

DEPARTMENT OF INFORMATION TECHNOLOGY

NUTAN MAHARASHTRA INSTITUTE OF ENGINEERING AND TECHNOLOGY TALEGAON DABHADE, PUNE

**SYNOPSIS**

This is to certify that the seminar synopsis entitled

“Fast email spam filtering methods”

Submitted by

Name :Nagraj Aajure Roll No:-BI101

Name :Aditya Bhute Roll No:-BI102

Name:Shubham Dhanorkar Roll No:-BI107

Is a bonafide work carried out by her under the supervision of **Prof.Dheeraj Patil** and it is submitted towards the fulfillment of the requirement of Savitribai Phule Pune University, Pune for the award of the degree of Bachelor of Engineering (Information Technology).

Prof.Dheeraj Patil Prof. Aditi Das

**Project Guide Project coordinator**

Prof.Nitin Dhawas Dr. Lalitkumar Wadhwa

**HOD Principal**

**Place:** Nutan Maharashtra Institute of Engineering and Technology, Talegaon.

**Date**:23/11/21

**SYNOPSIS**

**Title:** Fast Email spam filtering methods.

**Internal Guide:- Prof.Dheeraj Patil**

**Abstract:** The paper elaborates on how text analysis influences classification—a key part of the spam-filtering process. The authors propose a multistage meta-algorithm for checking classifier performance. As a result, the algorithm allows for the fast selection of the best-performing classifiers as well as for the analysis of higher-dimensionality data. The last aspect is especially important when analyzing large datasets. The approach of cross-validation between different datasets for supervised learning is applied in the meta-algorithm. Three machine-learning methods allowing a user to classify e-mails as desirable (ham) or potentially harmful (spam) messages were compared in the paper to illustrate the operation of the meta-algorithm. The used methods are simple, but as the results showed, they are powerful enough. We use the following classifiers: k-nearest neighbours(k-NNs), support vector machines (SVM), and the naïve Bayes classifier (NB). The conducted research gave us the conclusion that multinomial naïve Bayes classifier can be an excellent weapon in the fight against the constantly increasing amount of spam messages. It was also confirmed that the proposed solution gives very accurate results.

**Keywords:** classifiers; e-mail; ham; machine learning; spam

**Problem Statement:-**

Any “attempt to abuse, or manipulate, a techno-social system by producing and injecting unsolicited and/or undesired content aimed at steering the behavior of humans or the system itself, at the direct or indirect, immediate or long-term advantage of the spammer(s)”. Here, we focus on so-called junk e-mails. These are unwanted messages sent at large scale by e-mail. The term spam refers to the undesired (or even harmful) e-mails, while ham is used to indicate the valid and important messages desired by the recipient. Additionally, we assume the scenario where junk e-mails are sent by botnets and they are not aimed at specific users (contrary to, e.g., spear phishing).

**Objectives:**

* The increasing number of spam e-mails has created a strong need to develop more reliable and efficient anti-spam filters, including ones based on machine-learning tools.
* They are efficient, since they only require the preparation of a set of training samples,i.e., pre-classified e-mails. In recent years, various machine-learning methods have been successfully used to effectively detect and filter unwanted messages.
* The following classification methods are most commonly used for spam filtering: Support Vector Machine(SVM), Naïve Bayes classifier (NB), k-Nearest Neighbours (k-NN), Artificial Neutral
* Network (ANN), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR).
* The values are given at the end of the numerical study in separate table.
* The applicability of using different machine-learning methods to recognize spam e-mails was analyzed in. The SpamAssassin dataset, which contains 6000 e-mails with the spam rate 37.04% used in all experiments.
* Sharma and Arora in analyzed Bayes Net(BN), Logic Boost (LB), RT, JRip (JR), J48-based DTs, Multilayer Perceptron (MP), Kstar (KS),RF, and Random Committee (RC) machine-learning algorithms. The dataset with 4601instances and 55 spam base attributes downloaded from UCI Machine-Learning Repositorywere used in the performed research.
* Harisinghaney et al. applied the following three different algorithms: k-NN, NB, and DBSCAN-based clustering. The performance for the four metrics accuracy, precision, sensitivity, and specificity were calculated and compared.

**Introduction**:-

The spam problem is an ongoing issue: in 2018 14.5 billion spam e-mails were sent per day. According to the Internet Security Threat Report released in 2019 by Symantec,spam levels for their customers increased in 2018. What draws the attention is that small enterprises were attacked more often than large companies, and e-mail malware reached stable levels. Therefore, there is a need to tailor even simple tools for detection and filtering of spam in all organizations.

Identification of the best-performing machine learning-based classifiers and selection of the one with the leading parameters. The proposed solution solves the problem of fast recognition of the most interesting parameters. This allows for quick analysis of data of higher dimensionality. This is especially important if large datasets are to be analyzed and we want to assure the proper scalability of our system.

**LITERATURE SURVEY:**

**Title:**A Method for Fast Selection of Machine-Learning Classifiers for Spam Filtering.

**Author name:** Sylwia Rapacz , Piotr Chołda and Marek Natkaniec

**Description:**The paper[1] addresses development of system that deals with spam filtering methods on surveys and results of various tests conducted by them on readymade databases.

**Title:**Detecting Spam Email With Machine Learning OptimizedWith Bio-Inspired Metaheuristic Algorithms.

**Author name**: SIMRAN GIBSON 1, BIJU ISSAC 1, (Senior Member, IEEE),LI ZHANG 1, (Senior Member, IEEE), AND SEIBU MARY JACOB 2, (Member, IEEE)

**Description:**The paper[2] successfully implemented models combined with bio-inspired algorithms. The spam email corpus used within the project were both numerical as well as alphabetical.Approximately 50,000 emails were tested with the proposed models. The numerical corpuses (PU), had restrictions in terms of feature extraction as the words were replaced by numbers. But the alphabetical corpuses performed better in terms of extraction of the features and predicting the outcome.

**Title:**SMS Spam Filtering Using Supervised Machine Learning Algorithms.

**Author name:** Pavas Navaney, Ajay Rana, & Gaurav Dubey.

**Description:**This paper[3] is all about SMS spam filtering using Machine Learning Techniques. Various methods of spam filtering algorithms are discussed.

**Application:**

* Improved Spam filtering methods in E-mail.
* Next level security against phishes,virus bound email.
* Future system with common spam filtering for SMS and Emails.
* Safe systems.
* Virtual Machines
* Virus Detection
* Improved UI

**FUTURE SCOPE:**

New improvised technology that does spam detection in Messages and Emails which also automatically level up the field according to threats.