Progress Report 1

Capstone Project

on

SMS Spam Collection

|  |  |  |
| --- | --- | --- |
| Name | : | Ling Hsiu Ann Karen |
| Date | : | 9 June 2024 |
| Lecturer | : | Avinnaash Suresh |

Table of Contents

|  |  |  |
| --- | --- | --- |
|  | Overview |  |
|  | Objectives |  |
|  | Challenges |  |
|  | Obtaining the data |  |
|  | Cleaning the data |  |
|  |  |  |
|  |  |  |
|  |  |  |

**Overview**

The aim of the project is to point out the proliferation of mobile communication has led to an increase in unsolicited and potentially harmful messages, commonly known as spam.

SMS spam can be a nuisance to users and can also pose significant security risks, including phishing attacks and malware distribution. Effective spam detection systems are essential for improving user experience and protecting personal information. Effective spam detection systems are essential for improving user experience and protecting personal information.

The link below is the open link to public on the Capstone’s Project Website.

https://github.com/karenling99/Capstone-Project-SMS-Spam-Collection-2024

**Objectives**

1. Classify it as spam or not spam
2. Build a real-time spam detection system that can analyze incoming SMS messages
3. Investigate which features are most important for accurately detecting spam messages (e.g., word frequency, message length, presence of certain keywords)

**Challenges**

In the digital age, the risks of privacy invasion, security breaches, and resulting financial loss have become critical issues. This capstone project aims to address these challenges by developing robust security measures in Python applications by creating a model that can identify/ detect these SMS spam contents.

1. **Privacy Invasion**

Unauthorized access to personal data can lead to significant privacy breaches. Users' sensitive information, such as personal identification and browsing habits, is at risk of being exploited. Mitigation strategies include implementing strong encryption, secure authentication methods, and ensuring compliance with data protection regulations like GDPR.

1. **Security Risks**

Applications face various security risks, including SQL injection, cross-site scripting (XSS), and man-in-the-middle attacks. These vulnerabilities can lead to unauthorized access, data manipulation, and service disruptions. Regular security audits, secure coding practices, and robust security protocols are essential to protect against these threats.

1. **Financial Loss**

The financial implications of privacy invasions and security breaches are substantial. Direct financial impacts include theft of funds and fraudulent transactions, while indirect impacts cover loss of business, reputational damage, and legal fees. Investing in preventative measures, such as security training and advanced technologies, is crucial to mitigate these risks.

**Datasets**

The dataset used for this project is the SMS Spam Collection dataset, which contains a set of SMS labeled messages that have been collected for spam research.

The datasets obtain are from Kaggle datasets. Description of data attributes, such as the text of the message and its corresponding label (spam or ham).

**Cleaning the Data**

Steps to clean and preprocess the data, including tokenization, removal of stop words, and text normalization. Techniques for handling imbalanced data, such as resampling or using class weights.

Exploratory Data Analysis (EDA): -

Statistical analysis of the dataset to understand the distribution of spam and ham messages. Visualization of common words and phrases in spam vs. ham messages using word clouds or bar plots.

Model Selection: -

Justification for choosing specific machine learning models for the task, such as Logistic Regression.