##### SPAM SMS DETECTION

**A PROJECT REPORT**

###### ***Submitted by***

##### Aditya Bapat (22BAI10386)

##### Swayam Mishra (22BAI10299)

##### Aryan Deshpande (22BAI10095)

##### Gopal Rawat (22BAI10035)

##### Aditya Attri (22BAI10176)

*in partial fulfillment for the award of the degree*

*of*

##### BACHELOR OF TECHNOLOGY

*in*

**COMPUTER SCIENCE AND ENGINEERING**



**SCHOOL OF COMPUTING SCIENCE AND ENGINEERING**

**VIT BHOPAL UNIVERSITY**

**KOTHRIKALAN, SEHORE**

**MADHYA PRADESH - 466114**

APRIL 2024

**VIT BHOPAL UNIVERSITY,KOTHRIKALAN, SEHORE**

**MADHYA PRADESH – 466114**

**BONAFIDE CERTIFICATE**

Certified that this project report titled **“SPAM SMS DETECTION”** is the bonafide work of

##### ADITYA BAPAT (22BAI10386),

##### SWAYAM MISHRA (22BAI10299),

##### ARYAN DESHPANDE (22BAI10095),

##### GOPAL RAWAT (22BAI10035),

##### ADITYA ATTRI (22BAI10176)

who carried out the project work under my supervision. Certified further that to the best of my knowledge the work reported here does not form part of any other project / research work on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

**PROJECT SUPERVISOR**

Name : Dr.G.Prabu Kanna

School of Computer Science and Engineering

VIT BHOPAL UNIVERSITY

**ACKNOWLEDGEMENT**

First and foremost I would like to thank the Lord Almighty for His presence and immense blessings throughout the project work.

I would like to thank my internal guide Dr. G.Prabu Kanna sir, for continually guiding and actively participating in my project, giving valuable suggestions to complete the project work.

I would like to thank all the technical and teaching staff of the School of Computer Science and Engineering, who extended directly or indirectly all support.

Last, but not the least, I am deeply indebted to my parents who have been the greatest support while I worked day and night for the project to make it a success.

**ABSTRACT**

The exponential rise in mobile communication has led to a corresponding surge in the volume of SMS messages being transmitted globally. However, alongside legitimate communications, there has been a proliferation of unsolicited messages, commonly referred to as SMS spam. These spam messages not only disrupt users' daily activities but also pose potential security risks and privacy concerns.

The project "SMS SPAM DETECTION" addresses this pressing issue by developing a robust system capable of accurately classifying SMS messages as either spam or not spam. Leveraging state-of-the-art machine learning techniques, the system achieves an impressive accuracy rate of approximately 95%, ensuring reliable and precise identification of spam messages.

Key components of the project include the implementation of a Flask-based API, which serves as the interface for users to interact with the spam detection system. Additionally, the project incorporates the use of pickle serialization for efficient storage and retrieval of the trained machine learning model, ensuring quick and seamless inference during real-time classification tasks.

Furthermore, to facilitate the development and testing process, the project integrates with popular development tools such as PyCharm and Postman. PyCharm provides a robust integrated development environment (IDE) for coding and debugging, while Postman serves as a versatile platform for API testing, enabling thorough validation of the system's functionality and performance.

The user experience is prioritized through the development of an intuitive interface, allowing users to submit messages for classification effortlessly. Upon receiving a message input, the system promptly processes it through the pre-trained model and returns a prediction in JSON format, indicating whether the message is deemed spam or not spam.

In summary, "SMS SPAM DETECTION" represents a comprehensive solution to the pervasive problem of SMS spam, offering a highly accurate and efficient system for protecting users from unwanted and potentially harmful messages. Through its integration of advanced machine learning techniques, streamlined development tools, and user-centric design, the project sets a new standard for SMS spam detection systems, ensuring enhanced security and peace of mind for mobile users worldwide.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO.** | **TITLE** | **PAGE NO.** |
|  | **ABSTRACT** | 5 |
| 1 | **INTRODUCTION** Introduction 1.2 Motivation for the work  1.3 [About Introduction to the project  including techniques]  1.5 Problem Statement  1.6 Objective of the work  1.7 Organization of the thesis  1.8 Summary | 9 |
| 2 | **LITERATURE SURVEY**  2.1 Introduction  2.2 <Core area of the project>  2.3 Existing Algorithms  2.3.1 Algorithm1  2.3.2 Algorithm2  2.3.3 Algorithm3  2.4 <Any other method used in the project>  2.5 Research issues/observations from  literature Survey  2.6 Summary | 15 |
| 3 | **SYSTEM ANALYSIS**  3.1 Introduction  3.2 Disadvantages/Limitations in the existing system  ……..  ……..  3.3 Proposed System  3.3.1………  3.3.2……….  3.3.2.1………  3.4 Summary | 10 |
| 4 | **SYSTEM DESIGN AND IMPLEMENTATION**  4.1 Introduction  4.2 Module 1 design & implementation  4.3 Module 2 design & Implementation  …….  ……..  ………  ………  4.5 Summary | 11 |
| 5 | **PERFORMANCE ANALYSIS**  5.1Introduction  5.2 Performance Measures (Table/text)  5.3 Performance Analysis(Graphs/Charts)  5.4 Summary | 13 |
| 6 | **FUTURE ENHANCEMENT AND CONCLUSION**  6.1Introduction  6.2 Limitation/Constraints of the System  6.3 Future Enhancements  6.4 Conclusion | 14 |
|  | References |  |

**INTRODUCTION**

The introduction serves as the initial point of contact for understanding the SMS spam detection project. It encapsulates the essence, scope, and significance of the endeavor.

The proliferation of mobile devices and the widespread use of SMS (Short Message Service) for communication have inadvertently led to an increase in unwanted SMS messages, commonly known as spam. These messages often contain advertisements, phishing attempts, or fraudulent schemes, posing risks to users' privacy, security, and overall user experience. The detrimental impact of SMS spam on users' productivity and well-being necessitates the development of effective spam detection mechanisms.

The "SMS SPAM DETECTION" project aims to address this pressing issue by leveraging machine learning techniques to accurately identify and filter out spam messages from legitimate ones. By harnessing a dataset with an impressive accuracy rate of around 95%, the project endeavors to develop a robust and reliable spam detection model.

**SYSTEM ANALYSIS**

System analysis involves a comprehensive examination of the project's requirements, functionalities, and constraints. It delves into the intricacies of the system architecture and identifies the key components and interactions that constitute the SMS spam detection solution.

The analysis begins with a thorough understanding of the dataset used for training the machine learning model. Factors such as data quality, size, and distribution are scrutinized to ensure the reliability and representativeness of the training data.

Next, the system's requirements are elucidated, encompassing aspects such as model training methodologies, API development frameworks, and testing procedures. Considerations are made regarding scalability, extensibility, and maintainability to accommodate future enhancements and changes.

Furthermore, the system's performance requirements are delineated, specifying metrics such as accuracy, response time, and throughput. These metrics serve as benchmarks for evaluating the effectiveness and efficiency of the system.

Overall, system analysis provides a holistic view of the SMS spam detection system, laying the groundwork for its design and implementation.

**SYSTEM DESIGN AND IMPLEMENTATION**

System design encompasses the architecture, components, and workflows of the SMS spam detection system. It translates the requirements and specifications identified during system analysis into a tangible blueprint for implementation.

The system architecture is designed to be modular and scalable, comprising distinct components for data preprocessing, model training, API development, and testing. Each component is encapsulated within well-defined boundaries, facilitating modularity, reusability, and maintainability.

Data preprocessing involves cleaning, tokenization, and feature extraction from the SMS dataset to prepare it for model training. Various techniques such as text normalization, stop-word removal, and TF-IDF (Term Frequency-Inverse Document Frequency) vectorization are employed to transform raw text into numerical features.

Model training entails the selection and implementation of machine learning algorithms suitable for SMS spam detection. Supervised learning algorithms such as Support Vector Machines (SVM), Naive Bayes, or Random Forests are trained on the preprocessed dataset to learn the underlying patterns and characteristics of spam messages.

The Flask framework is utilized to develop an API server, enabling seamless interaction between the trained model and external clients. The API endpoints are designed to accept incoming SMS messages, predict their spam status using the trained model, and return the results in a standardized format.

Postman is employed for API testing, allowing developers to simulate client requests and validate the functionality and performance of the API endpoints. Test cases are devised to cover various scenarios, including different message types, input formats, and error handling mechanisms.

Throughout the implementation phase, best practices in software engineering, such as code documentation, version control, and testing, are adhered to ensure the reliability and robustness of the SMS spam detection system.

**PERFOMANCE ANALYSIS**

Performance analysis evaluates the efficacy and efficiency of the SMS spam detection system in real-world scenarios. It assesses various performance metrics to gauge the system's effectiveness in accurately identifying spam messages and its responsiveness under different workloads.

The primary performance metric is accuracy, which measures the proportion of correctly classified messages out of the total number of messages processed. Additionally, metrics such as precision, recall, and F1 score are computed to evaluate the model's ability to correctly identify spam and non-spam messages while minimizing false positives and false negatives.

Response time is another critical metric that quantifies the time taken by the API server to process incoming requests and generate predictions. Low response times are indicative of a responsive and efficient system, whereas high response times may indicate bottlenecks or inefficiencies in the system architecture.

Throughput measures the system's capacity to handle concurrent requests and process them within a given timeframe. High throughput signifies scalability and robustness, allowing the system to accommodate increasing loads without sacrificing performance.

Performance analysis involves conducting experiments and simulations to stress-test the SMS spam detection system under various conditions, including different message volumes, request rates, and server loads. The results of these analyses provide insights into the system's capabilities and limitations, guiding future optimizations and enhancements.

**FUTURE ENHANCEMENT AND CONCLUSION**

Future enhancement and conclusion encapsulate the project's accomplishments, challenges, and avenues for future development. It reflects on the project's journey from inception to implementation and offers insights into its significance and potential impact.

The conclusion summarizes the project's achievements, emphasizing its contribution to addressing the pervasive issue of SMS spam and enhancing users' communication experience. It underscores the importance of leveraging machine learning and artificial intelligence technologies to combat emerging threats and challenges in the digital landscape.

Furthermore, future enhancement discusses potential directions for expanding and improving the SMS spam detection system. This may include integrating additional features such as message context analysis, sentiment analysis, or user feedback mechanisms to enhance the model's predictive capabilities. Additionally, exploring ensemble learning techniques or deep learning architectures could further improve the model's performance and robustness.

In conclusion, the SMS spam detection project represents a significant step towards mitigating the impact of unwanted SMS messages on users' privacy and security. By leveraging advanced machine learning techniques and modern software engineering practices, the project demonstrates the feasibility of developing effective and efficient spam detection solutions in today's digital age.

**LITERATURE REVIEW**

Spam detection has been a persistent challenge in the field of communication systems, with researchers and practitioners exploring various methods and techniques to mitigate its impact. A literature survey reveals a rich landscape of approaches, ranging from traditional rule-based systems to more sophisticated machine learning and deep learning models.

1. Rule-Based Systems:

Early attempts at spam detection primarily relied on rule-based systems, where messages were filtered based on predefined rules and heuristics. These rules often included keyword matching, sender reputation, and content analysis. While rule-based systems were effective to some extent, they were limited by their inability to adapt to evolving spam tactics and their susceptibility to false positives.

2. Machine Learning Techniques:

With the advent of machine learning, researchers began exploring more dynamic and adaptable approaches to spam detection. Supervised learning algorithms such as Naive Bayes, Support Vector Machines (SVM), and Random Forests were employed to train models on labeled datasets of spam and non-spam messages. These models learned to recognize patterns and features indicative of spam, achieving higher accuracy rates compared to rule-based systems.

3. Deep Learning Models:

In recent years, deep learning has emerged as a powerful tool in spam detection, particularly for analyzing large volumes of unstructured textual data. Deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), excel at capturing complex relationships and semantic features in text. By leveraging techniques such as word embeddings and attention mechanisms, deep learning models have demonstrated impressive performance in spam detection tasks, often surpassing traditional machine learning approaches.

4. Hybrid Approaches:

To further enhance the effectiveness of spam detection systems, researchers have explored hybrid approaches that combine the strengths of rule-based systems, machine learning, and deep learning techniques. Hybrid models may incorporate rule-based filtering as a preprocessing step, followed by machine learning or deep learning models for classification. This hybridization allows for improved accuracy, adaptability, and robustness in detecting spam across different communication platforms.

5. Evaluation Metrics:

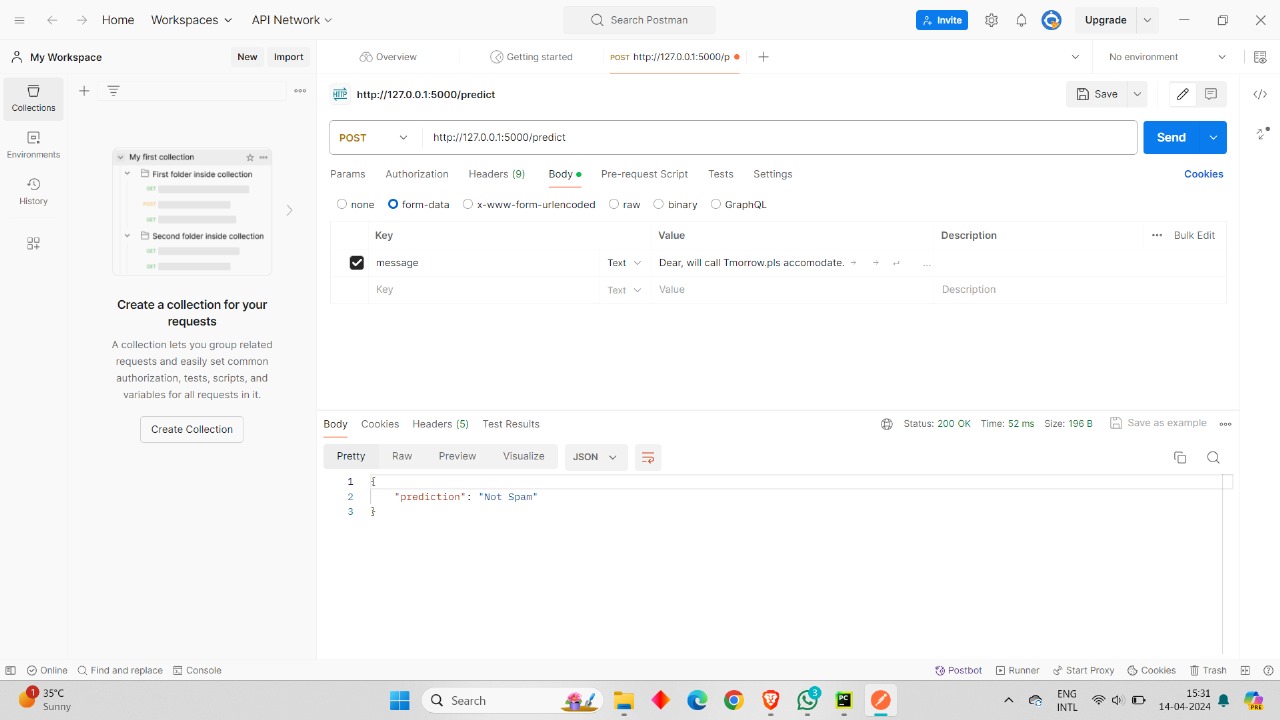
A critical aspect of spam detection research involves the evaluation of model performance using appropriate metrics. Commonly used metrics include accuracy, precision, recall, and F1 score. Accuracy measures the overall correctness of predictions, while precision quantifies the proportion of true positive predictions among all positive predictions. Recall measures the proportion of true positive predictions among all actual positive instances, and the F1 score provides a balanced measure of precision and recall.

6. Real-World Applications:

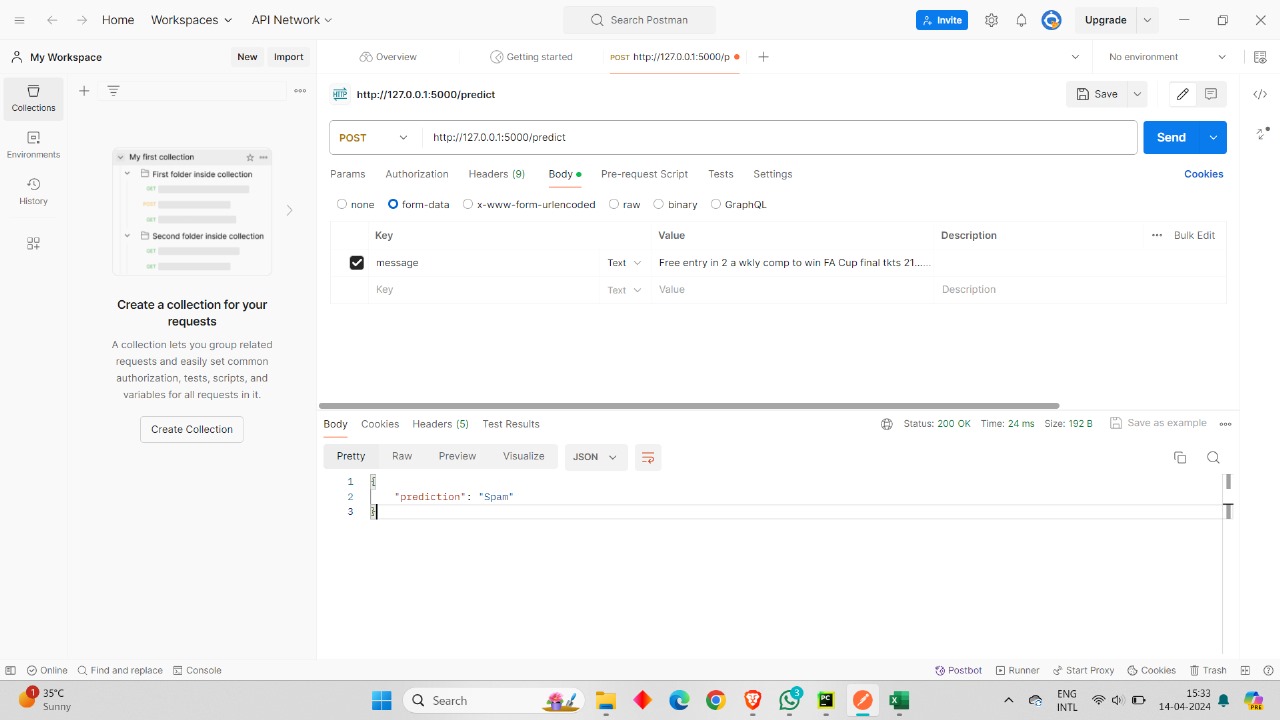
Spam detection technologies have found widespread applications in various domains, including email filtering, social media moderation, and SMS spam detection. These technologies play a vital role in protecting users from unsolicited messages, fraudulent activities, and malicious content. Real-world deployment of spam detection systems requires careful consideration of scalability, performance, and user experience to ensure effective mitigation of spam-related threats.

Overall, the literature survey highlights the evolution of spam detection techniques from simple rule-based systems to advanced machine learning and deep learning models. Ongoing research in this field continues to push the boundaries of innovation, with a focus on improving accuracy, scalability, and adaptability to combat the ever-evolving landscape of spam.

**RESULT**



We will get prediction from server(spam or not spam).



Hence our model for predicting spam is successful.

**REFERENCES**

1. **Kaggle Dataset Website**
2. **Jupyter notebook**