Internship Report

On

EMAIL SPAM DETECTION USING MACHINE LEARNING

Submitted by
S.Naga savitha
21691A31A9
Madanapalle Institute of Technology and Science

Submitted to
Mallika Srivastava
Head, Training Delivery
EISystems Services

&

Mayur Dev Sewak
Head, Internships & Trainings
EISystems Services

Student's Declaration

I, S. Naga Savitha, a student of B.Tech program, Roll No. 21691A31A9 of the Department of CSE(Artificial Intelligence),Madanapalle Institute of Technology and Science College do hereby declare that I have completed the mandatory internship in Eisystems Technologies under the faculty guideship of Mr.R.Ashok kumar, Department of CSE(Artificial Intelligence, Madanapalle Institute of Technology and Science.

S.Nagasavitha/15-07-2024

(Signature and Date)

Endorsements

SIGNATURE
Mr.R.Ashok kumar
CSE(Artificial Intelligence)
Madanapalle Institute of Technology and Science

SIGNATURE
Dr.K.Chokkanathan
CSE(Artificial Intelligence)
Madanapalle Institute of Technology and Science

Table of Content

List of Figures

Executive Summary

Nowadays communication plays a major role in everything be it professionalor personal. Email communication service is being used extensively because of itsfree use services, low-cost operations, accessibility, and popularity.

Emails have onemajor security flaw that is anyone can send an email to anyone just by getting theirunique user id. This security flaw is being exploited by some businesses andill-motivated persons for advertising, phishing, malicious purposes, and finally fraud.This produces a kind of email category called SPAM.Spam refers to any email that contains an advertisement, unrelated andfrequent emails. These emails are increasing day by day in numbers. Studies showthat around 55 percent of all emails are some kind of spam.

A lot of effort is beingput into this by service providers. Spam is evolving by changing the obvious markersof detection. Moreover, the spam detection of service providers can never beaggressive with classification because it may cause potential information loss toincase of a misclassification.To tackle this problem we present a new and efficient method to detect spamusing machine learning and natural language processing.

A tool that can detect andclassify spam. In addition to that, it also provides information regarding the textprovided in a quick view format for user convenience.

Overview of Organization

India's leader in workshops & trainings at IITs, NITs & top engineering colleges

EISystems Services is a leading Indian technology identity with operations across India. EISystems (We call it EISys) offers trainings in Cybersecurity, Machine Learning, Automobiles, Internet of Things, Robotics and Socialmedia for enterprises and student community. Till date we have trained approximately 50000 students and impacted around 2 lakhs students through our various outreach initiatives since our founding.

Our Presence

Some of the colleges where we had already felt our presence are given below:-

Indian Institute of Science, Bangalore
Indian Institute of Technology, Bombay
Indian Institute of Technology, Delhi
Indian Institute of Technology, Madras
Indian Institute of Technology, Kanpur
Indian Institute of Technology, Roorkee
Indian Institute of Technology, Guwahati
Indian Institute of Technology (Banaras Hindu University), Varanasi
Indian Institute of Technology, Indore
Indian Institute of Technology, Jodhpur
Indian Institute of Technology, Hyderabad
National Institute of Technology, Tiruchirappalli
National Institute of Technology, Warangal
National Institute of Technology, Calicut
National Institute of Technology, Patna
National Institute of Technology, Jalandhar
National Institute of Technology, Jaipur
National Institute of Technology, Durgapur
National Institute of Technology, Surat
National Institute of Technology, Allahabad
Indian Institute of Information Technology, Allahabad
ABV Indian Institute of Information Technology, Gwalior
PDP Indian Institute of Information Technology, JabalpuJawahar Lal Nehru Technological Univeristy, Hyderabad
College of Engineering, Guindy
Delhi Technological Univeristy, New Delhi
& around 100 engineering colleges.

Project Summary

Idea behind this project:

Spam email detection using machine learning involves training models to classify emails as either spam or non-spam (ham). The idea is to leverage patterns in email content, metadata, and other features to make accurate predictions. Here's a basic outline of the process:

1. **Data Collection:** Gather a large dataset of emails labeled as spam or ham. This dataset is used for training and testing the model.

2. **Feature Extraction**: Convert emails into a format suitable for machine learning. Common features include:
   - Text Features: Words and phrases in the email body, subject, and metadata.
   - Statistical Features: Frequency of certain words, presence of links, etc.
   - Metadata Features: Sender's email address, time of sending, etc.

3. **Preprocessing:** Clean and preprocess the text data by removing stop words, stemming/lemmatizing words, and transforming text into numerical representations (e.g., using TF-IDF or word embeddings).

4. **Model Training**: Use machine learning algorithms to train a model on the preprocessed data.
   - Support Vector Machines (SVM): Can handle high-dimensional data well.

5. **Model Evaluation:** Assess the model's performance using metrics like accuracy, precision, recall, and F1-score on a separate test set. Ensure the model is not overfitting and generalizes well to new data.

6. **Deployment:** Integrate the trained model into an email system to classify incoming emails in real-time. Continuously monitor and update the model as needed to maintain its effectiveness against evolving spam techniques.

7. Implement continuous learning mechanisms to keep the model updated with new spam patterns and improve its detection capabilities over time.

Using machine learning for spam detection allows for more accurate and adaptive filtering compared to rule-based systems, as the models can learn from patterns and anomalies in large datasets.

Software used:

Operating system : Windows 8/10.

IDE Tool : Google colaboratory

Coding Language : Python 3.8

APIs : Numpy, Pandas,PySpark, Matplotlib,tkinter,nltk data

Technical apparatus requirements:

Processor : Pentium i3 or higher.

RAM : 4 GB or higher.

Hard Disk Drive : 20 GB (free).

Peripheral Devices : Monitor, Mouse and Keyboard.

Result:

From the results obtained we can conclude that an ensemble machine learning model is more effective in detection and classification of spam than any individual algorithms. We can also conclude that TF-IDF (term frequency inverse documentfrequency) language model is more effective than Bag of words model in classification of spam when combined with several algorithms. And finally we can say that spamdetection can get better if machine learning algorithms are combined and tuned to needs.

Data Flow Diagram / Process Flow

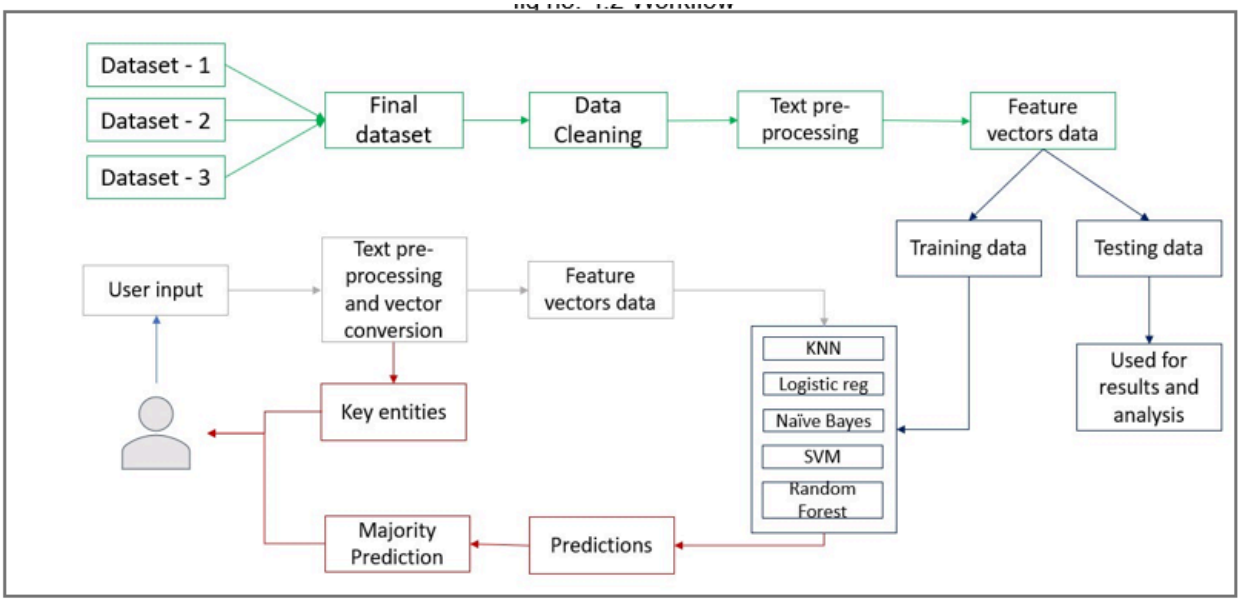

Figure 1 - Data Flow Diagram(DFD)

In the above architecture, the objects depicted in Green belong to a module called Data Processing. It includes several functions related to data processing, natural Language Processing. The objects depicted in Blue belong to the Machine Learning module. It is where everything related to ML is embedded. The red objects represent final results and outputs.

Code

```
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import MultinomialNB
from sklearn.pipeline import Pipeline
from sklearn import metrics

# Load the dataset
url =
'https://archive.ics.uci.edu/ml/machine-learning-databases/00228/smsspamcollec
tion.zip'
df = pd.read_csv(url, sep='\t', header=None, names=['label', 'message'])

# Convert labels to binary values: spam=1, ham=0
df['label'] = df['label'].map({'ham': 0, 'spam': 1})

# Split dataset into features and labels
X = df['message']
y = df['label']

# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Create a pipeline that combines a CountVectorizer with a Naive Bayes classifier
pipeline = Pipeline([
   ('vectorizer', CountVectorizer()),  # Converts text to token counts
   ('classifier', MultinomialNB())     # Naive Bayes classifier
])

# Train the model
pipeline.fit(X_train, y_train)

# Make predictions
y_pred = pipeline.predict(X_test)
```

```
# Evaluate the model
accuracy = metrics.accuracy_score(y_test, y_pred)
confusion = metrics.confusion_matrix(y_test, y_pred)
classification_report = metrics.classification_report(y_test, y_pred)

# Output the results
print("\nModel Evaluation:")
print(f'Accuracy: {accuracy:.2f}')
print('\nConfusion Matrix:')
print(confusion)
print('\nClassification Report:')
print(classification_report)

# Function to classify new messages
def classify_message(message):
    prediction = pipeline.predict([message])
    return 'spam' if prediction[0] == 1 else 'ham'

# Example of classifying new messages
new_messages = [
    "Congratulations! You've won a $1000 gift card. Call now to claim your prize!",
    "Hi there, can we meet tomorrow to discuss the project details?",
    "Limited time offer! Buy one get one free on all items.",
    "I hope you're doing well. Just checking in to see how the project is going."
]

print("\nClassifications of New Messages:")
for msg in new_messages:
    result = classify_message(msg)
    print(f"Message: {msg}")
    print(f"Predicted label: {result}")
    print("-" * 50)
```

Input / Output with Datasets & Supported Screenshots

Input dataset link:
'https://archive.ics.uci.edu/ml/machine-learning-databases/00228/smsspamcollection.zip'

output :

```
First few rows of the dataset:
   label                                            message
0    ham  Go until jurong point, crazy.. Available only ...
1   spam  Free entry in 2 a wkly comp to win FA Cup fina...
2    ham  U dun say so early hor... U c already then say...
3    ham  Nah I don't think he goes to usf. He lives aro...
4   spam  FreeMsg Hey there darling it's been 3 week's n...

Model Evaluation:
Accuracy: 0.98

Confusion Matrix:
[[946   8]
 [ 19 153]]

Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.99      0.98       954
           1       0.95      0.89      0.92       172

    accuracy                           0.98      1126
   macro avg       0.96      0.94      0.      1126
```

```
           0       0.98      0.99      0.98       954
           1       0.95      0.89      0.92       172

    accuracy                           0.98      1126
   macro avg       0.96      0.94      0.95      1126
weighted avg       0.98      0.98      0.98      1126

Classifications of New Messages:
Message: Congratulations! You've won a $1000 gift card. Call now to claim your prize!
Predicted label: spam
--------------------------------------------------
Message: Hi there, can we meet tomorrow to discuss the project details?
Predicted label: ham
--------------------------------------------------
Message: Limited time offer! Buy one get one free on all items.
Predicted label: spam
--------------------------------------------------
Message: I hope you're doing well. Just checking in to see how the project is going.
Predicted label: ham
--------------------------------------------------
```

References

[1] S. H. a. M. A. T. Toma, "An Analysis of Supervised Machine Learning Algorithms for Spam Email Detection," in International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI), 2021.

[2] S. Nandhini and J. Marseline K.S., "Performance Evaluation of Machine Learning Algorithms for Email Spam Detection," in International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020.

[3] A. L. a. S. S. S. Gadde, "SMS Spam Detection using Machine Learning and Deep Learning Techniques," in 7th International Conference on Advanced Computing and Communication Systems (ICACCS), 2021, 2021.

[4] V. B. a. B. K. P. Sethi, "SMS spam detection and comparison of various machine learning algorithms," in International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN), 2017.

[5] G. D. a. A. R. P. Navaney, "SMS Spam Filtering Using Supervised Machine Learning Algorithms," in 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2018.

[6] S. O. Olatunji, "Extreme Learning Machines and Support Vector Machines models for email spam detection," in IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), 2017.

[7] S. S. a. N. N. Kumar, "Email Spam Detection Using Machine Learning Algorithms," in Second International Conference on Inventive Research in Computing Applications (CIRCA), 2020.

[8] R. Madan, "medium.com," [Online]. Available: https://medium.com/analytics-vidhya/tf-idf-term-frequency-technique-easiest-explanation-for-text-classification-in-nlp-with-code-8ca3912e58c3.

[9] N. D. J. a. M. M. A. M. M. RAZA, "A Comprehensive Review on Email Spam Classification using Machine Learning Algorithms," in International Conference on Information Networking (ICOIN), 2021, 2021.

[10] A. B. S. A. a. P. M. M. Gupta, "A Comparative Study of Spam SMS Detection Using Machine Learning Classifiers," in Eleventh International Conference on Contemporary Computing (IC3), 2018.

[11] M. M. J. Fattahi, "SpaML: a Bimodal Ensemble Learning Spam Detector based on NLP Techniques," in IEEE 5th International Conference on Cryptography, Security and Privacy (CSP), 2021, 2021.

[12] Harika, "Analytics Vidhya," [Online]. Available: https://www.analyticsvidhya.com/blog/2021/07/an-introduction-to-logistic-regression/.

[13] İ. A. D. a. M. D. H. Karamollaoglu, "Detection of Spam E-mails with Machine Learning Methods," in Innovations in Intelligent Systems and Applications Conference (ASYU), 2018.

[14] M. N. U. a. R. K. H. F. Hossain, "Analysis of Optimized Machine Learning and Deep Learning Techniques for Spam Detection," in IEEE International IoT, Electronics and Mechatronics Conference (IEMTRONICS), 2021.

[15] H. Deng, "Towards Data Science," [Online]. Available: https://towardsdatascience.com/random-forest-3a55c3aca46d.

[16] j. Brownlee, "machinelearningmastery," 2017. [Online]. Available: machinelearningmastery.com/gentle-introduction-bag-words-model.

[17] d. AI, "deepai," [Online]. Available: deepai.org/machine-learning-glossary-and-terms/accuracy-error-rate.

Student *Self Evaluation of the Short-Term Internship*

Please rate your performance in the following areas:

1) Oral communication	**1**	**2**	**3**	•	**5**
2) Written communication	**1**	**2**	**3**	•	**5**
3) Initiative	**1**	**2**	**3**	**4**	•
4) Interaction with staff	**1**	**2**	**3**	**4**	•
5) Attitude	**1**	**2**	**3**	**4**	•
6) Dependability	**1**	•	**3**	**4**	**5**
7) Ability to learn	**1**	**2**	**3**	**4**	•
8) Planning and organization	**1**	**2**	**3**	**4**	•
9) Professionalism	**1**	**2**	**3**	**4**	•
10) Creativity	**1**	**2**	**3**	**4**	•
11) Quality of work	**1**	**2**	**3**	**4**	•
12) Productivity	**1**	**2**	**3**	•	**5**
13) Progress of learning	**1**	**2**	**3**	•	**5**
14) Adaptability to organization's culture/policies	**1**	**2**	**3**	**4**	•
15) OVERALL PERFORMANCE	**1**	**2**	**3**	**4**	•

Rating Scale: 5 will be Best while 1 will be Worst

S.Naga savitha

Signature of the Student

Annexure 1

Daily Activity Report

Week No: _____
(1/2/3/4/5/6/7/8/9/10/11/12/13/14/15/16)

Day & Date	Brief Description of Daily Activity	Learning Outcome	Person In-Charge
Day 1	Introduction to ML	NA	Robokwik Training
Day 2	Working of ML	NA	Robokwik Training
Day 3	Notes	NA	Robokwik Training
Day 4	Python:fundamentals	NA	Robokwik Training
Day 5	ML:fundamentals	NA	Robokwik Training

Annexure 2
Weekly Progress Report

Week No: 1-8
(1/2/3/4/5/6/7/8)

Week(s)	Summary of Weekly Activity
Week 1	Introduction to Machine Learning, How Machine Learning works?, Foundation of python , variable , constant& naming convention
Week 2	Print function and comments, starting with Datatypes-Number Datatypes, String datatypes, List
Week 3	List methods, Tuple, Dictionary, set, Boolean, user Input & Type casting, Control statement, project 1(Quiz game),loops in python
Week 4	File handling, Function, packages and mobiles ,exception handling, oops concept
Week 5	Numpy, pandas ,matplotlib Model demonstration, models &projects
Week 6	Project work
Week 7	Project work
Week 8	Project work