WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Email Classification Using Machine Learning

Harshada Tanpure[1], Sneha Hande[2], Prajakta Thorat[3], Prof. Bangar.A.P[4]

Jaihind College of Engineering Kuran, Pune-410511
tanpureharshada1@gmail.com[1] , snehahande2000@gmail.com[2], prajaktathorat633@gmail.com[3]

ABSTRACT: -

In today's word rate of exchange information via emails is increasing. These Emails are sent for a number of reasons: Extracting confidential information from individuals, promotion and marketing/advertising of products and services. Thus, keeping this mind, the importance to build a comprehensive system for Spam Classification based on semantics based text classification using various Machine Learning algorithms have been surveyed and the objective is to creating model with high performance and efficiency.

Keywords: - Email Classification, Machine learning, supervised learning , svm, naïve baised,random forest

I Introduction

Most of us should be intimate with spam emails. It defines it as unwanted junk email. Typically, spam is sent for commercial purposes. It can be sent in massive volume networks of infected computers. Therefore, spam email filtering is an essential feature for email services such as Gmail. Services providers are extensively using Machine learning techniques used to filter and classify them successfully.
Email classification work on some basic concept .By going to the text of the mail, we will use ML.In this paper we are discussed various Machine Learning Algorithm, which are Support Vector Machine, Naive Bayes and Random Forest. On the basis of this algorithm calculate the accuracy of the algorithm and find which give high accuracy and classify email are spam or not.

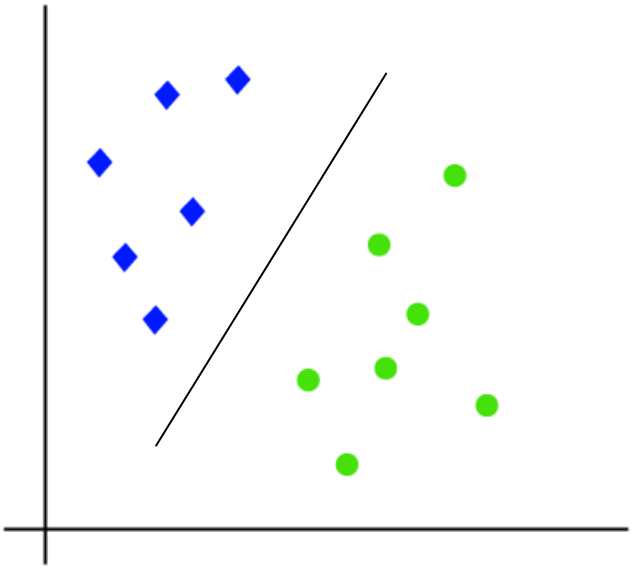
II Supervised learning

The ML problem is divided into two parts that is supervised learning and unsupervised learning. *Email Classification* is a Machine Learning problem that goes under the category of *Supervised Learning.* Supervised learning is an goal creating artificial intelligence, where a computer algorithm is trained on input data that has been labeled for a particular output.
Email Classification works on the same basic concepts. By going through the text of the email, we will use Machine Learning algorithms to predict whether the email has been written by one user or other. The supervised learning process involves input variables, which we call X, and an output variable, which we call Y. We can use an algorithms to learning the mapping function from the input to the output. In simple mathematical process, the output Y is a dependent variable of input X as illustrated by:
$Y = f(X)$
Here, our final goal is to try to approximate the mapping function f, so that we can predict the output variables (Y) when we have new input data X.

III SVM (support vector machine)

Support vector machine is supervised learning algorithms, which is used for classification as well as regression problems. The SVM classifier take and generates indices of each word. After training process the test cases emails are given input to test the accuracy of the model.
Support Vector Machine is a supervised machine learning algorithm which can be used for both classification or regression method. It is mostly used in classification problems. In the SVM algorithm, we can plot each data item as point in n-dimensional space (where n is a number of features) with the value of each feature being the value of a particular coordinatewe are going to classify emails into Spam and Ham . We have used SVM Machine Learning Model for that.

Linear SVM: It is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.The working of the SVM algorithm can be understood by using an example. Suppose we have a dataset that has two tags (spam and ham), and the dataset has two features x1 and x2. We want a classifier that can be classify the pair (x1, x2) of coordinates in either green or blue. Consider the below image:



Hence, the SVM algorithm helps to find the decision boundary; this best boundary or region is called as a *hyperplane.*

IV Naive Bayes

Naive Bayes spam filtering technique is a baseline method for deal with spam that can tailor itself to the emails needs of individual users and gives low false positive spam detection rates that are acceptable to users.

It is one of the ways of doing spam filtering.

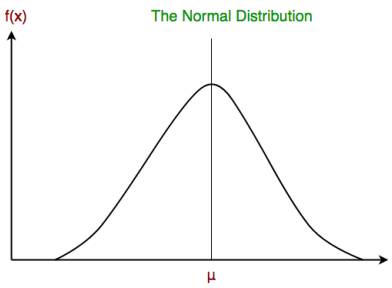
One of the simplest yet powerful classifier algorithms, Naive Bayes is based on Bayes' Theorem Formula with an assumption of independence among predictors. Given a Hypothesis A and evidence Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem stated mathematically as the following equation:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

- A, B = Events
- P(A|B) = Probability of A given B is true
- P(B|A) = Probability of B given A is true
- P(A), P(B) = The independent Probabilities of A and B

This theorem mainly used for classification techniques in data analytics. The Naive Bayes theorem calculator plays an important role in spam detection of emails.

Naïve Bayes classifiers is a collection of classification algorithms that is based on Bayes Theorem. It is not a single algorithm but a family of algorithms where all of them share a same principle, that is Every pairs of features being classified is independent of each others. In Gaussian Naive Bayes Therom, continuous values associated with each features are assumed to be distributed according to a Gaussian distribution. When plotted, it gives a bell shaped curve which is symmetric about the mean of the feature value as shown below:



The accuracy of naïve bayes for this particular problem in 0.9934640522875817.

V Random Forest

Random forests are Supervised Learning algorithm built on Decision trees. Random Forests are used for regression technique and classification technique. This algorithm takes its name from the random selection of features. This algorithm is also the most flexible and easy to use algorithm. A forest is comprised of trees. So the more trees it has, the more robust a forest is. Random forests generates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. It also provides a good indicator of the feature importance.We can use the Random Forests algorithm from the sklearn library on our dataset as *RandomForestClassifier()*. The accuracy of the algorithm is 1.0. The training time is 1.2s, which is reasonable but overall, it does not prove to be a good tool for our problem. The main reasons for the low accuracy is the randomness of feature selection, which is a property of random forests algorithm. The random forest model is a made up of many decision trees. Rather than just simply averaging the prediction of trees , so this model uses two key concepts that give it the name random: Random sampling of training data points when building trees.

VI Conclusion

This text classification of emails is performed using algorithms for comparison purpose. The algorithms used were Supervised Learning Random Forest, Naive Bayes, and Support Vector Machine. The algorithms created models and trained them using some of the data and tested the effectiveness of the model using a test subset of the data

References

1.V. Sri Vinitha, D. Karthika Renuka "Performance Analysis of E-Mail Spam Classification using different Machine Learning Techniques",IEEE 2020

2. Akash Junnarkar; Siddhant Adhikari; Jainam Fagania; Priya Chimurkar; Deepak Karia , "E-Mail Spam Classification via Machine Learning and Natural Language Processing ", IEEE 2021

3. Mansoor raza, nathali dilshani jayasnghe, muhana magboul ali muslam, "A Comprehensive Review on Email Spam Classification using Machine Learning Algorithms", IEEE 2021

4. Sri vinitha , D. karthika renuka, "performance analysis of e-mail spam classification using different machine learning techniques" , IEEE 2019

5. Sebastian romugomes ,sk golam saroar, a comparative approach to email classification using navie bayes classifier and hidden markov model, IEEE 2018

6. K sai prasanthi, t deepika, s anudeep, m sai koushik , " an efficient email spam detection using support vector machine", international journal of innovative technology and exploring engineering , 2019

7. K. Krombholz, H. Hobel, M. Huber, and E. Weippl Advanced Social Engineering Attacks", Journal of information security and applications 22 (2015) 113-12

8. E. Sorio, A. Bartoli, and E. Medvet Detection of Hidden Fraudulent URLs within Trusted Sites using Lexical Features 2013

9. M. Khonji, Y. Iraqi, and A. Jones Lexical url analysis for discriminating phishing and legitimate websites 2011

10. S. Sheng, B. Wardman, G. Warner, L. F. Cranor, J. Hong, and C. Zhang an empirical analysis of phishing blacklists.

11. Zhuang, L., Dunagan, J., Simon, D.R., Wang, H.J., Tygar, J.D., 2008. Characterizing Botnets from Email Spam Records, LEET'08 Proceedings of the 1st Usenix Workshop on LargeScale Exploits and Emergent Threats

12. Enrico Blanzieri, Anton Bryl, 2008. A survey of learning based techniques of email spam filtering, Technical Report DIT-06-056.