email and IoT platforms. Based on accuracy, precision, recall, etc., a thorough comparison of different methods is also made. Finally, thorough conclusions and potential future study directions are also covered.

Keywords— Machine learning, Naïve Bayes, support vector machine-nearest neighbor, random forest, bagging, boosting, neural networks.

1. Introduction

1.1 Introduction

In the era of information technology, information sharing has become very easy and fast. Users can exchange information on a variety of platforms with people all around the world. Email is the most straightforward, affordable, and quick method of disseminating information on a global scale. Emails are also susceptible to a variety of attacks due to their simplicity, with spam being the most prevalent and destructive. Nobody wants to receive emails that are not relevant to them because doing so wastes their time and resources. Additionally, these emails may contain malicious material concealed as attachments or URLs that could compromise the host system's security.

Spam is any irrelevant and unwanted message or email sent by the attacker to a significant number of recipients by using emails or any other medium of information sharing. As a result, there is a huge need for email system security. Spam emails could contain Trojans, rats, and viruses. Attackers primarily employ this strategy to entice people to use internet services. They might send spam emails with multi-file attachments and URLs jam-packed with harmful and spammy websites, which could result in identity theft, financial fraud, and data breaches. Many email service providers let their customers create keyword-based rules for email filtering. However, because it is challenging and users do not want to customise their emails, spammers attack users' email accounts as a result of this approach.

In this paper, we consider different machine learning algorithms for spam detection. Our contributions are delineated as follows:

I. The study discusses various machine learning-based spam filters, their architecture, along with their pros and cons. We also discussed the basic features of spam email.

II. Some exciting research gaps were found in the spam detection and filtering domain by conducting a comprehensive survey of the proposed techniques and spam's nature.

III. Open research problems and future research directions are discussed to enhance email security and filtration of spam emails by using machine learning methods.

IV. Several challenges currently faced by spam filtering models and the effects of those challenges on the models' efficiency are discussed in this study.

V. A comprehensive comparison of machine learning techniques and concepts that help understand machine learning's role in spam detection is provided.

1.2 Motivation

i. For the last decade, researchers have been trying to make email communication better than today.

ii. Spam emails are increasing day by day has become a common problem.

iii. It's a Fascinating topic.

iv. Numerous studies have been conducted in this field because to its expensive and significant impact in a variety of situations, including customer behavior and bogus reviews.

v. The possibility that anybody can leave an email or a message provides a golden opportunity for spammers to write spam message about our different interests.

1.3 Aim and Objective of the work

i. The aim of this literature survey is to study identify spam detection using machine learning algorithms as Spam fills inbox with number of ridiculous emails

ii. The study discusses various machine learning-based spam filters, their architecture, along with their pros and cons. We also discussed the basic features of spam email.

iii. To study several challenges currently faced by spam filtering models.

iv. The number of spam emails is rapidly increasing in marketing, chain communications, stock market tips, politics, and education.

Project objectives:

i. We knew that Machine learning is helpful for building Artificial intelligence systems based on tacit knowledge because it can help us to solve complex problems due to real word data.

ii. On the other side we knew that Knowledge engineering is helpful for representing expert's knowledge which people aware of that knowledge project.

iii. This must be highly focused and feasible and should address the more immediate project outcomes

iv. Email has been the most important medium of communication nowadays, through internet connectivity any message can be delivered to all aver the world. More than 270 billion emails are exchanged daily, about 57% of these are just spam emails.

2. Scope of ML in spam email detection

In recent years, internet has become an integral part of life. With increased use of internet, numbers of email users are increasing day by day. This increasing use of email has created problems caused by unsolicited bulk email messages commonly referred to as Spam. Email has now become one of the best ways for advertisements due to which spam emails are generated.

2.1 Specifications

Types of Spam Email

i. Email address harvesting

ii. Obfuscating message content

iii. Defeating Bayesian filters

iv. Spam-support services

v. Chain Letters

vi. Ads

vii. Malware Warning

2.2 Features

For Detection of Spam Emails

i. Email ID of sender

ii. Day of week when email was sent

iii. Subject of Email

iv. Number of Sum of all character length of words

v. Number of URLs in the email

vi. URL Based Image Source

vii. Matching Domains (From & Body)

viii. Keywords (bounty, winning, cash, money, credit, bitcoin, etc)

2.3 Illustration

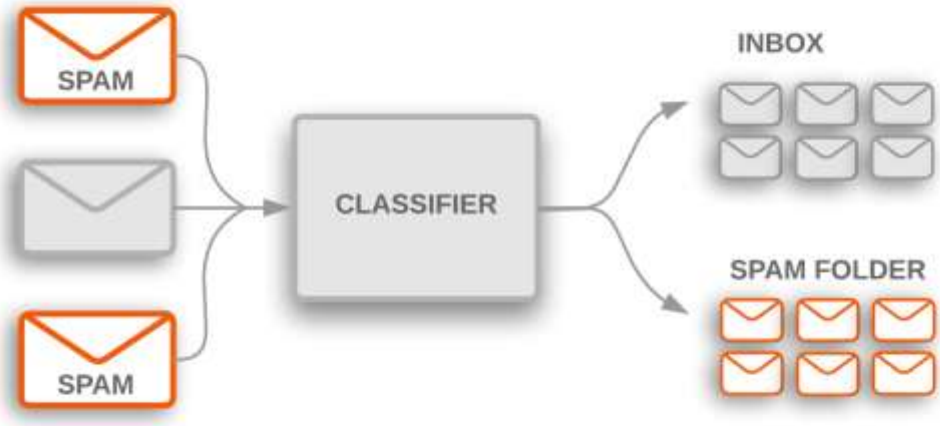


Figure 1. cycle of classification of email into spam non-spam

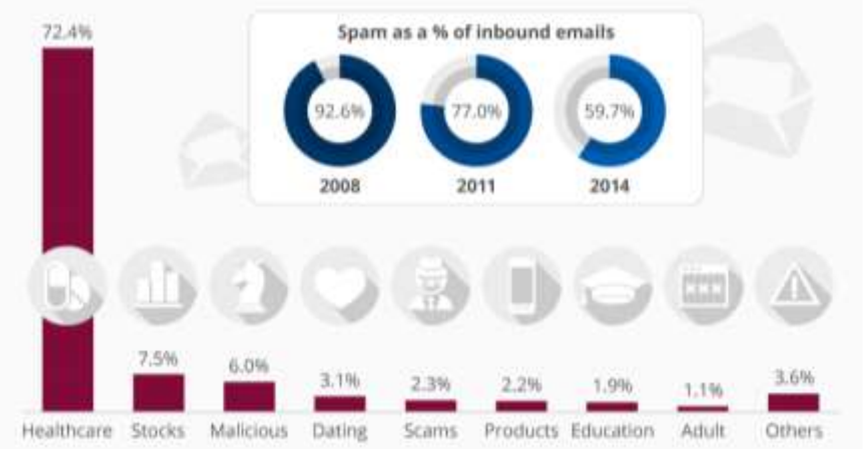


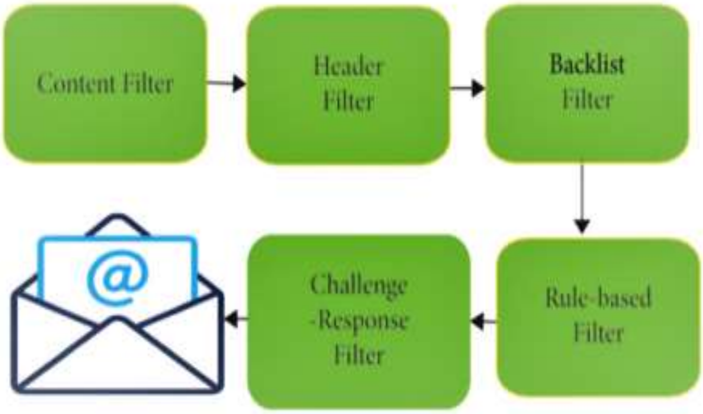
Figure 2. different fields where spam detection occurs

3. SPAM MESSAGES

3.1 Standard filtering

The email spam definition is ambiguous since everybody has their views on it. At present, email spam is getting the attention of everyone. Email spam ordinarily includes particular spontaneous messages sent in mass by individuals you do not know. In the era of technology, the dodger/spammer shows a story where the unfortunate casualty needs forthright financial help so that the fraudster can gain a lot bigger total of cash, which they would then share.

The fraudster will either earn a profit or avoid communication when the unfortunate victim completes the instalments. Currently, various companies develop different techniques and algorithms for efficient spam detection and filtering. We address some filtering strategies in this section to understand the filtering process. Standard spam filtering is a filtering system that implements a set of rules and works with that set of protocols as a classifier. Figure 1 illustrates a standard method for filtering spam. In the first step, content filters are implemented and use artificial intelligence techniques to figure out the spam. The email header filter, which extracts the header information from the email, is implemented in the second step. After that, backlist filters are applied to the emails to clinch the emails coming from the backlist file to avoid spam emails. After this stage, rule-based filters are implemented, recognizing the sender using the subject line and user-defined parameters. Eventually, allowance and task filters are used by implementing a method that allows the account holder to send the mail,



3.2 Use of Machine Learning for Spam Email Detection

Machine learning is one of the most important and valuable applications of artificial intelligence (AI), which gives computer systems the ability of automatically learning and enhancing their functionality without explicit programming. The primary purpose of machine learning algorithms is to build automated tools to access and use the data for training. Machine learning consists of three major kinds, used for numerous tasks. For the last decade, researchers have been trying to make email communication better than today. Spam filtering of emails is one of the most critical ways of protecting email networks. Many research articles have been published using various machine learning approaches to identify and process spam emails, but there are still some research gaps. Junk mail is one of the central, attractive research fields for filling the gaps. That is why, this paper is presented to make a summarized version of different existing machine learning models and approaches that are being used for email spam detection. This paper also evaluates the most common machine learning approaches like KNN, SVM, random forest, and Naive Bayes.

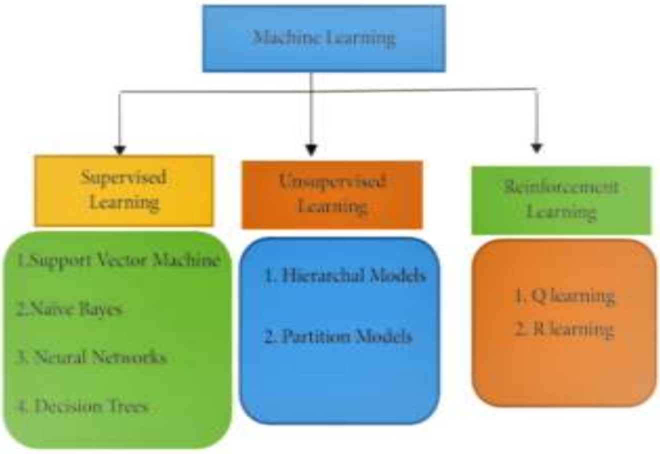


Figure 5.1: MACHIN LEARNING

Figure 3. Types of ML techquies

3.2.1 Machine Learning Models

Decision Tree-Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems but mostly it is preferred for solving Classification problems. It is a true-structured classifier, where internal nodes represents the features of the dataset, branches represent the decision rules and each leaf node represents the outcome.

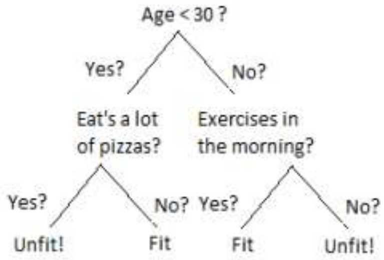


Figure 4. Decision Tree

Naive Bayes- Naive Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

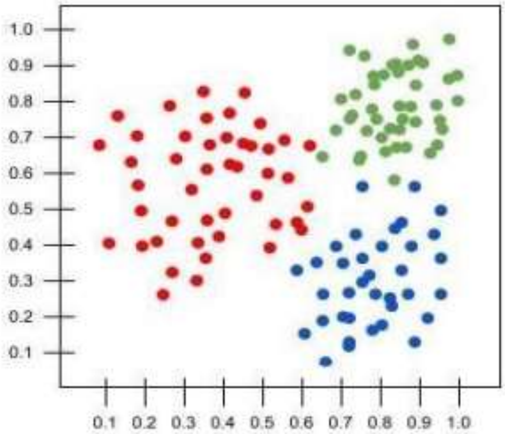


Figure 5. Naïve Bayes

Random Forest- As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

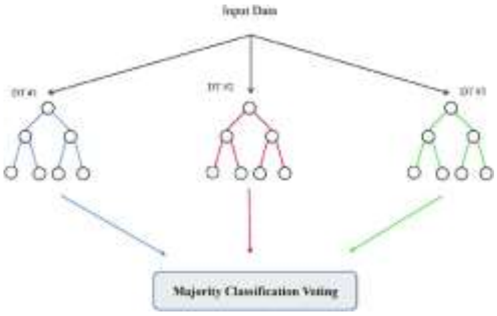


Figure 6. Random Forest

SVM-The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

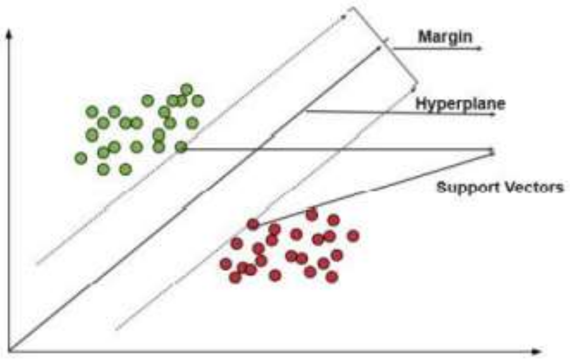


Figure 7. SVM

3.3 Overall Insights of Machine Learning Algorithms for Spam Detection

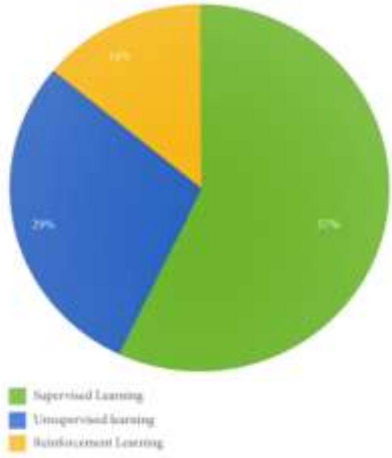


Figure below illustrates the percentage of work on email spam detection discussed in this survey. After discussing the literature, we observed that most of the datasets used to train, test, and implement different models are synthetically created. There is a lack of examples for analysis and the complexity of labeling all the supervised model data. So, the classifiers' results are not 100 used for the models' training. These are not representative of real-world spam reviews as vast numbers of machine learning models are currently used for email spam detection or filtering. The three learning algorithms, logistic regression, Naive Bayes, and support vector machine (SVM), are widely used, and they outperform the other learning algorithms in most of the discussed studies.

This survey paper elaborates the existing machine learning-based spam filtering techniques and models by exploring and observing numerous methods. The conclusions are discussed by the overview of several spam filtering techniques and summarizing the accuracy of different proposed approaches based on various parameters. We conclude that all the spam filtering techniques perform well. Some have outstanding results, while some are trying to use other methods to increase the accuracy level. Though all are effective, the spam filtering system still lacks some, which are the primary concern for researchers. They are trying to generate next generation spam. filtering processes that can work on multimedia data and prominently filter spam emails.

3.4 Enhancement

i. This algorithms used are efficient and effective in processing.

ii. This can be studied further to design an effective solution algorithm which can locate stops and optimize routes considering walking accessibility simultaneously.

iii. The data used in the experiment is generated randomly according to the proposed problem.

iv. As there are many such algorithms, by using them ML algorithms it can be improved.

v. There is high adoption of supervised learning algorithms for email spam detection

vi. There are existing methods are efficient for filtering spam emails. Some have successful results, and others are attempting to incorporate other ways to boost their accuracy performance.

RESEARCH GAPS AND OPEN RESEARCH PROBLEMS

This section discusses the research gaps and open research problems of the spam detection and filtration domain. In the future, experiments and models should be trained on real-life data rather than manually created datasets, because, in the various article, the models trained on artificial datasets perform very poorly on real life data. Moreover, future research should concentrate on the availability of standard labelled datasets for researchers to train classifiers and the addition of more attributes to the dataset to improve the accuracy and reliability of spam detection models, such as the spammer's IP address and the location. The following are some other future research directions and open research problems in the domain of spam detection.

I. Some studies considered header, subject of the email, and message body as a feature for spam classification. While these features are not enough for fully accurate results, manual feature selection and features should also be.

II. Fault tolerance, self-learning, and quick response time can be better by using comprehensive feature engineering and an accurate pre-processing phase.

III. Almost all researchers presented their results based on accuracy, precision, recall, etc., while the time complexity of machine learning models should be considered an evaluation metric.

IV. Deep learning models with dynamic updating of feature space are needed to implement for better spam classification. Most of the current filters cannot update their feature space.

V. The security of spam detection and filtration system is needed for better accuracy and reliable results.

CHALLENGES IN SPAM DETECTION

Some critical challenges faced by spam filters are discussed as follows:

I. The growing amount of data on the Internet with various new features is a big challenge for spam detection systems.

II. Features' evaluation from several dimensions such as temporal, writing styles, semantic, and statistical ones is also challenging for spam filters.

III. Most of the models are trained on balanced datasets, while self-learning models are not possible.

IV. Many spam detection models face adversarial machine learning attacks that will decrease their effectiveness. Adversaries can throw a variety of attacks during the training and testing of ML models. Adversaries can harm training data to cause a classifier to classify the data incorrectly (poisoning attack), create unfavorable samples during testing to evade detection (evasion attack), and obtain sensitive training data via a learning model (privacy attack)

CONCLUSION

In the last two decades, spam detection and filtration gained the attention of a sizeable research community. The reason for a lot of research in this area is its costly and massive effect in many situations like consumer behaviour and fake reviews. The survey covers various machine learning techniques and models that the various researchers have proposed to detect and filter spam in emails and IoT platforms. The study categorized them as supervised, unsupervised, reinforcement learning, etc. The study compares these approaches and provides a summary of learned lessons from each category. This study concludes that most of the proposed email and IoT spam detection methods are based on supervised machine learning techniques. A labelled dataset for the supervised model training is a crucial and time-consuming task. Supervised learning algorithms SVM and Naive Bayes outperform other models in spam detection. The study provides comprehensive insights of these algorithms and some future research directions for email spam detection and filtering.

References

[1] Kumaresan, T., and C. Palanisamy. "E-mail spam classification using S-cuckoo search and support vector machine." International Journal of Bio-Inspired Computation 9, no. 3 (2017): 142-156

[2] Kriti Agarwal, Tarun Kumar "Email Spam Detection using integrated approach of Naïve Bayes and Particle Swarm Optimization", Proceedings of the Second International Conference on Intelligent Computing and Control Systems (ICICCS), 2018.

[3] Olatunji, Sunday Olusanya. "Extreme Learning Machines and Support Vector Machines models for email spam detection." In Electrical and Computer Engineering (CCECE), 2017 IEEE 30th Canadian Conference on, pp. 1-6. IEEE, 2017.

[4] Mujtaba, Ghulam, et al. "Email classification research trends: Review and open issues." IEEE Access 5 (2017).

[5] D. Davino "Spam Detection by Machine Learning-Based Content Analysis". In journal Progress in Artificial Intelligence and Neural Systems (ISBN-978-981-15- 5092-8)

[6] Cihan Varol, Hezha M.Tareq Abdulhadi Comparison of String Matching Algorithms on Spam Email Detection, International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism Dec, 2018.

[7] Duan, Lixin, Dong Xu, and Ivor Wai-Hung Tsang. "Domain adaptation from multiple sources: A domaindependent regularization approach." IEEE Transactions on Neural Networks and Learning Systems 23.3 (2012).

[8] Mohammed Reza Parsei, Mohammed Salehi E-Mail Spam Detection Based on Part of Speech Tagging 2nd International Conference on Knowledge Based Engineering and Innovation (KBEI), 2015.