



Somaiya Vidyavihar University
K. J. Somaiya College of Engineering
Department of Computer Engineering

Batch: C3 Roll No.: 16010122818

Machine Learning IA 2

Spam Detection System Using Naïve Bayes Algorithm

Literature Survey:

1) A Comprehensive Review On Email Spam Classification Using Machine Learning Algorithms

Summary:

The paper primarily delves into spam classification using machine learning algorithms. It offers a comprehensive analysis and review of research conducted on different machine learning techniques and email features utilized in various machine learning approaches for spam detection. Additionally, it discusses future research directions and the challenges in the spam classification field, providing valuable insights for future researchers in this domain.

Paper Contributions:

M. RAZA, N. D. Jayasinghe and M. M. A. Muslam, "A Comprehensive Review on Email Spam Classification using Machine Learning Algorithms," 2021 International Conference on Information Networking (ICOIN), Jeju Island, Korea (South), 2021, pp. 327-332, doi: 10.1109/ICOIN50884.2021.9334020.

Paper Link: <https://ieeexplore.ieee.org/document/9334020>

2) Integrated Spam Detection For Multilingual Emails

Summary:

The paper proposes an integrated approach to enhance the efficiency of detecting and filtering such malicious emails. By combining various techniques, the goal is to improve the ability to identify and prevent fraudulent activities, ultimately enhancing cybersecurity in email communication.



Somaiya Vidyavihar University
K. J. Somaiya College of Engineering
Department of Computer Engineering

Paper Contributions:

A. Iyengar, G. Kalpana, S. Kalyankumar and S. GunaNandhini, "Integrated SPAM detection for multilingual emails," *2017 International Conference on Information Communication and Embedded Systems (ICICES)*, Chennai, India, 2017, pp. 1-4, doi: 10.1109/ICICES.2017.8070784.

Paper Link: <https://ieeexplore.ieee.org/document/8070784>

3) A Hybrid Approach For Spam Filtering Using Support Vector Machine and Artificial Immune System

Summary:

This research paper addresses the problem of spam emails, which waste resources and disrupt communication on the internet. It introduces a novel approach that combines the strengths of Support Vector Machine (SVM) and Artificial Immune System (AIS) algorithms to create a more effective spam filtering system. By leveraging the advantages of both techniques in a hybrid approach, the paper aims to overcome the limitations of traditional spam filtering methods. It discusses the shortcomings of existing techniques and highlights how the proposed approach offers improvements in spam detection and filtering.

Paper Contributions:

K. Jain and S. Agrawal, "A hybrid approach for spam filtering using support vector machine and artificial immune system," *2014 First International Conference on Networks & Soft Computing (ICNSC2014)*, Guntur, India, 2014, pp. 5-9, doi: 10.1109/CNSC.2014.6906699.

Paper Link: <https://ieeexplore.ieee.org/document/6906699>

Dataset Used: <https://github.com/karansanghvi/ML-Mini-Project/blob/main/spam.csv>



Somaiya Vidyavihar University
K. J. Somaiya College of Engineering
Department of Computer Engineering

Program:

```
# SPAM DETECTION SYSTEM USING NAIVE BAYES ALGORITHM IN MACHINE LEARNING

import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB

spam_df = pd.read_csv("./spam.csv")

print(spam_df.groupby('Category').describe())

spam_df['spam'] = spam_df['Category'].apply(lambda x: 1 if x == 'spam' else 0)
print(spam_df['spam'].head())

x_train, x_test, y_train, y_test = train_test_split(spam_df.Message,
                                                    spam_df.spam, test_size=0.25)
print(len(x_train), len(x_test))

cv = CountVectorizer()
x_train_count = cv.fit_transform(x_train.values)
print(x_train_count.shape)

model = MultinomialNB()
model.fit(x_train_count, y_train)

email_ham = ["cricket tickets later"]
email_ham_count = cv.transform(email_ham)
print(model.predict(email_ham_count))

email_spam = ["reward money click"]
email_spam_count = cv.transform(email_spam)
print(model.predict(email_spam_count))

x_test_count = cv.transform(x_test)
print(model.score(x_test_count, y_test))
```



Somaiya Vidyavihar University
K. J. Somaiya College of Engineering
Department of Computer Engineering

Output:

```
Message
count unique top freq
Category
ham 4825 4516 Sorry, I'll call later 30
spam 747 641 Please call our customer service representativ... 4
0 0
1 0
2 1
3 0
4 0
Name: spam, dtype: int64
4179 1393
(4179, 7536)
[0]
[1]
0.9863603732950467
```