INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 9, Issue 5 - V9I5-1190)
Available online at: https://www.ijariit.com

SMS spam detection in Machine Learning using Natural Language Processing

Thanniru Lakshman
lakshmantanniru30@gmail.com
Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh

Singarapu Sanjay Kumar
sanjaykumarsingarapu@gmail.com
Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh

Ulligaddala Satish Kumar
sateeshkumarusk2105@gmail.com
Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh

Yenikepalli Sri Sekhar
sreesekhar042@gmail.com
Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh

Yellamati Suresh
sureshyallamati@gmail.com
Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh

ABSTRACT

This paper presents identification of Spam and ham messages using supervised machine learning algorithms Random Forest Classifier, Logistic Regression algorithms and analyzes how each filter performs when detecting Ham and Spam. Spam message is a big issue in mobile communication to reduce this effective spam detection techniques should be building Preprocessing is done using NLTK library with various Stemming Algorithms, Word clouds used and tokenizing also performed. The data set divided into two categories for training and testing the classifiers. the results of this demonstrated that performance of Random Forest is better than Logistic Regression. Random forest achieved a better accuracy of 97%.

Keywords: *SMS, Machine Learning, Random Forest, Logistic Regression, NLP, Spam, Ham.*

1. INTRODUCTION

As the Internet continues to evolve, people are using free online services more and more. It is common for people to exchange personal information across several websites, but that information is also shared with various companies that send spam to market their services to individuals. For automated channels, SMS spam poses additional challenges. The amount of content that can be used to determine if a message is spam or a ham is reduced when SMS texts are frequently limited to 160 characters [1].

The character limit on SMS messages gives a better platform for detecting the undesired messages through search algorithms than it does for email spam communications. This is the main distinction between filtering SMS and email messages[2].Spam channels nowadays are made up of different modules that analyze different aspects of messages (such as the sender address, header, content, and so forth)[4].

One source of this flaw, which jeopardizes the security of sending mobile messages, is spam. Spam is one of the biggest problems with instant messaging and email systems. Trash mail is an email or communication that is spam. When given to users without their previous consent, spam emails and messages are unsolicited by the reciepents[3].

There are several ways to define spam and distinguish it from legitimate emails. "An undesired electronic mail" is the shortest definition of spam. One of the main issues with the implementation of spam filtering is that legitimate emails could be mistakenly classified as spam or over looked[5].

In this research paper, we proposed a spam detection approach that classifies spam and ham communications using machine learning methods like Random Forest and Logistic Regression. A dataset of SMS spam was employed for training and testing in the current work. Thirty percent of the data set was utilized for testing prediction models, while the remaining seventy percent was used for training. Performance evaluation measures like recall, F1-score, accuracy, and precision were taken into account when assessing the suggested study.

2. DATA SET

You may find the data set utilized in this work on Kaggle, a machine learning repository. The "SMS spam collection dataset" consists of 5572 occurrences and two attributes, v1 and v2. It is decided whether or not to classify the input messages, or v2. The expected label v1 has two classes: 1 spam and 0 non spam. The data contains 4900 non-spam samples and 672 spam samples repository. The dataset was divided into two halves, 30% and 70%, for the purposes of training and validating the predictive model.

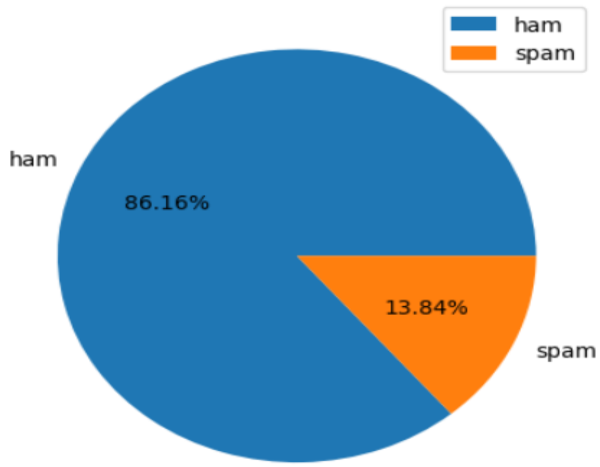

Fig-1: Data set Visualization

3. PREPROCESSING USING NLTK

Python's NLTK library is a well-known natural language processing library. It offers a range of resources and methods, Data in humans languages can be manipulated and manipulated using a range of resources and methods, such as tokenization, parsing, tagging, stemming, and more.

Text categorization, sentiment analysis, and part-of-speech tagging are among the tasks for which NLTK offers datasets, methods, and pre-trained models.

Tokenizing

The practice of dividing a text into discrete words, or tokens, is known as tokenization. Depending on the use case, these tokens can be words, phrases, or even characters. Usually the first stage in natural language processing (NLP), tokenization is essential for many text analysis tasks, including information retrieval and text categorization. Tokenizing a text can be accomplished using a variety of methods, including regular expressions and white-space tokenization.

Stemming

The practice of reducing words to their root or basic form is called stemming. This is accomplished by stripping words of their suffixes, which can aid in clustering related words together. One common method of text normalization that is used to increase task accuracy in text analysis is stemming. One of the most commonly used stemming algorithms is the Porter stemmer.

Word cloud

Words are shown in varying sizes according to how frequently they occur in the text in a word cloud, which is a visual representation of text data. A word appears larger in the cloud the more frequently it occurs. Data exploration and visualization can be facilitated by word clouds, which provide a brief summary of the terms that occur most frequently on a document.

4. METHODOLOGY

The methodology used in the proposed research,
It consists supervised machine learning algorithms which can be used best classifiers for prediction of target value

Data Collection

Gather a suitable dataset for the research. In this case, the kaggle dataset is chosen, which contains a 5574 spam and ham messages.

© 2023, www.IJARIIT.com All Rights Reserved

Data Preprocessing
Clean and preprocess the dataset using NLTK. This may include data normalization, handling missing values, and encoding categorical variables. Data splitting into training, validation, and testing sets is also done.

Model Training

Design the architecture of the hybrid deep learning model, which combines LSTM and GRU layers. Specify hyper parameters, such as the number of layers, units, learning rate, and dropout rates. Train the model on the training data, monitoring its performance on the validation set. Gets appropriate loss functions and optimization techniques.

Evaluation
Assess the model's performance using various evaluation metrics, such as accuracy, precision, recall, F1 score. The research might also employ confusion matrices to visualize the results.

UI Designing
In this using stream lit open-source library a simple web UI will be developed for users to check the SMS in convenient and easy way it will be accessed using a url.

Comparison
Compare the performance of Random Forest and Logistic Regression with existing methods or models. This step aims to demonstrate the potential improvements achieved by the proposed approach.

5. PROPOSED ARCHITECTURE OF SMS SPAM DETECTION SYSTEM:

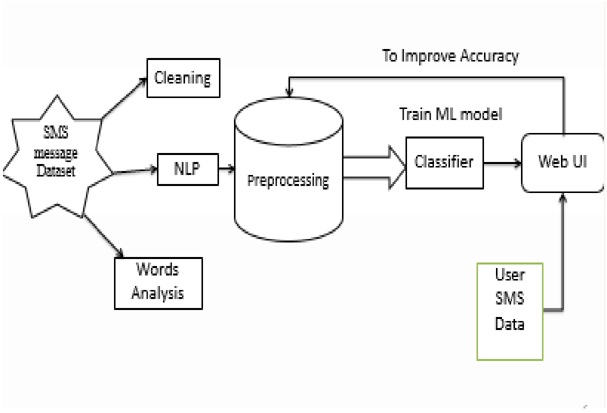

Fig-1: Architecture of Proposed System

6. ADVANTAGES OF THE SYSTEM

Improved Accuracy
The Effective spam detection models can be developed if the models have better accuracy rate and which helps the detector to detect correctly.

Convenient way to check
There is no convenient way to check SMS is spam or not so in this a web UI allows user to input the SMS and check whether it is spam or not the UI was developed using Streamlit.

7. OBJECTIVES OF SYSTEM

Perform Stemming in a Better way
In Pre-processing stage using the NLTK module of python stemming should be performed in a better way using Porter Stemmer and Lancaster Stemmer so that malicious words can be found easily.

Develop a Web UI
The goal of the system is to develop a UI that allows user to check the SMS. This is developed using Streamlit an open source frame work for hosting ML web applications.

8. ALGORITHM OF PROPOSED SYSTEM

Input: SMS
Step 1 Preprocess the dataset.
Step 2 Divide the data into sets for testing and training.

© 2023, www.IJARIIT.com All Rights Reserved

Step 3 Compile the model.
Step 4 Train the model.
Step 5 Evaluate the model on the test data.
Step 6 Output the trained model and evaluation results.
Output: Spam or Ham.

9. MODEL AND RESULT

Specific Model

Alphabets for Classification:
For the categorization of spam and ham, the machine learning methods mentioned below were taken into account.
Statistical Inference
For classification situations where the goal is to predict the likelihood that an instance will belong to a given class, supervised machine learning techniques termed logistic regression are typically utilized.
It gives probabilistic values, which range from 0 to 1, rather than the precise values, which are 0 and 1. Possible answers include True or False, 0 or 1, Yes or No, etc.
The well-known machine learning algorithm Random Forest is one of the supervised learning techniques. It may be applied to classification and regression-based machine learning issues. The idea of ensemble forms its foundation.
The construction of the random forest from the combination of N decision trees and the prediction of each tree produced in the first phase are the two steps of the Random Forest's operation.
Step 1: From the training set, choose K data points at random.
Step 2: Create the decision trees linked to the chosen data points (subsets)
Step 3: Select the number N for the decision trees you wish to construct.
Step 4: Carry out Steps 1 and 2.
Step 5: Locate each decision tree's predictions for the new data points, then allocate them to the group receiving the majority of votes.

Result

Metrics for Model Assessment Accuracy:

The proportion of the model's overall predictions that come true is known as accuracy. (Total Predictions / (True Positives + True Negatives)) is the formula.
Precision: Measures the model's ability to forecast successful outcomes. This is the ratio of verified affirmative forecasts to all positive forecasts.
The formula is True Positives / (True Positives + False Positives).
True Positive Rate of Recall, also known as Sensitivity, measures the model's ability to find every relevant incident in the dataset. It represents the ratio of true positives to all true positives. The formula is True Positives / (True Positives + False Negatives).
F1-Result: The F1-Score, which achieves a balance between the two measurements, is the harmonic mean of recall and accuracy. 2 * (Recall * Precision) / (Recall) is the formula.

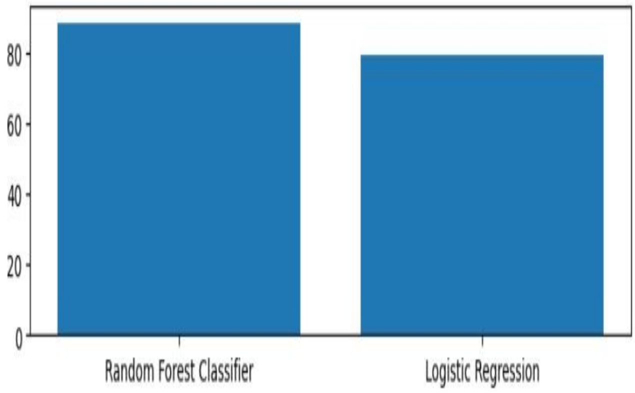

Fig-2: Performance Comparison

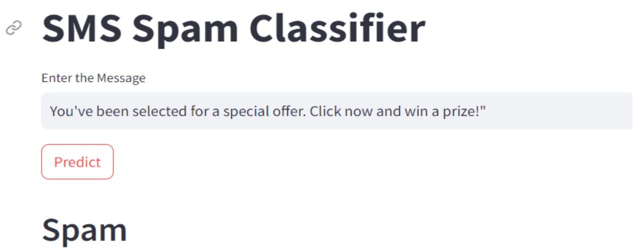

Fig-3: Web UI

© 2023, www.IJARIIT.com All Rights Reserved

10. CONCLUSION

For the purpose of protecting email and message exchanges, spam detection is crucial. There are a number of techniques available for detecting spam, but the precise identification of spam remains a significant challenge. It is, however, not possible to identify spam effectively and precisely using these techniques. We have suggested a technique for spam identification utilizing machine learning predictive It is, however, not possible to identify spam effectively and precisely using these techniques. In order to address this problem. The technique is used in order to identify spam. The experimental results demonstrate how well the suggested strategy can identify spam.

11. REFERENCES

[1]. Pavas Navaney;Gaurav Dubey;Ajay Rana 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence).
[2]. A. Alzahrani and D. B. Rawat, "Comparative Study of Machine Learning Algorithms for SMS Spam Detection," 2019 SoutheastCon, Huntsville, AL, USA, 2019, pp. 1-6.
[3]. Luo GuangJun , 1 Shah Nazir,2 Habib Ullah Khan,3 and Amin Ul Haq4.
[4]. M.Rubin Julis, S.Alagesan.
[5]. Ali Shafigh Aski a, * , Navid Khalilzadeh Sourati.
[6]. Suparna Das Gupta et al 2021 J. Phys.: Conf. Ser. 1797 012017.
[7]. Tiago A. Almeida School of Electrical and Computer Engineering University of Campinas Campinas, Sao Paulo, Brazil.
[8]. Mehul Gupta, Aditya Bakliwal, Shubhangi Agarwal & Pulkit Mehndiratta.
[9]. HoushmandShirani-Mehr, hshirani@stanford.edu
[10]. José María Gómez Hidalgo Universidad Europea de Madrid Villaviciosa de Odón 28670 Madrid, SPAIN 34 91 211 5670.

© 2023, www.IJARIIT.com All Rights Reserved