Review Web Spam Detection using Data Mining

Swathi Raj[1], **K. Raghuveer**[2]

[1]*Dept of CNE, The National Institute of Engineering, Mysuru, Karnataka, India*
[2]Associate Professor, Dept. of IS&E, *The National Institute of Engineering, Mysuru, Karnataka, India*

---***---

Abstract - *In today's digital world a large a part of verbal exchange, each professional and private, takes vicinity within the shape of electronic mails or emails. However, because of advertising companies and social networking websites most of the emails circulated include undesirable information which is not relevant to the user. Spam emails are a kind of electronic mail wherein the user receives unsolicited messages via email. Spam emails reason inconvenience and economic loss to the recipients so there is a want to clear out them and separate them from the valid emails. Many algorithms and filters were advanced to locate the spam emails but spammers continuously evolve and sophisticate their spamming strategies because of which the existing filters are becoming less effective. The approach proposed in this paper involves developing a unsolicited mail clear out the usage of binary and non-stop opportunity distributions. The algorithms carried out in building the classifier model are Naive Bayes and Decision Trees. The effect of over fitting on the performance and accuracy of selection bushes is analyzed. Finally, the higher classifier model is diagnosed primarily based on its accuracy to correctly classify unsolicited mail and non-unsolicited mail emails.*

***Key Words*: binary distribution, continuous distribution, decision tree, naive bayes, spam**

1. INTRODUCTION

Electronic mail, or email, is a way of changing virtual messages between human beings the usage of virtual gadgets such as computers, capsules and cell telephones. Internets being the predominant platform for communication in nowadays age, emails are taken into consideration as one of the fastest manner to change data. A big a part of communiqué in nearly every subject takes part in the form of these digital messages. Unsolicited Commercial Email, commonly referred to as Spam, is generated by using sending unsolicited industrial messages to many recipients. Spam messages bring about many problems such as decreased overall performance of the mail engines, profession of pointless space in the mailbox and destroying the steadiness of mail servers. In some cases they also contain viruses, Trojans and other substances that may be potentially dangerous for positive class of customers. Spam mails are a purpose why users spend quite a few times in doing away with undesirable correspondence and sorting incoming mails. An issue related to Spam Mail has been growing exponentially through the years. Email users, on a day by day Foundation, get hold of loads of spam messages with new content and new assets and these spams are generated automatically by means of robotic software program. To filter spam with conventional methods including black-white lists [1] (domain names, IP addresses, mailing addresses) is sort of not possible. The troubles associated with junk mail.

Mails are escalating as is using web. The truth that out of 80 billion emails acquired normal forty eight billion of them being junk mail highlights the significance and urgency of enforcing effective class techniques for emails [2]. This has led to the necessity of distinguishing junk mail and non-unsolicited mail emails in order that the ones categorized as junk mail emails can at once visit the unsolicited mail folder and no longer into the inbox. With the growing community bandwidth and enhancing era spam emails have turn out to be more sophisticated and it's miles vital to use advanced algorithms to create efficient junk mail filters. Despite the huge amount of research work that has taken location in this Sphere, there's no junk mail filter out which is one hundred% efficient. Hence, there may be a need to increase more sophisticated and correct classifier model to remove the problem of unsolicited mail emails.

The relaxation of this paper is framed as follows. Section II Describes the associated paintings in the subject of spam category. An evaluation about possibility distributions and category algorithms used are discussed in Section III. Section IV offers the effects acquired from the experiments conducted. Further dialogue and analyses of acquired effects is presented in section V. The paper is concluded with summary in Section VI. Section VII outlines the destiny work.

2. RELATED WORK

Exhaustive research has been carried out inside the discipline of junk mail category and many algorithms have been used for the equal. Support Vector Machine [3], Bayesian class for Extractions of features are the not unusual approaches utilized by researchers. The continuously changing behavior and attributes of spam emails has been a subject of interest. Many researchers have proposed numerous steps to enhance the overall performance of spam filters.

D. Wang, D. Irani and C. Pu [4] performed an extended-time period evolutionary examine at the Spam Archive dataset [5]. They studied the evolution of junk mail emails over a period of fifteen years, from 1998 to 2013. Their study and evaluation confirmed that even though the quantity of spam emails skilled a slight drop in later years (2009-2011) it took place best due to the fact spammers had grow to be greater capricious and complex and the filters were now not green sufficient to hit upon the junk mail emails. The authors of [6], proposed using general records preprocessing steps in junk mail filters. These steps included elimination of lacking and noisy values using information cleaning, statistics integration, records transformation and discount. They used data normalization previous to characteristic extraction in their analysis. Data mining strategies play an essential function in classifying junk mail and ham (non-unsolicited mail) emails. The authors of [6] cautioned that spam filtering results might be stepped forward by means of applying numerous preprocessing steps such as black list and white listing. In black listing a listing of domain names which can be typically used by using the spammers is created and all of the mails coming from that domain are blacklisted. In white listing a list of depended on domain names is created and emails from those domain names are labeled are valid.

Naive Bayes classifiers are taken into consideration to be very sturdy in relation to continuous beside the point attributes however they make strong assumptions on independence of probabilities [7]. Ron Kohavi [7], suggested a hybrid approach to mix each selection bushes and Naive Bayesian classifiers. According to his set of rules, a decision tree turned into created with splits at each node however the leaves were constructed using Naive Bayes classifiers. The set of rules became tested on UC Irvine repository and recorded an accuracy of eighty four. Forty seven%.

Various algorithms were used over the time frame to broaden content material-based totally junk mail filters. A. Saab, N. Mitri and M.Awad [8], carried out a comparative observe of Support Vector Machine, Local Mixture Support Vector Machine, Artificial Neural Networks and Decision Trees. Artificial Neural Networks recorded the best accuracy observed by way of Decision Trees.

Previous studies have targeted on developing spam filters the use of above stated algorithms. However no longer an awful lot research has been finished on growing junk mail filters the use of one-of-a-kind possibility distributions. This paper focuses on junk mail classification the use of Naive Bayes and Decision Trees with binary and continuous chance distributions.

3. BACKGROUND

A. Probability Distribution

Probability distribution, in records and opportunity theory, affords the chance of incidence of possible effects of an experiment. It describes a random phenomenon in phrases of chances for an event. Statistically, it is a function that describes the likelihoods and feasible values that a random variable can take. Probability distribution may be widely classified into discrete chance distribution and continuous chance distribution.

1) Discrete Probability Distribution: Discrete Probability Distribution is typically relevant to conditions wherein the set of all viable consequences of an event is discrete. Bernoulli distribution is a sort of discrete chance distribution with simplest feasible outcomes given with the aid of m=0 and m=1. Success chance (m=1) is denoted by way of p and the possibility of failure (m=zero) is denoted via q. mathematically, q = 1 - p, where p lies between 0 and 1. If Y is a random variable,

$$P(Y = 0) = 1 - P(Y = 1) = 1 - p = q$$

2) Continuous Probability Distribution: Continuous Probability Distribution is applicable to situations wherein all of the possible consequences of an test can have values in a non-stop variety. Some of the widely known continuous distributions consist of normal distribution and chi-squared distribution. If Y is a random variable, then its continuous probability distribution is given by way of:

$$P[p \le Y \le q] = \int_{qp} f(y)dy$$

B. Naive Bayes

Naive Bayes is a classifier model based on Bayes theorem Which applies robust independence assumptions among the Capabilities of the version. It assumes that given a category variable, the price associated with a few feature x is completely unbiased of the value related to some other characteristic y. Mathematically, the Bayes theorem, on which Naive Bayes classifiers stand, can be said

$$p(C_{k|y}) = p(C_{k})p(y/C_{k})p(y)$$

Where in y is the hassle instance which may be written in vector shape as :

$$y = y1,y2, ...yn$$

Here, p(C / y) is the posterior probability of a class(target) given the predictor(attribute), p(C) is the prior probability of the class, p(y / C) is the likelihood or the probability of the predictor given the class and p(y) is the prior probability of predictor.

C. Decision Tree

Decision tree is a predictive modeling approach which is used in machine learning, data mining and statistics. It creates a model which on the basis of several input variables predicts the value of target variable. It is a widely used algorithm which follows the greedy approach at each split and progressively builds a tree. Each node of a decision tree represents an experiment on an attribute, branches represent the result of the experiments and the leaf nodes contains the class labels. The decision tree splits are chosen such that they minimize impurity and maximize purity of the subset being constructed. Some impurity measures include:

1) *Entropy:* Entropy is used to calculate information gain which decided the feature to split on at each step of the decision tree. It is used by ID3, C4.5 and C5.0 decision tree generation algorithms. Entropy can be calculated.

2) Gini Index: Gini impurity (Gini Index) applies in a multiclass classifier context and it is a measure of misclassification. Here pj is the probability of a class j. Gini impurity is used by CART (classification and regression tree) algorithm.

3) ID3 Algorithm: The ID3 or Iterative Dichotomies three set of rules became invented by Ross Quinlan in 1986 for building a selection tree. The algorithm starts off evolved with the real dataset on the initial node or the foundation. The set of rules iterates thru a exclusive attribute of the set each time and the entropy for that attribute is calculated. The attribute with least entropy or maximum records benefit is chosen as the characteristic to be cut up on. The recursion continues till all leaf nodes incorporate natural subsets.

4) C4.5 Algorithm: C4.5, additionally invented by way of Ross Quinlan, is a succession to the ID3 algorithm. The algorithms trains a decision tree version within the identical changed into as ID3 the usage of entropy. However, it provides certain enhancements to the ID3 set of rules including handling missing values, operating with each continuous and discrete attributes, etc. The decision tree unsolicited mail classifier became built the usage of C4.5 set of rules.

5) Overfitting: Given a hypothesis space Hs, a hypothesis h in Hs is stated to overfit the training records if there exists a few alternative hypothesis h' in Hs, such that h has smaller error than h' over the education examples, but h' has a smaller blunders than h over the entire distribution of instances. Overfitting in a decision tree takes place while the tree is just too trained on the schooling dataset and as a end result does now not perform well on real-international unknown times. Small education datasets and noisy information can purpose overfitting and decrease the performance of a decision tree extensively [8].

4. EXPERIMENTAL RESULTS

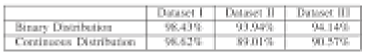
The Naive Bayes and Decision Tree classifier fashions in each binary and continuous distribution were tested on three well known datasets. The first dataset become accumulated from the Indiana State University repository which contained 5200 spam and non-unsolicited mail emails. The 2d dataset became taken from the Ling-Spam corpus with lemmatize and forestall-listing disabled (contained a complete of 2893 documents). The 0.33 dataset changed into additionally taken from the Ling-Spam corpus with lemmatizes enabled however forestall-list disabled (2893 junk mail and non-junk mail emails).

The use of Bernoulli distribution in this research painting is carried out to indicate the occurrence of attributes taken into consideration in each email. If an attribute takes place in an email the count related with that characteristic is 1 else zero. In case of non-stop distribution, the entire quantity of occurrences for each characteristic in an email is taken under consideration for building the classifier models.

The Naive Bayes classifier first calculates the probability chances of all attributes gift in the education set and then makes use of the earlier opportunity (both unsolicited mail or non-junk mail) to are expecting the label of a document. Once the model is skilled the classifier shows ten maximum representative and least representative words for junk mail together with their calculated possibilities. The model also produces the accuracy of the classifier at the side of the confusion matrix. The consequences obtained from Naive Bayes classifier are tabulated in Table I.

The creation of selection tree throughout education section became done the usage of entropy and statistics advantage as given by using C4.Five algorithm. The consequences acquired from choice tree classifier are tabulated below in Table II

TABLE II. DECISION TREE CLASSIFIER RESULTS



	Dataset I	Dataset II	Dataset III
Binary Distribution	98.43%	93.94%	94.14%
Continuous Distribution	98.62%	89.01%	90.57%

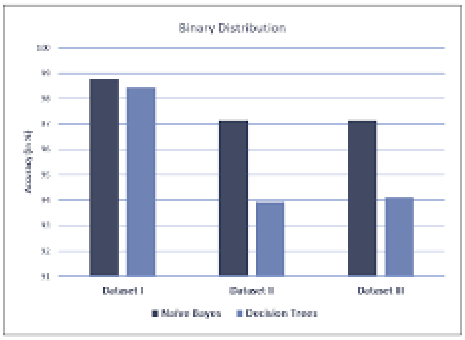


Fig. 1. Binary Distribution : Performance Analysis

5. DISCUSSION

A. Binary Distribution

The performance of Naive Bayes and Decision tree fashions in binary distribution is plotted in Figure 1. Evidently, the Naive Bayes classifier plays higher than the Decision Tree classifier here. A viable cause in the back of this more desirable Naive Bayes overall performance might be because of the reality that the version works on independence assumptions. Correct estimation implies correct prediction, but accurate prediction does now not mean accurate estimation.

B. Continuous Distribution

From Figure 2 it is obvious that Decision Tree version plays better than Naive Bayes model in non-stop distribution. The robustness and grasping approach of C4.5 set of rules could be a chief issue.

C. Comparative Study

As shown in Figure three, the classifier fashions carry out better with binary (Bernoulli's) chance distribution than with continuous opportunity distribution. Best effects are acquired with Binary Naive Bayes classifier. Continuous Naive Bayes performs poorest among the 4 with an aggregate accuracy of 90.22%.

D. Effect of Overfitting aspect

To analyze the impact of overfitting at the performance of decision trees an appropriate cost is selected because the overfitting thing. If the length of model report built in the training section is much less than the overfitting element then it may be concluded that the skilled model is too particular to the education dataset

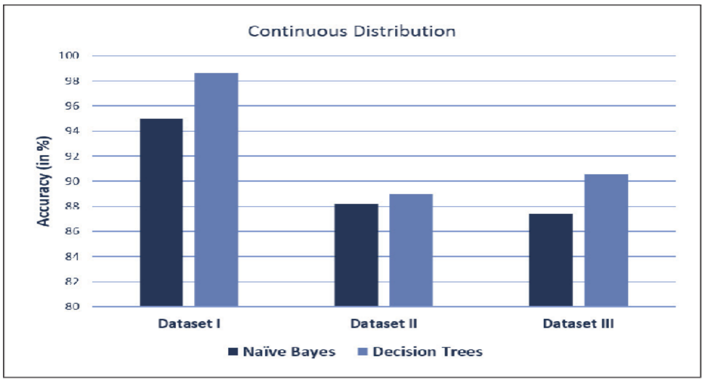


Fig. 2. Continuous Distribution : Performance Analysis

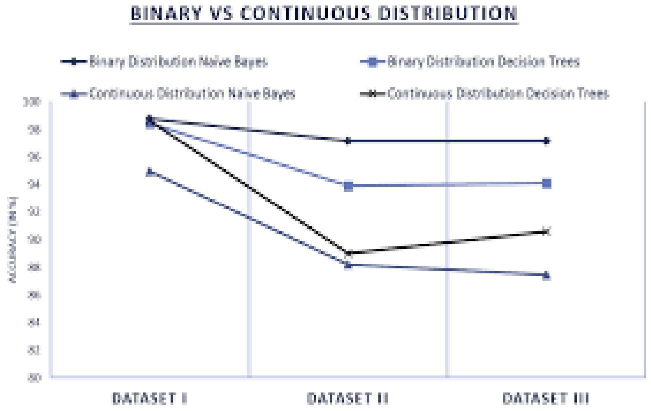


Fig 3. Continuous vs Binary Distribution

Figure four shows the effect of overfitting thing considered in the decision tree model on its accuracy. As the overfitting thing is decreased from 40 to twenty, the tree depth increases and a larger tree is built which leads to better category consequences to begin with. However, if the selection tree is grown beyond a sure limit, the tree will become too precise to the schooling dataset and plays poorly at the test records. This announcement is sincerely established via the dip inside the graph whilst the overfitting thing is reduced from 20 to ten.

6. CONCLUSIONS

This paper provides a method to unsolicited mail class Using Bernoulli's and non-stop possibility distribution. The Spam classifiers were examined on 3 benchmark datasets and the experimental consequences revealed that the classifier fashions carry out better with Bernoulli's chance distribution than with continuous chance distribution. It can also be concluded that there may be inherently no advanced classifier version between Naive Bayes and Decision Trees (hence proving the "no free lunch" theorem).

The overall performance of classifier fashions range and depends on several elements together with the opportunity distribution used, dataset and the problem involved. In Naïve Bayes, the classifier has to gain knowledge of by means of hand, at the same time as in Decision Tree the classifier choices the pleasant characteristic by using looking on the desk or version document. Over fitting in choice timber can Reduce its performance extensively. However, pruning the Selection tree at some stage in the schooling section can avoid it from overfitting the schooling dataset.

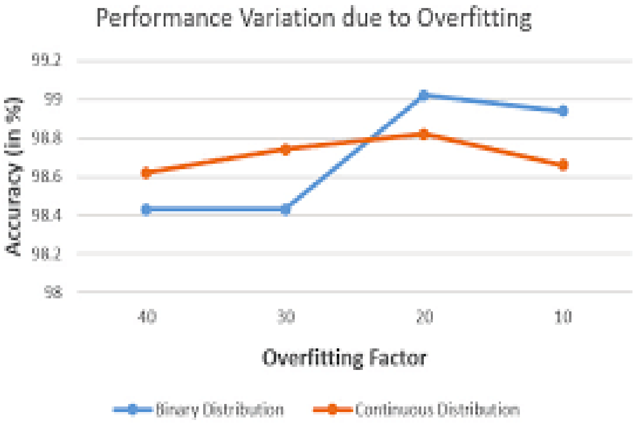


Fig 4. Effect of overfitting on decision trees

7. FUTURE WORK

Further studies must awareness on implementation of boosting algorithms to look if the accuracy of classifier fashions can be expanded. Implementation of Adaptive Boosting algorithm and its evaluation with the proposed approach of implementation can be an exciting prospect.

REFERENCES

[1] R. K. Kumar, G. Poonkuzhali, and P. Sudhakar, "Comparative study on email spam classifier using data mining techniques," in Proceedings of the International Multi Conference of Engineers and Computer Scientists, vol. 1, 2012, pp. 14–16.

[2] A. Harisinghaney, A. Dixit, S. Gupta, and A. Arora, "Text and image based spam email classification using knn, naive bayes and reverse dbscan algorithm," in Optimization, Reliabilty, and Information Technology (ICROIT), 2014 International Conference on. IEEE, 2014, pp. 153–155.

[3] B. Hssina, A. Merbouha, H. Ezzikouri, and M. Erritali, "A comparative study of decision tree id3 and c4. 5," International Journal of Advanced Computer Science and Applications, vol. 4, no. 2, 2014.

[4] H. Chauhan and A. Chauhan, "Implementation of decision tree algorithm c4. 5," International Journal of Scientific and Research Publications, vol. 3, no. 10, 2013.

[5] D. Wang, D. Irani, and C. Pu, "A study on evolution of email spam over fifteen years," in Collaborative Computing: Networking, Applications and Work sharing (Collaboratecom), 2013 9th International Conference Conference on. IEEE, 2013, pp. 1–10.

[6] R. Kohavi, "Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid." in KDD, vol. 96. Citeseer, 1996, pp. 202–207.

[7] S. A. Saab, N. Mitri, and M. Awad, "Ham or spam? a comparative study for some content-based classification algorithms for email filtering," in Mediterranean Electro technical Conference (MELECON), 2014 17th IEEE. IEEE, 2014, pp. 339–343.

[8] V. K. Pang-Ning Tan, Michael Steinback, Introduction to Data Mining. Pearson, 2007.