A Comparative Study for Spam Classifications in Email Using Naïve Bayes and SVM Algorithm

Ziyan Mohammed
Bearys Institute of Technology
Mangalore, India

Mohammed Farhaz J A
Bearys Institute of Technology
Mangalore, India

Mohammed Irshad M P
Bearys Institute of Technology
Mangalore, India

Mustafa Basthikodi
Dept.of CSE
Bearys institute of Technology
Mangalore, India

Ahmed Rimaz Faizabadi
Dept.of CSE
Bearys institute of Technology
Mangalore, India

Abstract— The Email has become a form of communication that's very reliable and people tend to use it to communicate for various purposes. Almost everyone among us who are into the technical world has an email address. Spams are no strange emails that is happened to almost all of us. To classify the Emails as spam or not spam we use Naïve Bayes and Support Vector Machine algorithms and then calculate their accuracy, precision, recall and F-measure. After getting these data we will compare them to know which algorithms are better performer in classifying the spam emails.

Keywords— Spam, Machine Learning Algorithms, Email Classification, Naïve Bayes, Support Vector Machine.

I. INTRODUCTION

The spam is the unwelcomed guest which is known to be unsolicited emails, junk emails or the illegal emails that is sent to user's emails without their consent. Now a days it's rare not to see a spam in our emails if we are not using any spam filtering options. Sometimes it may cost user's a lot if the user fall prey to the attackers who asks for credit card or other banking details by pretending to be a manager of the bank or someone known to be working for a particular type of banks. Email classification here involves the classification of both spam and not spam emails which is the main concern in avoiding spam emails. After classifying spam and not spam emails users can delete and even block those spammers from sending the emails to avoid that particular spammers from contacting the user. Mainly we are going to compare two techniques implemented by other people where one group of them [1] used Naïve Bayes Algorithm and other group [2] used Support Vector Machine Algorithms to detect the spam. After which the comparison is made based on the accuracy, recall, precision and F- measure.

II. EMAIL SPAM FILTERING

The emails which the users receive can contain spam which is completely useless in any way. It wastes users time, energy and many other basic things. To classify the spam emails and not spam emails we make use of two machine learning algorithms such as Naïve Bayes and SVM. These algorithms help us to filter the spam emails after classifying them individually. Since it's a comparative study we will use the data from the journals [1] and [2] which can be found in the references section.

III. CLASSIFICATION ALGORITHMS

A. Naïve Bayes Algorithm According to IBM, "The Naive Bayes classification algorithm is a probabilistic classifier. It is based on probability models that incorporate strong independence assumptions." [3] A Naive Bayes Classifier is a supervised machine-learning algorithm that uses the Bayes' Theorem, which assumes that features are statistically independent. [4] To calculate the probability that the email is spam or not spam the Naïve Bayes Algorithm uses the formula which is give below.

$$P(spam \mid word) = \frac{P(spam).P(word \mid spam)}{P(spam).P(word \mid spam) + P(not-spam).P(word \mid not-spam)}$$

Where the P(spam|word) is the probability that the email contains a particular word, which implies that the email is spam.

The P(spam) is the probability that the any of given email is a spam.

The P(word|spam) is the probability that the word which is known to be of spam present in the given email.

P(not-spam) is the probability that any of the particular word is not spam.

And P(word|not-spam) is the probability that the email contains a particular word, which implies that the email is not spam.

In the journal [1] we found that they have used three phases to achieve the objective of classifying the spam and non-

spam emails using Naïve Bayes Algorithm. The three phases are given below.

1. Pre-processing
2. Feature Selection
3. Naïve Bayes Classifier.

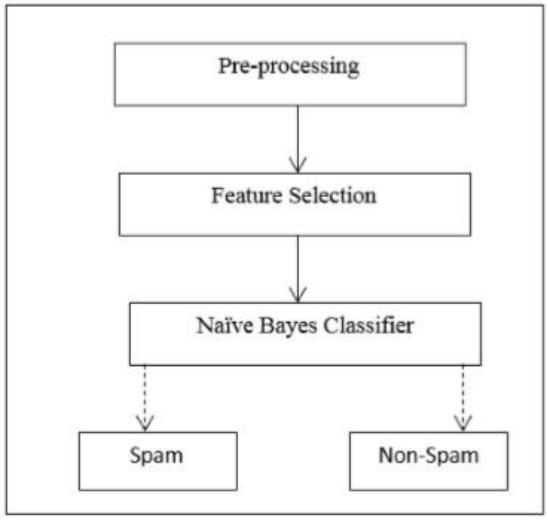

Figure 1. Process of E-mail spam filtering (From Nurul Fitriah Rusland et al 2017. [1])

The first step here is the Pre – processing step where the Pre – processing of E-mails is done where typical conjunction words and articles are removed for training filter. They have used the WEKA tool for facilitate of the experiments. In the second step they have applied the feature selection algorithm, in this case the algorithm which they have used was the Best first feature selection algorithm.[5] In the last and the final step, they used Naïve Bayes Classifier to classify the email to be spam or not-spam. And then they have found the accuracy, recall, precision and F-measure of two data sets namely spam data and SPAMBASE. The Spam Data [6] was used at first to test performance of the spam filter which is based on the Naïve Bayes algorithm. That dataset contained 9324 E-mails and 500 attributes. That dataset was obtained by them from the Usenet posts that existed in twenty Newsgroup collections and collect from lots of account e-mails which was located on various different e-mail servers. The second dataset, SPAMBASE was originally taken from the UCI machine learning repositories [7] and was actually created by Jaap Suermondt, George Forman, Erik Reeber, and Mark Hopkins. That dataset is known to have contained 4601 E-mail messages and 58 attributes. The SPABASE dataset collection of the non-spam email came from the personal e-mail, filled work and single e-mail account which is suitable to detect whether the email is spam or not. And also, it had 58 attributes. The accuracy, recall, precision and the F-measure can be calculated by using below formulas.

$$Accuracy = \frac{TN + TP}{TP + FN + FP + TN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$and\ the\ F - measure = 2.\frac{Precision.Recall}{Precision + Recall}$$

B. Support Vector Machine Algorithm

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labelled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two-dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side. [8] Here we have used the data from [2] journals where they have classified email as spam or not spam using another type of classification algorithm which is known as Support Vector Machine Algorithm.

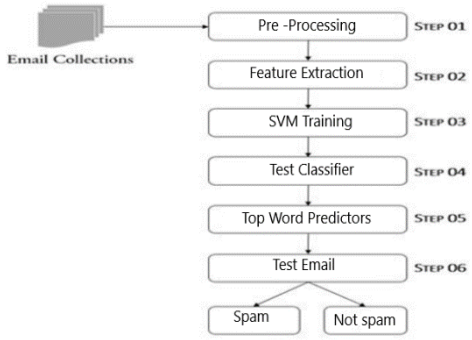

Figure 2: Workflow for Email Spam Classification. Found in [2]

This algorithm is different from the Naïve Bayes Algorithm. Their methodology (as in figure 2) has five steps which is given in detail below.

1. The pre-processing step was utilized to expel the noises from the email which are irrelevant and require not to be available. The pre-processing step incorporates a) Removal of Numbers b) Removal of Special Symbol c) Removal of URLS d) Stripping HTML e) Word Stemming.
2. Feature Extraction was utilized to separate the essential and important features from the email body. The feature transforms the email into 2D vector space having features number.
3. In the SVM Training step the email spams were utilized for the training necessity. The training dataset include content of spam and classifier were prepared utilizing it. Subsequent to training, the classifier was prepared to classify the spam emails.
4. The classifier was tested in the fourth step which is Test Classifier step with various training information to test the accuracy of the classifier.
5. In the fifth step which is Test Email step where after the training stage was finished, an example email was given as input to the classifier to characterize the email. The classifier produces output in the forms of 0 or 1, 1 implies it is spam and 0 implies it is not a spam.

IV EXPERIMENTAL RESULTS

Finally, after going through the journals [1] and [2] we found the accuracy of Naïve Bayes Algorithm and SVM Algorithm in classifying the spam and non-spam emails. But the accuracy is not sufficient in machine learning algorithms, the recall, precision and F-measure is also important. [9] So we referred another journal [10] for it. After learning through these three journals [1], [2] and [3] we found the below results.

Naïve Bayes Algorithm	Accuracy	Precision	Recall	F-Measure
Spam Data	91.13%	83%	83%	60%
SPAMBASE	82.54%	88%	86%	77%

Table 1: Accuracy, Precision, Recall and F-measure of two datasets, the Spam Data and SPAMBASE (from [1])

Algorithm	Accuracy	Precision	Recall	F-Measure
Naïve Bayes	97.44%	89.58%	92.06%	90.53%
SVM	98.3%	100%	87.7%	93.45%

Table 2: Comparison of NB and SVM algorithm in terms of Accuracy, Precision, Recall and the F-measure. (Obtained from [11].)

V. CONCLUSION

From the experimental results we came to know two main things regarding machine learning algorithms such as Naïve Bayes and Support Vector Algorithm. In the table 1, we used two datasets for Naïve Bayes Algorithm such as Spam Data and SPAMBASE. The accuracy is more if we use Spam Data but precision is what we should look for as it means it classifies the spam emails correctly if it is one hundred percent precise. In SPAMBASE, the precision is 88% which is more than the Spam Data. In the table 2, we came to know that the SVM algorithm is preferred over Naïve Bayes Algorithm as it is one hundred percent precise in [11] case for classifying non-spam emails but not always one hundred percent precise if we work on different datasets. Anyways no algorithms classify both spam and ham (not-spam) always without making a tiny mistake. So, the SVM algorithm is recommended as per my findings to classify the spam and not-spam emails as it outperforms Naïve Bayes Algorithm in precision.

REFERENCES

[1] Nurul Fitriah Rusland, Norfaradilla Wahid, Shahreen Kasim, Hanayanti Hafi 2017 IOP Conf. Ser.: Mater. Sci. Eng. 226 012091

[2] Shradhanjali, Prof. Toran Verma, "E-Mail Spam Detection and Classification Using SVM and Feature Extraction", International Journal of Advance Research, Ideas and Innovations in Technology, Rungta College of Engineering and Technology Dept. of Computer Science and Engineering Bhilai, Chhattisgarh, India.

[3] Naïve Bayes – IBM <https://www.ibm.com/support/knowledgecenter/SSLVMB_22.0.0/com.ibm.spss.statistics.cs/spss/tutorials/naivebayes_table.htm>

[4] Machine Learning Algorithms Explained - Naive Bayes Classifier, <https://blog.easysol.net/machine-learning-algorithms-4/>

[5] Rizky et al. "The Effect of Best First and Spread subsample on Selection of a Feature Wrapper with Naïve Bayes Classifier for Classification of Ratio of Inpatients". Scientific Journal of Informatics.

[6] Feng et al., "A Support Vector Machine based Naïve Bayes Algorithm for the spam filtering," (2016) IEEE 35th International Performance Computing and the Communications Conference, Las Vegas, NV, 2016.

[7] Tretyakov and K. "Machine Learning techniques in Spam filtering": A Data Mining problem - oriented Seminar at MTAT.

[8] Chapter 2 : SVM (Support Vector Machine)—Theory <https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72>

[9] Why accuracy alone is a bad measure for classification tasks, and what we can do about it. <https://tryolabs.com/blog/2013/03/25/why-accuracy-alone-bad-measure-classification-tasks-and-what-we-can-do-about-it/>

[10] Seongwook Youn and Dennis McLeod, "A Comparative Study for Email Classification", University of Southern California, Los Angeles, CA 90089 USA.

[11] Spam classification with Naive Bayes and Support Vector Machines. <https://www.kaggle.com/pablovargas/naive-bayes-svm-spam-filtering>.