Text and Image Based Spam Email Classification using KNN, Naïve Bayes and Reverse DBSCAN Algorithm

Anirudh Harisinghaney, Aman Dixit, Saurabh Gupta, Anuja Arora
CSE/ IT department,
Jaypee Institute of information Technology, Noida, India
anirudh.singhaney@gmail.com, aman_dix@yahoo.co.in , anuja.arora@jiit.ac.in

Abstract—Internet has changed the way of communication, which has become more and more concentrated on emails. Emails, text messages and online messenger chatting have become part and parcel of our lives. Out of all these communications, emails are more prone to exploitation. Thus, various email providers employ algorithms to filter emails based on spam and ham. In this research paper, our prime aim is to detect text as well as image based spam emails. To achieve the objective we applied three algorithms namely: KNN algorithm, Naïve Bayes algorithm and reverse DBSCAN algorithm. Pre-processing of email text before executing the algorithms is used to make them predict better. This paper uses Enron corpus's dataset of spam and ham emails. In this research paper, we provide comparison performance of all three algorithms based on four measuring factors namely: precision, sensitivity, specificity and accuracy. We are able to attain good accuracy by all the three algorithms. The results have shown comparison of all three algorithms applied on same data set.

***Key words*— Spam, Ham, KNN, Naïve Bayes, reverse DBSCAN, Image Spam**

I. INTRODUCTIONT

As number of internet users is increasing day by day, more people are finding email communication an inexpensive way to send their data and communicate with their peers. With pros also come some cons. Almost every website ask for email id so as to complete their registration, thus making users more and more prone to get affected by the spam mails. This is evident from the fact that spam emails have accounted for 68.8% of all email traffic in 2012[1].

The increasing numbers of spam emails not only wastes one's time but also wastes network resources significantly. Most importantly they expose users to scams such as phishing and virus attacks.

Spammers have now gone a step ahead and to prevent spam filters from detecting their mails, images containing the spam text are sent. This has increased the burden to detect these manifold spam emails. Thus, a solution for this menace is imperative. Keeping in mind “ Spam is in the eye of the recipient” approach, this paper proposes email spam filtering based on three algorithms-KNN, Naive Bayes and Reverse DBSCAN along with their accuracies.

The remainder of this research paper is organized as follows: related work is reviewed in subsequent section, section 2 is about the methodology used, followed by detailed description about the algorithms: KNN, Naive Bayes and Reverse DBSCAN, used for classification of spam emails based on text and image in section 4. Section 5 is about results and in last we concluded the work done.

II. RELATED WORK

Many researchers have earlier tried to solve this problem of spam filtering. The common approaches used by them are using of Support Vector Machines (SVM) [2], Bayesian classification or feature extraction. They have not applied dedicated pre-processing steps to identify spam mails. Pre-processing can help in improving results significantly.

Data mining plays an important role in separating spam mails from ham mails. Text classification is one of the text mining technologies, and is the basis of our work [3]. Only text mining isn't the solution, basic filtering techniques also help the cause that too faster. Some techniques are black listing and white listing [4].Using black lists and white lists can assist in blocking unwanted messages and allowing wanted messages to get through.

Black Listing: Black-listing is creating a list of domain names which are used by the spammers, when a mail comes from that specific domain which is black listed it is considered spam. No further processing is done.

White Listing: White list is a list of trusted domains and a mail from them is always ham. White listing is a method used to classify user's email addresses as legitimate ones.

But blacklisting and white listing is not always accurate. Therefore, to counter all these techniques employed by spam filters, spammers now send mails with embedded images containing the spam text. To extract the text out of these images is an arduous task. It must be done by sophisticated OCR tools and based on the high level, low level, and combination of both the features of image in a spam mail can be predicted [5].

We employed basic algorithms of data mining for the detection of spam mails. For this we only used the existing classifying algorithms like kNN & Naive Bayes but also developed and applied our own reverse DBSCAN algorithm. To detect spam mails containing images we employed Google's inbuilt open source OCR engine, 'Tesseract'[6, 10]. Tesseract is the one of the most accurate OCR engine. Tesseract is an open-source OCR engine that was developed at HP between 1984 and 1994. Tesseract began as a PhD research project in HP Labs, Bristol, and gained momentum as a possible software and/or hardware add-on for HP's line of flatbed scanners [6]. It is combined with the 'Leptonica Image Processing Library', and can read a wide variety of image formats and convert them to text.

978-1-4799-2995-5/14/$31.00©2014 IEEE

III. EMPIRICAL ANALYSIS

Many researches have been done in the field of spam detection and spammers have always had the upper hand. We have tweaked the algorithms so far proposed in research papers by applying them after some pre-processing in the database.

A. E-mail Dataset Used

Enron corpus datasets [7] have been used. Enron data set is a large set of email messages. The Enron corpus was made public during the legal investigation concerning the Enron Corporation. In the cleaned Enron corpus, there are a total of 200,399 messages belonging to 158 users with an average of 757 messages per user. But this is the one third the size of original corpus [8].

In our work, we picked very small set of Enron corpus data set. We picked out 2500 mails for training and another 2500 mails for testing our algorithms.

B. Pre-Processing

We maintain a database of all the words that occur in each mail with the frequency of the word stored in each column. So we converted them to their root form first by applying Porter Stemmer algorithm. Some steps of this algorithm are:

- Remove the plurals and –ed or –ing suffixes.
- Turn terminal y to i when there is another vowel in the stem.
- Deal with suffixes, -full, -ness etc.
- Take off suffixes -ant , -ence, etc.

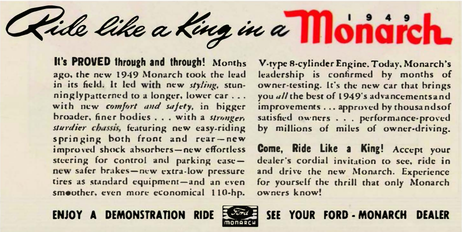
After we have prepared our database with the stemmed words, with each mail name in one column and the frequency of occurrence of words in other we move on to next phase.

C. Black listing and White listing

All those web pages and domains that are notorious for sending spam mails and are not trusted; go on the list of black list [9]. Thus, if a domain that matches from this list, the mail is predicted spam without any further processing. Further, spam is in the eye of the recipient, so a white list is maintained where users can mark those websites they want mails from whether they send “spam” or not. Thus no processing is done when a white listed domain matches.

D. Extracting words from Image

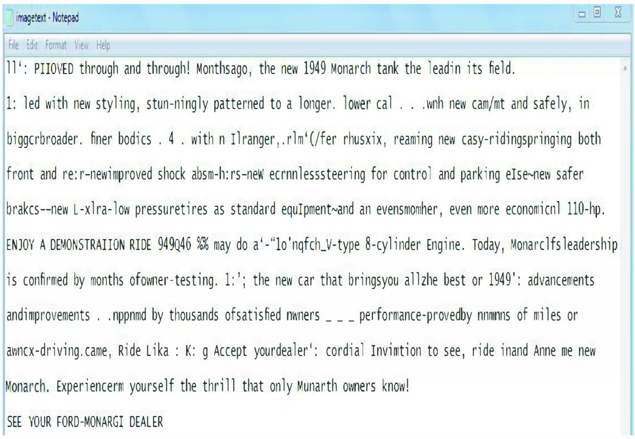
Users have an option of attaching image to their mails. The image is passed through the google's open source library Tesseract, and words are extracted from it. These words then pass through our different algorithms to predict our mail as spam or ham. Optimum accuracy is achieved for a clear resolution image and more popular fonts like Times New Roman as shown in figure 1(a) and figure 1(b). Captcha images are hard to detect.


Ride like a King in a 1949 Monarch

It's PROVED through and through! Months ago, the new 1949 Monarch took the lead in its field. It led with new *styling*, stunninglypatterned to a longer, lower car . . . with new *comfort and safety*, in bigger broader, finer bodies . . . with a *stronger, sturdier chassis*, featuring new easy-riding springing both front and rear—new improved shock absorbers—new effortless steering for control and parking ease—new safer brakes—new extra-low pressure tires as standard equipment—and an even smoother, even more economical 110-hp. V-type 8-cylinder Engine. Today, Monarch's leadership is confirmed by months of owner-testing. It's the new car that brings you *all* the best of 1949's advancementsand improvements . . . approved by thousandsof satisfied owners . . . performance-proved by millions of miles of owner-driving.

Come, Ride Like a King! Accept your dealer's cordial invitation to see, ride in and drive the new Monarch. Experience for yourself the thrill that only Monarch owners know!

ENJOY A DEMONSTRATION RIDE Ford MONARCH SEE YOUR FORD - MONARCH DEALER



```
11': PIIOVED through and through! Monthsago, the new 1949 Monarch tank the leadin its field.

1: led with new styling, stun-ningly patterned to a longer. lower cal . . .wnh new cam/mt and safely, in

biggcrbroader. finer bodics . 4 . with n Ilranger,.rlm'(/fer rhusxix, reaming new casy-ridingspringing both

front and re:r-newimproved shock absm-h:rs-neW ecrnnlesssteering for control and parking eIse~new safer

brakcs--new L-xlra-low pressuretires as standard equIpment~and an evensmomher, even more economicnl 110-hp.

ENJOY A DEMONSTRAIION RIDE 949Q46 %% may do a'-“lo'nqfch_V-type 8-cylinder Engine. Today, Monarclfsleadership

is confirmed by months ofowner-testing. 1:'; the new car that bringsyou allzhe best or 1949': advancements

andimprovements . .nppnmd by thousands ofsatisfied nwners _ _ _ performance-provedby nnmnns of miles or

awncx-driving.came, Ride Lika : K: g Accept yourdealer': cordial Invimtion to see, ride inand Anne me new

Monarch. Experiencerm yourself the thrill that only Munarth owners know!

SEE YOUR FORD-MONARGI DEALER
```

Fig. 1.(a) Embedded Image in a SPAM Email (b) Text extracted from Embedded Image from Spam E-Mail

IV.ALGORITHMS APPLIED FOR SPAM EMAIL CLASSIFICATION

A. *K- Nearest Neighbour or KNN algorithm*

The K-Nearest Neighbour algorithm is similar to the Nearest Neighbour algorithm, except that it looks at the closest K instances to the unclassified instance. The class of the new instance is then given by the class with the highest frequency of those K instances. We are choosing K by trial and error method, for which we obtain the optimal result. The proximity is calculated by finding the Euclidean distance i.e.

Euclidean $$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

We calculate the proximity of the users mail from our database of mails where k=20. Thus from the majority of the 20 mails, we predict a mail spam or ham. KNN gives a better accuracy than many algorithms, but it has a higher complexity as proximity from each mail is calculated.

B. Naive Bayes classification:

$$P_f = \frac{\frac{s}{t_s}}{\frac{s}{t_s} + \frac{kn}{t_n}}$$

Combine the probabilities of the N most interesting features using Bayes theorem, and closer the P is to 0, the more likely the message is non spam and the closer P is to 1, the more likely message is spam

$$P = \frac{P_{f_1}P_{f_2}P_{f_3}...P_{f_N}}{P_{f_1}P_{f_2}P_{f_3}...P_{f_N} + (1-P_{f_1})(1-P_{f_2})(1-P_{f_3})...(1-P_{f_N})}$$

C. ProposedAlgorithm- Reverse DBSCAN

DBSCAN is a density-based spatial clustering of applications with noise. It is a density based clustering algorithm because it finds a number of clusters starting from the estimated density distribution of corresponding nodes. We proposed a different approach that somewhat is reverse of DBSCAN.

As DBSCAN algorithm would separate objects into cluster, by calculating the distances, our approach takes cluster of ham and spam already defined. So when a mail comes its distance from each of this cluster is calculated. More the proximity better would be the chances of it being a spam or ham. We calculate the distance using the Euclidean formula as used in KNN.

V. RESULT

Accuracy of our algorithms with and without pre-processing is shown in table1. We are validating results with the help of following research questions:

RQ1: Can pre-processing helps in improving the result as compared to using data directly?

As shown in Table1, we are able to achieve almost 50% better accuracy using pre-processed data as compared to accuracy achieved using with pre-processed data in all three algorithms. KNN with pre-processing data getting 83% accuracy in text and image based spam filtering as compared to 45% which was without pre-processed data. Similarly, Using Reverse DBSCAN, we are attaining 74% accurate result using pre-processed data as compared to 48% accuracy without pre-processed data. And finally best accuracy achieved by Naive Bayes algorithm which is 87% accurate result which was just 47% without pre-processed data.

Hence, we are able to increase accuracy of result amazingly by applying defined pre-processing steps.

RQ2: Which algorithm is able to achieve better results?

We have achieved our best accuracy from Naive Bayes algorithm that is close to 87%.

TABLE I. ACCURACY FOR DIFFERENT ALGORITHMS

Algorithms	**Precision**	**Sensitivity**	**Specificity**	**Accuracy**
KNN (without pre-processing data)	0.1056	0.3191	0.4783	0.4526
KNN (with pre-processed data)	0.9538	0.7713	0.9429	0.8392
Reverse DBSCAN (without pre-processed data)	0.9739	0.4853	0.3275	0.4823
Reverse DBSCAN (with pre-processed data)	0.7379	0.7374	0.7490	0.7433
Naive Bayes (Pre Processed Data)	0.7445	0.4962	0.5163	0.4772
Naive Bayes (Processed Data)	0.8295	0.7919	0.9104	0.8683

VI. CONCLUSION

In this paper we applied three different algorithms to detect spam mails, one being a new approach to spam detection. We adapted the spam filter to each user's preferences and predicted a mail spam or not by text mining and text recognizing by OCR library TESSERACT.

Although methods used by us have many advantages, it certainly does come with some disadvantages. The disadvantage of text filtering is that they are time consuming. The OCR based detection also has disadvantages like, the recognition is not always perfect, and works for certain fonts only, cannot predict for CAPTCHA images and obviously are expensive. In future, as it is an adaptable and scalable project thus we would like to detect threats found in emails that are viruses.

REFERENCES

[1]. http://royal.pingdom.com/2013/01/16/internet- 2012-in-numbers/

[2]. N. Nhung and T. Phuong. "An Efficient Method for Filtering Image-Based Spam E-mail". Proc. IEEE International Conference on Research, Innovation and Vision for the Future (RIVF07), IEEE Press, Mar. 2007, pp. 96-102. doi: 10.1109/RIVF.2007.369141.

[3]. Liu, G., & Yang, F. (2012, August). The application of data mining in the classification of spam messages. In Computer Science and Information Processing (CSIP), 2012 International Conference on (pp. 1315-1317). IEEE.

[4]. A. Ramachandran and N. Feamster. Understanding the network-level behavior of spammers. In ACM SIGCOMM, 2006

[5]. Ketari, Lamia Mohammed, Munesh Chandra, and Mohammadi Akheela Khanum. "A Study of Image Spam Filtering Techniques." Computational Intelligence and Communication Networks (CICN), 2012 Fourth International Conference on. IEEE,2012.

[6]. R. Smith, "An Overview of the Tesseract OCR Engine," in Proc. International Conference on Document Analysis and Recognition, 2007

[7]. Enron corpus database http://archive.ics.uci.edu/ml/datasets/ Spambase

[8]. Klimt, Bryan, and Yiming Yang. "The enron corpus: A new dataset for email classification research." Machine learning: ECML 2004. Springer Berlin Heidelberg, 2004. 217-226.

[9]. Blacklist database collected from internet. http://www.blacklistalert.org/, http://www.joewein.de/sw/blacklist.htm, http://www.sdsc.edu/~jeff/spam/Blacklists_Compared.html

[10]. Breuel, Thomas M. "The OCRopus open source OCR system." DRR 6815 (2008): 68150.