|  |
| --- |
| **A Project Report on**  Email/ SMS Spam Detection  Submitted in partial fulfillment of award of BACHELOR OF TECHNOLOGY degree  in  **COMPUTER SCIENCE & ENGINEERING**  By  SHUBHAM SHARMA (1900820100145)  SHUBHAM (1900820100143)  SHIVANI (1900820100138)  HARSHITA NAILWAL (2000820109001)  TABISH ABSAR (1900820100153)  Session: 2019- 2023  DR. SATENDRA KUMAR  (Assistant Professor)  SUPERVISOR    C:\Users\Sharma\Downloads\transparent (1).png  **Department of Computer Science and Engineering**  **Moradabad Institute of Technology**  **Moradabad (U.P.)**  **MAY 2023** |

**CERTIFICATE**

Certified that the Project Report entitled **“EMAIL SPAM DETECTION”** submitted by **Shubham Sharma (1900820100145)**, **Shubham** **(1900820100143)**, **Shivani (1900820100138)**, **Harshita Nailwal (2000820109001)**, **Tabish Absar (1900820100153)** is their own work and has been carried out in our supervision. It is recommended that the candidates may now be evaluated for their project work by the university.

**Date:**

**Dr. Satendra Kumar**

Assistant Professor

(Project Supervisor)

Department of Computer Science and Engineering

Moradabad Institute of Technology, Moradabad (082)

**ABSTRACT**

The project titled “Email Spam Detection” is implemented using the CRISP-DM methodology. You will get to know Business understanding, Data Understanding (Data Description and Exploration), Data Preparation, Modelling, and Evaluation steps. Project is implemented using Python class object-based style. Email spam detection is done using machine learning algorithms Naive Bayes and SVM (Support vector machines). Further, it shows the complete program flow for Python-based email spam classifier implementation such as Data Retrieval Flow, Data Visualization Flow, Data Preparation Flow, Modeling, and Evaluation Flow. Also, including the section regarding Data ethics.

This whole project is divided into these steps: Data Cleaning (Remove unwanted/ useless data from the dataset), EDA (Exploratory data Analysis (EDA) is an analysis approach that identifies general patterns in the data by perform various operations), Text Preprocessing (In this step there are different steps we have perform like tokenization, lower case, remove special characters, remove stop words, and punctuations, etc.), Model Building (Build Machine learning model with machine learning algorithms), Evaluation, Improvements (Depending upon the evaluation), Website (in this step we perform various steps like study about streamlit, APIs, etc.), Deploy (In this step we perform various steps like study about Render and make this project machine independent).

Short\_Message Service (SMS), which allows users to send and receive messages, has become a multi-billion dollar industry as mobile phone usage has soared. The cost of messaging services has also decreased, which has led to an increase in the amount of spam that is delivered to mobile devices. Up to 40% of SMS messages in some regions of Asia were spam in 2012. Due to short message lengths, lack of reliable databases for SMS spams, informal language, and brief message characteristics, the current email filtering algorithms may not perform well in their. In this project, real SMS spam databases from the ML repository are used. Following feature extraction and preprocessing, On the databases, numerous machine learning methods are used. After comparing the results, the best algorithm for text message spam filtering is then presented. The results utilising that in this study decreases the total error rate of the best model in the original research referencing this. The following algorithms are used in this technique: Spam communications are categorised in mobile device communication using decision trees, K-Nearest Neighbour, and logistic regression The SMS spam collecting set is used to test the approach.

**ACKNOWLEDGEMENT**

Firstly, we would like to thank **Dr. Rohit Garg**, Director (MIT College) for permit us to make this project. Secondly, we would like to thank **Dr. Manish Gupta**, HOD (Department of CSE) for giving us opportunity to making this project. Thirdly, we would like to thank **Dr. Satendra Kumar (**student mentor), without their guidance this project seems far away from our reach. He helped us a lot regarding any technical issue and give their precious time and best advices which we have followed in this project. A paper is not enough for us to express the support and guidance we received from them almost all the work we did there. Although this project is a good example of team work and we are thankful for our team. Finally, we apologized all other unnamed who helped us in various ways to have a good training.

Last but not the least, we wish to thank our **Parents** for financing our studies in this college as well as for constantly encouraging us to learn engineering. Their personal sacrifice in providing this opportunity to learn engineering in gratefully acknowledged.

**SHUBHAM SHARMA (1900820100145)**

**SHUBHAM (1900820100143)**

**SHIVANI (1900820100138)**

**HARSHITA NAILWAL (2000820109001)**

**TABISH ABSAR (1900820100153)**

**TABLE OF CONTENTS**

**CONTENTS PAGE NO.**

**ABSTRACT……………………………………………………………………………………………………….........…............iii-iv**

**TABLE OF CONTENTS………………………………...………………………………………………………..….....vi-viii**

**LIST OF FIGURES………………………………...…………………………………………………………….….…...........ix-x**

**CHAPTER 1. INTRODUCTION………………………………………………………………………………………..11**

**1.1 Introduction………………………………………………………………..……………………………………………11-12**

**1.2 Aim & Objectives………………………………………………………..……………………………….……………......13**

**CHAPTER 2. Literature Survey……………………………………………………………….…...……………………14**

**2.1 Data Mining………………………………………………………………………………………………………………….14**

**2.1.1 Data Mining Survey…………………………………………………………………………………….……14-15**

**2.1.2 Stages in Data Mining………………………………………………………………………………………15-16**

**2.1.3 Techniques in Data Mining……………………………………………………………………...………17-18**

**2.1.4 Benefits of Data Mining……………………………………………………………………………………...…18**

**2.2 Existing System…………………………………….………………………………………………………………………19**

**2.3 Proposed System…………………………………………………………………..………………………,.…………..…20**

**2.3.1 Modules of Proposed System……………………………………………………………...……………20-21**

**2.4 Software Description…………..………………………………………………..……………………………………….21**

**2.4.1 Jupyter Notebook………………………………………………………………………….…………………21-22**

**2.4.2 Pycharm……………………………………………………………………………………………………………22-23**

**2.5 Python Library Description…………..…………………………………………..…………………………………24**

**2.5.1 Numpy……………………………………………………………………………………………………………………24**

**2.5.2 Pandas………………………………………………………………………………………………………………25-26**

**2.5.3 Matplotlib…………………………………………………………………………………………………………26-27**

**2.5.4 Scikit-Learn…………………………………………………………………………………………………...…27-28**

**2.5.5 Seaborn…………………………………………………………………………………………………….….……28-29**

**CHAPTER 3. Requirement Analysis……...………………..………………………………….……………………..30**

**3.1 Functional Requirements……………………………………………..………………………………….….……….30**

**3.2 Non-Functional Requirements………….……………………………………………………………..…….30-31**

**3.2.1 Accessibility…………………………………………………………………………………………...………………31**

**3.2.2 Maintainability………………………………………………………………………………...……………………31**

**3.2.3 Scalability………………………………………………………………………………………………………………32**

**3.2.4 Portability……………………………………………………………………………………………...………………32**

**3.3 Hardware Requirements………………..………………………………………..………....…………….………….32**

**3.4 Software Requirements………………..………………………………………..……………...………….………….32**

**CHAPTER 4. SYSTEM DESIGN…………………………………………..……………………………..…...………..33**

**4.1 Activity Diagram …………………………………………………….…..………………………………………….33-34**

**4.2 Class Diagram………….…………………...……………………..……………………………………………….….35-36**

**4.3 ER Diagram………….…………………...……………………..…………………………………………………...….37-38**

**4.4 Data Flow Diagrams………….…………………...……………………..………………………………………..…….39**

**4.4.1 Zero Level DFD Diagram……………………………………………………………………….……………40**

**4.4.2 First Level DFD Diagram………………………………………………………………………….……40-41**

**4.4.3 Second Level DFD Diagram………………………………………………………………………...…41-42**

**CHAPTER 5. IMPLEMENTATION…………………….………………...…..………………………..…………... 43**

**5.1 Data Cleaning..……………………………………….………….…..……………………………………….……….43-44**

**5.2 EDA………….…………………...……………...……………..……………………………...…………………………….44-45**

**5.3 Data Pre-proccessing..………………………………….………….…..………..…………………………….……….45**

**5.3.1 Lower Case……………………………………………………………………………………………………………46**

**5.3.2 Tokenization…………………………………………………………………….………………………………43-44**

**5.3.3 Removing Special Characters……………………………………...…………………………………47-48**

**5.3.4 Removing stop words & Punctuations……………………………..……………………………48-49**

**5.3.5 Stemming………………………………………………………………………………………………………….……49**

**5.4 Model Building………….……………...……………...……………..…………………………………………………….50**

**5.4.1 Logistic Regression………………………………………………………………….………………………50-51**

**5.4.2 Support Vector Machine…………………………………………………………………...……………51-52**

**5.4.3 Naïve Bayesian Classifier…………………………………………………………………..……………52-53**

**5.4.4 AdaBoost Classifier…………………………………………………………………………………………53-54**

**5.4.5 K Neighbours Classifier………………………………………………………….……………………………54**

**5.5 Evaluation………….……………...……………...……………..……………………………………………………….55-56**

**5.6 Improvement………….……………...……………...……………..……………….……………………………………….56**

**5.6.1 TF-IDF……………………………………………………………………………………...………………………56-57**

**5.6.2 Scaling……………………………………………………………………………………………………………………57**

**5.6.3 Voting…………………………………………………………………………………………………………..……57-58**

**5.6.4 Stacking…………………………………………………………………………………………………………….……58**

**5.7 Website………….……………...……………...……………..……………………...…………………………………….58-59**

**5.8 Deploy………….……………...……………...……………..………………………………………………..…………….59-60**

**CHAPTER 6. TESTING…………………….………………...…..………………………..………………………….. 61-62**

**6.1 Unit Testing..……………………………………….………….…..………………………………………….………. 62-64**

**6.2 Integration Testing..……………………………………….………….…..…………………..………….………. 64-65**

**6.3 White box Testing..……………………………………….………….…..……………………….……….………. 66-67**

**6.4 Black box Testing..……………………………………….………….…..………………………..……….………. 67-68**

**6.5 System Testing..……………………………………….………….…..………………………………….….………. 68-70**

**6.6 Testing Performed..………..……………………….………….…..………………………………….….………. 70-75**

**CHAPTER 7. OUTPUT SCREEN…………………….………………...…..……………………………..………….. 76**

**7.1 Home page output..……………………………………….………….…..…………………………………….………. 76**

**7.2 Spam Output..……………………………………….………….…..……………………………….………………….…. 77**

**7.3 Not Spam/ Ham Output..……………………………………….………….…..………………………………….... 78**

**CHAPTER 8. CONCLUSION…………………………………………………..………………………….………….…..79**

**CHAPTER 9. FUTURE SCOPE…………………………………………...…..………………………….……………..80**

**REFERENCES………………………………………………………………………………………………………………...81-84**

**APPENDIX……………………………………………………………………………………………………………………....85-92**

**Annexure**

1. **Research Paper**

**LIST OF FIGURES**

**S. No. Fig. No. Fig. Name PAGE NO.**

01. Fig.2.1 Data Mining.......................................................................................................................15

02. Fig.2.2 Stages in Data Mining................................................................................................16

03. Fig.2.3 Techniques in Data Mining....................................................................................18

04. Fig.2.4 Jupyter Notebook...........................................................................................................22

05. Fig.2.5 PyCharm...............................................................................................................................23

06. Fig.2.6 NumPy....................................................................................................................................25

07. Fig.2.7 Pandas.....................................................................................................................................26

08. Fig.2.8 Scikit Learn........................................................................................................................28

09. Fig.4.1 Activity Diagram...........................................................................................................34

10. Fig.4.2 Class Diagram..................................................................................................................36

11. Fig.4.3 ER Diagram.......................................................................................................................38

12. Fig.4.4 Zero Level DFD Diagram........................................................................................40

13. Fig.4.5 First Level DFD Diagram.......................................................................................41

14. Fig.4.6 Second Level DFD Diagram.................................................................................42

15. Fig.5.1 Dataset before cleaning.............................................................................................43

16. Fig.5.2 Dataset after cleaning..................................................................................................44

17. Fig.5.3 Pie chart for spam and ham messages............................................................45

18. Fig.5.4 Convert text to Lover Case.....................................................................................46

19. Fig.5.5 Tokenization......................................................................................................................47

20. Fig.5.6 Removing Special Character.................................................................................48

21. Fig.5.7 Removing Stopwords.................................................................................................49

22. Fig.5.8 Stemming.............................................................................................................................49

23. Fig.5.9 Accuracy and Precision.............................................................................................56

24. Fig.5.10 Local host website.........................................................................................................59

25. Fig.5.11 Deployment website....................................................................................................60

26. Fig.6.1 Testing Techniques......................................................................................................62

27. Fig.6.2 Unit Testing.......................................................................................................................64

28. Fig.6.3 Integration Testing........................................................................................................65

29. Fig.6.4 White Box Testing........................................................................................................67

30. Fig.6.5 Black Box Testing.........................................................................................................68

31. Fig.6.6 System Testing................................................................................................................70

32. Fig.6.7 Transform\_Text() Error............................................................................................71

33. Fig.6.8 Transform\_Text() Actual Output.......................................................................71

34. Fig.6.9 Transform\_Text() Error Resolved....................................................................72

35. Fig.6.10 Transform\_Text() Expected Output................................................................72

36. Fig.6.11 Dataframe Error.............................................................................................................73

37. Fig.6.12 Dataframe Error Resolved.....................................................................................73

38. Fig.6.13 Bad Visualization.........................................................................................................74

39. Fig.6.14 Better Visualization....................................................................................................74

40. Fig.6.15 Apply method Error....................................................................................................75

41. Fig.6.16 Apply method Error Resolve...............................................................................75

42. Fig.7.1 Home Page Output........................................................................................................76

43. Fig.7.2 Spam Output......................................................................................................................77

44. Fig.7.3 Non Spam/Ham Output.............................................................................................78