International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 04 | Apr-2018 www.irjet.net p-ISSN: 2395-0072

Spam Filter Classification

Smriti Bhat[1], Varsha Rao[2], Sheetal Pamecha[3], Tanvi Zakir Hussain[4], Anitha R[5]

[1,2,3,4] *Student, Dept. of CSE, NIE, Karnataka, India*
[2] *Associate Professor, Dept. of CSE, NIE, Karnataka, India*

Abstract-*Naive Bayes classifiers are most extensively used to filter spam emails. Their performance in accurately identifying spams is primarily limited due to assuming independence between the features. Another machine learning based algorithm, known as Support Vector Machine models the situation by creating a feature space. A feature space is a finite-dimensional vector space, each dimension of which represents a "feature" of a particular object are usually used for classification. We implement a system combining the approaches of both SVM and NB, to form the SVM-NB algorithm. This system is evaluated on the dataset obtained from Lingspam. A comparative study of the algorithms is made, based on the parameters f1 score and computational time.*

***Index Terms*- Naive Bayes, Support Vector Machine, Spam Filtering, Classification, Comparison, Stacking.**

1. INTRODUCTION

Communication is the basis for interactions between all species on this planet. In today's world, communication through internet, primarily Email, is the most important form of communication. The total number of business and consumer emails sent and received per day has reached nearly 269 billion and is expected to continue to grow at an average annual rate of 4.4% over the next four years. These statistics highlight the importance of emails[1].

The average number of email accounts per user is 1.9. Approximately half of the worldwide population uses email in 2017 [1]. A major problem or threat posed to email users is spam. Statisticians say 94 billion spam threats are sent each day throughout the world. In terms of a worldwide figure, £1.6 billion is lost in productivity annually due to unsolicited emails [2]. These unsolicited mails affect productivity cause inconvenience to the recipients and further lead to financial losses. Hence it is essential to filter and separate spam emails from the legitimate ones.

Mailing is the most cost effective form of commercial advertising, which in turn is the primary contributor of spam mails. Spamming harms communication by overwhelming the recipient's system with a vast amount of unwanted communication, which further consumes time of the recipient and thereby decreasing the productivity. Spam emails, now account for over 80% of all email traffic. According to IBM 2017 X-Force researchers[3], It is not just the number of spam emails that's rapidly increasing, it is also the number of spam emails containing malicious attachments which is dramatically increasing.

Various spam filtering techniques and algorithms such as the blacklist and whitelist filtering techniques [4], decision tree based approaches [5], [6], and machine learning based methods [7], [8] are currently available. Amongst these solutions, machine learning based algorithms provide high accuracy in detecting spams. Most of the existing machine learning based solutions are based on either the support vector machine (SVM) or Naive Bayes (NB) methods.

Combining the characteristics of both SVM and NB methods, we implement an innovative spam filtering algorithm, namely Support Vector Machine - Naive Bayes (SVM-NB)[3]. This algorithm first makes attempt to classify training samples/emails using the original NB algorithm. Then, it constructs a hyperplane in the space containing all training samples so that the training set is divided into two parts. For the samples around the hyperplane, their corresponding classification results (i.e., whether spam or not) will be further checked. For a particular sample, if its nearest neighbor (with the shortest Euclidean distance in the sample space) has a different classification result, this sample will be rejected. This refined training set will be used by the NB algorithm to learn the classification model and thus detect spam emails. Based on the dataset provided by Lingspam, we carried out intensive experiments and found that the proposed SVM-NB system outperforms benchmark algorithms, NB-based algorithm and is equivalent to SVM-based algorithm, in terms of f1 score.

2. PRELIMINARIES

The rest of this paper is organized as follows. In section 2, we introduce the related work. In section 3, we briefly discuss the preliminaries. In section 4, we discuss the proposed system. We tell about the implementation of the SVM-NB with the graphs showing

3. RELATED WORKS

Over the last decade, several studies have been conducted on spam filtering. Most of the Internet service providers (ISPs), as well as email service providers, provide junk email detection and filtering service. Many algorithms have been proposed for classifying spam and legitimate emails. These classification strategies are categorised into groups:

In the first group, spam emails are detected based on sender's' identifications [4]. Blacklisting is a technique that identifies IP addresses that send large amounts of spam. These IP addresses are added to a Domain Name System-Based Blackhole List and future email from IP addresses on the list are rejected. In a white list based email system, user marks and adds the IP address into the white list. The emails from the address other than the ones in the whitelist are treated as spam or junk. One limitation of this type of approach is that the sender may change his/her identity by using dynamic IP, IP spoofing techniques and IP proxy . Many more such as detecting bulk emails, scanning message headings, etc. The success ratio of machine learning algorithms over non-machine learning algorithms is much more.

Second group is decision tree based technique, rule-based approach. The earliest decision tree based learning system was developed by Hunt, dating back to 1966 [5].

The last group among which we will be discussing some in this column are Machine Learning Based Technique, Support Vector Machine[7][8], Multi-Layer Perceptron, Naive Bayes Algorithm and many more. Using Machine Learning techniques, data can be extracted from the already existing set of emails then the obtained information is used to classify the newly received emails.
Classification principle, efficiency and accuracy. We will first discuss the basic principle and then propose the combined version of both the algorithms.

Classification Algorithms - Classification models are those which draw some conclusion from the observed values. There are two types of classification techniques, supervised and unsupervised. Dataset with labels is fitted into supervised model and dataset without label is fitted into unsupervised model.

Naive Bayes - The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. Bayes theorem provides a way of calculating the posterior probability, P (c|x), from P(c), P(x), and P (x|c). Naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence [9].

Support Vector Machine - In machine learning, support vector machines are supervised learning models with associated learning algorithm that analyze data and recognize patterns, used for classification and regression analysis. But it's usually used for classification. Given two or more labeled classes of data, it acts as a discriminative classifier, formally defined by an optimal hyperplane that separates all the classes. New examples that are then mapped into that same space can then be categorized based on on which side of the gap they fall.

4. PROPOSED SYSTEM

Comparison of support vector and naive bayes is to be made on the basis of the parameters such as f1 score and computational time. The proposed system takes the advantages of both SVM and NB methods, an innovative spam filtering algorithm, namely support vector machine - naive Bayes (SVM-NB)[10]. The SVM-NB algorithm first makes attempt to classify training samples/emails using the original NB algorithm. Then, it constructs a hyperplane in the space containing all training samples so that the training set is divided into two parts.
Preprocessing:

Document preprocessing is the process of absorbing a new text document into text classification system. It used to represent the document efficiently by removing useless keywords which further improves retrieval performance. It consists of following stages:

a. Lexical analysis
b. Stop word elimination
c. Lemmatization

a Lexical analysis

The functionality of lexical analyzer is to extract keywords from text document by using tokenizer. Words are determined from text document by separating the input alphabet into characters (the letters a-z) and separators (space, newline, tab). Digits and punctuation marks are removed since they are insignificant in making decision in text classification.[11]

b. Stop Word Elimination

Stop words, such as articles, prepositions, conjunctions, pronouns and possibly some verbs, nouns, adverbs, are removed from text document to improve the performance of text classifier.

c. Lemmatization

Morphological variants of words have similar semantic interpretations and are considered to be equivalent. Stemming Algorithms are used to reduce a word to its root form. For example, the words "writing", "written", and "write" are reduced to the root word, "write".

Feature Selection:

Feature selection is classic refinement method in classification. It is an effective dimensionality reduction technique to remove features that are considered irrelevant for the classification [12]. Feature selection is necessary to make large problems computationally efficient. Further, well-chosen features can improve classification accuracy substantially, or equivalently, reduce the amount of training data needed to obtain a desired level of performance [13].

Classifier - Using the combination of Naive Bayes and Support Vector we will classify the emails as ham or spam. We will look into the implementation part in next section.

5. SPAM FILTERING WITH SVM-NB

Naive Bayes has the limitation of assuming independent feature vectors from training dataset. SVM overcomes this limitation as it tries to maximize the distance between the boundaries of different categories.

Taking SVM and NB as base models and again taking NB as the meta model. SVM eliminates the wrongly classified samples. The predictions of the base models are used to form a new dataset. The meta model is fitted on this new dataset and prediction is based on the new dataset.

The algorithms were implemented in python using scikit-learn machine learning library. The graphs were plotted using matplotlib plotting library.

The dataset is randomly split into three parts, namely training, validation and testing dataset. They are in the ratio of 50:20:30. Total 2892 mails are used for implementation.

We use the following parameters to perform the comparative analysis of the three algorithms.

The F measure (F1 score or F score) is the average of recall and precision.

$$F1 = 2 * \frac{\text{(precision * recall)}}{\text{(precision + recall)}}$$

Precision is the ratio of correctly predicted positive observations to all positive observations. It is also referred to as positive predictive value.

Recall rate is also known as sensitivity and true positive rate. It is the ratio of correctly predicted positive observations to all the observations in the actual class.

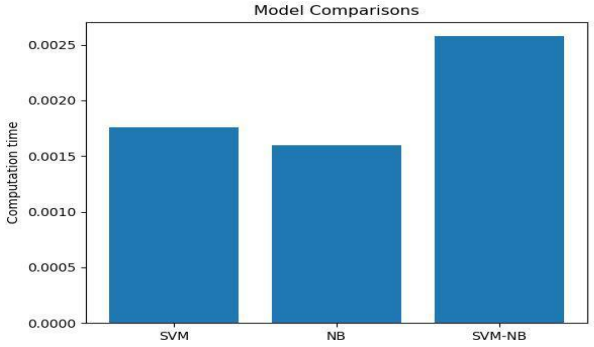

Figure 1 Comparison of computational time of models.

In figure 1 the computation time for naive bayes is the least. SVM-NB has the highest computation time. Even when dataset size increases the computation time of NB and SVM-NB are least and highest respectively (table 1). The computations were performed on a system with 16 GB RAM and Intel Core i7 2.40 GHz CPU.

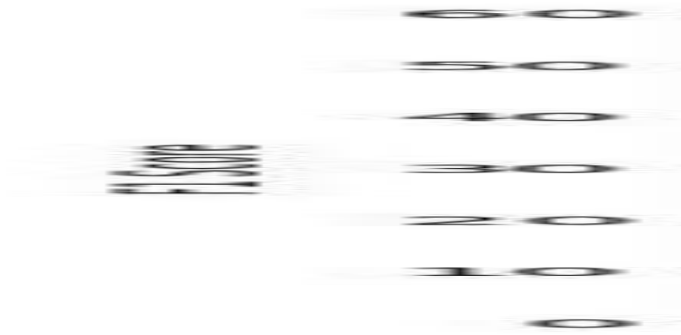

Figure 2. Comparison of F1 Score of different models.

In figure 2 we observe f1 score for SVM-NB is the best. F1 score increases with the dataset size for all the models, according to table 1. As the dataset size is increased the f1 score of SVM-NB is slightly better than SVM.

Table 1 Comparison table of different models based on parameters.

Model	Dataset Size	F1 score	Computational time
SVM-NB	22	0.57143	0.0026
SVM		0.42857	0.0018
NB		0.28571	0.0016
SVM-NB	2892	0.85294	12.9810
SVM		0.83527	10.1081
NB		0.75038	0.3985

CONCLUSION

We did a comparative analysis of NB, SVM and SVM-NB. SVM and SVM-NB provides better results with large data set. Computational time for Naive Bayes is least among the three.

FUTURE WORK

To achieve better results k-folds technique can be used and implemented on large dataset.

REFERENCES

[1] Email Statistics Report, 2015-2019. Ed. Sara Radicati. The Radicati Group, Inc https://www.radicati.com/wp/wp/content/uploads/2017/01/Email-Statistics-Report-2017 2021-Executive-Summary.pdf

[2] Fastnet SA https://www.mailcleaner.net/blog/spam-world-ne ws/how-much-does-spam-cost-the-world/

[3] https://securityintelligence.com/ibm-x-force-iris-uncovers-active-business-email-compromise-campaign-targeting-fortune-500-companies/

[4] J. W. Yoon, H. Kim, and J. H. Huh, "Hybrid spam filtering for mobile communication," computers & security, vol. 29, no. 4, pp. 446–459, 2010.

[5] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Transactions on knowledge and data engineering, vol. 21, no. 9, pp. 1263–1284, 2009.

[6] S. Ruggieri, "Efficient c4. 5 [classification algorithm]," IEEE transactions on knowledge and data engineering, vol. 14, no. 2, pp. 438–444, 2002.

[7] B. Sch Ikopf, S. Mika, C. Burges et al., "Input space versus feature space in kernel-based method," IEEE Trans Neural Networks, pp. 1000–1017.

[8] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," Machine learning, vol. 29, no. 2-3, pp. 131–163, 1997.

[9] A. McCallum, and K. Nigam, A comparison of event models for naïve Bayes text classification, in Proceedings of AAAI-98 workshop on learning for text categorization, pp. 41–48, 1998.

[10] Weimiao Feng, Jianguo Sun, Liguo Zqing Yanghang, Cuiling Cao proposed "A Support Vector Machine based Naive Bayes Algorithm for Spam Filtering"-2016

[11] J. F. Pang, D. Bu and S. Bai , "Research and Implementation of Text Categorization System Based on VSM,"Application Research of Computers, 2001.1

[12] P. Domingos, A few useful things to know about machine learning, Communications of ACM, vol. 55, no. 10, pp. 78–87, 2012.

[13] G. Forman, An extensive empirical study of feature selection metrics for text classification, Journal of Machine Learning Research, vol. 3, pp. 1289–1305, 2003.