ISSN 2395-1621

Ham or spam? A content based classification for Email filtering

[#1]Pooja Badhe, [#2]Vrushali Durge, [#3]Manisha Jangale, [#4]Rohini Shirsath,
[#5]Prof. D. S. Zingade

[1]poojabadhe999@gmail.com
[2]vrushalidurge7@gmail.com
[3]manishajangale94@gmail.com
[4]rohinishirsath15@gmail.com

[#1234]Department of Computer Engineering, Savitribai Phule Pune University
Kennedy Road, Near R.T.O., Pune, 411 001, India

ABSTRACT

Spam emails are widely spreading to constitute a significant share of everyone's daily inbox. It is been a source of financial loss and inconvenience for the recipients, spam emails have to be filtered and separated from legitimate ones. This paper presents a filtering algorithm Naïve Bayes that rely on text classification to decide or calculate whether an email is spam or not. A comparison among emails is performed on the Spam Base dataset to identify the best classification algorithm in terms of computational time, accuracy and precision/recall rates. The proposed Naïve Bayes algorithm was found to be extremely fast in recognition and with good error rates. It can be used as baseline learning element, in terms of accuracy, CPU time and against other learning elements of agent i.e. algorithms such as Support Vector Machine and Decision tree.

Keywords: Ham, Spam, Spam Filtering, Content Based Filtering, Spam email detection, Nearest neighbour classifier, Data mining, Classification, Naive Bayes, Machine Learning, Pattern recognition, Learning Agent

ARTICLE INFO

Article History

Received :4th May 2016

Received in revised form :

6th May 2016

Accepted : 12th May 2016

Published online :

17th May 2016

I. INTRODUCTION

Spam mails are largely known for unwanted and unsolicited emails sent with the purpose of financial gain i.e. fraud or simply causing harm i.e. harass or irritate users. They may be used to distribute fake announcements or viruses that cause responders an average loss of 29 USD per reply. It has been estimated that 1 out of 40,000 users reply to spam emails unknowing. Moreover, the fact that 48 billion of the 80 billion emails daily send are spam so that both the importance and urgency of developing effective classification procedures for received emails. Filtering spam is one of the important applications of pattern recognition and data mining advancement as heavy research has been conducted to write algorithms capable of recognizing spam from legitimate i.e. legal emails. Emails are filtered based on their content, which includes images and textual data or their header fields which provide information about the sender who are intended for communication. In our project, the spam problem is treated as a classification problem, which is known as pattern recognition problem as well. The user needs to decide only whether an email is spam or not.

So an intelligent agent will learn from his decisions dependent on his past and present calculations to sort out whether a future email which is spam or ham. In this paper, the specialized Naive Euclidean model is explained for the email spam problem.

II. ASSOCIATION RULE USING NAÏVE BAYES CLASSIFIER

Research on Text Classification Using Naive Bayes Classifier is used to classify text and the dependability of the Naïve Bayes Classifier with Associated Rules. But this method the negative calculation is ignored for any specific class determination in some cases may fall with accuracy. As e.g. to classify a text it will just calculates the probability of different classes with the probability values of the matched set while ignoring training sets of the unmatched sets of. As rule set in a result if test set matches with a test cases, which has weak or less probability to the

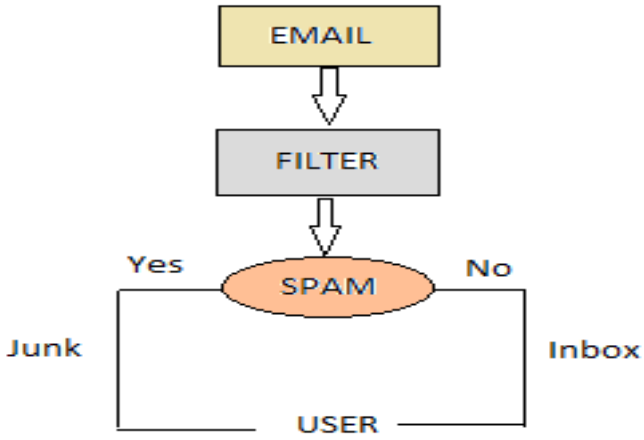
actual class, may cause wrong or inaccurate classification i.e. it may give wrong classification.
Steps observed to classify text with Naïve Bayes classifier using association rule are:

- Each abstracts which are used to train is suppposed as a transaction in the text data.
- The text data is cleaned further by removing unnecessary words i.e. text data is filtered and related to subject words are collected likewise whitespaces. Association rule mining is applied to the set of transaction data where each frequent interval.
- Word set is what from each abstract is considered as a single transaction.
- A large word set is generated or created with their occurrence frequency determined in training.
- After that, Naïve Bayes classifier is used for probability calculation for spam words.
- Before classifying a new document the text data (abstract), target class of which is to be determined as training data set, is again pre-processed similar to the process applied to training data repeatedly.
- Frequent words are viewed as word sets for better result.
- In the list of word sets matching words set(s) or its subset (more than one) collected from training data with that of subset(s) of frequent word set of new document is searched further.
- The corresponding probability values of matched word set(s) for each target class are collected and result of probability is calculated.
- Last of all the probability values for each target class from Naïve Bayes classification algoritm are calculated and the corresponding class of a new document is determined to generate final result value.

III.FIGURES AND TABLE

List of figures:

1. **Block diagram**



IV. IMPLEMENTATION MODULE

a) User Module:

. First user have to register to create his /her own account with username and password.

.After that user can login into the system.

.User can send or receive the mails to or from another registered users.

b) Admin Module:

.Authentication .

. Store the message in database.

. Apply filtering.

Then fetch data.

c)Send Mail :

.Registered users can send mail to each other.

d) Bayes Classifier:

.Calculate product of all feature probabilities

.Calculate probability that the features can be classified as the category given.

.Feature is classified as the available category.

$$P(c/d) = P(c).P(d/c) / P(d)$$

e)**Classification Module:**

.Maintain classified feature set

. And then calculate probability.

. Depending upon probability new classification is constructed.

f)Blocking :

In this modules, the mails which are classified as a spam mails are blocked.

V. CONCLUSION

.In this project we have build software system which can be efficiently separate out e-mails as ham or spam making appropriate use of Naive-bays. This work promises to enhance the spam filtering domain in future.

VI. ACKNOWLEDGEMENT

We are glad to present the preliminary project report on ‘Ham or spam? A content based classification for Email filtering’. I would like to thank my internal guide Prof. Mrs. D.S.Zingade for giving me all the guidence and help when we needed. I am really grateful to them for their such kind support. Their valuable suggestions were very helpful. I am also grateful to Prof. S.N.Zaware, Head of Computer Engineering Department, All India Shri Shivaji Memorial Society’s Institute of Information Technology for her indispensable support, suggestions on time. In the end our

special thanks to Prof. S.P.Pimpalkar for providing various resources such as laboratory all with needed software platforms, Internet connection, for Our Project. Badhe Pooja Durge Vrushali Jangale Manisha Shirsath Rohini (B.E. Computer Engg.)

REFRENCES

[1] Provost, J. Nave-Bayes vs. Rule-Learning in Classification of Email.

[2] Miszalska, I., Zabierowski, W., Napieralski, A. Selected Methods fors Spam Filtering in Email.

[3] Zhang L., Zhu J. ”An evaluation of statistical spam filtering techniques”

[4] Chan, T. Y., Ji, J., Zhao, Q. Learning to Detect Spam: Naive-Euclidean Approach.

[5S. Dumais,] M. Sahami, D. Heckerman and E. Horvitz. A Bayesian Approach to Filtering Junk E-Mail. In Learning for Text Categorization.

[6] A. Perkins. The Classification of Search Engine

[7] Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques, Second Edn.

[8] Lam H.-Y. and Yeung, D.-Y.”A Learning Approach to Spam Detection based on Social Networks”

[9] Youn, S., McLeod, D. A comparative study for email classification.

[10] Khorsi A.,”An Overview of Content-Based Spam Filtering Techniques”

