SPAM MAIL DETECTION USING PYTHON

FINAL PROJECT SUBMISSION

NAME: ABHISHEK ATTRI

TOOLS USED: PYTHON, MATPLOTLIB/SEABORN, GITHUB

DATE: 10 JUNE 2025

INTRODUCTION

- • WHAT IS SPAM MAIL?
- - UNSOLICITED, UNWANTED EMAILS USUALLY SENT IN BULK.
- • WHY DETECT SPAM?
- - PROTECT USERS FROM SCAMS, MALWARE, AND PHISHING.
- - IMPROVE EMAIL SYSTEM EFFICIENCY.

OBJECTIVE OF THE PROJECT

- • BUILD A MODEL TO CLASSIFY EMAILS AS SPAM OR NOT SPAM.
- • USE VISUALIZATIONS TO EXPLORE:
- - WORD DISTRIBUTIONS
- - FREQUENCY OF SPAM KEYWORDS
- - CORRELATION OF FEATURES
- • IDENTIFY THE MOST IMPORTANT FEATURES FOR CLASSIFICATION.

DATASET OVERVIEW

- • SOURCE: [E.G., UCI ML REPOSITORY, KAGGLE]
- • RECORDS: [E.G., 5,572 EMAILS]
- • FIELDS:
- - LABEL (SPAM/HAM)
- - MESSAGE CONTENT
- • PREPROCESSING:
- - TEXT CLEANING
- - LOWERCASING, PUNCTUATION REMOVAL
- - TOKENIZATION & VECTORIZATION

VISUALIZATIONS CREATED

- 1. BAR CHART – COUNT OF SPAM VS. HAM EMAILS
- 2. WORD CLOUD – MOST COMMON WORDS IN SPAM MESSAGES
- 3. HISTOGRAM – MESSAGE LENGTH DISTRIBUTION
- 4. HEATMAP – FEATURE CORRELATION
- 5. (OPTIONAL): PIE CHART OF LABEL PROPORTION

CHART TYPES & WHY THEY WERE CHOSEN

- • BAR CHART: CLEAR CATEGORY COMPARISON
- • WORD CLOUD: EASY KEYWORD SPOTTING
- • HISTOGRAM: UNDERSTAND CONTENT LENGTH
- • HEATMAP: CHECK RELATIONSHIPS AMONG FEATURES
- • ALL CHOSEN TO MATCH DATA NATURE

STORYTELLING THROUGH DATA

- • SPAM MESSAGES TEND TO BE SHORTER AND USE PROMOTIONAL KEYWORDS.
- • WORDS LIKE “FREE”, “WIN”, “CLICK” APPEAR FREQUENTLY IN SPAM.
- • VISUAL PATTERNS HELP UNDERSTAND MODEL FEATURE IMPORTANCE.
- • SPAM DETECTION BECOMES MORE INTERPRETABLE.

INTERACTIVITY (IF APPLICABLE)

- • INTERACTIVE FEATURES:
- - HOVER FOR EXACT VALUES
- - FILTERS FOR SPAM/HAM
- - ZOOMABLE WORD FREQUENCY
- • ENABLED WITH TOOLS LIKE PLOTLY, POWER BI, TABLEAU

ACCURACY & VALIDATION

- • DATA CROSS-VERIFIED FOR MISSING VALUES
- • USED STRATIFIED SAMPLING FOR CLASS BALANCE
- • SCALED FEATURES FOR BETTER CORRELATION
- • VALIDATED PLOTS USING SEABORN/MATPLOTLIB

GITHUB REPOSITORY

- • LINK: [INSERT YOUR GITHUB REPO LINK HERE]

- INCLUDES:
- • SOURCE CODE
- • VISUALIZATIONS
- • DATASET
- • README WITH OBJECTIVES & USAGE

CONCLUSION

- • SUCCESSFULLY BUILT AND VISUALIZED A SPAM MAIL DETECTION SYSTEM.
- • DISCOVERED CLEAR PATTERNS IN SPAM MESSAGES USING CHARTS.
- • GITHUB REPO READY WITH COMPLETE DOCUMENTATION.