**PROJECT REPORT**

*on*

**Spam Classifier in Python/R**

(IV Semester Mini Project)

Submitted in partial fulfillment of the requirement for the IV semester

**Bachelor of Computer Science and Engineering**

**A picture containing text, scene, room, gambling house

Description automatically generated**

***Guided By: Submitted By:***

Ms. Garima Sharma Ruudra Amola

(Resource Person) Univ R.No. : 2016972

CSE-IV-SEM

Section : G (12)

Session : 2022-23

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

**Text

Description automatically generated with medium confidence**

**CANDIDATE’S DECLARATION**

I hereby certify that the work which is being presented in the dissertation entitled **“Spam Classifier using Python/R”** in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineering, submitted in the Department of Computer Science and Engineering of the Graphic Era Deemed to be University, Dehradun is an authentic record of my own work carried out under the supervision of Ms. Garima Sharma, Assistant Professor, Department of Computer Science and Engineering of the Graphic Era Deemed to be University, Dehradun (Uttarakhand).

**Name- Ruudra Amola**

**Univ. Roll no.- 2016972**

**CSE-Core (IV Semester)**

**PROBLEM STATEMENT:**

To create a simple email spam classifier which is trained in a way that it predicts if the entered email as a spam or not.

**SOFTWARES AND TOOLS/MODULES USED:**

* **Language used**- Python (version-3.10.2), HTML
* **Flask Framework for python**
* **Modules used**- pickle module, numpy module, Flask module
* **IDE used** - Visual Studio Code by Microsoft

**MOTIVATION:**

The main motivation and idea behind taking up this project was to have an experience with a new programming language which is not part of our main course curriculum and to gain knowledge about Machine Learning.

My other motivation was to know and implement a system which has a vast real time and practical application in the industry, so I chose the domain of classifier systems in ML as my topic.

A spam filter or classifier is a program used to detect unsolicited, unwanted and [virus](https://www.techtarget.com/searchsecurity/definition/virus)-infected emails and prevent those messages from getting to a user's inbox. Like other types of filtering programs, a spam filter looks for specific criteria on which to base its judgments.

For example, one of the simplest and earliest versions of spam filtering, like the one that was used by Microsoft's Hotmail, was set to watch out for particular words in the subject lines of messages. An email was excluded from the user's inbox whenever the filter recognized one of the specified words. More sophisticated programs, such as [Bayesian filters](https://www.techtarget.com/whatis/definition/Bayesian-filter) and other [heuristic](https://www.techtarget.com/whatis/definition/heuristic) filters, identify spam messages by recognizing suspicious word patterns or word frequency. They do this by learning the user's preferences based on the emails marked as spam.

Overall, it was a great experience for me to learn something out of the books and implement it myself.

**METHODOLOGY AND ABOUT THE PROJECT:**

The project is developed in **Python language**. Python is an interpreted high-level general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant indentation. Its language constructs as well as its object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects. Python consistently ranks as one of the most popular programming languages.

**HTML** (**H**yper **T**ext **M**arkup **L**anguage) is the code that is used to structure a web page and its content. For example, content could be structured within a set of paragraphs, a list of bulleted points, or using images and data tables.

It is a markup language that defines the structure of your content. HTML consists of a series of elements, which you use to enclose, or wrap, different parts of the content to make it appear or act a certain way. The enclosing tags can make a word or image hyperlink to somewhere else, can italicize words, can make the font bigger or smaller, and so on.

**FLASK** stands for Flux Advanced Security Kernel. [Flask](http://flask.pocoo.org/) is a small and lightweight Python web framework that provides useful tools and features that make creating web applications in Python easier. It gives developers flexibility and is a more accessible framework for new developers since you can build a web application quickly using only a single Python file. Flask is also extensible and doesn’t force a particular directory structure or require complicated boilerplate code before getting started. Flask uses the Jinja template engine to dynamically build HTML pages using familiar Python concepts such as variables, loops, lists, and so on. Flask is a backend technology.

Advantages of using Flask framework are: There is a built-in development server and a fast debugger provided within the framework.

**Following python modules used in the code:**

*from* flask *import* Flask,render\_template,request

Importing flask module in the project is mandatory. An object of Flask class is our **WSGI** application.

Flask constructor takes the name of **current module (\_\_name\_\_)** as argument.

[*render\_template*](https://github.com/pallets/flask/blob/master/src/flask/templating.py) is a [Flask](https://www.fullstackpython.com/flask.html) function from the flask.templating package. It is used to generate output from a [template file based on the Jinja2 engine](https://www.fullstackpython.com/template-engines.html) that is found in the application's templates folder.

*import* pickle

Pickle in Python is primarily used in serializing and deserializing a Python object structure. Python pickle module is used for serializing and de-serializing python object structures. The process to converts any kind of python objects (list, dictionaries, etc.) into byte streams (0s and 1s) is called pickling or serialization.

We can convert the byte stream (generated through pickling) back into python objects by a process called as unpickling.

**Pickle is a useful Python tool that allows you to save your models, to minimise lengthy re-training and allow you to share, commit, and re-load pre-trained machine learning models.**

*import* numpy

**NumPy** (Numerical Python) is an open-source library for the Python programming language. It is used for scientific computing and working with arrays.

Apart from its multidimensional array object, it also provides high-level functioning tools for working with arrays.

**In order to train the model on the basis of the dataset, I downloaded and used a random dataset from the internet.**

**The Naïve Bayes Algorithm:**

All the major classifier project are built on some specific machine learning theorems. Particularly for my project that is the Spam classifier, a very extensive machine learning algorithm called the Naïve Bayes Algorithm is used.

**The dataset I’ve used from the internet is already built over the Multinomial Naïve Bayes Model of machine learning.** The naïve bayes algorithm as the name suggests very much, is based upon the Bayes theorem of probability of two events. Now the Naïve Bayes theorem trains the machine in such a way that on the basis of the previous mails received it creates a dictionary of all the words in the mails and creates a multidimensional array to store the status of the words.

Then later on the to check or predict that if a new mail is a spam or not, it predicts on the basis of the probability of occurrence of the words in the new mail which is to be checked from the multi-d array it formed earlier.

The Naïve Bayes algorithm is of three types-

Gaussian Naive Bayes: It is used in classification and it assumes that features follow a normal distribution.

Multinomial Naïve Bayes: It is used for discrete counts. For example, let’s say, we have a text classification problem. Here we can consider Bernoulli trials which is one step further and instead of “word occurring in the document”, we have “count how often word occurs in the document”.

Bernoulli Naïve Bayes: The binomial model is useful if your feature vectors are binary (i.e. zeros and ones).

**The Multinomial Naïve Bayes is particularly used for Spam classification models.**

**In the code:**

I have 2 parts of my code where in one, I have made Html web page to implement my problem using the flask deployment tool.

The other part of the code is that where I have imported the files containing my datasets as encrypted and created the main flask framework to call the HTML code to implement the problem.

**CONCLUSION:**

This project helped me to gain knowledge about the python language and its basic features. After and during my work on the project I was able to understand the concept of classifier systems in ML. The simplicity and vastness of the python language and its developer friendly syntax was very interesting to study. Using the Flask framework for the first time was a bit complicated for me so I just limited myself just to the project requirements. HTML made it very easy and simple for me to create a web interface in order to implement my problem.

The knowledge of Python and its libraries which the project provided, would now enable me to efficiently write codes and help me in future.

Some wide applications of a spam classifier system in an organization are -

* Fraud Detection
* Text classifications
* Medical Diagnosis
* Marketing
* Spam filtering
* News categorization

Throughout the course of my project, I was very much active on some major programming sites like w3schools.com and geeksforgeeks.com , javatpoint.com for assistance and learning as this was a first-time experience in machine learning AI for me.

**THANKYOU!!**