

MES COLLEGE OF ENGINEERING, KUTTIPPURAM
DEPARTMENT OF COMPUTER APPLICATIONS
20MCA245 – MINI PROJECT

PRO FORMA FOR THE APPROVAL OF THE THIRD SEMESTER MINI PROJECT

(Note: All entries of the pro forma for approval should be filled up with appropriate and complete information. Incomplete Pro forma of approval in any respect will be rejected.)

Mini Project Proposal No : _____1_____
(Filled by the Department)

Academic Year : 2021-2022
Year of Admission : 2020

1. Title of the Project : **SPAM SMS FILTERING USING MACHINE LEARNING**
2. Name of the Guide : **Mr. Vasudevan T V**
3. Number of the Student: **1**
4. Student Details (in BLOCK LETTERS)

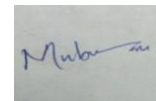
Name

Roll Number

Signature

MUBASHIRA A

29



Date:

Approval Status : Approved / Not Approved

Signature of
Committee Members }

Comments of The Mini Project Guide

Dated Signature

Initial Submission :

First Review :

Second Review :

Comments of The Project Coordinator

Dated Signature

Initial Submission:

First Review

Second Review

Final Comments :

Dated Signature of HOD

SPAM SMS FILTERING USING MACHINE LEARNING

MUBASHIRA A

Introduction:

Due to its convenient, economical, fast, and easy to use nature messaging systems is a vital revolution taking place over traditional communication systems. A main obstruction in electronic communications is the vast publicizing of unwanted, harmful messages known as spam messages. Lots of time of client is being wasted for sorting approaching these messages and erasing undesirable correspondence, so there is a need for spam detection so that its outcomes can be reduced. The main aim is to development of suitable filters that can appropriately detect those messages and results in a high-performance rate. In this project, Spam Detection aims to differentiate between spam and authorized messages. Here, the evaluation of it is done by using a Machine Learning algorithm named SVM. Machine learning algorithms, especially Support Vector Machine (SVM), can play a major role in spam detection. In this project, the classification is done by defining feature vectors calculated by TF-IDF values.

Objectives:

Using cell phones and smart systems has grown more and more in recent years, and short message service has become one of the most significant communication means. Eightieth of the world's active users use mobile phone sms as a communication method. Unwanted SMS messages made for the following purposes are among some of this large variety of short messages. Sending low quality SMS and plenty of incredibly low value cell SMS kit operators. Quick message has developed into a business-trade of many billion bucks. The creation of inappropriate messages for the purpose of advertisement is harassment and the source of these messages on SMS has become the main challenge during this service. Especially SMS spam is more irritating then email spam. We are inclined to use RNN to distinguish ham and spam sequences of variable length. SVM is the most suitable method for this purpose.

Problem Definition:

To communicate with each other the Internet has become an indispensable method, because of its popularization, low cost, and fast delivery of messages. In recent years, there has been a dramatic growth in spam along with the growth of the messaging systems. Spam can arise from any location across the globe where internet access is available. Spamming is the misuse of electronic messaging systems to send voluntary bulk messages or to promote products or services, that are almost universally undesired. The Spam problem is currently of serious and surge concern, and it is challenging to develop spam filters that can effectively eliminate the increasing volumes of unwanted SMS automatically before they enter a user's inbox. If people have to spend time and effort on identification SMS every day it is evident that their work efficiency and their emotions will be influenced.

Basic functionalities:

Automatic distinguishing of spam has important meaning and applying value. Automated SMS filtering using Machine Learning (ML) is one popular solution. One of the most used techniques as the base classifier to overcome the spam problem is the Support Vector Machine. SVM Classify spam that is to differentiate between spam and authorized messages. Hence application of this SVM classifier with SVC model is studied in this paper. First Dataset containing 1000 messages samples containing both spam and ham(non-spam) type messages is used. Where 70% messages are used for training and remaining 30% are used for testing. Later vocabulary is built, which contains a set of most frequently words chosen from the training/testing set. Vocabulary is then used to calculate the TF-IDF values, where each messages will be represented on the basis of the importance of word in the entire dataset which is a N dimensional vector. This vector is feature vector. A Machine learning algorithm, Support Vector Machine(SVM) is trained to classify the given messages. Each of these messages belongs to only one of two classes. The idea of SVM classification is find a linear separation boundary that correctly classifies training samples. And later this model is used to predict new messages given, which is the main aim or the system.

SOFTWARE REQUIREMENT SPECIFICATION

HARDWARE SPECIFICATION

- Processor : i3
- Hard Disk : 500 GB
- RAM : 4 GB

SOFTWARE SPECIFICATION

- Language : Python
- Front End : Python-Django
- Back end : SQLite
- Dataset : Spam harm dataset from Kaggle website
- Algorithm : SVM
- IDE : Visual Studio Code
- OS : Windows/Linux