VISVESVARAYA TECHNOLOGICAL UNIVERSITY
BELAGAVI - 590 018, KARNATAKA

An Internship Report on

"EMAIL SPAM CLASSIFIER USING AI- ML"

Submitted in partial fulfilment of the requirements for the degree of s

BACHELOR OF ENGINEERING
IN
COMPUTER SCIENCE AND ENGINEERING

For the Academic year

2022 - 2023

Submitted By

SHETTY DISHA ASHOK
4SH19CS062

Internship carried out at
Zephyr Technologies &Solutions Pvt. Ltd.

Internal Guide
Ms. Raksha Puthran
Assistant Professor
Shree Devi Institute of Technology

External Guide
Mr. Vedanth Shenoy
Chief Technology Officer
Zephyr Technologies

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SHREE DEVI INSTITUTE OF TECHNOLOGY
MANGALURU- 574 142

SHREE DEVI INSTITUTE OF TECHNOLOGY

KENJAR, MANGALURU- 574142

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

Certified that the Internship work entitled **"EMAIL SPAM CLASSIFIER USING AI-ML"** was carried out at **Zephyr Technologies & Solutions Pvt. Ltd.** By **SHETTY DISHA ASHOK (4SH19CS062),** bonafide student of VIII semester B.E, Computer Science and Engineering, Department of Shree Devi Institute of Technology, Mangaluru-574142, in partial fulfilment for the award of Bachelor of Engineering in Computer Science and Engineering of the Visvesvaraya Technological University, Belagavi during the academic year 2022-2023. It is certified that all corrections/suggestions indicated by the guide have been incorporated in the report. The internship report has been approved as it satisfies the academic requirements in respect of the internship prescribed for the said degree.

Signature of the Guide **Prof. Nishmitha M R** **Dept. of CSE, SDIT**	Signature of the HOD **Prof. Anand S. Uppar** **Dept. of CSE, SDIT**

EXTERNAL VIVA

Name of the Examiners	**Signature with date**
1. ----------	**----------**
2.----------	**----------**

SHREE DEVI INSTITUTE OF TECHNOLOGY

KENJAR, MANGALURU – 574142

Department of Computer Science and Engineering

DECLARATION

I **SHETTY DISHA ASHOK** bearing USN **4SH19CS062** student of Eighth semester Bachelor of Engineering, Computer Science and Engineering, Shree Devi Institute of Technology, Mangalore declare that the project work entitled **"EMAIL SPAM CLASSIFIER USING AI-ML"** has been duly executed by us under the guidance of Nishmitha M R**,** Asst. Professor, Department of Computer Science and Engineering and Shree Devi Institute of Technology, Mangalore and submitted in partial fulfillment of the requirements for the award of **Bachelor of Engineering** in **Computer Science Engineering** during the year 2022-23.

Date: **SHETTY DISHA ASHOK[4SH19CS062]**

Place: Mangalore

EXECUTIVE SUMMARY

This report is regarding the internship carried out at Zephyr Technologies & Solutions Pvt. Ltd., Mangalore. In this comprehensive report, the major aspects of the company and work done are highlighted, as observed, and perceived during the internship program.

The details of company since incorporation till date, along with their services, products and research and development has been discussed in this report. The major work done during the internship was on learning to implement Artificial Intelligence and Machine Learning from basics to advanced concepts. All the results have been thoroughly analysed under the guidance provided by the company and internal guide.

The report consists of seven chapters. The first chapter includes the company profile, the second chapter starts with the introduction to all domains and the third covers all the professional and technical take away's from the company, the fourth chapter consists of the tasks performed and the projects carried out, the fifth chapter presents results and outcomes. Finally, the sixth chapter reflects on the technical and non-technical outcomes and the last chapter gives a closing note to this report.

ACKNOWLEDGEMENT

I, **Shetty Disha Ashok** express my deep gratitude to **Dr. K E Prakash**, Director and Principal of Shree Devi Institute of Technology, Kenjar, Mangaluru for providing all the facilities for the timely completion of the internship.

I am grateful to **Prof. Anand S. Uppar**, Head of the Department, Computer science and Engineering, for his support and encouragement.

I would like to offer my earnest gratitude to our External Guide, **Mr. Vedanth Shenoy**, Chief Technology Officer, Zephyr Technologies and Solutions Pvt. Ltd . and my internal guide, **Prof. Nishmitha M R,** Assistant Professor, Department of Computer Science and Engineering, SDIT, Mangaluru. This work would not have been possible without their guidance and support.

I wish to express my sincere gratitude to all the Faculty and Technical staff of the Dept. of Computer Science and Engineering, SDIT, Mangaluru for their valuable help and support.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 BACKGROUND

Zephyr Technologies Mangalore is a Software Pvt Ltd Company that was founded in 2005 by three friends, Karun lal, Samah and Musthafa. With abilities in branding, website development, graphic design, and digital media content, we are on a mission to alter the advertising and social media market. Brand identity, website design, packaging design, and marketing communications design are the services offered by Zephyr Technologies. They create a vision and set of principles. The Company turn thoughts and concepts into quantifiable realities.

1.2 ABOUT ZEPHYR TECHNOLOGIES

Figure 1.2.1 Zephyr logo

Zephyr Technologies is a software firm that provides on-time delivery of high-quality, cost effective, and reliable web and e-commerce solutions to a global clientele. Professionalism, competence, and knowledge are the instruments utilised to make the web work for company, resulting in the highest potential return on investment in the shortest amount of time. For its very demanding and online clients located around the globe, Zephyr has delivered its best on IT projects of different difficulties. Developed unique online solutions that boost business efficiency and competitive advantage while also providing satisfaction to end customers. Professionalism, abilities, and expertise are the tools that convert into high-quality work at every stage of any project. The organisation gives customers an advantage by providing intellectual property protection for source codes created expressly for them. The company provides an edge with protection of intellectual for the source codes developed specifically for business. The company does not sell the source codes to the third parties and all elements that they create for the web solutions that belongs to the clients. Zephyr Technologies' project managers and business analysts

place great value for building a clean communication link with their clients as they consider it the key ingredient for the success of any project in hand.

The company's objectives are as follows:

- Planning, comprehensive, composite artifact that gathers all information required to manage the project.
- Analysis, requirements analysis, also called requirements engineering, is the process of determining user expectations for a new or modified product.
- Design, transform user requirements into some suitable form, which helps the programmer in software coding and implementation.
- Development, process of conceiving, specifying, designing, programming, documenting, testing, and bug fixing involved in creating and maintaining applications.

Zephyr offers courses in:

- Web development
- Android development
- IoS development
- Artificial Intelligence and Machine Learning
- Python
- Java
- Digital Marketing

Certifications:

- MCA(Ministry of Corporate Affairs) approved company

1.3 CONTACT DETAILS

Head Office: Gs2, Heavenly Plaza, Suite No.352, Kakkanad, Kochi, Kerala - 682 021

Registered Office: Door No 18/208 D3 III Floor, Golden Chambers, Kandamkulam P.O, Calicut, Kerala - 673002

Regional Office: Oberle Tower, Above Cafe Coffee Day, 2nd Floor, Balmatta, Mangalore 575002

Regional Office: VP Towers, Opp. League Office, Kasaragod

Email: mail@zephyrtechnologies.co

Contact number:

+91 8111843307

+91 7994082021

+91 8129664492

Chapter 2

INTRODUCTION

Technology is advancing at a high rate. A few decades back, the only source of communication was the letters, which turned into telegrams, and in recent times it is in various forms like emails, phone calls, SMS, etc. An average person sends 72 messages per day, as texting is also the most common cell phone activity. Almost 300 billion emails are exchanged per day, and half of them are spam emails. 'Spam Mail' is basically undesired and unwanted emails that are sent to many of recipients that is just filling up all the inboxes. Most of these messages are product buying links, which would consume our personal data or could be some links and attachments. Sometimes carelessness from some users can cause significant damage to their personal data. Spam mails not only fill your inbox with junk mails but also cause email traffic. Spam messages accounted for 45.1% of email traffic in March 2021. In short, such mails can be frustrating and dangerous at the same time.

Inboxes are 85% filled with Spam mails and due to which the valuable and important emails are ignored. Many researchers are developing various techniques to find the solution for such problems and secure to communication. Since the unsolicited emails are termed 'Spam', important and valuable ones are termed 'Ham'. There are many techniques developed to classify such spam and ham mails. One such technique is by using Natural language Processing and Machine Learning. With the help of Text classification methods like stemming, lemmatization, vectorization, etc., it is possible to classify the mails and train the model, which will be able to detect unwanted mails.

In this study, we have come up with our model that would classify emails and messages into either spam or ham. The evaluation metrics for performance such as accuracy were considered evaluating the proposed study. The results obtained from experiments confirmed that the proposed research achieved high accuracy.

Chapter 3

SYSTEM REQUIREMENT SPECIFICATION

3.1 Software Requirements

- **Python3:**

  Python is a high-level, interpreted, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation. Python supports multiple programming paradigms, including structured object-oriented and functional programming.

- **Pandas:**

  Pandas is a python package that provides fast, flexible, and expressive data structure designed to make working with “relational” or “labelled” data both easy and intuitive.

- **Scikit-learn:**

  Sklearn is a free software machine learning library for the python programming language. It features various classification, regression and clustering algorithms algorithms including support-vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries Numpy and SciPy.

- **Matplotlib:**

  Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general purpose GUI toolkits like Tkinter, wx-python, Qt, or GTK.

3.2 Hardware Requirements

These are the system specifications on which the software will be built and tested, it may vary and may support systems with lower specifications

- Ram: 8GB
- CPU: Intel Core, AMD
- Disk: 1TB

Chapter 4

LITERATURE SURVEY

In this paper (Pandey & Yadav, 2020), the author proposed a model where deep neural networks are exploited for detecting spam mails using Tensor Flow. This model uses a linguistic approach, demonstrating the advantage of automatically neural networks. This paper also surveyed various publicly available datasets and noted the basic structure of the model. They have also revealed plentiful of open research problems related to spam filters.

Chapter 5

PROBLEM STATEMENT

The email spam classifier using logistic regression is to develop a machine learning model that can accurately predict whether an incoming email is spam or not. The model should be trained on a dataset of labelled emails, where each email is classified as either spam or not spam.

Logistic regression is a supervised learning algorithm that can be used for binary classification problems such as this one. The model will be trained on features extracted from emails, such as the presence of certain keywords or phrases, the sender's email address, and the email's subject line. The goal is to create a model that can accurately classify new, unseen emails as either spam or not spam based on these features.

The ultimate objective of the email spam classifier is to improve the user's email experience by automatically filtering out unwanted spam messages, thereby reducing the risk of phishing attacks, scams, and other fraudulent activities.

Chapter 6

OBJECTIVES

The objective of this project is to build a predictive model that can accurately distinguish between spam and non-spam emails. The logistic regression algorithm is a supervised learning algorithm that can be trained on a labelled dataset of emails, where each email is labelled as either spam or ham.

The logistic regression model will learn the relationship between the features of the email and the probability that the email is spam. The model will output a probability score for each email, which can be used to clarify it as spam or ham based on predefined threshold.

Chapter 7

METHODOLOGY

7.1 DATA COLLECTION

- In the initial step, we are visiting download the dataset from Kaggle (https://www.kaggle.com/datasets/shantanudhakadd/email-spam-detection-datasetclassification) which contains 5572 records of two columns i.e., Message and Category. we'd wish to import the required python libraries to perform operations such as NumPy, Pandas, and sklearn then the downloaded dataset has got to be uploaded by the method “read_ csv” in pandas’ library.

7.2 LABEL ENCODING

- Later, Label encoding must be done and it's defined as the process of converting labels into numerical values to machine-readable form. Spam is labeled as “0” and Ham Mail is labelled as “1”.

7.3 FEATURE EXTRACTION

- Now the info in the spam dataset is categorized into Training data and Testing data in the ratio of 80:20 and then feature extraction is done using tf-idf vectorizer which id Term frequency-inverse document frequency. this is often a very common algorithm to transform the text into a meaningful representation of numbers which is used to fit machine algorithms for prediction. TF-IDF Vectorizer may be a measure of the originality of a word by comparing the number of times a word appears in the document with the number of documents the word appears in.

7.4MODEL TRAINING

- In this model, we are employing a Logistic Regression Classifier for predicting spam mail. Logistic regression is one among the most popular machine learning algorithms, which comes under supervised machine learning algorithms. it's used for predicting the specific dependent variable using a given set of independent variables. the outcome must be a discrete or categorical value. It is frequently either Yes or No, 0 or 1, true or false, etc.,

7.5 MODEL EVALUATING

- Model evaluation is the process of using different evaluation metrics to understand a machine learning model’s performance, also as its strengths and weaknesses. Model evaluation is vital to assess the efficacy of a model during initial research phases, and it also plays a task in model monitoring.

Chapter 8

LOGISTIC REGRESSION

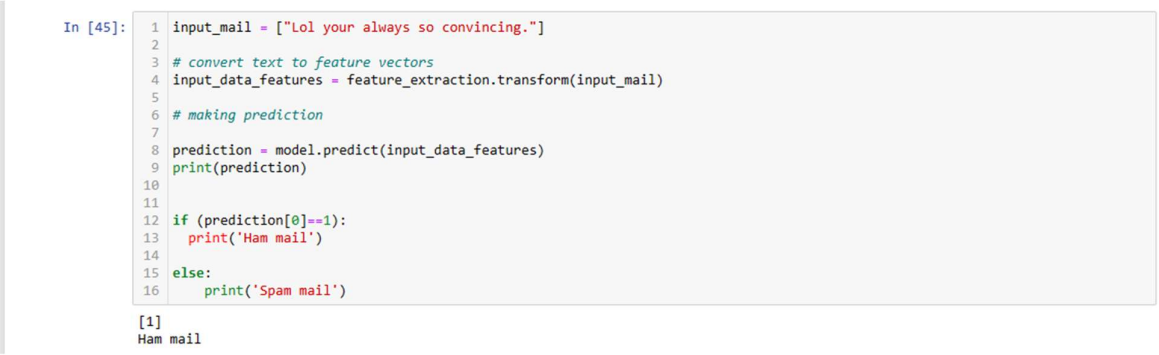
Logistic Regression is used to estimate discrete values (usually binary values like 0/1) from a set of independent variables. It helps predict the probability of an event by fitting data to a logit function. It is also called logit regression. These methods listed below are often used to help improve logistic regression models.

- Include interaction terms
- eliminate features
- regularize techniques
- use a non-linear model

Chapter 9

RESULT

This project uses Logistic Regression algorithm to classify the spam and ham mails. To train, validate and test the model, a dataset consisting of labelled emails. These emails were taken from various resources like Kaggle, google. By using dataset and logistic regression algorithm, a model is built. By using this model, classification of emails is performed.



```
In [45]:
input_mail = ["Lol your always so convincing."]

# convert text to feature vectors
input_data_features = feature_extraction.transform(input_mail)

# making prediction

prediction = model.predict(input_data_features)
print(prediction)


if (prediction[0]==1):
  print('Ham mail')

else:
    print('Spam mail')
```

```
[1]
Ham mail
```

Fig 9: Result

Chapter 10

CONCLUSION

With the increase usage of emails, this study focuses on using automated ways to detect spam emails. The study uses various machine learning and deep learning algorithms to detect them. In the study, a translated emails datasets including spam and ham emails is generated from Kaggle, which is pre-processed for various approaches. Accuracy is used as comparative measures to examine performance. The creation of actual dataset of emails can be considered as a viable.

10.1 FUTURE WORK

In the future, we can use neural network and deep learning models to predict a given message is spam or not. Deep learning works very well for natural language processing; however it requires a vast amount of data to given accurate results and to outperform other traditional machine learning algorithms. Since Natural Language Processing is a relatively underdeveloped area for research, further enhancements can be made to the proposed system for spam detection and email filtering in the field of online security.

REFERENCES

[1] Sjarif, Nila, & Amir, N. (2019). SMS Spam Message Detection using Term Frequency-Inverse Document Frequency and Random Forest Algorithm. Procedia Computer Science , 509-515..

[2] Shankar, S. (2018). Advanced Detection of Spam And Email Fitering using NLP algorithms. IJARIT.

[3] Kaggle Email Spam Detection