IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SPAM AND FAKE SPAM MESSAGE DETECTION FRAMEWORK USING MACHINE LEARNING ALOGORITHM

[1]D.Pallavi,[2]G.Harshavardhan, [3]Dr.A.Althaf Ali

[1]PG Research Scholar, [2]PG Research Scholar, [3]Assistant Professor.

[1, 2,3]Department of Computer Applications,

[1, 2,3]Madanapalle Institute of Technology and Science, Madanapalle, Andhra Pradesh, India, 517350.

Abstract: The period of time now occurring social media has generated incomparable amounts of social data. It is a popular communication and also known as knowledge exchange media. Data could be of any type of text, numbers, figures or statistics that are accessed by a computer. Now-a-days, many people depends on contents available in social media in their decisions. Sharing of information with peoples has also attracted social spammers to exploit and spread spam messages to promote personal web logs, advertisements, promotions, phishing, scam, frauds and so on. In this project, we propose a machine learning-based spam detection system that uses a set of machine learning techniques such as Logistic Regression and Random Forest,and Decision Tree to identify whether a certain message in the dataset is spam or fake spam.

Key words **- Machine Learning, Data Preprocessing, Logistic Regression, DT, RF.**

I. INTRODUCTION

Spam is an unwanted, uninvited digital message that is sent in large quantities. Spam is frequently delivered via email, but it can also be delivered via text messages, phone calls, or social media like YouTube, Facebook etc. Spam produces a variety of issues, such as squandering the user's time, memory, and network bandwidth. Spam poses a financial risk to both organisations and users. Fake spam means “it contains original content”. Fake spam is also known as “Ham”. Spam and Fake spam Message Detection is detecting the spam or ham in the certain dataset. Tester here uploading the dataset into the Jupyter Anaconda Tool, and importing packages and Libraries. And applying the algorithms like Logistic Regression, Random Forest, Decision Tree and manually analysis of all algorithms and displaying the best algorithm based on the parameters like Accuracy, F1-Score, Precision, Recall. This System is Developed by using Python Language. Tester can easily know the which are spam message and Fake Spam message Transactions by all algorithm performance.

AN OVERVIEW OF MACHINE LEARNING

Machine learning: Terminology and Definition ML methodologies typically involve a learning process with the goal of learning to perform a task from "experience" (training data). In machine learning, data is made up of examples [1]. Individual examples are typically described by a set of attributes, also known as features or variables [15]. Nominal (enumeration), ordinal (e.g., A+ or B-), binary (i.e., 0 or 1), or numeric features are all possible (integer, real number, etc.)[5][13]. A performance metric that improves with experience is used to assess the ML model's performance in a specific task. Various statistical and mathematical models are used to calculate the performance of ML models and algorithms. The trained model can then be used to classify, predict, or cluster

new examples (testing data) based on the experience gained during the training process [7]. Figure 1 depicts a typical machine learning approach.

II LITERATURE REVIEW

Literature survey is the most important step in software development process. Before developing the tool it is necessary to determine the time factor, economy and company strength. Once these things are satisfied, ten next steps are to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior programmers, from book or from websites. Before building the system the above consideration are taken into account for developing the proposed system.

III PROPOSED METHODOLOGY:

The proposed system is Spam message and Fake Spam Message Detection Framework Using ML Algorithms. It is used for detecting the spam message and fake spam message in accurate manner and to overcome the limitations of Existing system. we use three algorithms like Logistic Regression, Random Forest and Decision Tree to detect the spam messages or Fake Spam Messages. The aim is predicting the accuracy of all three algorithms and displaying the best algorithm in accuracy of spam message and fake message. This project has been developed using Python language in Jupyter Anaconda Tool.

Advantages:

- This Proposed System overcomes all the disadvantages of existing system.
- The Main advantages are easy and quickly detecting the Spam and Fake Spam Messages of YouTube Comments by analysing the Three algorithms and compare those algorithms and choose best algorithm to show Spam and Fake Spam it is.
- It Provides More Accurate.
- It Reduces the Complexity to predict Spam and Fake Spam.
- It provides more Efficiency in Proposed System.

IV RESULTS & DISCUSSION

- Firstly, we have collected Spam related dataset [7].
- Later we will load the collected dataset to our working environment [6].
- Necessary pre-processing steps will be completed here before building our required model [12].
- Dividing the data into train and test splits [18].
- Perform building machine learning model in a flask environment using Python [15].
- The model has been built with Logistic regressions [20].
- Designed as, the system delivers the prediction results to the user depending on the inputs entered [9].

MODULES

3.1.1 Import Dataset

3.1.1.1 Data collection

I constructed a dataset for detecting spam on YouTube by gathering comments from various categories such as education, sports, and cooking videos, then combining them with another dataset that I acquired from the internet which is related to YouTube comments.

3.1.1.2 Dataset Description

The dataset Contains Two columns, they are Content and Label. Content contains YouTube Comments. Label contains 0 and 1.Ham or Fake Spam represented by 0 and 1 and Spam Message represented by 1.Dataset contains 999 rows.

3.1.1.3 Split the dataset

After you've imported the data set into Jupyter Anaconda. Train Test Split is a machine learning activity that measures the performance of machine learning algorithms when they are used to forecast new data that hasn't been used to train the model. To split the data into train test sets, use the train test split() method in the sklearn library.

1.Train Data:

A subset of the dataset used to train the machine learning model, the output of which we already know.

2.Test Data:

A subset of the dataset is used to test the machine learning model, and the model predicts using the test set.

3.1.2 Data Pre-Processing

Preprocessing refers to the transformations that are applied to the data before it is sent to the algorithm. Data preprocessing is a technique used to transform raw data into a clean dataset. That is, whenever data is collected from different sources, it is collected in a raw format that is not useful for analysis. In Data Pre-Processing, I used Tokenization.

Tokenization is the process of breaking a text into a series of meaningful parts. These parts are called tokens. For example, you can divide a section of text into words or sentences. Depending on the task, you can define your own conditions for splitting the input text into meaningful tokens.

Count Factorization techniques are used in this project to convert the string to float values. Count Vectorization comes under the data Pre-processing.

3.1.3Data Visualization

Data visualization is a graphical representation of information and data in a diagram or graph format (e.g., charts, graphs, maps). Data visualization tools provide an accessible way to find and understand trends, patterns of data, and outliers. In this project, Data visualization was done in various forms like Describe, head, tail, Slicing, Graph etc.,

3.1.4 Classification

Applying the algorithms on dataset individually.

3.1.4.1 Logistic Regression

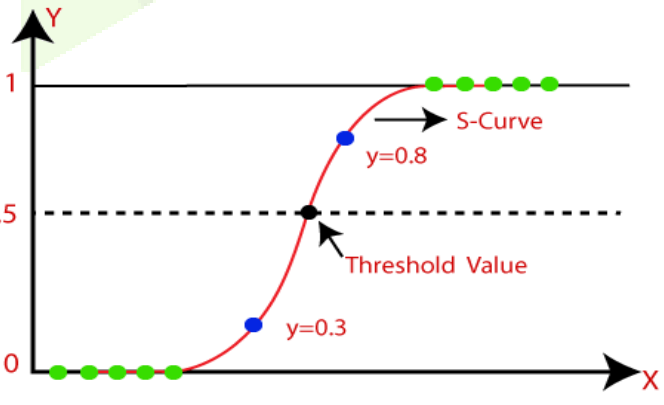
Logistic regression is a supervised classification technique that predicts the probabilities of binary dependent variables from a dataset's independent variables. This is a logistic regression that predicts the probability of an outcome with two values: 0 or 1, yes or 1. No, false or true. Logistic regression is similar to linear regression in that it produces a straight line, whereas linear regression produces a curve. The use of one or more predictors or independent variables is based on predictions. Logistic regression produces a logistic curve that represents a value between 0 and 1. Regression is a regression model in which the dependent variable is categorical and analyzes the relationships between several independent variables. There are many types of logistic regression models, including binary logistic models, multi-logistic models, and bionomical logistic models. The binary logistic regression model is used to estimate the probability of a binary response based on one or more predictors.

$$y = e^{(b0 + b1*x)} / (1 + e^{(b0 + b1*x)})$$

The above equation is the final equation for logistic regression.

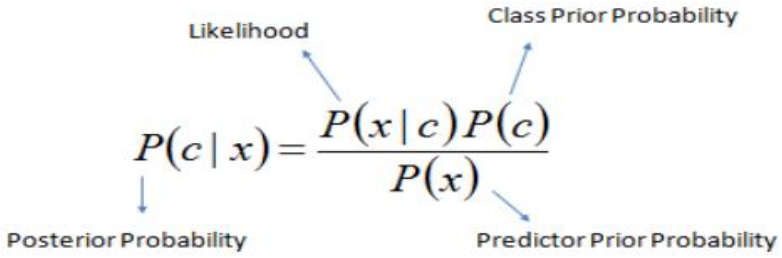
Where y is the predicted output, b0 is the bias or intercept term, and b1 is the coefficient of the single input value (x). Each column of input data is associated with a coefficient b (a constant real number) that needs to be learned from the training data.

The actual representation of the model stored in memory or file is the coefficients of the equation (beta values or b`s).



3.1.4.2 Random Forest

Random forest is a classification and regression technique. In summary, this is a collection of decision tree classifiers. Random forests are better than decision trees because they avoid the tendency to fill up the training set. To train each tree before building the decision tree, a subset of the training set is randomly sampled and each node is split from a random subset of the entire feature set into randomly selected features. Training in a random forest is very fast, even for large datasets with many features and data instances, because each tree is trained independently of the other trees. Random forest algorithms have been found to provide good estimates of generalization errors and tolerate overfitting.



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

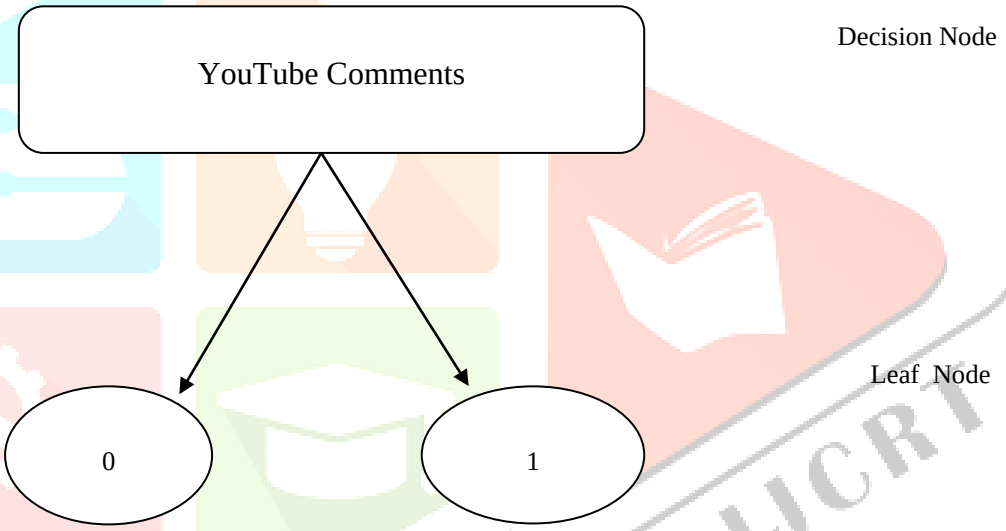
Random Forest is a natural approach to rank the relevance of variables in a regression or classification task.

3.1.4.3 Decision Tree

The decision tree is a graphical representation for providing solutions to classification and regression issues that are based on various conditions. It is tree-structured which helps in classification and regression problems but generally, it is utilized for classification problems. It has two nodes:

Decision Node: When a major node divides into several nodes and makes decisions then that node is known as a decision node.

Leaf Node: Leaf nodes are the outcome nodes and do not include more branches is called a Leaf node. The following Figure shows the Decision Tree of Spam Message and Fake Spam Message Detection :



Basic Structure of Decision Tree

Some Important terminologies of Decision Tree:

- Root Node: It is the node from which the whole decision tree starts and divides into two or more sets.
- Splitting: Division of nodes into various sub-nodes is known as splitting.
- Pruning: Removal of unwanted sub-nodes is known as pruning.
- Parent/Child Node: The major node of the Decision tree is also known as a parent node. The nodes which succeed the parent node is called child node.
- Branch/Sub Tree: Separate tree formed by the process of splitting is called a subtree.

3.1.5 Analysis

After applying all Three Algorithms on YouTube Spam message and Fake spam message dataset and analysis the Accuracy, Precision, Recall, Confusion Matrix of all algorithms and compare all metrices and choose the best accurately classified spam or fake spam of algorithms.

3.1.5.1 Confusion Matrix:

The Confusion Matrix is a table that displays the number of correct and incorrect predictions. It's used to figure out how well a classifier performs. True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) are all part of the Confusion Matrix (FN).

Actual	Predicted	
	Positive Class	Negative Class
Positive Class	True Positive(TP)	False Negative (FN)
Negative Class	False Positive (FP)	True Negative (TN)

a) True Positive (TP): The true value and predicted value are the same, i.e., the true value is positive and so is the forecasted value.

b) True Negative (TN): The true value is negative, hence the predictive value is negative as well.

b) False Positive (FP): The true value is different from the expected value, for example, the true value is negative while the predicted value is positive. Also referred as a Type 1 mistake.

d) False Negative (FN): When the true value is positive but the predicted value is negative, this is known as FN. Also referred as a Type 2 mistake. To evaluate the categorization model's performance, a few performance measures are used. A Few performance metrics are utilized to assess the performance of the classification model. The following are performance measurements:

- **Accuracy:** Accuracy helps calculate the total performance of the classifier.
- **Precision:** Precision calculates that out of all positive predictions, how many times fraud cases are there.
- **Recall:** Recall is the opposite of the Precision metric. The recall is the proportion of how often positive classes are correctly predicted.
- **F1 Score:** F1 Score is an overall measure of Precision and Recall and calculates the balance between them.
- **ROC Curve**: ROC is Receiver Operating Characteristics Curve. ROC Curve shows the graph of True Positive and False Positive Rate.

4.6 View Upload Dataset:

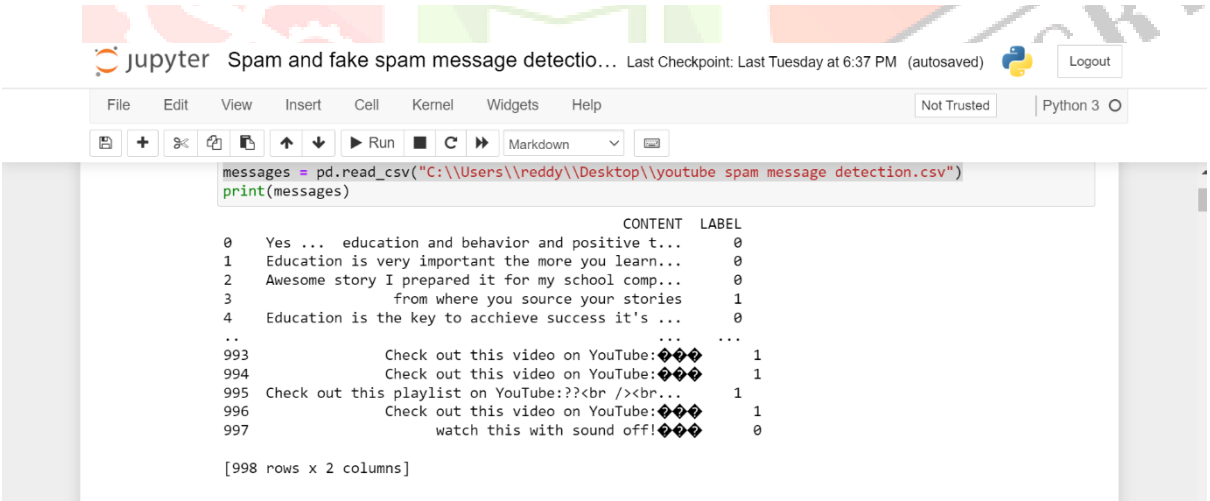


Figure 4.6: Dataset

4.7 Graphical representation of seasons:

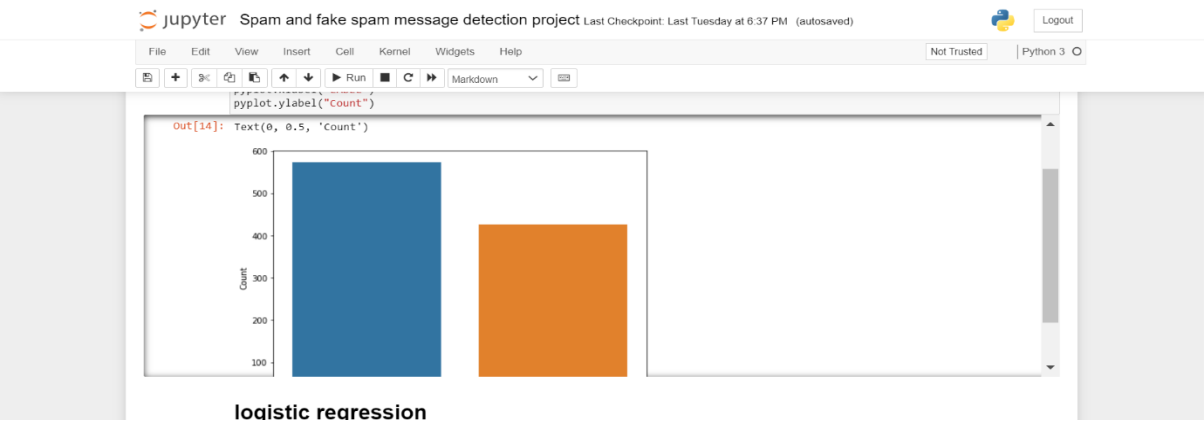


Figure 4.7: Diagram for Graphical (Bar Chart) representation of Seasons.

V CONCLUSION

Now-a-days, users are mostly used social media like Facebook, YouTube, etc., in huge manner. Spammers are sending spam messages easily through social media. Spam messages are allowed to loss of money and Collecting users' data without user permission. Spam messages are look like original messages but it is dangerous when it clicked link of spam by users. This project helps to detect the spam messages or fake Spam messages in the given data set. First we build the model using some machine learning algorithms such as logistic regression, Random Forest and decision tree, and detecting the spam or ham messages accurately.By analysis of all Parameters like Accuracy,Precision,Recall, F1-score to all algorithms.The Accuracy of Random Forest and Decision Tree is almost similar.Amoung all Algorithms random forest give best f1-score. So random forest shows best result for detecting spam and fake spam messages.

VI REFERENCES

1. J. Han, M. Kamber, and J. Pei, Data mining : concepts and techniques, 2nd ed. Elsevier Inc., 2006.
2. Nur'Ain Maulat Samsudin[1],Cik Feresa binti Mohd Foozy[2],Nabilah Alias[3],Palaniappan Shamala[4],Nur Fadzilah Othman[5],Wan Isni Sofiah Wan Din[6], Youtube spam detection framework using naïve bayes and logistic regression, "Indonesian Journal of Electrical Engineering and Computer Science", Vol. 14,No. 3, June 2019, pp. 1508~1517
3. https://towardsdatascience.com/nlp-spam-detection-in-sms-text-data-using-deep-learning-b8632db85cc8
4. https://github.com/sahilbhange/YouTube-comments-Spam-Detector
5. https://www.kaggle.com/code/sarthakniwate13/email-spam-classification-random-forest/notebook
6. Gomatham Sai Sravya, G Pradeepini, Vaddeswaram, Guntur, Mobile Sms Spam Filter Techniques Using Machine Learning Techniques, Mobile Sms Spam Filter Techniques Using Machine Learning Techniques," International Journal of Scientific & technology research", Volume 9, ISSUE 03, MARCH 2020.