IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 5, May 2024

SMS Spam Detection using Machine Learning

Phanirama Prasad[1], G Shivaraj[2], J M Renuka[3], Janardan Kulkarni[4], H Aishwarya[5]
Professor, Department of Computer Science & Engineering[1]
Students, Department of Computer Science & Engineering[2,3,4,5]
Ballari Institute of Technology and Management, Ballari, India

Abstract: *Emails are a crucial instrument for achieving the interconnectedness of machines and people worldwide in the global age we live in, when the world is becoming smaller and smaller and everyone is connected to one another. Since the development of communication technology and the exponential increase in internet usage in recent years, emails have become an integral part of our everyday lives. It has also, regrettably, led to the emergence of numerous online scams that employ phishing tactics to trick people into disclosing their personal information. These scams can result in serious financial theft, identity theft, character assassination, and other malicious activities that could have dire repercussions for internet users. Addressing the problem of spam and phishing emails becomes crucial as a result of these problems. Therefore, this project is about detecting phishing and spam e-mails using Natural Language Processing (NLP) techniques. Let's explore methods to identify malicious emails using NLP. This assists in detecting and classifying potentially harmful and spam emails, ultimately preventing harm to user data. This project involves exploration of how different methods are implemented to detect the spam e-mails and the work being done to improve upon the current and possible future scenarios because as the scammers keep evolving and we need to develop and rise up accordingly*

Keywords: Spam detection, NLP, Naïve bayes, Spam SMS, Scams

I. INTRODUCTION

Small Message once-over (SMS) is the majority often and extensively used message medium. The term "SMS" is used intended for together the user activity and all types of small text messaging inside a lot of parts of the world. It has developed into a medium of announcement as well as endorsement of products, banking updates, agricultural information, flight updates and internet offers. SMS works in direct marketing recognized as SMS marketing, from time-to-time SMS marketing is a matter of trouble to users. These kinds of SMS are called spam SMS. Spam is one or additional unwanted messages, which is unwanted to the users, sent or posted as part of a better collection of messages, every one having considerably matching content. The purposes of SMS spam be announcement and marketing of a variety of products, sending political issues, dispersal unsuitable adult content and internet offers, so asto is why spam SMS flooding has become a serious problem. All over world SMS spamming gain reputation over additional spamming approaches like electronic mail with SMS owing to the rising popularity of SMS Communication

II. LITERATURE REVIEW

SL No	Author	Proposed Methodology	Observations
[1]	Pumrapee Poomka*etal* (2019)	Long - Short Term Memory, Gated Recurrent Unit	This Machine Learning model proposes a method for detecting SMS spam using Natural Language Processing and Deep Learning, achieving an accuracy of 98.18%. It particularly excels in detecting spam messages with 90.96% accuracy, with only a 0.74% misclassification rate for normal messages.
[2]	Haslina Md Sarkan(2019)[2]	Term frequency-inverse document frequency (TF-IDF),Random	This study uses TF-IDF and the Random Forest Algorithm on SMS spam data, where Random Forest achieves an impressive 97.50% accuracy.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 5, May 2024

		Forest Algorithm	
[3]	Shah Nazir, *etal.* (2020)[3]	Logistic Regression, Decision Tree, K Nearest Neighbour	Proposes a machine learning-based spam detection method for mobile messages. Testing on the SMS spam collection dataset demonstrates high accuracy, reaching 99%. Our method also outperforms existing state-of-the-art methods.
[4]	Sridevi Gadde *et al.*(2021)[4]	LSTM and Data set from UCI	Machine learning and deep learning techniques to detect SMS spam, achieving a high accuracy of 98.5% with our LSTM model, using a dataset from UCI and implementing our approach in Python
[5]	*Houshmand Shirani-Mehr,.* (2019)[5]	SVM , Naïve Bayes, KNN	The rise of mobile phones has fueled a lucrative SMS industry, accompanied by a surge in spam messages. This project tackles spam detection using real SMS data, applying diverse machine learning techniques to find the most effective algorithm, resulting in a significant reduction in error rates.
[6]	Suparna Das Gupta *etal.* (2021)[6]	TF-IDF, NaiveBayes.	Form the above discussion and experimentation authors have concluded that machine learning algorithm can play a vital role in identifying spam sms. The accuracy obtained in this work is more than 95% in the both cases

III. IMPLEMENTATION MODULES

Modules:

1. Data Acquisition
2. Data Preprocessing
3. Machine Learning.
4. Flask.
5. User Login and User Signup.
6. Sentiment Analysis

Module Description:

- Data Acquisition: pandas is used for loading and manipulating SMS data in tabular formats (CSV, Excel) . NumPy is used for numerical computations and array operations , often used for data cleaning and feature engineering.
- Data Preprocessing: NLTK (Python): A comprehensive toolkit for natural language processing tasks, including text normalization (lowercase, stemming, lemmatization), tokenization (splitting text into words), stop word removal (eliminating common words like "the," "a").

Machine Learning Models:

- Scikit-learn (Python): Provides awide range of machine learning models for classification tasks, including:
- Naive Bayes: A probabilistic classifier well-suited for text classification due to its simplicity and efficiency.

Deployment:

- Flask (Python): Web frameworks for creating a user interface where users can input messages and receive spam/ham classifications.
- User Login: The user can login. The user enters their username & password and logs into the webpage.
- User sign-up: User can create an account with username and required password.
- Sentiment Analysis: This module analyses the sentiment of the not spam messages and returns if the sender is happy or sad.

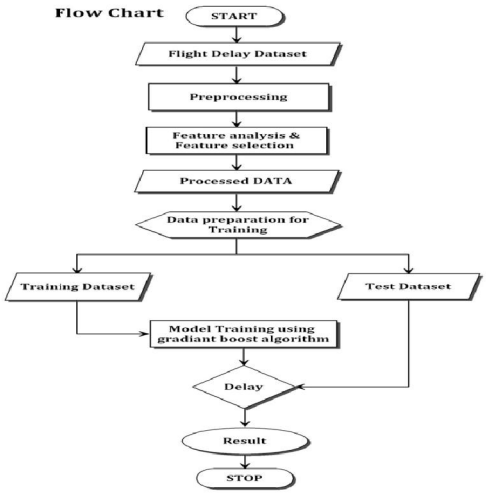
IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 5, May 2024

IV. COMMON COMPONENTS

- Actors: The users that interact with a system. An actor can be a person, an organization, or an outside system that interacts with your application or system. They must be external objects that produce or consume data.
- System: A specific sequence of actions and interactions between actors and the system. A system may also be referred to as a scenario.
- Goals: The end result of most use cases. A successful diagram should describe the activities and variants used to reach the goal.

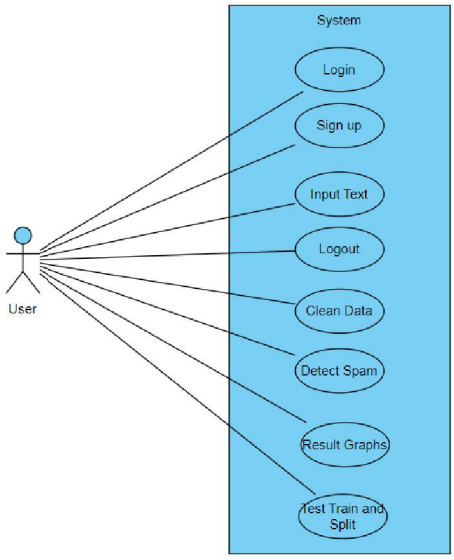
V. DIAGRAMS

Flowchart:



A flowchart is a type of diagram that represents an algorithm, workflow or process. Flowchart can also be define as a diagrammatic representation of an algorithm (step by step approach to solve a task). The flowchart shows the steps as boxes of various kinds, and their order by connecting the boxes with arrows.

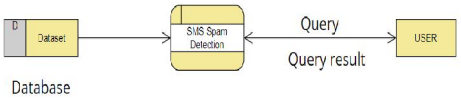
Use-Case Diagram:



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 5, May 2024

An effective use case diagram can help your team discuss and represent:
Scenarios in which your system or application interacts with people, organizations, or external systems
Goals that your system or application helps those entities (known as actors) achieve the scope of system

Data-Flow Diagram:



A Data Flow Diagram (DFD) is a traditional visual representation of the information flows Within a system. A neat and clear DFD can depict the right amount of the system requirement graphically. It can be manual, automated, or a combination of both. It shows how data enters and leaves the system, what changes the information, and where data is stored. The objective of a DFD is to show the scope and boundaries of a system as a whole.

VI. ACKNOWLEDGMENT

The authors gratefully acknowledge the support from the corresponding guide and concerned faculty of the CSE department, Ballari Institute of Technology and Management, Ballari.

VII. CONCLUSION

The SMS spam detection project implemented comprehensive feature extraction to distinguish between spam and non-spam messages effectively. Features such as message length, number of digits, special characters, frequency of common words, grammar and spelling errors, URLs or phone numbers, contextual coherence, call-to-action phrases, and time of day were utilized. This approach enabled accurate spam detection while routing non-spam messages for sentiment analysis, providing valuable insights into legitimate communications. The dual method enhanced both security and content analysis, presenting a robust framework for handling SMS data with improved accuracy and insight.

REFERENCES

[1] H. Shirani, "SMS Spam Detection using Machine Learning approach," *Stanford,* p. 4, 2021.

[2] W. P. N. K. a. K. K. Pumrapee Poomka, " SMS Spam Detection Based on Long Short-Term Memory," *International Journal of Future Computer and Communication,,* p. 6, 2019.

[3] N. F. M. A. S. C. H. M. Nilam Nur Amir Sjarif*, "SMS Spam Detection using Term Frequency Inverse Document Frequency," *ScienceDirect,* p. 7, 2019.

[4] 1. N. H. U. K. a. A. U. H. Luo GuangJun, "Spam Detection Approach for Secure Mobile Message using MAchine learning algorithms," *Hindawi Security and Communication Networks,* p. 6, 2020.

[5] S. S. S. k. D. Superna Das Gupta, "SMS Spam Detection using Machine Learning," *Journal of Physics: Conference Series,* p. 9, 2017.