Internship Report On

"Spam Ham Classifier"

A Dissertation submitted in partial fulfillment of the requirement
for the award of degree of

MASTER OF COMPUTER APPLICATIONS
of
Visvesvaraya Technological University, Belagavi

By

JAYANTH L
1RN19MCA20

Carried out at
NASTECH

Under the Guidance of

Internal guide:
Dr. Rajani Narayan
Associate Professor
Dept. of MCA
RNS Institute of Technology
Bengaluru – 560 098.

External Guide:
Mr. Azib Hasan
Subject Matter Expert
NASTECH
Mumbai,
Maharashtra – 400 608.

ESTD:2001
An Institute with a Difference

Department of Master of Computer Applications

RNS Institute of Technology

Dr. Vishnuvardhan Road, Channasandra, Bengaluru-560 098
APRIL 2022

“Spam Ham Classifier”

A Dissertation submitted in partial fulfillment of the requirement
for the award of degree of

MASTER OF COMPUTER APPLICATIONS
of
Visvesvaraya Technological University, Belagavi

By

JAYANTH L
1RN19MCA20

Carried out at
NASTECH
Under the Guidance of

Internal guide:
Dr. Rajani Narayan
Associate Professor
Dept. of MCA
RNS Institute of Technology
Bengaluru – 560 098.

External Guide:
Mr. Azib Hasan
Subject Matter Expert
NASTECH
Mumbai,
Maharashtra – 400 608.

ESTD:2001
An Institute with a Difference

Department of Master of Computer Applications

RNS Institute of Technology

Dr. Vishnuvardhan Road, Channasandra, Bengaluru-560 098
APRIL 2022

ESTD:2001
An Institute with a Difference

Department of Master of Computer Applications

RNS Institute of Technology

Dr. Vishnuvardhan Road, Channasandra, Bengaluru-560 098

CERTIFICATE

*This is to certify that **Mr. Jayanth L**, student of 6th semester MCA, bearing the USN: **1RN19MCA20** has completed his final semester internship/project work entitled **"Spam Ham Classifier"** as a partial fulfillment for the award of Master of Computer Applications degree, during the academic year 2022 under our joint supervision.*

Internal Guide

Dr. Rajani Narayan
Associate Professor
Department of MCA
RNS Institute of Technology
Bengaluru – 560 098.

External Guide

Mr. Azib Hasan
Subject Matter Expert
Nastech
Mumbai,
Maharashtra – 400 608.

Head of the Department

Dr. N P Kavya
Professor & HoD
Department of MCA
RNS Institute of Technology
Bengaluru – 560 098.

Principal

Dr. M K Venkatesha
Principal
RNS Institute of Technology
Bengaluru – 560 098.

DECLARATION

I, **Mr. Jayanth L**, student of 6th MCA, RNS Institute of Technology, bearing USN: **1RN19MCA20** hereby declare that the project entitled **“Spam Ham Classifier”** has been carried out by me under the supervision of External Guide **Mr. Azib Hasan**, Subject Matter Expert, and Internal Guide **Dr. Rajani Narayan**, Associate Professor, submitted in partial fulfillment of the requirements for the award of the degree of Master of Computer Applications by the Visvesvaraya Technological University during the academic year 2022. This report has not been submitted to any other Organization / University for any award of degree or Certificate.

Signature
Jayanth L

ACKNOWLEDGEMENT

The successful completion of any task would be incomplete without the mention of the people who made it possible, whose constant support and encouragement has crowned my efforts.

I take this opportunity to acknowledge the guidance I have received from different individuals and place on record my appreciation and thanks.

I express my sincere gratitude to **Dr. R N Shetty**, Founder and **Sri**. **Satish R Shetty**, Managing Director, RNSIT for providing us wonderful academic environment.

My deep sense of gratitude to our Principal **Dr. M K Venkatesha**, for his kind support.

I am grateful to Dr**. N P Kavya**, Head of the Department of MCA, RNSIT for nurturing our technical skills and contributing towards the success of this project.

I would also express my heartfelt thanks to my internal guide **Dr. Rajani Narayan**, Associate Professor, Department of MCA, RNSIT for her continuous guidance and valuable suggestions for this internship work.

It’s my pleasure to thank **Nastech** for providing me the best platform to complete the internship work and glad to thank the external guide **Mr. Azib Hasan**, Subject Matter Expert.

I also express my heartfelt thanks to all the teaching and non-teaching staff members of MCA department for their encouragement and support throughout this work.

JAYANTH L
1RN19MCA20

ABSTRACT

Recently unsolicited commercial or bulk e-mail also known as spam, become a big trouble over the internet. Spam is waste of time, storage space and communication bandwidth.

Machine learning field is a subfield from the broad field of artificial intelligence, this aims to make machines able to learn like human.

In e-mail pre-processing the content of email is received through our software, the information is extracted then as mentioned above, then the information (Feature) extracted is saved into a corresponding database.

Every message was converted to a feature vector with 21700 attributes (this is approximately the number of different words in all the messages of the corpus).

An attribute n was set to 1 if the corresponding word was present in a message and to 0 otherwise. This feature extraction scheme was used for all the algorithms.

TABLE OF CONTENTS

List of Figures

Chapter – 1

INTRODUCTION

1.1 Aim

The goal of this project is to construct an email spam filter using machine learning techniques.

1.2 Project Description

The objective is to implement a Naïve Bayesian anti-spam filter to segregate spam from ham and measure its efficiency using various cost effective measures.

A supervised learning approach is used to enable the filter to differentiate between spam and ham. The filter is trained on 70% off spam & ham corpus that requires Feature Extraction and calculation of spam probability of the extracted feature, fi, using a naïve Bayes.

1.3 Scope

In this module the extracted spam text and the ham text, then produce feature dictionary and feature vectors as input of the selected algorithm, the function of feature extraction is to train and test the classifier [9]. For the train part, this module account frequency of words in the email text, we take words which the time of appearance is more than three times as the feature word of this class. And denote every email in training as a feature vector.

Spam classification through the steps above, we take standard classification email documents as training document, pre-treatment of email, extract useful information, save into text documents according to fix format, split the whole document to words, extract the feature vector of spam document and translate into the form of vector of fix format. We look for the optimal classification using the selected algorithm which is constructed using the feature vector of spam documents.

Chapter - 2

COMPANY PROFILE

2.1 Organization structure

- NASTECH is formed with the purpose of bridging the gap between Academia and Industry.

- NASTECH is one of the leading Global Certification and Training service providers for technical and management programs for educational institutions. We collaborate with educational institutes to understand their requirements and form a strategy in consultation with all stakeholders to fulfill those by skilling, reskilling and upskilling the students and faculties on new age skills and technologies.

- We offer industry and project oriented training programs which not only expose students to hands-on training experience but also make them practical oriented towards the industry-readiness expected in today's time.

- We take pride that all our programs are mapped to a certain Global Certification Exams i.e. after the students are done with their training, they will prove themselves on a global level via a global certification exam.

- We lead from the front in terms of costing of our overall global certification and training programs.

- We know that learning is easier when you have an excellent trainers. That's why most of our educators have achieved an advanced degree in their field. Our instructors are passionate about the subjects they teach and bring this enthusiasm into their training workshop.

2.2 Different departments and functions

- Our Data Science programs starts from basic and takes students to the level where they develop relevant programming abilities. Demonstrate statistical analysis of data and assess data based models. Mapped to Global Certification Exam from Microsoft.
- Machine Learning using Python
- Machine learning is touted as one of most in- demand concept in todays' world. After undergoing our course students will good understand of machine learning concepts with hands-on experience on different datasets thereby knowing challenges in machine learning, data, model selection, model complexity, etc. Mapped to Global Certification Exam from Microsoft. Ethical Hacking

- To be a cyber-security personnel, "you have to think like a hacker", keeping this phrase in mind we have developed this program in sync with industry requirements. Students will understand the loopholes present in the current system. how to resolve those and safeguard their personal and professional data from being getting hacked.

- Mapped to Global Certification Exam from Oracle
- Other Technical & Management Courses
- Apart from the courses mentioned above we execute different technical programs as per the interest of students and requirement of colleges.

Course Names:

- IOT with AWS Cloud (Online)
- Cloud Computing using Azure
- Cloud Security
- Python Programming with advance concepts
- Business Analytics
- Power BI (Business Intelligence)
- Advance Excel
- Digital Marketing and many more.

2.3 Job process / Services / Facilities

- Industry oriented trainings mapped to global certification exams.
- NASTECH has taken pledge to skill maximum students pan India on the new age skills to make them industry ready. We are scaling this through our Global Certification Programs which are mapped to different departments of universities/colleges.

Chapter – 3

TOOLS AND TECHNOLOGY

3.1 Tools/technology used by company

- Exploratory Data Analytics
    - ✓ Exploratory data analytics refers to the critical process of performing initial investigations on data.
- Titanic Dataset
    - ✓ The titanic and titanic2 data frames describe the survival status of individual passengers on the Titanic.
- Matplotlib
    - ✓ Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy.
- Numpy
    - ✓ NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
- Pandas
    - ✓ Pandas is a software library written for the Python programming language for data manipulation and analysis.
- Multiple Linear Regression
    - ✓ Multiple linear regression is also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable.
- Logistic Regression
    - ✓ Logistic regression is a process of modeling the probability of a discrete outcome given an input variable.

3.2 Tools learned in company

- Matplotlib
  - ✓ Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy.
- Numpy
  - ✓ NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
- Pandas
  - ✓ Pandas is a software library written for the Python programming language for data manipulation and analysis.
- Multiple Linear Regression
  - ✓ Multiple linear regression is also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable.

Chapter - 4

INTERNSHIP WORK

4.1 Task assigned

We learnt about Python basics, Exploratory Data Analytics, Matplotlib, Numpy, Pandas, Computer Vision, Multi Linear Regression, Sentiment Analysis.

Based on the topics which have been covered in Internship by using those technologies develop a simple project and demonstrate it and also prepare a project report on it.

4.2 Applications developed using modern tools

- Spamdrain- email spam filter
- Spamhound SMS spam filter

4.3 Professional learning (Discipline, attitude, planning, groupwork, self-assessment, etc)

- Professional dialogue with colleagues, other professionals, parents, and learners.
- Focused professional reading and research.
- Leading or engaging in practitioner enquiry/action research.
- Experiential, action or enquiry-based learning.
- **Discipline** involves time management, self-control and dedication.
- An **attitude** is a negative or positive evaluation of an object which influence human's behaviour towards that object.
- **Planning** is the process of thinking regarding the activities required to achieve a desired goal.
- **Group work** refers to a collaborative learning environment where students work through problems and assessments together
- **Self-assessment** is the process of looking at oneself in order to assess aspects that are important to one's identity.

Chapter - 5

IMPLEMENTATION

5.1 Screen shots

The figure 5.1 shows main screen of spam ham

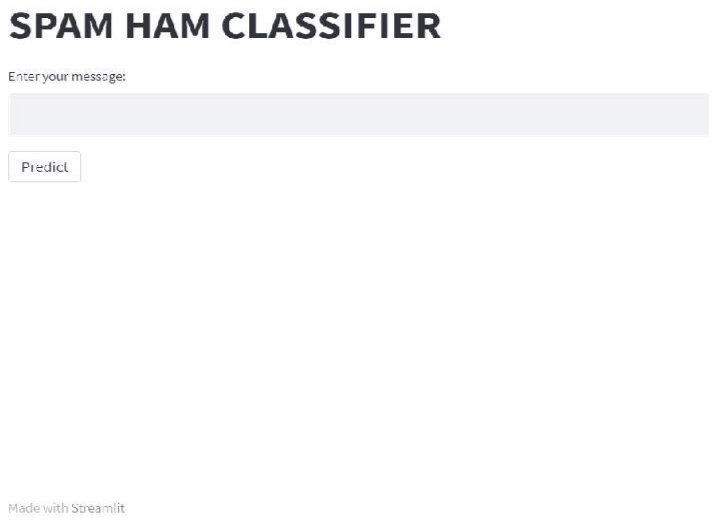
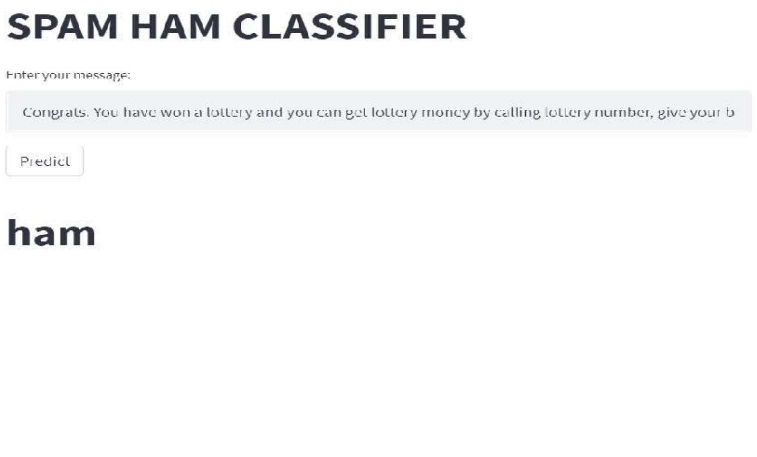


Fig 5.1 Main Screen

The figure 5.2 shows ham classifier

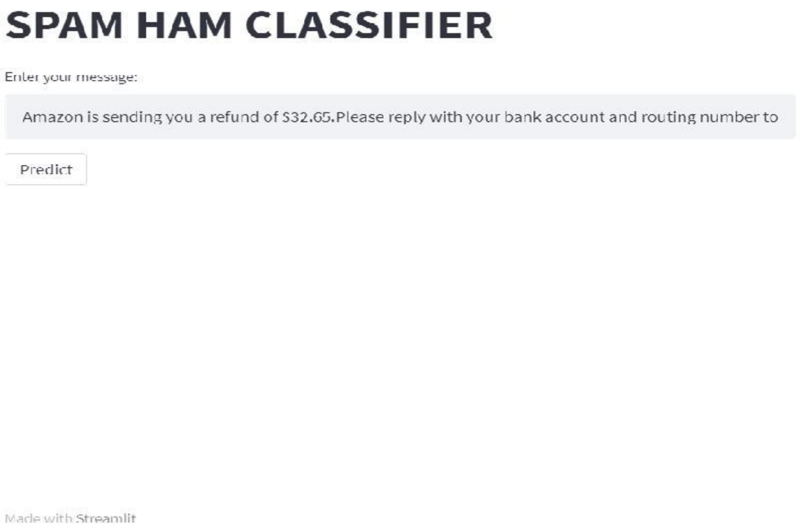


Congrats. You have won a lottery and you can get lottery money by calling lottery number, give your b

ham

Fig 5.2 Ham Classifier

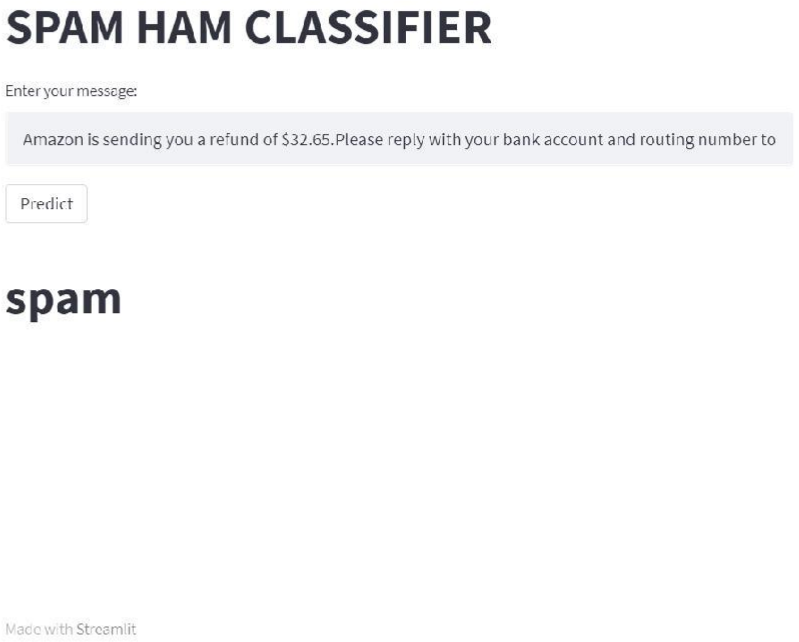
The figure 5.3 shows spam ham classifier



Amazon is sending you a refund of $32.65.Please reply with your bank account and routing number to

Fig 5.3 Spam Ham Classifier

The figure 5.4 shows spam ham classifier



Amazon is sending you a refund of $32.65.Please reply with your bank account and routing number to

spam

Fig 5.4 Spam Classifier

Chapter - 6

CONCLUSION AND FUTURE WORK

CONCLUSION

In this study, we reviewed machine learning approaches and their application to the field of spam filtering. A review of the state of the art algorithms been applied for classification of messages as either spam or ham is provided. The attempts made by different researchers to solving the problem of spam through the use of machine learning classifiers was discussed. The evolution of spam messages over the years to evade filters was examined. The basic architecture of email spam filter and the processes involved in filtering spam emails were looked into. The paper surveyed some of the publicly available datasets and performance metrics that can be used to measure the effectiveness of any spam filter.

By using spam classifier easily we can easily detect the spam messages which are present in the spam folder and also it is very much helpful so that misleading of data or personal details would not be done easily.

REFERENCES

[1] https://colab.research.google.com/drive/1X5r80Ib1kjQOhMUcfvn13hi3wBFIIa-_

[2] https://www.sciencedirect.com/science/article/pii/S2405844018353404

[3] https://myblindbird.com/spam-email-detection-using-machine-learning-projects-beginners-python/

[4] https://realpython.com/beautiful-soup-web-scraper-python/

[5] https://facebook.github.io/prophet/

[6] https://www.analyticssteps.com/blogs/6-major-branches-artificial-intelligence-ai

[7] https://www.ibm.com/in-en/topics/computer-vision

[8] https://numpy.org/doc/stable/user/basics.html

[9] https://pandas.pydata.org/docs/user_guide/index.html#user-guide