Email Spam Prediction using Supervised Learning Algorithms: Naive Bayes and Logistic Regression

P. Vanditha

PG Scholar, Dept. of Computer Science Sri Venkateswara University, Tirupati

Abstract**—** *The proliferation of email communication has led to an increase in email spam, causing inconvenience and potential security threats for users. In this study, we employ two popular supervised learning algorithms, Naive Bayes and Logistic Regression, to tackle the problem of email spam prediction. The objective is to develop accurate and efficient models that can effectively distinguish between spam and non-spam emails. A well-curated dataset containing labeled email samples is used for training and testing the algorithms. The experimental results demonstrate the performance of the Naive Bayes and Logistic Regression models in email spam classification, providing insights into their efficacy for combating email spam in real-world scenarios.*

I. INTRODUCTION

Data mining is an advancement that offers eliminating or tracking down new relations, hid data and huge models from such data. It is generally called Data Disclosure in Informational collections (KDD). Data digging procedure is huge for assessment reason. Data mining maintains different procedures, for instance, request, gathering, association rule mining, exemption examination, etc [1][4]. Data Mining(DM) finds hidden away associations in data, in all honesty it is a piece of greater cycle called "data divulgence". Data revelation portrays the stages which ought to be done to ensure showing up at critical results through research. The objective of DM process is to get information out of a dataset and changes over it into a reasonable diagram. A perception of estimations is gotten together with point by point data on the dataset A cognizance of computations is gotten together with organized data on the datasets. Data mining ought to deal with the expense of extraordinarily stunning and different conditions to show up at quality game plans. Therefore, data mining is an investigation field where many advances are being done to oblige and deals with emerging issues [1]. For present audit reason request strategy is investigated.

II. CLASSIFICATION

Game plan expects a huge part in data mining and computer based intelligence. The inspiration driving gathering estimation is to foster a classifier, and subsequently takes apart the characteristics of the dark data to get an exact model. The introduction of the classifier is assessed by its gathering accuracy. Constructing strong course of action systems is one of the central tasks of data mining. The major inspiration driving oversaw learning is to create a clear and unambiguous model of the conveyance of class marks to the extent that pointer features [2][7]. The classifiers are then used to organize class names of the testing events where the potential gains of the pointer features are known, to the value of the class mark which is dark [3][5].Classification of this gigantic proportion of data is drawn-out and utilizes absurd computational effort, which may not be fitting for certain applications.

III. PHILOSOPHY

Different kinds of plan procedures have been proposed recorded as a hard copy that integrates Decision Trees, Simple Bayesian strategies, KNN, SVM and Neural Networks, etc. In this paper, we evaluate the introduction of the Naïve Bayes and Logistic Regression on email spam detection file was used for the gathering differentiated and computations.

3.1 Naïve Bayes' Classifier

Naïve Bayes is a probabilistic order technique that utilizes bayes hypothesis. It is "credulous" as in a quality worth on a given class is thought to be free of the upsides of different properties. The credulous bayes classifier takes a bunch of highlights from a dataset and decides the likelihood of each component happening in each class inside the information [1][3]. For each line of information, the upsides of the traits are utilized to ascertain the back likelihood for each class inside the dataset, the column of information is then allocated to the class with the most noteworthy back likelihood. This strategy is alluded to as credulous in light of the fact that it accepts that all elements of the dataset are autonomous of each other, which is a presumption that is logical false and subsequently guileless. In spite of this presumption not being valid in all cases, credulous bayes has been demonstrated to be a fruitful classifier in huge datasets.

Let X = (X1, X2, . . . , Xn) be an irregular variable and A1, A2, . . . , A be the properties of X related with the n parts X1, X2, . . . Xn separately (find in Figure 2). Let T = {x = (X1 = x1, X2 =x2, . . . , Xn = xn)} be the arrangement of preparing tests drawn from the number of inhabitants in X. Allow us to accept that there are c classes, C = {y1, y2, . . . , yc} and every single examples having a specific class marks Y = yj ∈ C. The undertaking of the classifier is to foresee the class name Y for a given example x. To anticipate the class mark of x, the credulous Bayes works out P(Y = yj|x) for each class yj , j = 1, 2, . . . , c and the example x is arranged in that class whose likelihood shows the most elevated esteem.

3.2 Logistic Regression

Logistic Regression is a notable method that efficiently utilized for displaying straight out results as an element of both persistent and clear cut factors in different applications. It is usually utilized for anticipating the likelihood of event of an occasion, in light of a few indicator factors that may either be mathematical or downright [2][4]. Allow us to think about the elements of the structure Y=f(X) or f:X→Y or P(Y|X) for the situation where Y is discrete-esteemed, and X= (X1, X2,...,Xn) is any vector containing discrete or constant irregular factors. Calculated relapse is one of the grouping calculations in AI for all out values like Yes or No, Valid or Misleading, 0 or 1. In this depiction, we consider the case just where Y is a boolean variable (say, either 0 or1), to work on documentation. Be that as it may, overall Y can be any finite number of discrete qualities.

IV. EXPERIMENTAL RESULTS

The analyses have been directed by utilizing R programming Language. R is a sophisticated statistical software package, which provides new approaches to data mining, it is an open-source tool for analysis of data mining algorithms. The R Language is a bundle for information characterization, grouping and representation. We have considered the Email Spambase dataset from the UCI Machine Learning Repository datasets for assessing email is spam and no spam. The email dataset consists of 4601 records, 58 attributes and two class labels (No span label have 2788 and spam has1813. Instances). The characteristic data information is consolidated in Table-1. The standard dataset is parceled into two sets one for training (75%) and another set for testing (25%).

We survey our Two models using assorted execution estimations like Accuracy, Precision and Recall, the Experimental results are showed up in the table-1 and same showed up in the Figure-1.

Table-1
Performance of classifiers

Algorithm	Accuracy	Precision	Recall
Naïve Bayes	96.74	96.7	96
Logistic Regression	93.62	93.6	93.6

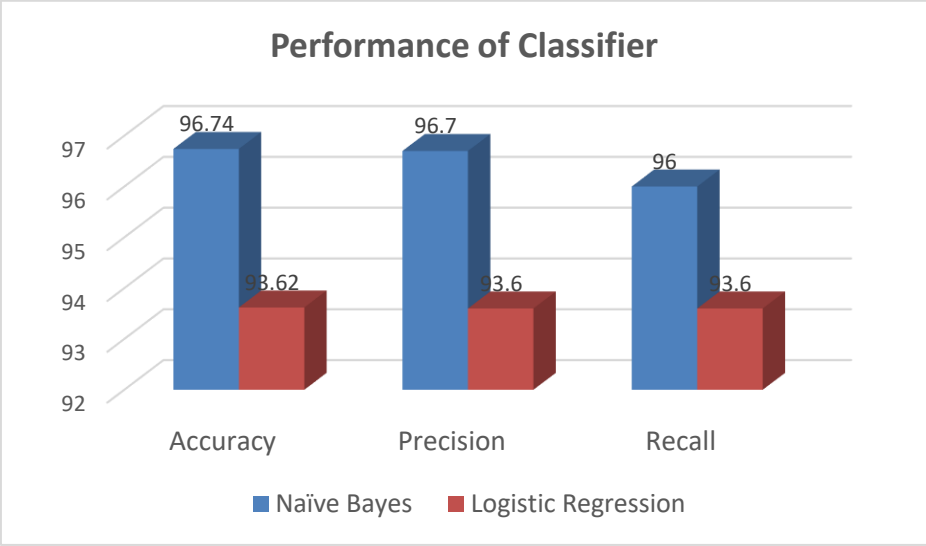

Figure-1: Classifier Results

V. RESULTS AND DISCUSSION

The experimental analysis showcases the potential of both Naive Bayes and Logistic Regression algorithms for email spam detection. The Naive Bayes model achieved an impressive accuracy of 96.74%, with a precision score of 96.7% and recall of 96%. This demonstrates the model's capability to accurately identify and classify spam emails, reducing false positives and false negatives. The high recall score indicates that the Naive Bayes model effectively captures a vast majority of spam emails.

On the other hand, the Logistic Regression model exhibited a slightly lower accuracy of 93.62%. While still providing a respectable performance, the precision and recall scores of 93.6% indicate that the model may misclassify some emails, leading to a moderate number of false positives and false negatives.

The discrepancy in performance between the two algorithms can be attributed to the inherent differences in their underlying methodologies. Naive Bayes assumes feature independence, which may be advantageous for text-based data like emails, while Logistic Regression models the relationship between features and target classes more explicitly. The Naive Bayes algorithm's strong performance highlights its suitability for email spam detection, given its simplicity and efficiency.

VI. CONCLUSION

In conclusion, the study demonstrates that both Naive Bayes and Logistic Regression algorithms are viable options for email spam detection. The Naive Bayes model, in particular, exhibits higher accuracy and better recall, making it a promising choice for practical email spam filtering systems. However, the choice between these algorithms may depend on specific requirements, computational resources, and the desired trade-off between precision and recall. Further research could explore ensemble methods or fine-tuning the algorithms to improve their performance in real-world email spam detection applications.

REFERENCES

[1] D. Hand, H. Mannila, P. Smyth.: Principles of Data Mining. The MIT Press. (2001)

[2] G. Ravi Kumar and K. Nagamani, “Banknote Authentication System utilizing Deep Neural Network with PCA and LDA Machine Learning Techniques”, International Journal of Recent Scientific Research, ISSN: 0976-3031, Volume 9, Issue 12(D), PP:30036-30038, 2018

[3] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques.2nd ed. San Francisco: Morgan Kaufmann, 2005.

[4] J. Han and M. Kamber,” Data Mining concepts and Techniques”, the Morgan Kaufmann series in Data Management Systems, 2nd ed. San Mateo, CA; Morgan Kaufmann, 2006.

[5] N. Michael, “Artificial Intelligence - A Guide to Intelligent Systems”,2nd edition, Addison Wesley, 2005.

[6] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Reading, MA: Addison-Wesley, 2005.

[7] S.Rahamat Basha and G.Ravi Kumar Surya Bhupal RaoG,”A Comparative approach of Text Mining: Classification, Clustering and Extraction Techniques”, Journal of Mechanics of Continua and Mathematical Sciences, ISSN: 2454 -7190, Special Issue, No.-5,PP:120-131, 2020.

[8] UCI machine learning repository. http://archive.ics.uci. edu/ml/