RESEARCH ARTICLE OPEN ACCESS

Detection of Spam Mail Using Naïve Bayesian Theorem without Vector Method

Ankur Sharma[1] Sohit Agarwal[2]
M.TechScholar, Department of CEIT, Suresh Gyan Vihar University, Jaipur
Associate Professor Department of CEIT, Suresh Gyan Vihar University, Jaipur
Corresponding author: Ankur Sharma

ABSTRACT: In this paper we are discussing the classification technique of separation of ham and spam. The problem at hand is to classify and email as spam or non spam given features of the email message and we have seen a second what kind of features can be used to qualify and email which would make it distinctive for a class spam classifier. So, the probability of spam given features of a message is what we have to estimate, and this is actually the posterior probability of the class spam. By base rule as the likelihood of the features of the email message given then it was a spam message times the prior probability of spams divided by the features of the message

Keywords— Bayesian, Spam, Probability,Tokenization

Date Of Submission: 30-09-2019 Date Of Acceptance: 14-10-2019

I. INTRODUCTION

When the community escalation is not in place the characters easily and quickly through the address index to help publish themselves without a very low value of the huge amount, the number of sender's users with any staff. Nowadays there are several ways to filter the variety of spam used. H XS Txawm Academia and its processing application Some spam tracking techniques.

Naxos' theory of virtualization and expression, which can be accessed in such a way that they are presented to him: and in the opinion of a good cause of testimony. A lot of research has been conducted to improve the performance of this workbook. I know what time it makes more business cards to learn the spam filter developer. Paul Graham applied Bayeevan's approach to spam [1] for access to training before accessing the fragmented database at low arithmetic. According to the test to find out why the public is available, spam. As for the rest, set in the allotted space, because a piece of paper.

There is no doubt what elements in hand, e-mail or profile are provided due to spam and spam messages. So, along with one of the best news from it, the characteristics of spam have been provided because we have a reason to be able to do it: this is a potential back-to-back spam.

This can break down the rigid propaganda lines of the e-mail message to be given to the spam message in a first harmful probability separated from the correct probability of the general message divided by the simple Kents situation and we can suppose that the different characteristics of the word and conditional independent data let's know about the family now It's spam So, there is a rule that Bayes is deaf to the presumption of freedom of conditional words

II. BACKGROUND AND REALTED WORK

Recently, review mining has become a hot research topic. Hu and Liu[1] used association rule mining based on the Apriori algorithm to mine product features from reviews and extract the adjective near the product feature as opinion word. They identified the semantic orientation of an opinion word by the set of seed adjectives with known orientation and WordNet. Popescu et al[2] first extracted nouns andnoun phrases from reviews as candidate product features and assesse those by computing the Point-wise Mutual Information scores between the candidate product features and merony my discriminators associated with the productclass and then applied manual extraction rules between inorder to find the opinion words. Zhao and Zhou[5] firstextracted the templates of POS tags between product featuresand the corresponding opinion words from training corpusand then identified product features and the correspondingopinion words from test corpus based on these templates andthe set of seed product features and opinion wordsiteratively. Somprasertsri and Lalitrojwong[6] mined productfeatures and opinion words based on the dependencyrelationships. Konstantin Tretyakov et al., [6] have evaluated several most popular machine learning methods i.e., Bayesian classification, k-NN, ANNs, SVMs and of their applicability to the problem of spam-filtering. In this work, the author proposed most trivial sample implementation of the named techniques and the comparison of their performance on the PU1 spam corpus dataset is presented. The author used extracting feature to convert all messages to vectors of numbers (feature vectors) and then classify these vectors. This is because most of the machine learning algorithms can only classify numerical objects like vector.

III. METHODOLOGY

To achieve the objective, the research and procedure is conducted in three phases. The phases involved are as follows:
(i) Phase 1: Pre-processing
(ii) Phase 2: Feature Selection
(iii) Phase 3: Naive Bayes Classifier
The following sections will explain the activities that involve in each phases in order to develop this project. Figure 1 shows the process for e-mail spam filtering based on Naive Bayes algorithm.

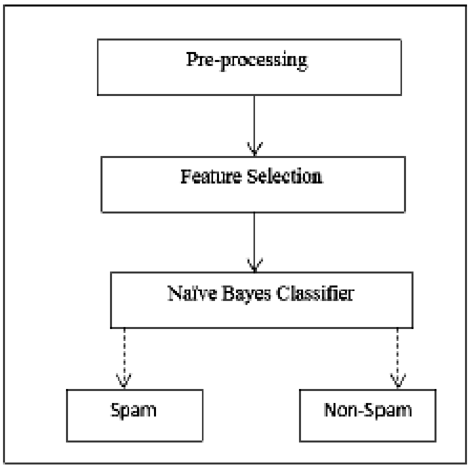

Figure 1 Process for e-mail spam filtering

Naive Bayesian [2] [3] is a great way to use technology to use spam problems with other techniques [4]. Paul Graham intends to use this idea [1]. This process has been revised [5]. The most important Baisy filter made the difference in different words [6]. There are many that continue to apply to the algorithm [7].
X and the category with the highest likelihood of probability is the target group, for example X.

$$\mathrm{P}\ (\mathrm{C_i l\, X}) = \frac{\mathrm{P(C_i)P(x_i I\, C_i)}}{\sum \mathrm{P(x_i)}}$$

$$\mathrm{P}\ (\mathrm{C_i l\, X}) = \frac{\mathrm{P(C_i)P(x_1 I\, C_1)}}{\mathrm{P(C_i)P(x_1 I\, C_1) + \ P(C_2)P(x_2 I\, C_2)}} \geq \tau$$

IV. DATA PROCESSING AND MODELLING

It has 4601 example cases, 39.4 percent of that a spam, the rest are non-spamIt has 4601 cases, 39.4 percent spam, and the rest is not random mailings. Each e-mail is represented as a relative frequency and we see how it is measured. Frequency associated with 48 keywords these keywords are words "free" and "cash", these are the keywords that often appear in spam emails in order to increase the probability that email is an unwanted message. We delete all special characters so that they have 6 characters and 3 attributes for the length of execution. So, we will not keep in mind race in features and we will focus only on the first 54 features.
The representation of e-mail (Figure 2) is important, because it will use the correct use of naive naiveté.

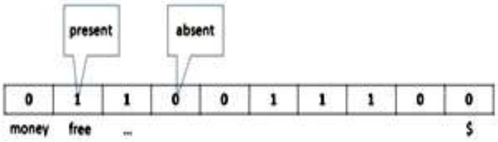

Figure 2. Format ofE-mail

He has the responsibility of overseeing the train at the train by using force. We have used the library only through scratch and nothing else, and the rest have been written down in the following code.

Then we will see how to estimate the divorce of the chapter. Let's say we have an email and we have a word w that we want to know if the word w has spam.

In our database, each email is represented as a vector, and each site displays the number of messages in email. Let's say we talk about some words w so vectors will have the time appear in an email divided by all the details of the whole number and multiplied by 100 to the percentage. So what's happening here is that if there are 100 words in the email, how we use this information to calculate this. So, this should display all the information spam in the file on the right and get the same eighth emails per page once per email. So what we do is think about all the spam in our database, and then these numbers are average. In fact, we share the numbers we get after calculating an average of 100 to the proportion of 0 to 1 in the probability and we have other costs.

V. RESULTS AND DISCUSSION

First we separate the class 0 and class 1; so class 0 represents that the email is not spam, class 1 represents the email is spam. First, we broke the first chapter and two chapters. So, device 0 is equipped email, not a registration form, and the first is spam. After that we count the average column. As seen in the last kick, we calculated the finances of each column. This is why the process is 0 and distributed by 100 to the percentage increase to the maximum. So, we need to balance the data gap without using the cover as seen in the slide.

We can see that we are developing more productive products through a variety of works. Thus, each of these numbers is less than one, often, smaller, and possibly able to conclude with a low number who can not represent fairly. So this causes problems and problems in line. So we do not get hit. We turn all the challenges into an accessible way to guess the information and add more fancy lines to it. This method takes the value of 0 and 1. What we need to do if the difference is 0, so in each position we calculate the result of the attribute. If this function, we calculate less than the actual, so the function does not exist.

Figure 3 shows the accuracy of the system. For the accuracy average, the difference total of two datasets is 8.59% which Spam Data get91.13% while SPAMBASE get 82.54% . On the other hand SPAMBASE get the highest percentage with 88% while Spam Data 83% for the average ofprecision. It means SPAMBASE get almost correctly prediction for spam e-mail.

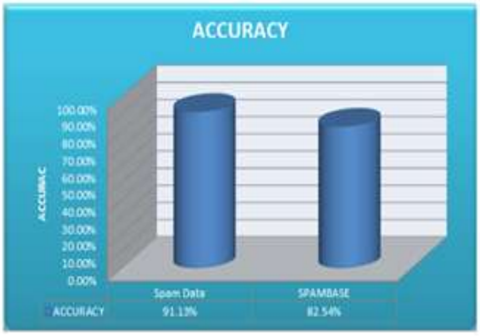

Figure 3. Accuracy of the system.

VI. CONCLUSION

E-mail spam filtering is an important issue in the network security and machine learningtechniques; Naive Bayes classifier that used has a very important role in this process of filteringe-mail spam. The quality of performance Naive Bayes classifier is also based on datasets thatused. As can see, dataset that have fewer instances of e-mails and attributes can give good performance for Naive Bayes classifier.We have design the adaptive email and spam classification model. This model is working very well with 91% accuracy. This model can be used in twitter,email or Facebook account.

REFERENCE

[1]. J. Clark, I. Koprinska and J. Poon, "Linger - A Smart Personal Assistantfor E-Mail Classification", in International Conference on ArtificialNeural Networks, 2003, pp. 274–277.

[2]. S. Wasi, S. Jami and Z. Shaikh, "Context-based email classificationmodel", Expert Systems, vol. 33, no. 2, pp. 129-144, 2015.

[3]. I. Alsmadi and I. Alhami, "Clustering and classification of emailcontents", Journal of King Saud University - Computer and InformationSciences, vol. 27, no. 1, pp. 46-57, 2015.

[4]. J. Rennie, "ifile : An Application of Machine Learning to E-MailFiltering", in Proceedings of the KDD (Knowledge Discovery inDatabases) Workshop on Text Mining, 2000.

[5]. S. Sayed, "Three-Phase Tournament-Based Method for Better EmailClassification", International Journal of Artificial Intelligence &Applications, vol. 3, no. 6, pp. 49-56, 2012.

[6]. M. Fuad, D. Deb and M. Hossain, "A trainable fuzzy spam detectionsystem", in 7th International Conference on Computer and InformationTechnology, 2004.

[7]. S. Youn and D. McLeod, "Spam Email Classification using an AdaptiveOntology", JSW, vol. 2, no. 3, 2007.

[8]. M. Aery and S. Chakravarthy, “eMailSift: Email Classification Based onStructure and Content,” Data Mining, Fifth IEEE Int. Conf., pp. 18–25,2005.

[9]. S. Chakravarthy, A. Venkatachalam, and A. Telang, “A graph-basedapproach for multi-folder email classification,” Proc. - IEEE Int. Conf.Data Mining, ICDM, pp. 78–87, 2010.

[10]. T. Ayodele, S. Zhou, and R. Khusainov, “Email Classification UsingBack Propagation Technique,” Int. J., vol. 1, no. 1, pp. 3–9, 2010.

Ankur Sharma" Detection of Spam Mail Using Naïve Bayesian Theorem without Vector Method" International Journal of Engineering Research and Applications (IJERA), vol. 9, no. 10, 2019, pp 46-48