IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 24, Issue 3, Ser. II (May. –June. 2022), PP 01-06
www.iosrjournals.org

Short Message Service Classifier Application using Naïve Bayes algorithm

Ghosh Madhumita*, Ravi Gor
*Research Scholar, Department of Mathematics, Gujarat University, Ahmedabad-380009
Department of Mathematics, Gujarat University, Ahmedabad-380009

Abstract
Short Message Service (SMS) is one of the quickest and most usable communication channels among all available channels. It is used for both commercial and personal purposes. Many messages are received in daily life and among these some are spam. To detect spam messages machine learning is a good technique. In machine learning, Supervised Learning technique is used to identify whether the message is spam or ham. During classification Naïve Bayes outperforms Decision Tree algorithm. Also, SMS classifier application is created using Naïve Bayes algorithm.
Keywords: *Supervised learning, Naïve Bayes, Decision Tree, Flask, SMS data*

Date of Submission: 29-05-2022 Date of Acceptance: 10-06-2022

I. Introduction

SMS is used to share information for both personal and business communication. SMS is better than all communication channels. Instant Messaging (IM) application like Telegram, Skype, WeChat, etc. offer lower price for communication, but the higher price of SMS communication reduces spam because the sender has to pay for sending the message. Spam messages are tricks to get personal details and money from the people by offering attractive false deals or malicious links.

Spam is type of unwanted and unsolicited message. These messages are not coming from another phone. They are generated from a computer and sent to mobile phone via an instant messaging account or email address. Spammers need few numbers of responses to justify the efforts, so they often send bulk messages to randomly select or automatically generate numbers.

In the past few years, the ratio of message scandals has been increased so spam detection is very important in mobile message communication. Different kind of machine learning techniques and algorithms are used to create automatic spam detection devise. These types of SMS classifier devises effectively classify the large number of data sets in an appropriate time frame with adequate accuracy. SMS classification tool identifies spam messages and prevents scam.

II. Literature Review

Radhakrishnan and Vaidhehi (2017) applied two significant algorithms: Naïve Bayes and J48 Decision Tree to classify emails as spam or ham. They calculate the weight score of text using TF-IDF. They also tested both algorithms with different feature size. From the tested results they found that J48 Decision Tree gives 96% accuracy in classifying emails as spam or ham with a minimum feature size of 400 attributes and classification time 0.06 seconds.

Pandey et al. (2018) applied Multinomial Naïve Bayes algorithm to classify the product in anonymous marketplaces. In this experiment they used the product catalogues database from Amazon, Flipkart, Snapdeal and Paytm. They classified the products into three classes namely toy car, game and computer. From this experiment they conclude that this algorithm gives 70% accurate results.

Ponmalar and Krishnaveni (2018) classified medicinal plant leaves using a Random Forest classifier. The researchers used 816 images of leaves from 30 different medicinal plant species. The leaf photos are pre-processed before being used. For each medicinal plant leaf, the Morphological shape features and Local vector Patterns are computed and saved as a training feature data set. Five leaf photos from each plant species are selected as test images from the training set. They discovered that the random forest classifier performed better, with 99% accuracy, based on the experimental data.

Chowdhury and Schoen (2020) classified the research paper abstract by using four machine learning method: Support Vector Machine (SVM), Naïve Bayes, K-Nearest Neighbor (KNN) and Decision Tree. They classified different publications into three fields: Science, Business and Social Science. They also calculated

accuracy, precision, recall and F1-score. From these assessments they conclude that SVM gives better result, KNN and Naïve Bayes perform comparatively well.

Chaudhary (2020) analysed different Supervised Machine Learning (ML) methods and found the optimum algorithm based on the data set, number of variables, and features. To identify the best algorithm, they used the Diabetes dataset. They employed eight independent variables, including pregnancy number, oral glucose tolerance test for 2 hours with ration as plasma glucose, and diastolic blood pressure, among others. Mean Absolute Error (MAE) is also calculated to verify model accuracy. According to the findings, Support Vector Machine (SVM) was discovered to be a very accurate and precise method. Naive Bayes and Random Forest classification algorithms were found to be more accurate after SVM.

Ahmed and Ahmed (2021) used Naïve Bayes, Logistic Regression (LR), Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) algorithms to classify online news articles. They classified the data into eight categories: Crime, Entertainment, World News, Politics, Sports, Business, Media and Tech. They compared the results of the four algorithms based on accuracy, precision, recall and F1-score. From the comparison they found that Naïve Bayes performs best while KNN performs worst.

Siddique et al. (2021) used Naïve Bayes, Convolutional Neural Network (CNN), Support Vector Machine (SVM) and Long Short-Term Memory (LSTM) to detect spam emails written in Urdu. They examined the comparative performance of models by calculating the accuracy, precision, recall, F-measure, ROC-AUC, and model loss. They conclude that deep learning model LSTM obtained better result with accuracy of 98.4%.

Gupta and Vanmathi (2021) predicted quality of wine using Decision Tree, Random Forest, Support Vector Machine, MP5 (Multiple Regression Model) and K-Nearest Neighbor algorithms. The Red Wine and White Wine datasets were utilised to train these models. There are 1599 red wine samples and 4898 white wine samples. Acidity, sugar content, chlorides, sulphur, alcohol, pH and density are taken as independent variable. They also calculated the accuracy of the models and discovered that the MP5 model outperformed the rest.

Sandra et al. (2021) used Artificial Neural Network (ANN), Naïve Bayes, Logistic Regression, Support Vector Machine and Decision Tree algorithms to predict the success of students' performance. The 11 research publications were chosen from 2753 articles in the IEEE Access and Science Direct databases that were published between 2019 and 2021. They classified the student success into two or three categories: pass/fail; or fail/pass/excellent and concluded that ANN performs better than others.

Pandey and Maurya (2022) developed a classification model to forecast a student's job prospects as an undergraduate. They proposed k-nearest neighbour, support vector machine, stochastic gradient descent, decision tree, logistic regression, and neural network as the six most common machine learning classification techniques. The input variable was student's grade point average in 10th, 12th, B.Tech/Diploma, communication skills, etc. Government Job, M.Tech/ME/MS, MBA, Others, and Private Job were the output variables. The accuracy of each algorithm is assessed, and it is determined that K - Nearest Neighbor outperforms the others.

Bhavsar and Gor (2022) predicted restaurant ratings with the help of Machine Learning Model. Information such as Restaurant id, Country, categories for dining, cost, currency, online delivery option, aggregate rating, rating, votes were provided to the Artificial Neural Network model. The ratings were classified in 5 different categories form poor to Excellent. Results of three different optimizers Adam, Adamax and Nadam were compared, where Nadam shows best accuracy.

III. Model Explanation

Naïve Bayes:

Naïve Bayes is a simplest probabilistic machine learning algorithm. It can handle both continuous and discrete data problems. Naïve Bayes classifier algorithm is widely used to solve variety of classification tasks. It is based on Bayes Thorem.

$$P(y|X) = \frac{P(X|y) * P(y)}{P(X)}$$

This equation solve the probability of y using input features X.

Where, X = input variables or dependent feature

y = output variable or class variable

$P(X|y) = P(x_1|y) * P(x_2|y) * ... * P(x_n|y)$

Therefore,

$$P(y|X) = \frac{P(x_1|y)P(x_2|y) \dots p(x_n|y) * P(y)}{P(x_1)P(x_2) \dots P(x_n)}$$

When solving for y, denominator P(X) remains constant which means that it can be removed from the equation

$$P(y|X) \propto P(X|y) * P(y)$$

$$P(y|X) \propto P(y) * \prod_{i=1}^{n} P(x_i|y)$$

Now, choose the y with the maximum probability

$$y = argmax_y[P(y) * \prod_{i=1}^{n} P(x_i|y)$$

$P(x_i|y)$ = conditional probability and $P(y)$ = class probability
Where, argmax is an operation that gives the maximum value of target function.

Decision Tree:

Decision Tree is a machine learning algorithm which can be used for regression and classification both. Decission Tree has two nodes: Decision Node and Leaf Node.

Decision node : Decision node is also known as root node, which are used to make any decision and the tree have two or more branches.

Leaf node : Leaf nodes are the final output node, and the tree do not contain any further branches.

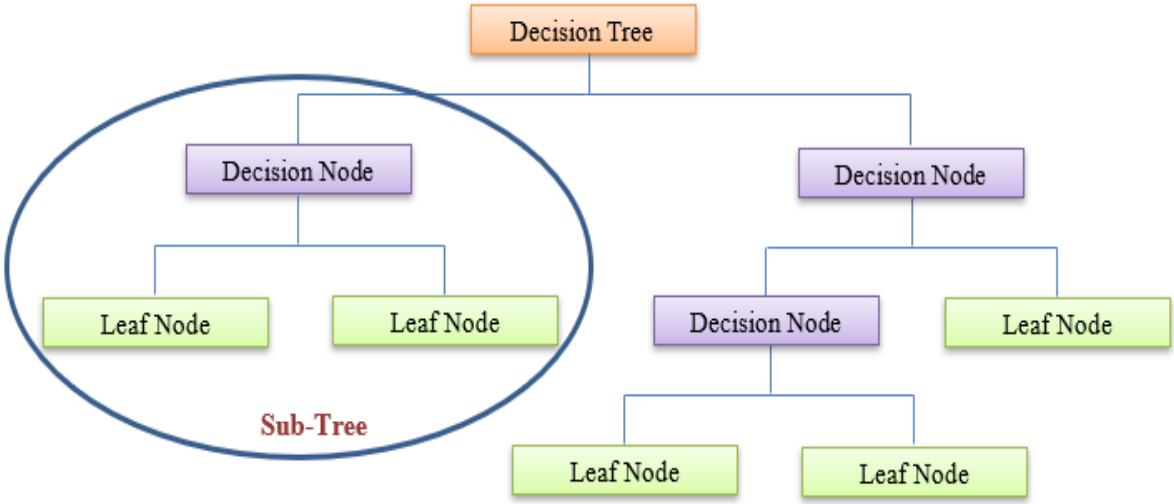

Figure:1 Structure of Decission Tree algorithm

Step-1 Start with the root node, which contains the complete data set.
Step-2 Find the best attribute using Attribute Selection Measure (ASM).
Two popular techniques for ASM are as follows

- Information Gain:

$$Information\ Gain = Entropy(s) - [(Weighted\ Avg) * Entropy(each\ Feature)]$$

Where, s = Total number of samples

$$Entropy = -\sum_{i=1}^{n} p_i * \log(p_i)$$

- Gini Index:

$$Gini\ index = 1 - \sum_{i=1}^{n} ({p_i}^2)$$

Step-3 Divide the root node into subsets that contain Information Gain values for the best attributes.
Step-4 Generate the decision tree node, which contains the best feature.
Step-5 Recursively construct new decision trees until a stage is reached where the decision tree has all leaf nodes.

Flask:

Flask is a web framework, which is used to develop ideal web application using python. It is implemented on Jinja2 template engine and Werkzeug toolkit. It also provides a devlopment server and debugger.

Step-1 Create a new folder to run a flask and inside it set up four folders: data, models, static and templates.
Step-2 Create a index.html file, results.html file and app.py file.
Step-3 Create a POST route for the index page and POST route called /predict/
Step-4 Create a css/styles.css file on index.html
Step-5 Acquiring (receiving) the inputs from the HTML form.
Step-6 Performing prediction with the acquired (obtained) data.

Step-7 Show result in a HTML page.

IV. Methodology Used In The Paper

The main aim of this paper is to build an SMS spam detection web application. In this paper, Naïve Bayes and Decission Tree classification models are used to detect whether a message is a spam or not.

The data has been taken from the Kaggle is shown in table 1. Then, the dataset are cleaned by removing unnamed columns. After cleaning the data, number of spam and not spam messages are visulized by bar graph, where around 87.3% are ham messages and 12.7% are spam messages. In these models, Count Vectorizer tool is used to transform a given text into a vector on the basis of the frequency of each word that occur in the entire text. Which means that each message and each word are labelled by number of occurrences in a text. At the end, Naïve Bayes and Decission Tree algorithms are applied to classify the data.

smstype	sms
ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
ham	Ok lar... Joking wif u oni...
spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's
ham	U dun say so early hor... U c already then say...
ham	Nah I don't think he goes to usf, he lives around here though
spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, å£1.50 to rcv
ham	Even my brother is not like to speak with me. They treat me like aids patent.
ham	As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your friends Callertune
spam	WINNER!! As a valued network customer you have been selected to receivea å£900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only.
spam	Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile Update Co FREE on 08002986030
ham	I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried enough today.
spam	SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info
spam	URGENT! You have won a 1 week FREE membership in our å£100,000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C www.dbuk.net LCCLTD POBOX 4403LDNW1A7RW18
ham	I've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will fulfil my promise. You have been wonderful and a blessing at all times.
ham	I HAVE A DATE ON SUNDAY WITH WILL!!
spam	XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> http://wap. xxxmobilemovieclub.com?n=QJKGIGHJJGCBL
ham	Oh k...i'm watching here:)
ham	Eh u remember how 2 spell his name... Yes i did. He v naughty make until i v wet.
ham	Fine if thatåÕs the way u feel. thatåÕs the way its gota b
spam	England v Macedonia - dont miss the goals/team news. Txt ur national team to 87077 eg ENGLAND to 87077 Try:WALES, SCOTLAND 4txt/ì¼1.20 POBOXox36504W45WQ 16+

Table:1 Short Message Services data

80% of data is used for training and 20% of data is used for testing purpose. Then both the results are compared with each other. After comparing the result we can conclude that Naïve Bayes gives the better result with accuraccy 98.4%.

Now, web application is created by using Flask with Naïve Bayes algorithm. As Naïve Bayes gives better classification result.

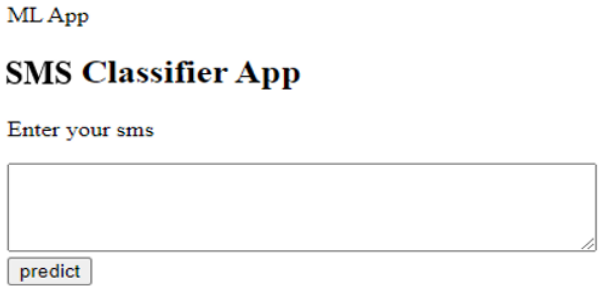

Figure:2 User Interface of Classifier Application to write a message

V. Result And Discussion

Naïve Bayes and Decision Tree algorithms are used to classifiy the spam and non-spam message. The accurcy of models are calculated to compare both the results. Accuraccy ratio of train and test is depicted in Figure:3. From the comparison, Naïve Bayes outperforms the Decision Tree.

Models	**Accuraccy**
Naïve Bayes	98.4%
Decission Tree	93.5%

Table 2: Accurccy obtained by models

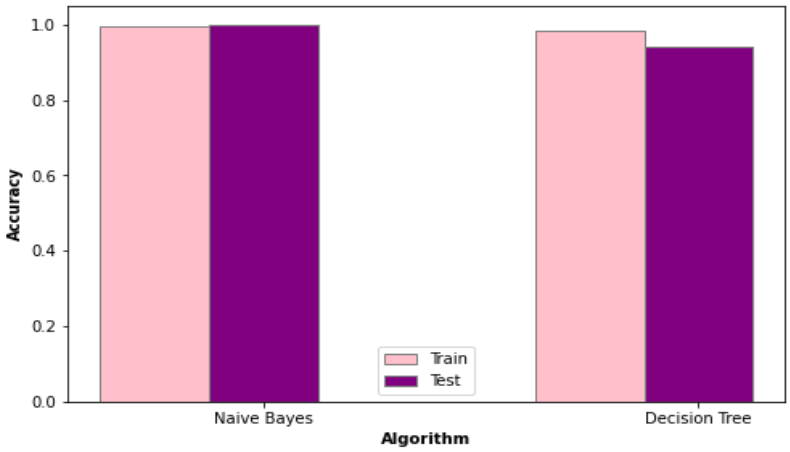

Figure:3 Graph of accuracy obtained by model

Also, web appliction is created using Naïve Bayes algorithm. As shown in fig. 4 the message can be written in the box. After clicking on the forecast button, it is seen that the application predicts the reply that the message is ham or spam.

Figure:4 User Interface of Classifier Application for prediction

VI. Conclusion

In supervised learning, there are many algorithms which are used for classifiction and regression both. These algorithms are used for fraud detection, spam classification etc. Detection of spam messages is necessary to stop such kind of fraud. Here, Naïve Bayes and Decision Tree algorithms are used to identify the spam and ham text message. Then, comparison of both algorithms shows that Naïve Bayes performs better than Decision Tree algorithm. The result acquired from comparison is that Naïve Bayes achieves the highest accuraccy of 98.4%. Therefore, SMS classifier web application is created by using Naïve Bayes algorithm.

In future, this type of classification problems can be solved with other supervised learning techniques and other web developer tool.

Refferences:

[1]. A. Radhakrishnan and Vaidhehi.V, "Email Classification Using Machine Learning Algorithms", *International Journal of Engineering and Technology*, 9(2), pp.2319-8613, 2017

[2]. S. Pandey, Supriya M and A. Shrivastava, "Data Classification Using Machine Learning Approach", *Intelligent Systems Technologies and Applications*, 683(85), pp.112-122, 2018

[3]. K.Ponmalar and Dr.K.Krishnaveni, "Random Forest Classification of Medicinal Plant Leaves Using Shape and Texture Features*", JASC: Journal of Applied Science and Computations*, 5(8), pp.561-569, 2018

[4]. S. Chowdhury and M. Schoen, "Research Paper Classification using Supervised Machine Learning Tchniques", *Intermountain Engineering, Technology and Computing*, 478(56), 2020

[5]. J. Ahmed and M. Ahmed, "Online News classification Using Machine Learning Techniques", *International Islamic University Malaysia Engineering Journal*, 22(2), 2020

[6]. S. Chaudhary, "Supervised Machine Learning Algorithms: Classification and Comparison", *GRD Journals- Global Research and Development Journal for Engineering*, 5(6), 2020

[7]. Z. Siddique, M. Khan, I. Din, A. Almogren, I. Mohiuddin and S. Nazir, "Machine Learning-Based Detection of Spam Emails", *Hindawi*, 2021(3), pp.1-11, 2021

[8]. M. Gupta and Vanmathi. C, "A Study and Analysis of Machine Learning Techniques in Predicting Wine Quality",*International Journal of Recent Technology and Engineering*, 10(1), 2021
[9]. L. Sandra, F. Lumbangaol and T. Matsuo, "Machine Learning Algorithm to Predict Student's Performance: A Systematic Literature Review*", TEM Journal-Technology Education Management Informatics*, 10(4), pp.1919-1927, 2021
[10]. A. Pandey and and L. Maurya, "Career Prediction Classifiers based on Academic Performance and Skills using Machine Learning", *SSRG International Journal of Computer Science and Engineering*, 9(3), pp.5-20, 2022
[11]. Bhavsar Shachi and Ravi Gor, "Predicting Restaurant Ratings using Back Propagation Algorithm" *International Organization of Scientific Research Journal of Applied Mathematics (IOSR-JM)*, 18(2), pp.5-9, 2022.
[12]. https://www.kaspersky.com/resource-center/preemptive-safety/how-to-stop-spam texts#:~:text=What%20are%20spam%20texts%3F,address%20or%20instant%20messaging%20account.
[13]. https://www.analyticsvidhya.com/blog/2021/05/sms-spam-detection-using-lstm-a-hands-on-guide/
[14]. https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

Ghosh Madhumita. "Short Message Service Classifier Application using Naïve Bayes algorithm." *IOSR Journal of Computer Engineering (IOSR-JCE)*, 24(3), 2022, pp. 01-06.