A

SYNOPSIS REPORT ON

ON

SMS SPAM DETECTION

Submitted in partial fulfilment of the requirements of the degree of

MCA Department of Trinity Academy of Engineering

Submitted by

Aniket Sarjine (A207)

Gaurav Tilekar(A217)

Department of MCA Engineering

TRINITY ACADEMY OF ENGINEERING, PUNE

2024-2025

Department of Master of Computer Applications
Project Based Learning

Project Group No	Academic Year 2024-25	Sem-IIi
Roll No.	**Name**	**Email-id**
A207	**Aniket pravin Sarjine**	aniketsarjine11@gmail.com
A217	**Gaurav Rajendra Tilekar**	Gauravtilekar04@gmail.com

Title of Project	**SMS SPAM DETECTION**
Operating System	**Windows 10**
Processor	**Intel core i5**
Front End	**Python(Streamlit)**
Back End	**Python**

1 About Project

1.1 Problem Statement

A number of major differences exist between spam-filtering in text messages and emails. Unlike emails, which have a variety of large datasets available, real databases for SMS spams are very limited. Additionally, due to the small length of text messages, the number of features that can be used for their classification is far smaller than the corresponding number in emails. Here, no header exists as well. Additionally, text messages are full of abbreviations and have much less formal language that what one would expect from emails. All of these factors may result in serious degradation in performance of major email spam filtering algorithms applied to short text messages.

1.2 Objective

Prediction of SMS spam has been an important area of research for a long time. the goal is to apply different machine learning algorithms to SMS spam classification 15 problem, compare their performance to gain insight and further explore the problem, and design an application based on one of these algorithms that can filter SMS spams with high accuracy. The current work proposes a gamut of machine learning and deep learning-based predictive models for accurately predicting the sms spam movement. The predictive power of the models is further enhanced by introducing the powerful deep learning-based long- and short-term memory (LSTM) network into the predictive framework.

1.3 Scope of System

1. The Proposed mode is based on the study of sms text data and technical indicators. Algorithm selects best free parameters combination for LSTM to avoid over-fitting and local minima problems and improve prediction accuracy.

2. Our dataset consists of one large text file in which each line corresponds to a text message. Therefore, preprocessing of the data, extraction of features, and tokenization of each message is required. After the feature extraction, an initial analysis on the data is done using label encoder and then the models like naive Bayes (NB) algorithm and LSTM are used on next steps are for prediction.

3. The two methods used to predict the spam messages that are Fundamental and technical analyses. .

1.4 METHDOLOGY

1. Problem Definition

Objective:

- To classify SMS messages into two categories: "Spam" and "Not Spam."

2. Data Collection

Data Sources:

- Use publicly available datasets such as the SMS Spam Collection Dataset or collect your own data.

Dataset Characteristics:

- Typically includes columns like `label` (spam or ham) and `message` (the SMS content).

3. Data Preprocessing

Text Cleaning:

- **Lowercasing:** Convert all text to lowercase to ensure uniformity.
- **Tokenization:** Split the text into individual words or tokens.
- **Removing Punctuation:** Remove punctuation characters as they are generally not useful for text classification.
- **Removing Stopwords:** Eliminate common words that do not contribute to the meaning (e.g., "the", "is").
- **Stemming/Lemmatization:** Reduce words to their base or root form (e.g., "running" to "run").

4. Feature Extraction

TF-IDF (Term Frequency-Inverse Document Frequency):

- Converts the text data into numerical vectors that can be used by machine learning algorithms.
- Reflects the importance of a word in a document relative to its frequency in the entire corpus.

5. Model Selection and Training

Choosing Algorithms:

- **Naive Bayes:** Often used for text classification due to its simplicity and effectiveness.
- **Logistic Regression:** A straightforward model that can perform well on text data.
- **Support Vector Machines (SVM):** Effective in high-dimensional spaces.
- **Random Forest or Gradient Boosting:** Can be used but may require more tuning.

6. Model Evaluation

Evaluation Metrics:

- **Accuracy:** The proportion of correctly classified messages.
- **Precision:** The proportion of true positives among all positive predictions.
- **Recall:** The proportion of true positives among all actual positives.
- **F1-Score:** The harmonic mean of precision and recall.

7. Model Tuning and Optimization

Hyperparameter Tuning:

- Adjust model parameters to improve performance using techniques like Grid Search or Random Search.

8. Deployment

Building a User Interface:

- **Web Application:** Use frameworks like Flask or Streamlit to build an interface for users to input SMS messages and get predictions.

9. Monitoring and Maintenance

Monitoring:

- Continuously monitor the model's performance to ensure it remains accurate and effective over time.

Updating:

- Periodically update the model with new data to maintain its relevance and accuracy.

1.5 Application & Future Scope

Applications:

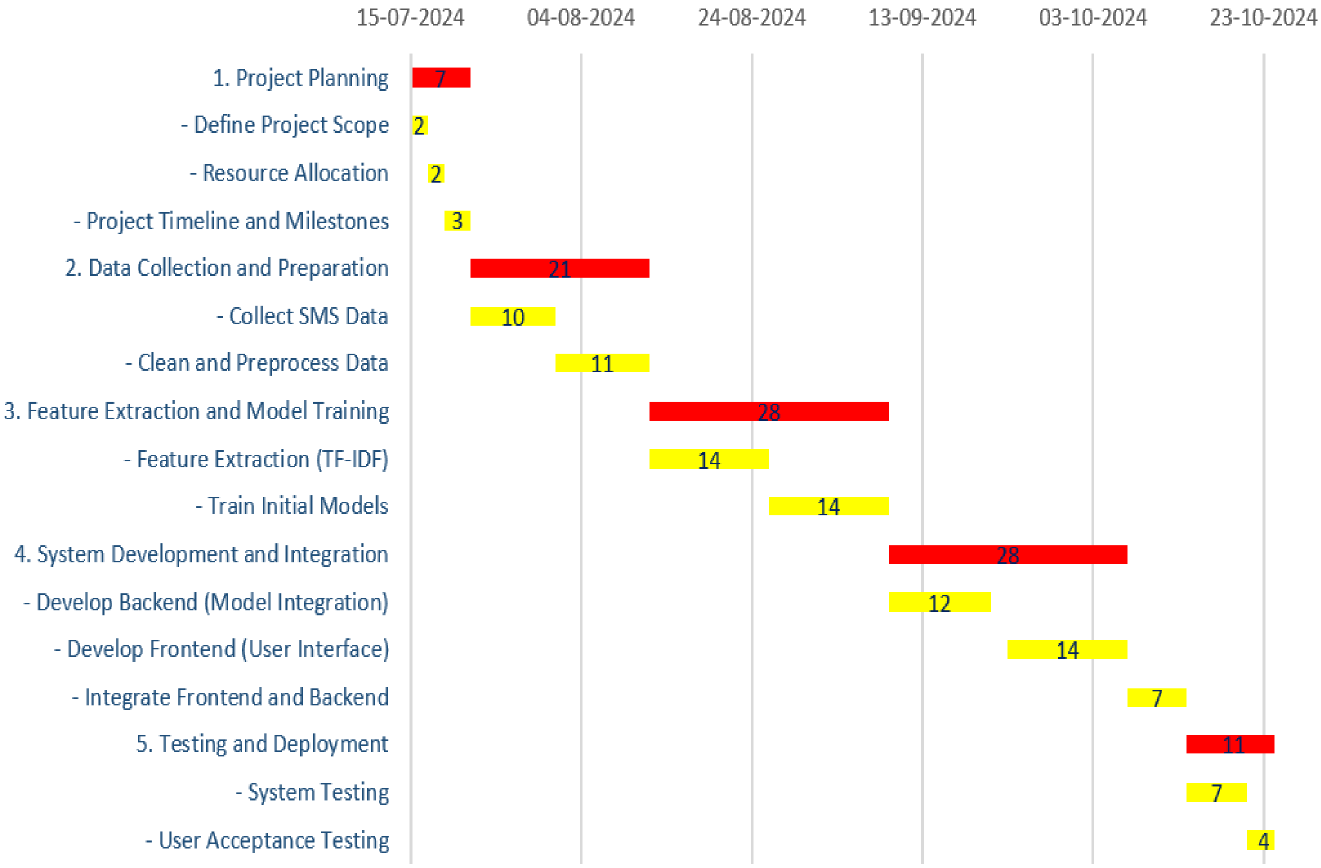
1. **Personal Communication Protection:**
   - **Use Case:** SMS spam detection systems can be integrated into mobile phones and messaging apps to filter out unwanted or harmful messages, protecting users from spam, scams, and phishing attempts.
   - **Benefit:** Enhances user experience and security by ensuring that users only receive legitimate and relevant messages.
2. **Telecommunications Industry:**
   - **Use Case:** Telecom companies can deploy spam detection systems to protect their networks from spam messages, thereby improving service quality and customer satisfaction.
   - **Benefit:** Reduces the operational burden on customer support and minimizes the negative impact of spam on network performance.

3. **Financial Sector:**
    - **Use Case:** Financial institutions can use spam detection to filter out fraudulent SMS related to banking and financial services, such as phishing attempts and fake transaction alerts.
    - **Benefit:** Enhances the security of financial transactions and protects customers from financial fraud.
4. **Healthcare:**
    - **Use Case:** Hospitals and healthcare providers can use spam detection systems to ensure that patients and healthcare professionals receive only relevant health-related information.
    - **Benefit:** Improves communication efficiency and helps in maintaining patient privacy.

Future Scope:

1. Advanced Machine Learning Models
2. Real-Time Analysis and Response
3. Contextual and Personalized Filtering
4. Integration with Artificial Intelligence
5. Privacy and Security Enhancements
6. Cross-Platform Integration

1.6 Project Duration (GANTT CHART):



1.7 References

- 1. *Research Papers*: -

https://archive.ics.uci.edu/ml/datasets/sms+spam+collection

https://ieeexplore.ieee.org/document/XXXXXX

https://arxiv.org/abs/XXXXXX

- 2. *Books*:

https://www.oreilly.com/library/view/introduction-to- machine/9781449369880/

https://www.manning.com/books/deep-learning-with-python