🏛️ Acharya Narendra Dev College

Under Guidance

Prof .SharanJeetKaur

Submitted By

MO SAIF

Ishan Srivastava

📧 EMAIL SPAM DETECTION PROJECT

Data Mining Project

🔗 Project Source

- **Dataset:** [UCI Data Mining Repository – SMS Spam Collection]

📜 Problem Statement

Build a model to classify emails as **Spam** or **Not Spam**.

📊 Dataset Overview

- **Total Emails:** 5574
- **Spam Emails:** 747
- **Not Spam Emails:** 4827

Example:

- *"You've won! Claim your prize now!"* → **Spam**

🛠️ Workflow

Data Collection → Preprocessing → Feature Extraction → Model Training → Evaluation → Prediction

---

🧰 Techniques and Methods

- **Feature Extraction:**
    - TF-IDF (Term Frequency-Inverse Document Frequency)
    - Bag of Words (BoW)
- **Model:**
    - Naive Bayes Classifier
- **Validation:**
    - 75%-25% Train-Test Split

📈 Data Visualization

- **Top 20 Common Words in Spam Messages**
  (Shown using a **Horizontal Bar Graph**)

📊 Model Performance

- **Confusion Matrix**
- **ROC Curve:** AUC = 0.996
- **Classification Report:**
  (Precision, Recall, F1-Score shown)
- **Performance Bar Chart:**
  (Comparison of Precision, Recall, F1-Score)

🔍 Feature Importance

- **Top Predictive Words:**
  "free", *"win"*, *"cash"*, *"urgent"*, *"claim"*

❗ Error Analysis

- **False Positives:** 1
- **False Negatives:** 11

Example of Misclassification:

- *"Meeting reminder"* → Incorrectly classified as **Spam**

✅ Conclusion

The **Naive Bayes model** successfully classifies emails with **near-perfect accuracy**.

🔮 Future Work

- Implement **Deep Learning** approaches (e.g., LSTM Networks)
- **Multilingual** Spam Detection
- **Real-time Filtering Systems**

👥 Team Contribution

- **Ishan Srivastava:** Data Preparation & Model Building
- **Mo. Saif:** Evaluation, Reporting & Visualization

🔗 Access Project

MO SAIF

Github link :https://github.com/MOSAIF-dev/Data-Mining-Project-/tree/main/Data%20Mining%20Emails%20Detection

LinkedIn Profile : https://www.linkedin.com/in/mo-saif-461768289/

Ishan Srivastava

Github Link : https://github.com/ishan261204

LinkedIn Profile :https://www.linkedin.com/in/ishan-srivastava-0b88842b0?utm_source=share&utm_campaign=share_via&utm_content=profile&utm_medium=android_app

Poster Overview

Data Mining Project

EMAIL SPAM DETECTION PROJECT

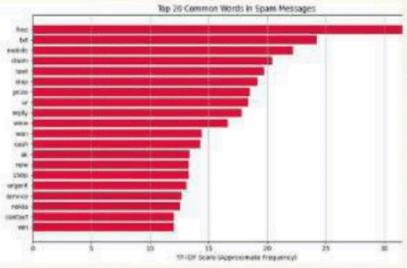
Project Source : This dataset is taken from UCI DATA MINING Respository: https://archive.ics.uci.edu/dataset/228/sms+spam+collection

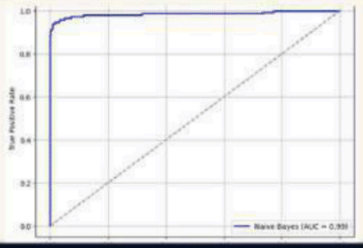
PROBLEM STATEMENT

We aim to create a Data Mining model to classify emails as spams or not spams

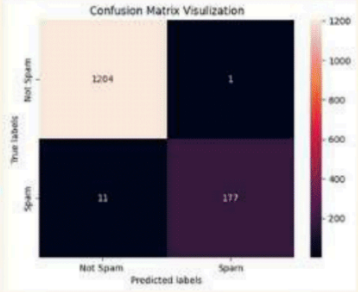
DATA VISUALIZATION EXAMPLES

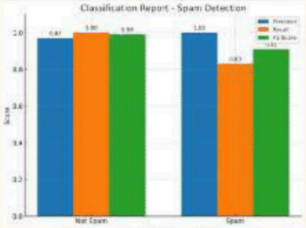
Top 20 Common Words in Spam Messages





- Feature Extraction (TE-IDF / Bag of Words)
- Feature extraction,Naive bayes .





Classlification Report - Spam Detect

Classification Report

SUMMARY

Label	Precision	Recall	F1-Score	Support
ham	0.99	1.00	1.00	1205
spam	0.99	0.94	0.97	188
Accuracy			0.99	1393
Macro Avg	0.99	0.97	0.98	1393
Weighted Avg	0.99	0.99	0.99	1393

Classification Report – Spam

CONCLUSION

The naive bayes model accurately classifies email messages as spam with high performance.

Ishan Srivastava **Mo. Saif**

✍️ Students

- **Ishan Srivastava**
- **MO SAIF**

QR Link