ISSN No:-2456-2165

Spam Message Detection Using Logistic Regression

NIKHIL KUDUPUDI[1], SHILPA NAIR[2]
U.G. Student, School of Engineering, Ajeenkya DY Patil University Pune, India -412105[1]
U.G. Student, School of Engineering, Ajeenkya DY Patil University Pune, India -412105[2]

Abstract:- The use of the internet is increasing day by day, and the spammers who consistently try to spam people by sending fraud mails and SMS. Mails and SMS are one of the most important and most used means of communication, because of which 2.4 billion messages are sent every one second. With the rise of such exchange of emails and messages, some find it an opportunity to fill other's inbox with preposterous messages that reduce internet speed and plunders our personal data. However, due to recent advancements in technology, it is possible to find solutions to all such problems easily. With the help of Natural Language Processing and Machine Learning, we can quickly detect spam messages. One of the crucial aspects of research in the world of machine learning applications is "NLP". In this paper, we have proposed a model where emails would be classified into the categories of Spam or Ham.

Keywords:- *Spam-Detector, Natural Language Processing, Logistic Regression.*

I. INTRODUCTION

Technology is advancing at a high rate. A few decades back, the only source of communication was the letters, which turned into telegrams, and in recent times it is in various forms like emails, phone calls, SMS, etc. An average person sends 72 messages per day, as texting is also the most common cell phone activity. Almost 300 billion emails are exchanged per day, and half of them are spam emails. 'Spam Mail' is basically undesired and unwanted emails that are sent to many of recipients that is just filling up all the inboxes. Most of these messages are product buying links, which would consume our personal data or could be some links and attachments. Sometimes carelessness from some users can cause significant damage to their personal data. Spam mails not only fill your inbox with junk mails but also cause email traffic. Spam messages accounted for 45.1% of email traffic in March 2021. In short, such mails can be frustrating and dangerous at the same time.

Inboxes are 85% filled with Spam mails and due to which the valuable and important emails are ignored. Many researchers are developing various techniques to find the solution for such problems and secure to communication. Since the unsolicited emails are termed 'Spam', important and valuable ones are termed 'Ham'.

There are many techniques developed to classify such spam and ham mails. One such technique is by using Natural language Processing and Machine Learning. With the help of Text classification methods like stemming, lemmatization, vectorization, etc., it is possible to classify the mails and train the model, which will be able to detect unwanted mails.

In this study, we have come up with our model that would classify emails and messages into either spam or ham. The evaluation metrics for performance such as accuracy were considered evaluating the proposed study. The results obtained from experiments confirmed that the proposed research achieved high accuracy.

II. LITERATURE SURVEY

In this paper [1], (Omay, 2010)the author mentioned the history and explained the concept of logistic Regression. He also explained types of logistic Regression like Binary Logistic Regression, Multinomial Logistic Regression, and Ordinal Logistic Regression; however, he gave detailed information on binary logistic Regression. The primary purpose of this paper is to assess the combination of independent variable's influence on dependent variables. For this, the author conducted a study on 200 students from Ankara University, and the dependent/target variable was critical thinking. The author found that an increase of one unit in scientific thinking led directly to a 14.4 percent increase in critical thinking, and a rise of one unit in epistemological belief resulted in a 4.9 percent increase in high critical thinking.

In this paper [2], (Lei, 2018) author 'Liu Lei' showed how logistic Regression could be used quickly and efficiently to detect Breast Cancer. He applied a logistic regression model to the breast cancer dataset. The author got the most accurate results with an accuracy of 96.5% when 'Maximum Texture' and 'Maximum Perimeter' were chosen as input to the model. In contrast, he got an accuracy of 90.48% when he took 'Mean Texture' and 'Mean Radius' as input to the model. Therefore, choosing a better feature combination will give more accurate results.

In this paper [3], (Radulescu, M.Dinsoreanu, & R.Potolea, 2014) the main goal is to detect spam comments. This was achieved by considering unclear comments with increased punctuation marks, new lines stop words, non-ASCII characters, new lines, capital letters, and offensive words and converting them into vectors to classify them into spam or non-spam comments. Next, they added word duplication ratio as spam comments tend to have repeated words and stop words ratio, which is the count of stop words divided by the total count of words in the comment. This increased the accuracy of classification. Finally, they added

post-comment similarity and topic similarity to remove comments unrelated to specific context. The authors also showed decision tree classifier works better with their spam detection model.

The authors of this paper [4] (Qaiser, Shahzad, & Ali, 2018) authors explained what is Term Frequency Inverse Document Frequency, how does TI-IDF works. They also discussed strengths and weakness of TI-IDF and how to overcome them. First, they collected data from different domains and removed stop words from data, then they applied TI-IDF on the processed data and displayed the results. The displayed results showed keywords and their TF-IDF value of different domains. Top keywords from '.biz', '.com', '.edu' and '.org' domain were parts, presidential, years and Marketing respectively.

The Authors[5] (Sjarif, Nila, & Amir, 2019) of this paper used Term frequency Inverse Document Frequency and Random Forest to detect spam messages. The data was collected from UCI Machine Learning Repository. Before applying the TI-IDF, they did some preprocessing like removing stop words as these messages contain special symbols, pronouns, and prepositions, which do not help in spam identification. After applying TI-IDF, the authors used multiple classification algorithms and found that Random Forest gave better Accuracy, Precision and F-measure compared to other classification algorithms.

In this paper[6] (Pandey & Yadav, 2020), the author proposed a model where deep neural networks are exploited for detecting spam mails using Tensor Flow. This model uses a linguistic approach, demonstrating the advantage of automatically neural networks. This paper also surveyed various publicly available datasets and noted the basic structure of the model. They have also revealed plentiful of open research problems related to spam filters.

Spam filters' sole purpose is analyzing the incoming data into unwanted(Spam) or wanted(Ham). Many researchers have come up with various types of filters. [7] (Shankar, 2018)The Model proposed in this paper uses Natural Language processing and Naïve Bayes. This Bayesian Spam Filter is trained, and a database is maintained to store and track the spam and ham messages. The messages are split into tokens and messages can be analyzed once the token database is created aby the filter. The model also introduces a threshold counter that helps to maintain the spam filter efficiency.

Different Spam classification methods are used to classify data into groups.[8] (Emmanuel, Gbengadada, & Joseph, 2016) Some of such types include Random Decision Tree, probabilistic Method, Support Vector Machine, Artificial Neural networks, etc. These classification techniques have been shown in the literature to be useful for spam mail filtering when combined with a content-based filtering strategy that recognizes specific features (keywords frequently utilized in spam emails). The likelihood for each feature in the email is determined by the frequency with which these qualities appear in emails, which is then compared to a threshold value. Spam is defined as email messages that exceed a specified number of recipients.

The dataset was subjected to a series of experiments based on Natural Language Processing (NLP) principles such as label encoding, tokenization, stemming, stop word removal, and generating features before being subjected to an ensemble approach - voting classifier. [9] (Pragna & .RamaBai, 2019)All of the trials in the model correctly categorize the data set. The algorithms used in this study produced good accuracy results. However, Support Vector Classifier, with an accuracy of 98.49 percent, is the best predictor of spam messages among the numerous trials conducted. Other methods have a comparable level of precision, with a variance of around 3%.

III. MATERIALS AND METHODS

Dataset

The main aim of our project was to detect spam messages accurately. For this, we have taken the "SMS Spam Collection Dataset" from Kaggle.com. The dataset contains 5574 messages with tags either legitimate/Ham or spam. There are 5574 messages in the dataset, out of which 4825 legitimate messages and 747 spam messages. The text messages were compiled from various accessible research sources like 425 spam messages were manually selected from the "Grumbletext" website. 3375 messages were chosen at random from the National University of Singapore SMS Corpus (NSC). 450 ham messages were collected from "Caroline Tag's" Ph.D. Thesis and 1324 messages were gathered from "SMS Spam Corpus v.0.1 Big" out of which 1002 were spam messages, and 322 were legitimate messages

Packages

To work on our project, we have imported different packages. The "pandas" package was imported to read the dataset and to convert categorical data into indicator variables like 0 and 1 using "get_dummies" function. "nltk" package was used to get functions like "stopwords", "porterstemmer" and "tfidvectorizer" to work on the test processing. "re" package (Regular Expression Operations) was also used for processing text data. "sklearn" package was imported to get "train_test_split" and "logisticRegression" function. "train_test_split" function was used to split the data into training and testing dataset while, "logisticRegression" function was for prediction model. "seaborn" and "matplotlib" packages were imported to plot confusion matrix of our final result. "joblib" package was imported to save the model and use it again without repeating every process to make predictions.

IV. PROPOSED ANALYSIS APPROACH AND RESULTS

1. Data Preprocessing

Dataset had five columns, out of which three had no values, and all the columns did not have a proper name. We removed those three columns as they were of no use and gave the other two columns proper names. The column with

"spam/ham" categorical values were converted into numeric values as machine learning algorithms work well with numeric data. This was done using the "get_dummies" function of the "pandas" package. "get_dummies" function converts a given column into two or more new columns with values in 0's and 1's based on categorical values present in the old column.

Label Encoding refers to change the value in numeric form so it can be Machine-readable. After conversion, Machine Learning Algorithms can decide how to operate with those labels. This is an essential step for Supervised Machine Learning.

1.1.Stop Words Removal

For the Machine to understand, analyze and operate Natural Language Processing on the data, the texts (emails in the dataset) should be readable. Machines do not understand human language, so we need to preprocess the data to make our data understandable by machines. To be pristine, we need to clear out useless data from the dataset. Such useless words are known as 'Stopwords'.

Some common examples of stopwords are 'is', 'are', 'a', 'as', etc. Stopwords are commonly used in NLP and even in text mining to eliminate useless information.

1.2. Stemming

Stemming refers to reducing the word to its root word, mostly by removing the suffix. It shortens the vocabulary space, which in turn helps to speed up the process. It is one more method to normalize sentences for machines.

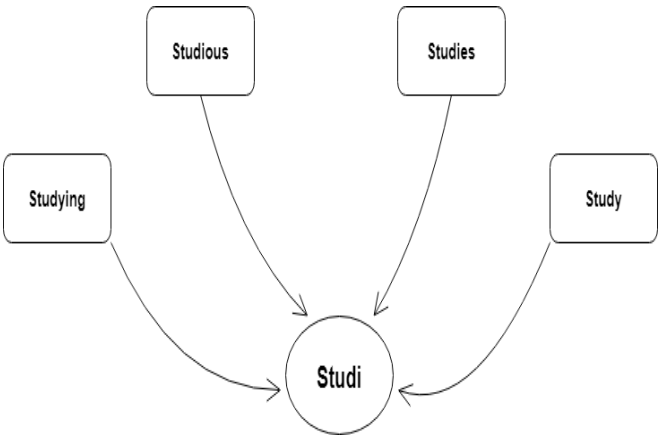

Fig 1: Stemming

1.3.TF-IDF

Now we need to convert text data into vectors as the machine learning algorithm works only on numeric data. For this, we will use Term Frequency-Inverse Document Frequency (TF-IDF).

Term frequency (TF) is used to measure the frequency of a word in a document. It is found by dividing the frequency of a word by the total number of words in that document. Let us suppose we want to find the TF of the word 'Health', which occurs 20 times in a document of 1000 words long. Therefore, the TF of Health in that document will be 0.02. Term Frequency alone will not give a good idea as some insignificant words might occur multiple times in a document but do not have much weightage. As Term Frequency treats every word equally, but every word has a different significance, Inverse Document Frequency (IDF) is used to tackle this issue. IDF helps reduce the weightage of terms that are very common in a set of documents. IDF is calculated by taking the log of the total number of documents divided by the number of documents in which that specific term is present. Let us suppose a word A1 is present in 10 documents out of 100 and word A2 is present in 60 documents out of 100 therefore, the IDF of A1 and A2 will be log(100/10) = 1 and log(100/60)=0.22 respectively. Term Frequency – Inverse Document Frequency is obtained by multiplying Term Frequency(TF) and Inverse Document Frequency(IDF).

2. Implementation of Algorithm

As cleaning and preprocessing of the dataset is done, we can use "train_test_split" function to divide the dataset into training and testing data. To implement the training data on the model and predict whether the text is spam or not, we need to import Logistic Regression algorithm from the "scikit-learn" library and performance metrics. In our project, we have used the Logistic Regression algorithm for classification purpose. Logistic Regression is an excellent predictive modeling algorithm that models probabilities for classification problems with two or more possible outcomes. Logistic Regression is similar to Linear Regression, where we get an S-shaped line to get output in either 0's or 1's instead of a straight line. To get this S shape curve, Logistic Regression uses the sigmoid function. The sigmoid function gives probabilities between 0 and 1. In our model, logistic Regression will give us whether the message is spam or not. Where if it's 1, it would be spam else it would be ham if the value is 0.

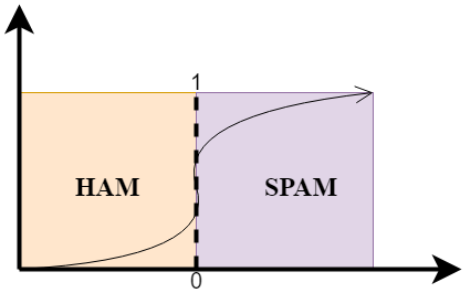

Fig 2: Logistic Regression

Let's Suppose you get the following message on your phone:
"CONGRATULATIONS!! Your email address has won a lottery sum of USD 2,500,000.00. To claim your prize, please contact our office via email claim3464@yahoo.com.hk or call +44 704 675 12446"
Here keywords are [lottery, prize, office, email]
The given weight vector is w = [0.3, 0.3, −0.1, −0.04]^T
The probability that the email is spam will be:

$x = [1,1,1,2]^T$
$w^T x = 0.3 * 1 + 0.3 * 1 - 0.1 * 1 - 0.04 * 2 = 0.42 > 0$

$$\Pr(y = 1|x) = \sigma(w^T x) = \frac{1}{1 + e^{-0.42}} = 0.603$$

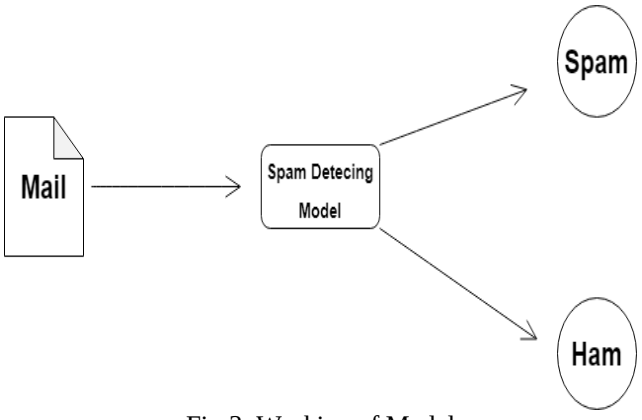

Fig 3: Working of Model

In order to test the accuracy of our model, an accuracy score metric is used. This metric compares the predicted results with the actual results. After running the code, we got 96% accuracy. We have also plotted a heat map to get an idea of how accurate our predicted values are compared to actual values.

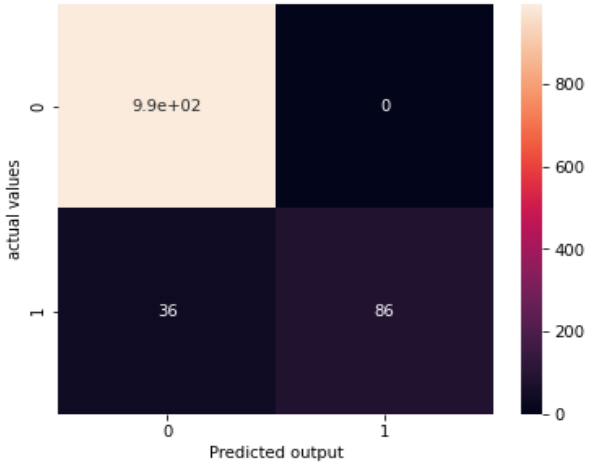

Fig 4: Heat Map

V. CONCLUSION AND FUTURE SCOPE

In this study, we looked into the general applications of spam detecting using NLP. We also reviewed the step-by-step process of the algorithm and how it classifies the mail into spam and Ham. The dataset we used in this paper was publicly available, and performance metrics was also implanted to check the model's accuracy. In the future, we can use neural network and deep learning models to predict a given message is spam or not. Deep learning works very well for natural language processing; however, it requires a vast amount of data to give accurate results and to outperform other traditional machine learning algorithms. Since Natural Language Processing is a relatively underdeveloped area for research, further enhancements can be made to the proposed system for spam detection and email filtering in the field of online security

REFERENCES

[1]. Sjarif, Nila, & Amir, N. (2019). SMS Spam Message Detection using Term Frequency-Inverse Document Frequency and Random Forest Algorithm. *Procedia Computer Science* , 509-515.
[2]. Shankar, S. (2018). Advanced Detection of Spam And Email Fitering using NLP algorithms. *IJARIT* .
[3]. Radulescu, C., M.Dinsoreanu, & R.Potolea. (2014). Identification of Spam Comments using Natural Language Processing Techniques. ICCP..
[4]. Qaiser, Shahzad, & Ali, R. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. *International Journal of Computer Applications* , 25-29.
[5]. Pragna, B., & .RamaBai, M. (2019). Spam Detecting using NLP Techniques. *IJRTE* .
[6]. Pandey, S., & Yadav, R. (2020). Email Spam Detection using Machine Learning and Deep Learning. *IJRASET* .
[7]. Omay, C. (2010). Logistic Regression: Concept and application. 2-3.
[8]. Lei, L. (2018). Research on Logistic Regression Algorithm of Breast Cancer Diagnose Data by Machine Learning. *ICRIS*, (pp. 3-4).
[9]. Emmanuel, Gbengadada, & Joseph. (2016). Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon*