Mining Massive Data Sets Midterm Report

1st 522H0036 - Luong Canh Phong
Faculty of Information Technology
Ton Duc Thang University
Ho Chi Minh City, Vietnam
522H0036@student.tdtu.edu.vn

2nd 520H0341 - Nguyen Thai Bao
Faculty of Information Technology
Ton Duc Thang University
Ho Chi Minh City, Vietnam
520H0341@student.tdtu.edu.vn

3rd 522H0030 - Le Tan Huy
Faculty of Information Technology
Ton Duc Thang University
Ho Chi Minh City, Vietnam
522H0030@student.tdtu.edu.vn

4th 522H0008 - Dao Minh Phuc
Faculty of Information Technology
Ton Duc Thang University
Ho Chi Minh City, Vietnam
522H0008@student.tdtu.edu.vn

5th 522H0136 - Nguyen Nhat Phuong Anh
Faculty of Information Technology
Ton Duc Thang University
Ho Chi Minh City, Vietnam
5220136@student.tdtu.edu.vn

***Abstract*—Spam emails are a common problem seen on the internet as it is an annoyance in daily life and a cyber security risks to any sensitive and important data of a person or an organization/business. With the number of spam emails increasing more and more significantly over the past few years, many more algorithms are created and improved in spam detection efficiency. Overall, this paper goes through the basic understanding of spam emails, understanding the necessity of a spam classification algorithm, and learn more about the methodologies, its effectiveness and usefulness when detecting spam emails.**

I. Introduction

Since the birth of the Internet, spam email has been a common occurrence. Along with the rapid growth and widespread of the Internet, the frequency has been increasing significantly, especially over the past decade. In addition to being nuisances, a waste of time and email storage, spam emails can be sent with malicious intent of stealing information, hijacking devices by storing malware within the content of the email itself. And with the nature of email spam being sent by botnets, it isn't easy to avoid the situation due to a new bot can be easily created in case another one got blocked or banned on the site. A common way how most platforms (such as Gmail, Yahoo!, Outlook) handle these spams is to develop a Machine Learning (ML) model to detect and get rid of the spam emails, lowering the number of spams getting into the inbox.

II. Importance of Spam Classification

To understand why spam classification is important to our lives, we must first understand the spam emails and its impact on daily life and businesses.

A. Different Types of Spam Emails

There are various forms of spam, sent with different intentions and purposes. But they're commonly grouped into:

- Phishing Emails: (TBA)
- Email Spoofing: (TBA)
- Tech support scams: (TBA)
- Current event scams: (TBA)
- Marketing/advertising email: (TBA)
- Malware scam: (TBA)

B. Problems with Spam Emails

According to statistics report in 2023, 160 billion spam emails are sent every day, which is 46% of the 347 billion emails sent on a daily basis. Out of which, the most common type being marketing/advertising emails, which take up around 36%, followed up with promotional of adult content, around 31.7% of total spam emails. Despite scam and fraudulent emails is the least common type, over 70% of them are phishing emails which is still over 6 billion phishing emails are being sent to users daily.

A single spam email carbon emission is almost 0.03g of CO2e, with the amount of spam being sent daily, it can easily get nearly 5 tonnes of CO2e being released every day. Additionally, two-thirds of spam receivers have been reported to have their mental health affected due to the number of spam or phishing scams.

For businesses, this spam can be sent as a way to get businesses to invest in nonexistent organizations under the disguise of an investment and promised the payback would be worth the money spent, for individuals it would be under the form of bitcoin investment or for a chartiable cause. Once the money is received, the sender would delete all traces and block the recipient contact.

III. Methodologies for Spam Classification

Many methods to prevent spam are applied; a commonly used method is using Machine Learning models like Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), Naive Bayes (NB) or Deep Learning models such as Artificial Neural Network (ANN), (Explantion of used algorithms)

IV. Recent Advances in Spam Classification

(TBA)