

ACKNOWLEDGEMENT

Working on the project was very inserting and really enhanced my knowledge.

I would like to express my sincere gratitude to my beloved **Principal Bhagawan Prasad** without whose permission, I would not be able to do my project and for taking keen interest for the students of Diploma in providing useful guidelines and giving all the necessary facilities.

I extend thanks to my Guide **Mr. Gagandeep M N, Lecturer and Project Guide** under Computer Department, Government Polytechnic Bantwal, for his valuable guidance and constant encouragement, which helped me in successfully completing my project.

Finally, I extend thanks to my parents and friends who were directly or indirectly involved in the completion of my project.

Place: BANTWAL

Date: May 03, 2024

.....

PAWAN

RegNo : 163CS21034

EXECUTIVE SUMMARY

An overview of my internship experience is given in this executive summary, which also highlights the important abilities, information, and successes I acquired while working for the company.

Internship gave me the chance to work in a stimulating and demanding atmosphere while contributing to a variety of projects and engaging with experts in the field. Through practical experience, I was able to learn practical skills that complemented my academic knowledge and a firm awareness of industry practises.

Technical expertise was one of the main things I needed to improve during my internship. I got to work with a variety of software tools and technologies, like Flask as front-end, Jupyter notebook as back-end. By honing my technical abilities and using them in practical situations, this experience helped me to improve my problem-solving abilities.

Throughout the internship, I actively engaged in various projects and successfully contributed to their completion. This included '**AMAZON REVIEW CLASSIFICATION**' and '**SPAM MAIL CLASSIFICATION**', which allowed me to apply my knowledge, demonstrate initiative, and deliver results. These achievements not only contributed to the organization's goals but also enhanced my confidence and professional development.

This project report describes about two projects, on which I have worked during internship. The first project is regarding '**AMAZON REVIEW CLASSIFICATION**', which helped to learn how to build a simple sentiment analysis application. The second project is about '**SPAM MAIL CLASSIFICATION**', in that I have developed a website to see whether the mail is spam or not.

INTERNSHIP CERTIFICATE ISSUED BY THE ORGANIZATION



02-05-2024

CERTIFICATE OF COMPLETION

This is to certify that **MR. PAWAN**, a bonafide student of **GOVERNMENT POLYTECHNIC, BANTWAL**, (Registration Number 163CS21034) has commendably completed an internship in "**ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING**" at **CodeLab Systems** from 01st January 2024 to 30th April 2024.

We acknowledge and commend **MR. PAWAN**, for his significant contribution and commitment throughout the internship period at **CodeLab Systems**. We wish his continued success in all his future endeavors.

Mr. Mohammed Mehroos,
Lead - T & D,
CodeLab Systems,
Mangalore.

CODELAB SYSTEMS
Ground Floor, Light House Condominium,
Light House Hill Road,
Mangaluru, Karnataka - 575001

codelabsystemsindia@gmail.com
Light House Condominium, Light House Hill Road, Bavutagudda
Mangaluru Karnataka 575001. Ph: +91 7349350390

| Sl. No | Contents | Page No |
|-------------------|--|--------------------|
| 1. | COMPANY PROFILE 1.1 OVERVIEW OF THE COMPANY 1.2 VISION AND MISSION OF THE ORGANIZATION 1.3 ORGANIZATION STRUCTURE 1.4 ROLES AND RESPONSIBILITIES OF PERSONNEL IN THE ORGANIZATION 1.5 PRODUCTS AND MARKET PERFORMANCE | 8-13 |

| | | |
|----|---------------------------------------|-------|
| 2. | ASSESSMENT OF ON JOB TRAINING – 1 & 2 | 14-36 |
| | 2.1 INTRODUCTION | |
| | 2.2 AIM | |
| | 2.3 OBJECTIVE | |
| | 2.4 PURPOSE | |
| | 2.5 SCOPE | |
| | 2.6 ADVANTAGES | |
| | 2.6.1 ADVANTAGES | |
| | 2.7 ACTIVITY DIAGRAM | |
| | 2.8 MODULE | |
| | 2.9 REQUIREMENT SPECIFICATION | |
| | 2.9.1 HARDWARE REQUIREMENTS | |
| | 2.9.2 SOFTWARE REQUIREMENTS | |
| | 2.9.3 LANGUAGES USED | |
| | 2.10 CONCLUSION | |

| | | |
|----|---|-------|
| 3. | USE CASE-1 AND USE CASE-2 | 37-39 |
| 4 | RESUME | 40 |
| 5 | PHOTO GALLERY | 41 |
| 6 | FUTURE SCOPE AND FURTHER ENHANCEMENT OF THE PROJECT | 42-44 |
| 7 | REFERENCE | 44 |

LIST OF FIGURES

| FIGURE NO. | PAGE NO. |
|-----------------------------------|----------|
| Figure 1.1 Organization structure | 10 |
| Figure 2 Activity diagram | 17 |
| Figure 3 Module diagram | 18 |
| Figure 2.2 activity diagram | 23 |
| Figure 2.3 Module diagram | 24 |
| Figure 4.1 Use case-1 diagram | 40 |
| Figure 4.2 Use case-2 diagram | 41 |

| SL NO. | ABBREVIATIONS/ NOTATIONS | DESCRIPTION |
|--------|-----------------------------|---------------------------|
| 1. | VS Code | Visual studio code |
| 2. | JS | Java script |
| 3. | HTML | Hypertext Markup language |
| 4. | PD | Pandas |
| 6. | LE | LabelEncoder |
| 6. | OJT | On JOB Training |
| 7. | CSS | Cascading Style sheet |
| 8 | CV | TfidVectorizer |

LIST OF TABLES

| TABLE NO. | TABLE NAME | PAGE NO. |
|-----------|--------------------------------|----------|
| 1.2 | Product And Market performance | 12 |

CHAPTER – 1

COMPANY PROFILE

COMPANY PROFILE

1.1 Overview of the company

Codelab System is a rapidly growing company in the field of computer application Implementation, solutions and services. Codelab System is a service provider of Web-based Development & Web based Software Development Solutions, Mobile Application Development, Graphic Design and Windows Applications. Codelab Systems is headquartered in Mangalore, with the Business development in UAE, Saudi Arabia and Qatar in a short span of 8+ years, our products as well as services & Solutions have been widely accepted by the global market. Today, Codelab Systems has the experience to undertake any IT development or deployment works on a single point responsibility basis.

Our efficient and experienced team is greatest resource Intellect's infrastructure Houses A-team of young and competitive professionals having experience n Web Designing and Software Development who are dedicated to providing high-end Solution to our clients. We develop software and web-based applications with Latest Technologies. For web development projects, we also provide hosting And, domain Facility for customers, so they don't need to bother about that. Our, products and services are user friendly with easy controls and are of Superior specifications. We are always proactive to fulfill client's needs and requirements to the best possible extent of their satisfaction. We manage interactive sessions with clients throughout the Project development.

1.2 VISION AND MISSION OF THE ORGANIZATION



VISION:

To help people and businesses throughout the world realize their full potential. Codelab inspiring vision statement seeks to support people. You can see its intention isn't about business; it's about people and giving those people the services to be their best selves. With this aim, Codelab has numerous initiatives. It's a big supporter of inclusivity, diversity, environmental issues, and corporate responsibility. We are on a journey to be the trusted performance leader that unleashes the potential of data.

MISSION:



“To enable people and businesses throughout the world to realize their full potential and to organize the world's information and make it universally accessible and useful.”

Codelab Systems provides customized package to suit the needs of every client and take into consideration the needs and requirements of each client's and plan different ideas to improve client's business strategies. Every customer satisfaction is our business and we pay special attention to each client. To provide best services. The main goal of our company is to provide best and innovative products that will help to drive potential customers to their businesses

1.3 ORGANIZATION STRUCTURE

An organization is a group of people who work together, like a neighborhood association, a charity, a union, or a corporation. You can use the word organization to refer to group or business, or to the act of forming or establishing something. Organizational structure (OS) is the systematic arrangement of human resources in an organization so as to achieve common business objectives. It outlines the roles and responsibilities of every member of the organization so that work and information flow seamlessly, ensuring the smooth functioning of an organization.

Types of Organizational Structure

- Hierarchical
- Flat
- Flatarchy
- Functional
- Divisional
- Matrix

In a flatarchy, there are little to no levels of management. A company using this structure could have only one manager in between its executive and all other employees. It is called a flatarchy because it is a hybrid of a hierarchy and a flat organization. This type of organizational structure is used more by smaller companies since they have fewer employees, though it can be used in companies of all sizes. While some companies grow out of this organizational structure, others continue to use it. Codelab systems have a Matrix organization structure, where teams report to multiple leaders. The matrix design keeps open communication between teams and can help companies create more innovative products and services. Using this structure prevents teams from needing to realign every time a new project begins.

1.4 ROLES AND RESPONSIBILITIES OF PERSONNEL IN THE ORGANIZATION

We have an expertise team that offer unique solutions. All the members of our team are professional, experienced and have in depth knowledge of the technology. Codelab Systems provides customized package to suit the needs of every client and take into consideration the needs and requirements of each clients and plan different ideas to improve client's business strategies. The

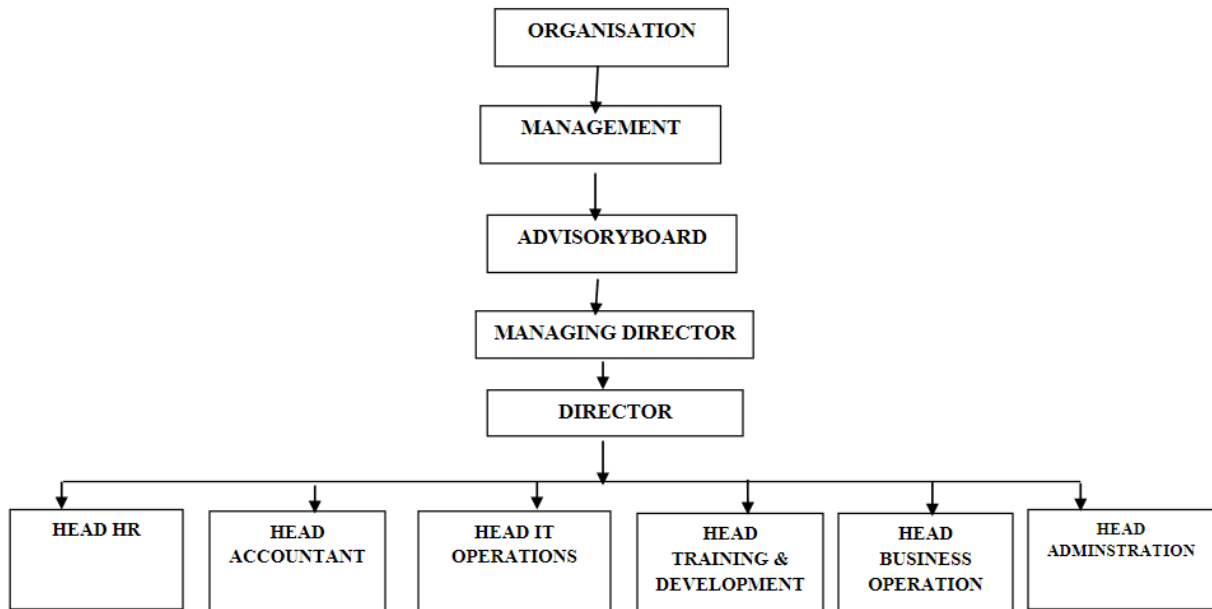


Figure 1.1 Organization Structure

Department of it operation:

Role: Head IT operation

Responsibilities:

- Development of clients line project
- Assigning tasks to the subordinate developers
- Managing client meetings and maintaining a good relationship on stakeholder

Department of Training and Development:

Role: Head of Training and Development

Responsibilities:

- Training to interns and newly joined employees & Work with new projects and domain

Department of HR:

Role: Head of HR

Responsibilities:

- Maintaining Employees data & payroll calculations
- Employee leave management & Interns internship program management.

Department of Account:

Role: Head of Account.

Responsibilities:

- Keep track of daily Account & Maintaining Balance sheet and Income tax procedure

Department of administration:

Role: Head Administration

Responsibilities:

- All the administration work such as file management, print, maintains data of computers and items. Arrangement of training program schedule.

Department of Business operation:

Role: Head Administration

Responsibilities:

- Conducting market research, Contact and approach clients for live projects.
- Communication with new clients and maintaining and managing social

1.5 PRODUCTS AND MARKET PERFORMANCE

- MSS LODGE(INDIA): MSS LODGE is a budget property located in the beautiful city of Ujire.
- SESCO: SESCO is one of the first enterprises in the electrical equipment sector,
- QACADEMIA: Q-Academia providing wide range of career oriented IT Courses.

SERVICES: We believe in quality services

| | |
|---|---|
| Web Development ◎ CMS ◎ E-Commerce ◎ Web Applications | Promotion ◎ SMO/SEO ◎ RANKING ◎ E-MARKETING |
| Professional Website ◎ Re-Design & Solution ◎ Design & Maintenance ◎ E-Commerce & Forums. | Graphics Designing ◎Banner,Poster& Brochure. ◎ Business Card & Identity Card. ◎ Logo, Letter <i>Head & Envelope</i> . |
| Mobile Application Platforms ◎ Android Applications ◎ Windows Applications ◎ App Store Optimization | Others ◎ CSS Conversion ◎ Old website to new generation. ◎Compitable with different Browser. |

table 1.2 Product And Market performance

CHAPTER- 2

ASSESSMENT OF ON JOB TRAINING – 1

CASE-1 ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

1.1 INTRODUCTION:

As the commercial site of the world is almost fully undergone in online platform people is trading products through different e-commerce website. And for that reason reviewing products before buying is also a common scenario. Also now a day, customers are more inclined towards the reviews to buy a product. So analyzing the data from those customer reviews to make the data more dynamic is an essential field nowadays. In this age of increasing machine learning based algorithms reading thousands of reviews to understand a product is rather time consuming where we can polarize a review on particular category to understand its popularity among the buyers all over the world.

The objective of this paper is to categorize the positive and negative feedbacks of the customers over different products and build a supervised learning model to polarize large amounts of reviews. A study on amazon last year revealed over 88% of online shoppers trust reviews as much as personal recommendations. Any online item with large amounts of positive reviews provides a powerful comment of the legitimacy of the item. Conversely, books, or any other online item, without reviews puts potential prospects in a state of distrust. Quite simply, more reviews look more convincing. People value the consent and experience of others and the review on a material is the only way to understand others impression on the product. Opinions, collected from users' experiences regarding specific products or topics, straightforwardly influence future customer purchase decisions. Similarly, negative reviews often cause sales loss. For those understanding the feedback of customers and polarizing accordingly over a large amount of data is the goal. There are some similar works done over amazon dataset. In did opinion mining over small set of datasets of Amazon product reviews to understand the polarized attitudes towards the products.

In our model, we used both manual and active learning approach to label our datasets. In the active learning process different classifiers are used to provide accuracy until reaching satisfactory level. After getting satisfactory result we took those labeled datasets and processed it. From the processed dataset we extracted features that are then classified by different classifiers. We used combination of two kinds of approaches to extract features: the bag of words approach and tf-idf & Chi square approach for getting higher accuracy.

1.2 AIM:

The world we see nowadays is becoming more digitalized. In this digitalized world e-commerce is taking the ascendancy by making products available within the reach of customers where the customer doesn't have to go out of their house. As now a day's people are relying on online products so the importance of a review is going higher. For selecting a product, a customer needs to go through thousands of reviews to understand a product. But in this prospering day of machine learning, going through thousands of reviews would be much easier if a model is used to polarize those reviews and learn from it. We used supervised learning method on a large scale amazon dataset to polarize it and get satisfactory accuracy.

1.3 OBJECTIVE:

- The objective of this research is to develop a supervised learning model that efficiently categorizes customer feedback into positive and negative sentiments for various products. In an era dominated by online commerce, where consumer decisions are heavily influenced by reviews, the need to streamline the analysis of vast amounts of feedback is paramount. By automating the sentiment polarization process, this study aims to provide consumers with valuable insights for informed purchasing decisions and assist businesses in understanding the reception of their products in the market.
- To achieve this objective, a combination of manual and active learning approaches will be utilized for labeling datasets. Active learning involves iteratively employing different classifiers to improve accuracy until a satisfactory level is attained. The labeled datasets will then undergo processing to extract relevant features. Feature extraction will utilize two approaches: the bag of words approach and tf-idf & Chi-square approach. These features will serve as input for various classifiers, ensuring accurate sentiment classification.
- Ultimately, the goal is to develop a robust model capable of efficiently categorizing customer feedback, thereby empowering both consumers and businesses in their decision-making processes within the dynamic landscape of online commerce.

1.4 PURPOSE:

- The purpose of this research is to address the pivotal role of online product reviews in influencing consumer behavior and purchasing decisions in the digital marketplace. With the increasing reliance on e-commerce platforms for trading goods, the significance of customer feedback has surged, as shoppers often turn to reviews to guide their buying choices. However, the sheer volume of reviews poses a challenge for effective analysis and interpretation.

- This study seeks to develop a supervised learning model that can automatically categorize customer feedback into positive and negative sentiments for diverse products. By leveraging machine learning algorithms, the aim is to streamline the process of sentiment analysis, making it more efficient and scalable. The ultimate objective is to empower both consumers and businesses with actionable insights derived from large-scale review data.
- Through a combination of manual and active learning approaches, the research endeavors to label datasets accurately, ensuring the training of robust classifiers. Active learning methodologies will be employed iteratively to enhance classification accuracy, culminating in a model capable of effectively polarizing sentiments across a wide array of products. Feature extraction techniques, including the bag of words approach and tf-idf & Chi-square approach, will be utilized to capture the nuanced characteristics of customer feedback and improve classification performance.
- By understanding and categorizing the sentiments expressed in customer reviews, this study aims to provide valuable intelligence to consumers, enabling them to make informed purchasing decisions. Likewise, businesses stand to benefit from insights into the reception of their products in the market, facilitating strategic decision-making and product development efforts. Ultimately, the research endeavors to contribute to the optimization of online shopping experiences and the enhancement of consumer satisfaction in the digital age.

1.5 SCOPE:

The scope of this research encompasses the development and implementation of a supervised learning model aimed at categorizing customer feedback into positive and negative sentiments across various products within the realm of e-commerce. Given the widespread reliance on online platforms for trading goods and the growing importance of customer reviews in influencing purchasing decisions, the study focuses on analyzing large volumes of review data to derive actionable insights.

1.6 ADVANTAGES:

1.6.1 ADVANTAGES:

- The script preprocesses the raw text data by removing stop words and cleaning the reviews, which helps in improving the quality of the input data for analysis.
- Through bar charts and pie charts, the script provides a visual representation of the distribution of positive and negative sentiments in the dataset, allowing for quick insights into sentiment proportions.
- The word cloud visualizations generated for both positive and negative sentiment categories offer a concise representation of the most frequent words used in reviews, aiding in understanding the key themes and sentiments expressed by customers.

- By using TF-IDF vectorization, the script transforms text data into numerical vectors, capturing the importance of words in reviews relative to the entire corpus. This approach helps in representing text data effectively for machine learning algorithms.
- The logistic regression model is a simple yet effective algorithm for binary classification tasks like sentiment analysis. It offers interpretability and can handle large feature spaces efficiently.
- The script evaluates the performance of the sentiment analysis model using accuracy score, providing a quantitative measure of how well the model performs on unseen data.
- The trained logistic regression model is saved using pickle, allowing for easy deployment and reuse without the need for retraining.
- The script utilizes the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance in the dataset, ensuring better generalization of the model by generating synthetic samples for the minority class.

1.8 ACTIVITY DIAGRAM:

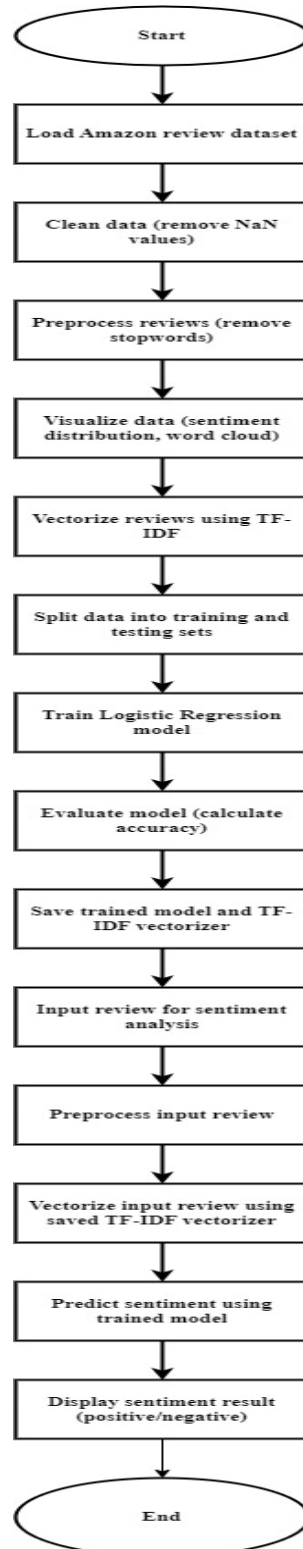


Figure 2 Activity diagram

1.9 MODULE:



Figure 3 Module

1.10 REQUIREMENT SPECIFICATION:

1.10.1 HARDWARE REQUIREMENTS

- RAM: 8 GB or higher
- Storage: 256 GB SSD or higher
- Network: Ethernet/Wi-Fi for internet connectivity
- Display: 15-inch monitor or larger
- Processor: Intel Core i5 or equivalent

1.10.2 SOFTWARE REQUIREMENTS

- Operating System: Windows 10 or Ubuntu 20.04 LTS
- Web Browser: Google Chrome or Mozilla Firefox
- Integrated Development Environment (IDE): Visual Studio Code, Python 3.8 or higher installed.

1.10.3 LANGUAGES USED

- Front-end: HTML, CSS, JavaScript
- Back-end: Python, Flask

1.11 CONCLUSION:

In conclusion, as online commerce continues to dominate global trade, the significance of customer reviews in influencing purchasing decisions has never been greater. Analyzing vast amounts of feedback to categorize and polarize sentiments towards products is essential in this landscape. With over 88% of online shoppers trusting reviews as much as personal recommendations, the volume and sentiment of reviews directly impact consumer trust and purchasing behavior. This study aimed to develop a supervised learning model to categorize positive and negative feedback, leveraging both manual and active learning approaches to label datasets. By employing various classifiers and feature extraction techniques, such as the bag of words and tf-idf & Chi-square methods, the model achieved satisfactory accuracy levels. Understanding and effectively polarizing customer feedback is crucial for businesses in maintaining consumer trust and competitiveness in the online marketplace.

ASSESSMENT OF ON JOB TRAINING - 2

CASE-2 ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

1.1 INTRODUCTION:

In recent years, internet has become an integral part of life. With increased use of internet, numbers of email users are increasing day by day. This increasing use of email has created problems caused by unsolicited bulk email messages commonly referred to as Spam. Email has now become one of the best ways for advertisements due to which spam emails are generated. Spam emails are the emails that the receiver does not wish to receive. a large number of identical messages are sent to several recipients of email. Spam usually arises as a result of giving out our email address on an unauthorized or unscrupulous website. There are many of the effects of Spam.

Fills our Inbox with number of ridiculous emails. Degrades our Internet speed to a great extent. Steals useful information like our details on you Contact list. Alters your search results on any computer program. Spam is a huge waste of everybody's time and can quickly become very frustrating if you receive large amounts of it. Identifying these spammers and the spam content is a laborious task. Even though extensive number of studies have been done, yet so far the methods set forth still scarcely distinguish spam surveys, and none of them demonstrate the benefits of each removed element compose. In spite of increasing network communication and wasting lot of memory space, spam messages are also used for some attack. Spam emails, also known as non-self, are unsolicited commercial or malicious emails, sent to affect either a single individual or a corporation or a bunch of people. Besides advertising, these may contain links to phishing or malware hosting websites found out to steal confidential information. to solve this problem the different spam filtering techniques are used. The spam filtering techniques are accustomed protect our mailbox for spam mails.

1.2 AIM:

Nowadays, a big part of people rely on available email or messages sent by the stranger. The possibility that anybody can leave an email or a message provides a golden opportunity for spammers to write spam message about our different interests. Spam fills inbox with number of ridiculous emails. Degrades our internet speed to a great extent. Steal's useful information like our details on our contact list. Identifying these spammers and also the spam content can be a hot topic of research and laborious tasks. Email spam is an operation to send messages in bulk by mail. Since the expense of the spam is borne mostly by the recipient, it is effectively postage due advertising. Spam email is a kind of commercial advertising which is economically viable because email could be a very cost-effective medium for sender. With

this proposed model the specified message can be stated as spam or not using Bayes theorem and Naive Bayes Classifier and Also IP addresses of the sender are often detected.

1.3 OBJECTIVE:

- Develop strategies or tools to effectively manage and filter spam emails to reduce clutter and ensure important messages are easily accessible.
- Investigate methods to mitigate the impact of spam on internet speed, whether through network optimization or more efficient email protocols.
- Implement measures to prevent the theft of sensitive information through spam emails, such as educating users about phishing techniques and enhancing email encryption.
- Research and deploy solutions to prevent spam emails from influencing search engine results and potentially leading users to malicious websites.
- Develop efficient spam identification and filtering techniques to reduce the time users spend dealing with spam emails and improve productivity.
- Explore new approaches or technologies to more accurately identify spammers and their tactics, making it easier to combat spam at its source.
- Continuously refine and innovate spam filtering methods, incorporating machine learning, artificial intelligence, and other advanced technologies to stay ahead of evolving spamming techniques.

1.4 PURPOSE:

The purpose of this passage is to highlight the growing problem of spam emails and its detrimental effects on internet users. It discusses how spam inundates inboxes, slows down internet speeds, compromises privacy by stealing information, and alters search results. Additionally, it emphasizes the frustration and time wastage caused by dealing with spam. The passage also mentions the challenges in identifying and filtering spam effectively despite various studies and techniques. Overall, it aims to raise awareness about the seriousness of the spam issue and the need for effective spam filtering techniques to protect users' mailboxes.

1.5 SCOPE:

The passage primarily focuses on the issue of spam emails and its consequences on internet users. It covers various aspects of spam, including its definition, how it is generated, its effects such as filling inboxes, degrading internet speed, stealing information, and altering search results. The passage also briefly mentions the use of spam for malicious purposes such as phishing and distributing malware. Additionally, it touches upon the challenges associated with identifying and filtering spam effectively despite the existence of numerous studies and techniques. However, the passage does not delve deeply into specific technical details of spam filtering methods or the intricacies of spam-related cyber-attacks. It primarily

aims to inform readers about the problems caused by spam and the importance of implementing effective spam filtering techniques.

1.6 ADVANTAGES:

1.6.1 ADVANTAGES:

- Customized the preprocessing steps, such as removing punctuations and stop-words, which can improve the quality of features extracted from the text data.
- By visualizing the data distribution using bar charts, pie charts, and word clouds, you gain insights into the characteristics of spam and non-spam emails, which can inform feature engineering and model selection.
- Using TF-IDF vectorization, you transform the textual data into numerical features, capturing the importance of words in each document relative to the entire corpus. This can improve the model's ability to distinguish between spam and non-spam emails.
- Addressing class imbalance using SMOTE helps to alleviate the problem of having significantly more examples of one class than the other, which can lead to biased models.
- Choosing Decision Tree Classifier for its simplicity and interpretability. Decision trees are easy to understand and can handle both numerical and categorical data.
- By serializing the trained model using pickle, you can save it to disk and reload it later for making predictions without the need to retrain.
- Providing a user-friendly interface for users to input email messages and receive predictions on whether they are spam or not-spam.

1.7 ACTIVITY DIAGRAM

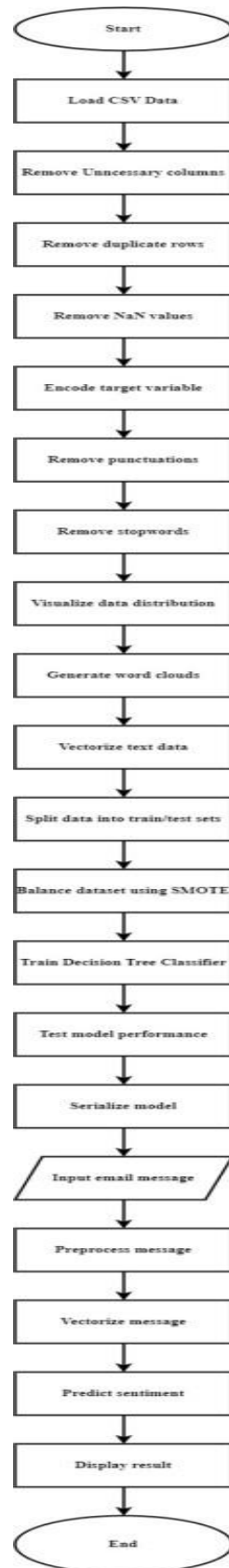


Figure 2.2 Activity diagram

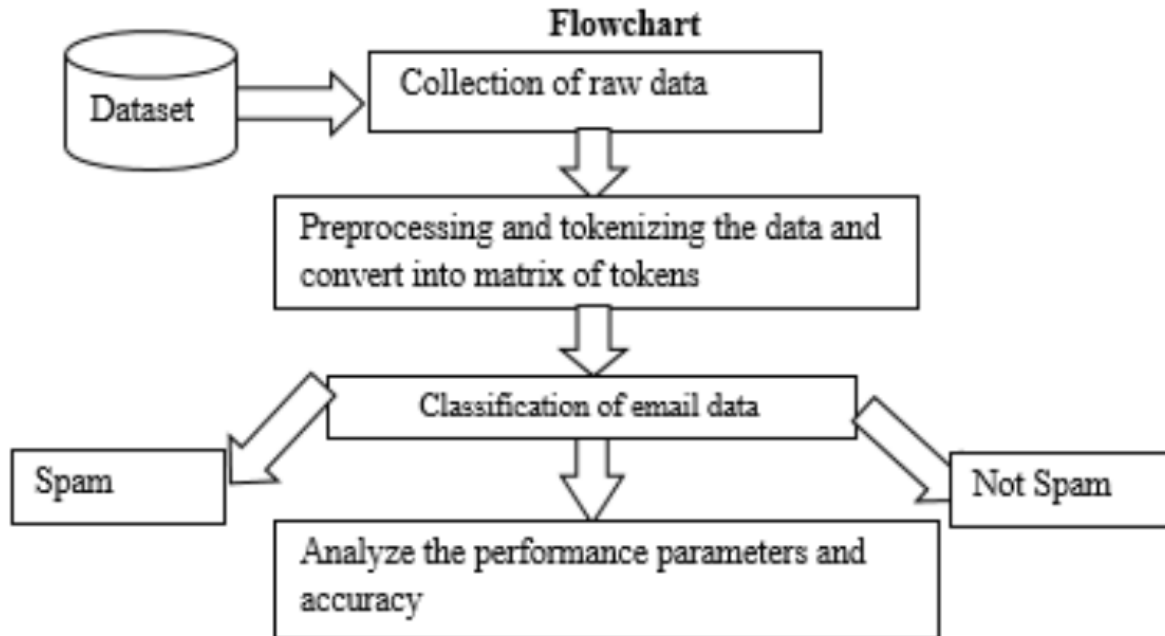
1.8 MODULE:

Figure 2.3 Module

1.10 REQUIREMENT SPECIFICATION:**1.10.1 HARDWARE REQUIREMENTS**

- RAM: 8 GB or higher
- Storage: 256 GB SSD or higher
- Network: Ethernet/Wi-Fi for internet connectivity
- Display: 15-inch monitor or larger
- Processor: Intel Core i5 or equivalent

1.10.2 SOFTWARE REQUIREMENTS

- Operating System: Windows 10 or Ubuntu 20.04 LTS
- Web Browser: Google Chrome or Mozilla Firefox

- Integrated Development Environment (IDE): Visual Studio Code, Python 3.8 or higher installed

1.10.3 LANGUAGES USED

- Front-end: HTML, CSS, JavaScript
- Back-end: Python, Flask

1.9 CONCLUSION:

In conclusion, the proliferation of spam emails has become a significant issue in today's digital landscape. From inundating our inboxes with ridiculous messages to potentially compromising our personal information, the effects of spam are wide-ranging and detrimental. Not only does it degrade internet speed and waste valuable memory space, but it also poses serious security risks, as spam messages can serve as vectors for phishing attempts and malware distribution.

Despite efforts to combat spam through various filtering techniques, including content-based analysis and sender reputation systems, the problem persists. Identifying and mitigating spam remains a challenging task, requiring ongoing research and innovation in cybersecurity.

In the face of this persistent threat, it is crucial for individuals and organizations to remain vigilant and employ robust spam filtering measures to protect themselves from unwanted solicitations and potential security breaches. By staying informed about the latest developments in spam detection and prevention, we can work towards creating a safer and more secure online environment for all users.

IMPLEMENTATION

IMPORT LIBRARY

```
import string
import warnings
warnings.filterwarnings('ignore')
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
import matplotlib.pyplot as plt
from wordcloud import WordCloud
import pickle
from sklearn.preprocessing import LabelEncoder
import nltk
nltk.download('punkt')
nltk.download('stopwords')
from nltk.corpus import stopwords
from imblearn.over_sampling import SMOTE
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score
```

READ DATASET

```
data = pd.read_csv('spam.csv')
data
```

OUTPUT:

| | v1 | v2 | Unnamed: 2 | Unnamed: 3 | Unnamed: 4 |
|------|------|---|------------|------------|------------|
| 0 | ham | Go until jurong point, crazy.. Available only ... | NaN | NaN | NaN |
| 1 | ham | Ok lar... Joking wif u oni... | NaN | NaN | NaN |
| 2 | spam | Free entry in 2 a wkly comp to win FA Cup fina... | NaN | NaN | NaN |
| 3 | ham | U dun say so early hor... U c already then say... | NaN | NaN | NaN |
| 4 | ham | Nah I don't think he goes to usf, he lives aro... | NaN | NaN | NaN |
| ... | ... | ... | ... | ... | ... |
| 5567 | spam | This is the 2nd time we have tried 2 contact u... | NaN | NaN | NaN |
| 5568 | ham | Will Ì_b going to esplanade fr home? | NaN | NaN | NaN |
| 5569 | ham | Pity, * was in mood for that. So...any other s... | NaN | NaN | NaN |
| 5570 | ham | The guy did some bitching but I acted like i'd... | NaN | NaN | NaN |
| 5571 | ham | Rofl. Its true to its name | NaN | NaN | NaN |

5572 rows × 5 columns

DROPPING UNNECESSARY COLUMNS

```
data.drop(columns=['Unnamed: 2', 'Unnamed: 3', 'Unnamed: 4'], inplace=True)
data
```

OUTPUT:

| | v1 | v2 |
|------|------|---|
| 0 | ham | Go until jurong point, crazy.. Available only ... |
| 1 | ham | Ok lar... Joking wif u oni... |
| 2 | spam | Free entry in 2 a wkly comp to win FA Cup fina... |
| 3 | ham | U dun say so early hor... U c already then say... |
| 4 | ham | Nah I don't think he goes to usf, he lives aro... |
| ... | ... | ... |
| 5567 | spam | This is the 2nd time we have tried 2 contact u... |
| 5568 | ham | Will Ì_b going to esplanade fr home? |
| 5569 | ham | Pity, * was in mood for that. So...any other s... |
| 5570 | ham | The guy did some bitching but I acted like i'd... |
| 5571 | ham | Rofl. Its true to its name |

5572 rows × 2 columns

```
data.info()
```

OUTPUT:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5589 entries, 0 to 5588
Data columns (total 2 columns):
#   Column  Non-Null Count  Dtype
---  -
0    v1      5589 non-null    object
1    v2      5589 non-null    object
dtypes: object(2)
memory usage: 87.5+ KB
```

```
data.dropna(inplace=True)
```

```
data.info()
```

OUTPUT:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5589 entries, 0 to 5588
Data columns (total 2 columns):
#   Column  Non-Null Count  Dtype
---  -
0    v1      5589 non-null    object
1    v2      5589 non-null    object
dtypes: object(2)
memory usage: 87.5+ KB
```

LABEL ENCODING A COLUMN

```
le = LabelEncoder()
```

```
data['v1'] = le.fit_transform(data['v1'])
```

```
print(data['v1'])
```

OUTPUT:

```
0      0
1      0
2      1
3      0
4      0
      ..
5584    1
5585    1
5586    1
5587    1
5588    1
Name: v1, Length: 5589, dtype: int32
```

REMOVING STOPWORDS

```
stp_words = stopwords.words('english')
def clean_message(message):
    clean_message = " ".join(word for word in message.split() if word not in stp_words)
    return clean_message

data['v2'] = data['v2'].apply(clean_message)
print(data.v2)
```

OUTPUT:

```

0      Go jurong point, crazy.. Available bugis n gre...
1                      Ok lar... Joking wif u oni...
2      Free entry 2 wkly comp win FA Cup final tkts 2...
3          U dun say early hor... U c already say...
4          Nah I think goes usf, lives around though
...
5584    Congrats 843136XXXX, Rs 23,650 credited Ludo A...
5585    Hi 843136XXXX, Get Full Body checkup 81 Tests ...
5586    Good News 8431364807, Transaction successfully...
5587    Best Health Insurance Plan For You! Get Covera...
5588    TrxnDate My name Edna, urgently need workers 1...
Name: v2, Length: 5589, dtype: object

```

VALUE COUNTS TYPICALLY REFERS TO A METHOD USED TO COUNT THE OCCURRENCES OF UNIQUE VALUES WITHIN A DATASET OR A SPECIFIC COLUMN

```
data['v1'].value_counts()
```

OUTPUT

```

v1
0    4825
1     764
Name: count, dtype: int64

```

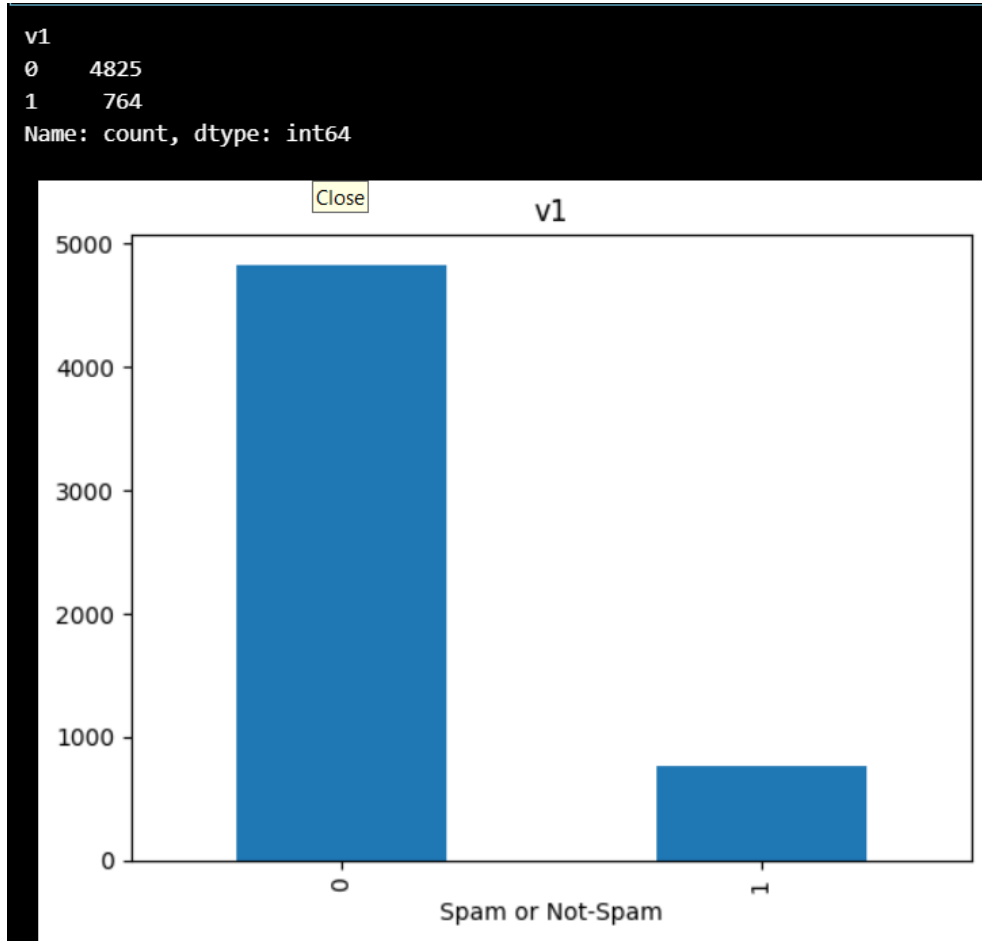
BAR GRAPH

```

trans = data['v1'].value_counts()
print(trans)
trans.plot.bar()
plt.title('v1')
plt.xlabel("Spam or Not-Spam")
plt.show()

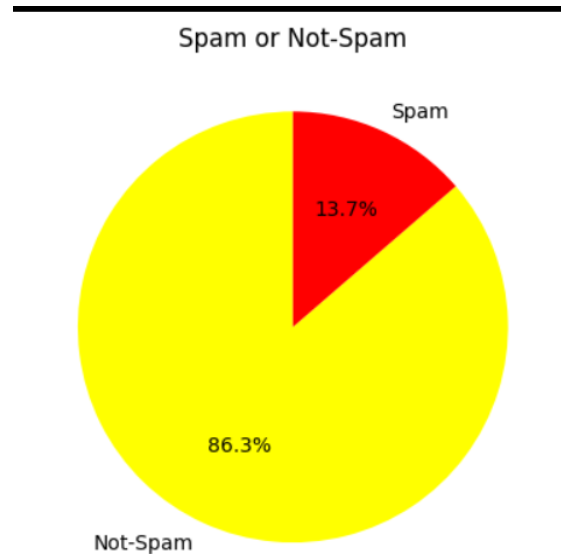
```

OUTPUT:



PIE CHART

```
x = data['v1'].value_counts()
y = 'Not-Spam', 'Spam'
plt.pie(x, labels = y, autopct = '%1.1f%%', startangle = 90, colors = ['yellow', 'red'])
plt.title('Spam or Not-Spam')
plt.show()
```

OUTPUT:**WORD CLOUD**

```
consolidated = ''.join(word for word in data['v2'][data['v1'] == 0].astype(str))
wordCloud = WordCloud(width = 1600,height = 800,max_font_size = 110)
plt.figure(figsize = (15,10))
plt.imshow(wordCloud.generate(consolidated),interpolation = 'bilinear')
plt.axis('off')
plt.show()
```

OUTPUT:

```
consolidated = ''.join(word for word in data['v2'][data['v1'] == 1].astype(str))
```

```
wordCloud = WordCloud(width = 1600,height = 800,max_font_size = 110)
plt.figure(figsize = (15,10))
plt.imshow(wordCloud.generate(consolidated),interpolation = 'bilinear')
plt.axis('off')
plt.show()
```

OUTPUT:



```
cv = TfidfVectorizer(max_features=2500)
X = cv.fit_transform(data['v2']).toarray()
print(X)
```

```
Y = data['v1']
```

```
with open('cv.pkl','wb') as file:
    pickle.dump(cv, file)
```

OUTPUT:

```
[[0.      0.      0.      ... 0.      0.      0.      ]
 [0.      0.      0.      ... 0.      0.      0.      ]
 [0.      0.      0.      ... 0.      0.      0.      ]
 ...
 [0.      0.      0.      ... 0.      0.      0.      ]
 [0.27618075 0.24230061 0.      ... 0.      0.      0.      ]
 [0.      0.      0.      ... 0.      0.      0.      ]]
```

```
#Initialize SMOTE with a sampling strategy (you can adjust it as needed)
smote = SMOTE(sampling_strategy = 'auto',random_state = 42)
```

```
#Apply SMOTE to resample the dataset
X_resampled,y_resampled = smote.fit_resample(X,Y)
y_resampled.value_counts()
```

OUTPUT

```
v1
0    4825
1    4825
Name: count, dtype: int64
```

```
print(x_train)
print(x_test)
```

OUTPUT:

```
[[0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 ...
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]]
[[0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 ...
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]]
```

```
print(x_train.shape)
```

```
print(x_test.shape)
```

OUTPUT:

```
(4191, 2500)
(1398, 2500)
```

TRAINING AND PREDICITNG

```
model = DecisionTreeClassifier()
```

```
#Model fitting
```

```
model.fit(x_train,y_train)
```

```
#testing the model
```

```
pred = model.predict(x_test)
```

```
#model accuracy
```

```
print(accuracy_score(y_test,pred))
```

```
print(pred)
```

```
import pickle
```

```
pickle.dump(model,open('model_save.pkl','wb'))
```

```
model = pickle.load(open('model_save.pkl','rb'))
```

OUTPUT:

```
0.9635193133047211
[0 0 0 ... 0 0 0]
```

```
def predict_sentiment(message_text):
```

```
    #Preprocess the input review
```

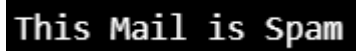
```
    cleaned_message = clean_message(message_text)
```

```
    #Transfrom the review using the TF-IDF vectorizer
```

```
transformed_message = cv.transform([cleaned_message]).toarray()
#Predict sentiment using the trained model
prediction = model.predict(transformed_message)

if prediction[0] == 1:
    return 'Spam'
else:
    return 'Not-Spam'

#Now you can use the predict_sentiment function to classify reviews
input_message = input('Enter the Mail::')
result = predict_sentiment(input_message)
print(f'This Mail is {result}')
```

OUTPUT:

CHAPTER -4

USE CASE-1 AND USE CASE-2

USE CASE-1 DIAGRAM:

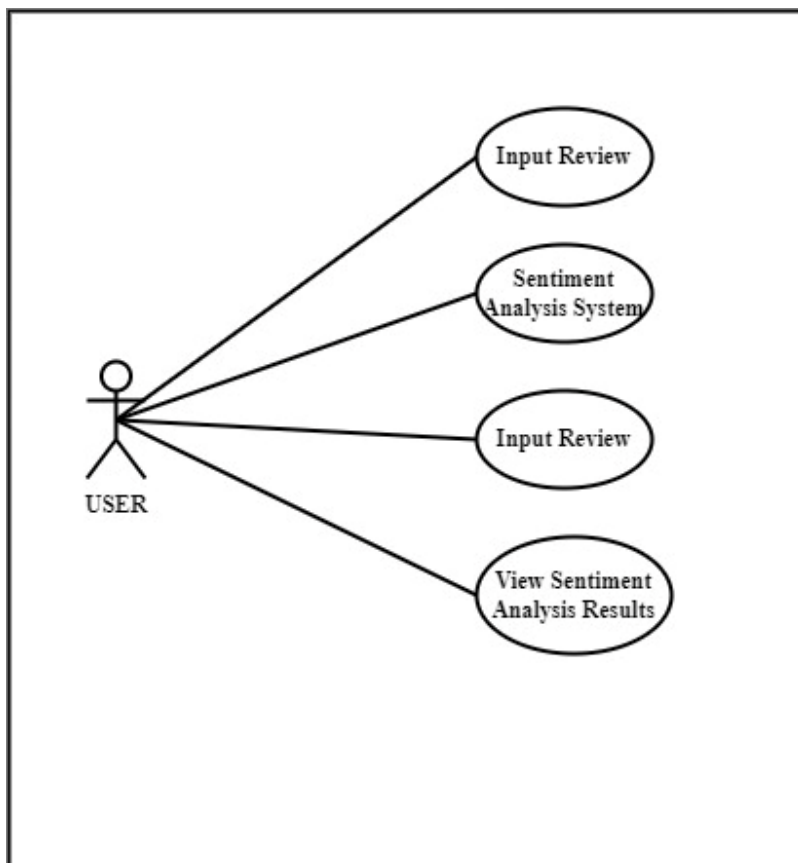


Fig 4.1: USE CASE-1

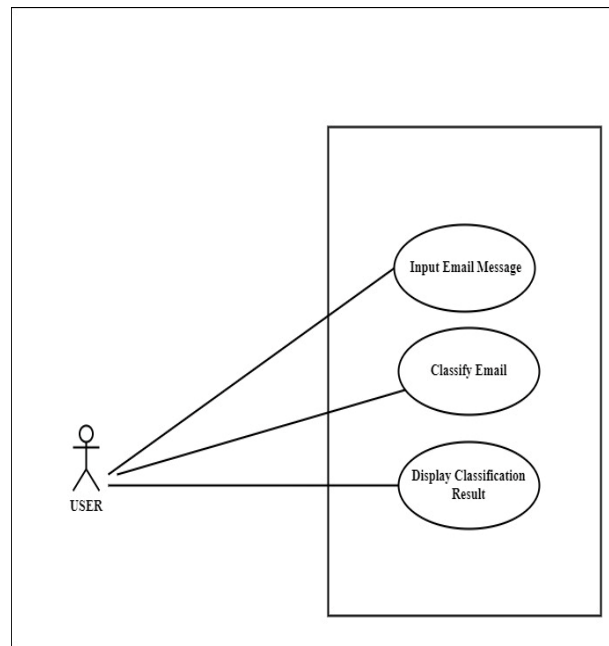
USE CASE-2 DIAGRAM

Fig 4.2: USE CASE-2

RESUME



PAWAN

Chandthimar House, Bantwal Post and Taluk, Dakshina
Kannada, Karnataka – 574211

8431364807 | pawanbangera561@gmail.com

TECHNICAL SKILLS

- ❖ AI & ML
- ❖ Python
- ❖ Java
- ❖ HTML
- ❖ Project Management
- ❖ Computer Network
- ❖ MS Word
- ❖ PowerPoint

PERSONNEL PROFILE

- ❖ Father's Name : Damodar
- ❖ Mother's Name : Geetha
- ❖ Date of Birth :08/02/2006
- ❖ Gender : Male
- ❖ Fluent : Kannada ,Hindi ,
English & Tulu

CERTIFICATES

- ❖ Introduction to Artificial Intelligent from Infosys Spring Board
- ❖ Data visualization using Python from Infosys Spring Board
- ❖ Explore Machine Learning using Python from Infosys Spring Board

Place : BANTWAL

Date :

OBJECTIVE

Enthusiastic and dedicated AI & ML Diploma student seeking opportunities to apply academic knowledge and skills in a real-world environment. Eager to contribute to innovative projects and learn from experienced professionals in the field.

ACADAMIC QUALIFICATION

- ❖ Holy Saviour English Medium School,Agrar
SSLC
76.96%
2021
- ❖ Government Polytechnic Bantwal
Diploma in Computer Science & Engineering
77.80% (up to 5th sem)
2024

PROJECT

Mini Project : Gold Price Prediction

In this Mini Project we have started the program by exploring the data and analyzing it by using preprocessing method and graphically represented the dataset. After preprocessing we divided the dataset into train & test set and also we applied Linear Regression, RandomForestClassifier and ANN models to find the best accuracy and compared it by Bar graph. In this comparison we got the best accuracy in ANN model about 99%.

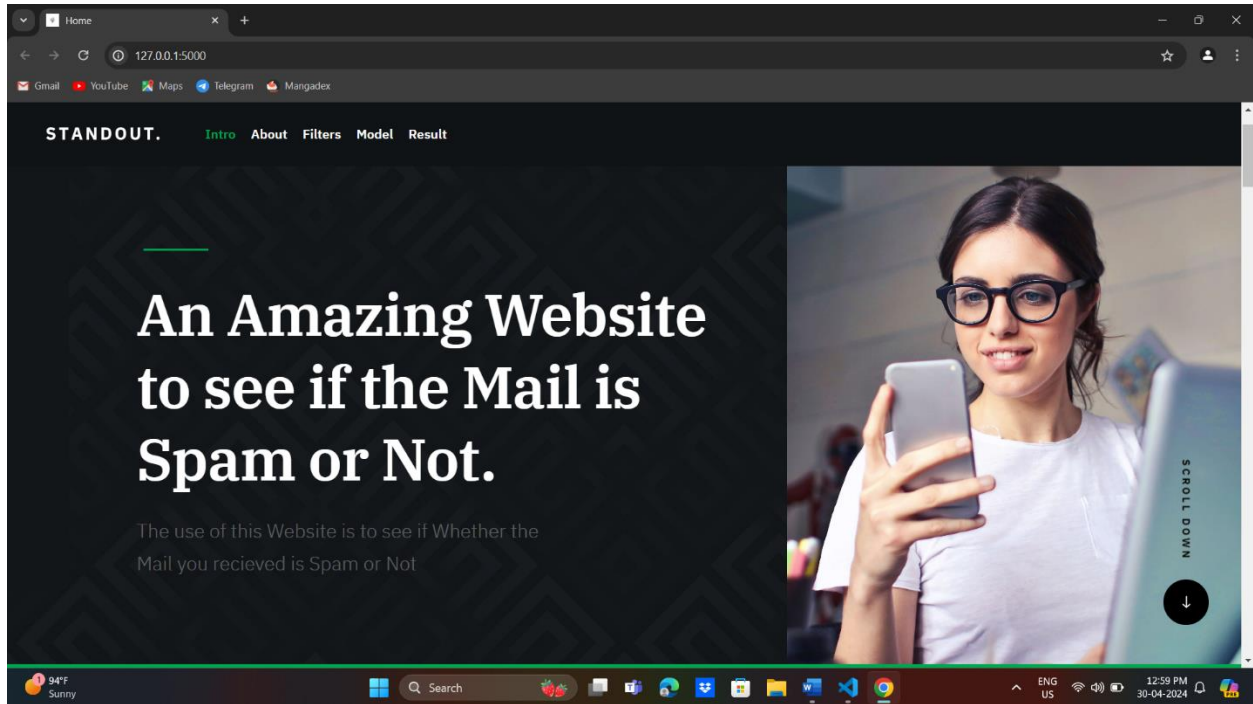
DECLARATION

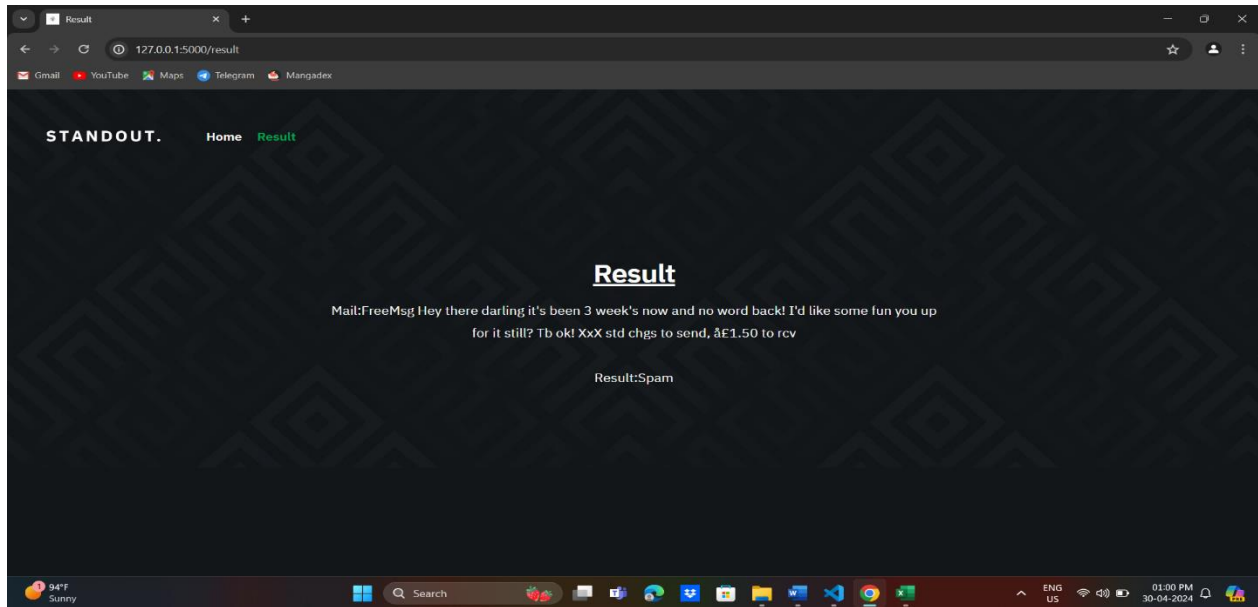
I here by declare that above given particulars are true to the best of my knowledge.

(PAWAN)

PHOTO GALLERY

SNAPSHOTS OF OJT – 2





FUTURE SCOPE AND FURTHER ENHANCEMENT OF THE PROJECT

FUTURE SCOPE:

- **Model Performance Monitoring:** Implement a system to monitor the performance of your deployed model over time. This could include tracking metrics such as accuracy, precision, recall, and F1-score, and setting up alerts for any significant changes in performance.
- **Model Performance Monitoring:** Implement a system to monitor the performance of your deployed model over time. This could include tracking metrics such as accuracy, precision, recall, and F1-score, and setting up alerts for any significant changes in performance.
- **Deployment Scalability:** Design your deployment infrastructure to be scalable, allowing for easy expansion as the volume of incoming emails or messages increases.
- **Security Measures:** Implement security measures to protect against potential attacks such as adversarial samples or model inversion. This could involve techniques like input sanitization, anomaly detection, or model robustness testing.

- **Continuous Learning:** Explore techniques for continuous learning, where the model can adapt to changes in the data distribution over time. This could include approaches like online learning or incremental model updates.
- **Multi-language Support:** Extend your model to support multiple languages by training it on multilingual datasets or incorporating language detection mechanisms.
- **Integration with Email Clients:** Integrate your spam classification model directly into email clients or messaging platforms to provide real-time classification of incoming messages.
- **User Interface Enhancements:** Improve the user interface of your application to make it more intuitive and user-friendly. This could include features like search functionality, message filtering, or customizable classification thresholds.
- **Privacy Considerations:** Ensure that your system complies with privacy regulations and best practices for handling sensitive user data. This may involve implementing data anonymization techniques or encryption methods.
- **Collaborative Filtering:** Explore collaborative filtering techniques to leverage the collective feedback from multiple users for better spam classification results.

FURTHER ENHANCEMENT OF THE PROJECT:

- **Explore Different Models:** Try experimenting with other classification algorithms such as Random Forest, Naive Bayes, or Support Vector Machines (SVM) to see if they provide better performance.
- **Hyperparameter Tuning:** Perform hyperparameter tuning for your chosen model(s) using techniques like grid search or random search to optimize model performance.
- **Feature Engineering:** Explore additional features that could improve the model's performance. For example, you could extract features like message length, presence of specific keywords, or punctuation usage.
- **Ensemble Methods:** Consider using ensemble methods such as Bagging or Boosting, which combine multiple models to improve performance.
- **Evaluate Imbalance Handling Techniques:** Since the dataset might be imbalanced (more examples of one class than the other), explore different

techniques for handling class imbalance such as oversampling, undersampling, or using different sampling algorithms like SMOTE.

- **Advanced Text Processing Techniques:** Experiment with advanced text processing techniques such as n-grams, word embeddings (Word2Vec, GloVe), or deep learning-based methods (LSTM, CNN) for text classification tasks.
- **Cross-Validation:** Implement cross-validation techniques to get a better estimate of the model's performance and ensure its generalization to unseen data.
- **Deployment:** Consider deploying your model as a web service or integrating it into an application for practical use. Tools like Flask or FastAPI can be helpful for creating APIs.
- **Feedback Loop:** Implement a feedback loop mechanism where user feedback on classification results is used to continuously improve the model over time.
- **Security Measures:** Ensure that the deployed model is robust against adversarial attacks and follows best practices for handling sensitive data, especially in the case of email content.

REFERENCE

[1] <https://www.ijert.org/email-based-spam-detection>

[2]

https://www.academia.edu/51233708/IJERT_Email_Spam_Detection_and_Data_Optimization_using_NLP_Techniques

[3]

https://www.irjmets.com/uploadedfiles/paper/issue_4_april_2023/36685/final/fin_irjmets1682217866.pdf