EasyChair Preprint

№ 4107

Analysis of Machine Learning Algorithms for Email Classification Using NLP

Mithun Das, Hardik Patel and Subrata Samanta

August 30, 2020

Analysis of Machine Learning Algorithms for Email Classification Using NLP

Mithun Das
Cisco Systems Private Limited
Bangalore, India
mithun.rcciit@gmail.com

Hardik Patel
Cisco Systems Private Limited
Bangalore, India
hardpat2@cisco.com

Subrata Samanta
Accenture Solution
Bangalore, India
myemail.subrata@gmail.com

***Abstract*— In the exponentially growing world, people are using email across all areas of industries including Educational field. Therefore, it is very much important to differentiate between legitimate and spam email. In this paper, we have preprocessed emails using natural language processing and applied several machine-learning algorithms to analyze their performance on email classification. The performance observed here is accuracy and F_1 score. The result shows that ANN outperforms the other algorithms. The ANN best accuracy is 98.80% and F_1 score is 0.977778.**

Keywords—Natural Language processing; Machine learing; spam classification; emails

I. INTRODUCTION

Now a day, we have been using email in all areas of industries. Email is the cost effective medium of communication over the globe. We share our important documents as well as personal details including images via email. We send inform via email in no time. Managing incoming email is critical job to many, as via email we receive many information like work messages, invitation, admission letter to any school or college etc.

Sometimes we get emails that are not useful to us in Inbox folder, as well as sometimes we found important emails in Spam folder. Occasionally we get emails that seeks our personal details by faking you (i.e., you won a lottery of 10 crores, though you did not participate in any contest) from unknown source. Sharing details with them may risk you. These spam emails can cause many threats to network security [1]; it can steal your information (i.e., online banking details). It is found that 66.34% of total email traffic was spam during the first quarter of 2014 [2] and day by day they are increasing tremendously. So, improvement of classification between legitimate email and spam mail is still needed.

In [3], Authors have come out with a solution using Integrated Particle Swarm Optimization and Decision Tree, which shows 98.3% accuracy; there they have used a dataset of 4601 emails. In [4], Chae and Sasikumaran proposed a classifier based on context based email classification as mail algorithm complimented by information gain calculation to increase the accuracy of classifier. In [5], Author used fuzzy logic techniques for email clustering. Same keyword goes into same cluster and if a new word comes, a new cluster is formed for that. In [6] multi-level mail filtering scheme proposed which is based on NLP. It's worked on black and white filter list.

Here we will assessed several machine-learning classifiers that includes Random Forest, Decision Tree, Logistic Regression, Artificial Neural Network and K- Nearest Neighbors. We will be using Natural Language processing for preprocessing the dataset as well as for feature extraction. We have used python to implement these classifiers and found that ANN has better accuracy than other classifier.

The rest of the paper is described as follows. Section II describes different Classification Techniques. In section III, we have described Design and implementation of spam email classifier. In section IV, we analyze each classifier and finally we conclude our work in section V.

II. CLASSIFICATION TECHNIQUES

In this section, we have described various machine-learning algorithms we have used for email type classification.

A. Logistic Regression

Logistic regression is a probabilistic classifier of the form $p(y, \mathbf{X})$, where the value of y depends on $\mathbf{X}$, where $\mathbf{X} \in R^n$, n denotes number of independent features.

$$p(y, \mathbf{X}) = \frac{1}{1+e^{-\sum w_i x_i + \mathbf{b}}}, \quad \text{where } 0 \le y \le 1$$

Where w_i is the parameter for i^{th} feature and b is bias. Using training dataset, LR predicts these parameters, to obtain the maximum accuracy for new dataset.

B. K Nearest Neighbor Classifier

KNN Algorithm is a classification technique to classify the email based on training features space. Given a training set, classifier, classifies coordinates into groups identified by a feature. In the k-Nearest neighbor classification, an email class is identified by the majority votes of its neighbor.

C. Naïve Bayes

A Naïve Bayes classification technique, based on the assumption of Bayes theorem. It assumes that all the features are independent of each other, means they are unrelated to

each other. Parameter estimation for Naïve Bayes uses the method of maximum likelihood or using Bayesian methods.

However, the assumption that input features are independent to each other is unrealistic for real data, still this technique works well for complex problem.

D. *Decision Tree*

Decision Tree classifier makes its prediction based on some organized series of questions and condition in tree structure. Series of questions have been structured based on the dependent features (**X**). The leaf node of decision tree contains the decision/the-dependent output variable (y).

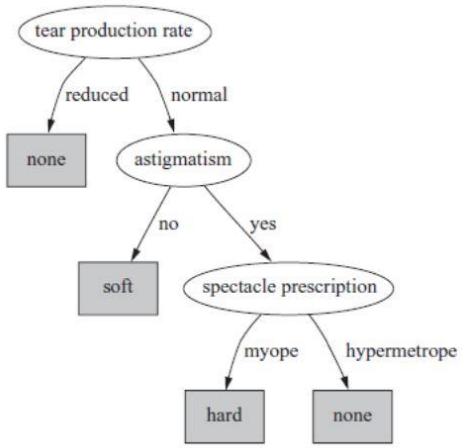

Figure 1: Decision Tree [9]

Decision Tree has a high variance and can predict the output more accurately when used in an ensemble. The Decision Tree example has shown in Figure 1.

E. *Random Forest*

Random Forest is an ensemble learning method for classification, where a large number of decision trees are created. RF classifier construct a multitude of Decision trees during training time and output the class that is majority of the classes of different decision Tree.

F. *Artificial Neural Network*

In neural network, there is three layers: input layer, hidden layer and output layer. In this paper, we have two (2) hidden layers with 256 activation units. ANN uses sigmoid for activation function.

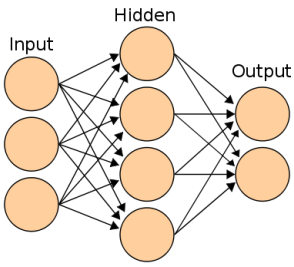

Figure 2: Artificial Neural Network

The learning algorithm used for this ANN is, forward propagation (FP) and backward propagation (BP). Example of an ANN is shown in Figure 2.

G. *SVM (Support Vector Machine)*

SVM is categorized as supervised learning algorithm. It classify the output based on the optimum hyperplane determined by the given training set.

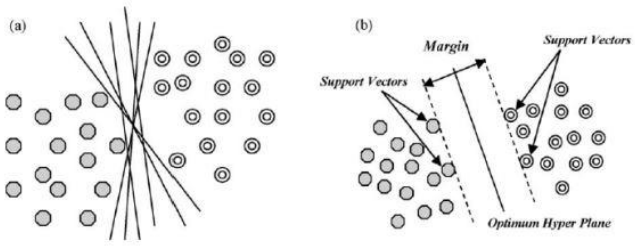

Figure 3: Hyper planes for linearly separable data (a). Optimum hyper plan and support vectors (b) [8]

As shown in Figure 3, there is only one hyper plane that provides maximum margin between two classes. For nonlinear equations, the data mapped into a higher dimensional space (H) through some nonlinear mapping functions. Kernel function is used to solve classification function. There are four basic kernels functions [7].

- Linear : $K(x_i, x_j) = x_i^T x_j$
- Polynomial : $K(x_i, x_j) = (yx_i^T x_j + r)^d$, $y > 0$
- RBF : $K(x_i, x_j) = \exp(-y\| x_i - x_j\|^2)$, $y>0$
- Sigmoid : $K(x_i, x_j) = \tanh(yx_i^T x_j + r)$

RBF is radial basis function. The y, r, and d are kernel parameters.

In this paper, we have used linear SVM and RBF SVM as a kernel to get better accuracy.

III. Design and Implementation

The design for comparing algorithms is described in Figure 4. Algortims compared use same dataset in order to have fair comparison result.

Figure 4: Design Step and Implementation

There are several stages to be done for this comparison: Email Preprocessing, Feature Extraction and Classification using different ML algorithms. For Email Preprocessing and Feature Extraction we have used NLP.

A. *Email Preprocessing*

Before starting with NLP to preprocess the data, it is good to have a look at an example from the dataset. Figure 5 shows a sample email which consist of URLs, email address, numbers and few spelling mistakes.

Congratulations Your email account has won 5 cror. To claim plz mail your details to cocacola_win@egroups.com . For further information please visit http://www.winmoney.com or call us on 9434XXXXXX.
To unsubscribe yourself from this mailing list, send an email to:
groupname-unsubscribe@egroups.com

Figure 5: Sample Email

While many emails can contain similar types of attributes like email address, URLs and numbers, but these things would vary mail to mail. Hence, we can apply "normalization" to these values, for example, we can consider them same. All email address can be replaced with "emladdr" to indicate an email address is present. Similarly, URLs can be treated as "urladdr" and all numbers can be re placed to "number". Chances are high, URLs, Email Address, Numbers etc. will be different mail to mail. Therefore, to replace them with common string will increase performance of spam classification.

In this paper, different NLP concepts are used, which are discussed below:

1) ***Lowercasing:*** The entire email is converted into lower case, so that capitalization is ignored (e.g., CaPital is treated the same caPITAL).

2) ***Normalization:*** All email address are replaced with the string "emladdr", all URLs are treated as "urladdr" and all numbers are represent as "number".

3) ***Word Stemming:*** Words are diminished to their root/stemmed form. For example 'includes', 'included' and 'including' all are replaced with "includ".

4) ***Removal of Non-Words:*** Non-words, punctuation are removed including white spaces, tab etc.

5) ***Removal of Stop-Words:*** For processing of English, we do not need stop words like I, on, the etc.

congratulation you email account win number claim mail you detail emladdr further information please visit urladdr call number unsubscribe you mail list send email emladdr

Figure 6: Preprocessed Sample Email

The result of the preprocessing steps shown in Figure 6. Now this preprocessed data is much easier to work with to perform feature extraction.

B. Feature Extraction

After preprocessing the data, we have list of words for each email. For feature extraction, we must choose which words we are going to use for classification process. Words that are rarely present in an email are not helpful to classify our system. Therefore, we have created a Bag of Words (Corpus), which consist of most frequent words in the email set. Having the BOG, now we can map each word in the processed email into the list of word indices that content the index of the word in the BOG.

Now we will convert each email into a vector $X \in R^n$ with features $x_1, x_2, \ldots, x_n$ where n = # words present in the BOG. Specifically i^{th} feature for any email correspond to the i^{th} word in the BOG. If xi is present in the email, it correspond to 1, otherwise 0. After feature extraction for each email we will have vectors $X_1, X_2, \ldots, X_m$ where m = # email in the dataset.

And each vector has a label 'l' from the set of labels L = {Spam, Legitimate}.

At this stage, different machine learning techniques have been used to analyze their performance.

C. Classification

We have already discussed different ML techniques on section II. ML algorithm does not need any human intervention. These algorithms use training set. Training set are labeled example after analyzing data manually. Based on the training set, ML algorithm computes some parameters. A test set is unlabeled data with same number of features of the form $x_1, x_2, \ldots, x_n$. Now based on the test set and precomputed parameters, algorithm will predict a label from the set of labels L.

We have applied our training set to different ML algorithms to train our model and we found that ANN is providing the best result out of all the classifier discussed above. To measure the performance of these classifier we have used error matrices discussed below:

	Legitimate	Spam
Legitimate	a (True Negative)	b (False Positive)
Spam	c (False Negative)	d (True Positive)

Where a, b, c and d are the number of legitimate classified as legitimate, number of legitimate classified as spam, number of spam classified as legitimate, number of spam classified as spam respectively.

The accuracy, precision, recall and F_1 score are calculated as follows:

Accuracy = (a+d)/(a+b+c+d)
Precision = (d)/(b+d)
Recall = (d)/(c+d)

F_1 Score $= \frac{2 \times Precision \times Recall}{Precision + Recall}$

The F_1 score is the harmonic average of the precision and recall, where an F_1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.

Requirement for the implementation are a training set and a test set. The dataset used for this experiment is collected from Spam Assassin Public Corpus. After preprocessing the dataset, we made a split of 85% and 15%. 85% data has been used as a training set and 15% as a test set. The efficiency of the classifier depends on the training set. Irrelevant training set leads to degradation of the classifier.

IV. Results and Discussion

The confusion matrix for each algorithm is obtained. From the confusion matrix we have calculated precision, recall and F1 score for each classifier. The summary of precision, recall and F1 score is shown in Table I. From the Table I, we can observe that ANN classifier has the highest F1 score (0.977778). In addition, logistic regression, linear SVM and Random Forest have quality F1 score 0.970297, 0.950739 and 0.92545.

Observing all the classifier, we can conclude that ANN has the highest performance level in terms of F1 score.

Table I. Comparison of Precision, Recall & F_1 Score

Classifier	**Precision**	**Recall**	**F_1 Score**
Random Forest	0.967742	0.886700	0.925450
Decision Tree	0.873171	0.881773	0.877451
Logistic Regression	0.975124	0.965517	0.970297
Naïve Bayes	0.672535	0.940887	0.784394
KNN	0.712177	0.950739	0.814346
RBF SVM	0.965517	0.827586	0.891247
Linear SVM	0.950739	0.950739	0.950739
ANN	0.980198	0.975369	0.977778

Now, we will analyze accuracy of each algorithm. The summary of accuracy for each classifier is shown in Figure 7.

From the Figure 7, we can observe that ANN and Decision Tree have 100% accuracy on Training set. However, in Test set ANN has 98.8% accuracy and DT has 91.67% accuracy. Therefore, we can conclude DT suffers from overfitting problem as it can classify Training set with no error, but it fails to produce the same outcome for the test set. In case of K-nearest neighbors, Naïve-Bayes, they are having accuracy of 88.29% and 84.52% on training set and 85.32% and 82.50% on the test set. It means these two algorithms are not able to classify both its training set and test set accurately, as they are suffering from under fitting, to overcome this situation we need more features.

Logistic regression has a quality performance for both Training set(99.98%) and Test set(98%) , close to ANN. Linear SVM and Random forest classifier have same accuracy for both training set and test set. However, linear SVM has better F_1 score than Random forest.

Based on accuracy and F_1 score we have found that ANN has better performance than other classifier.

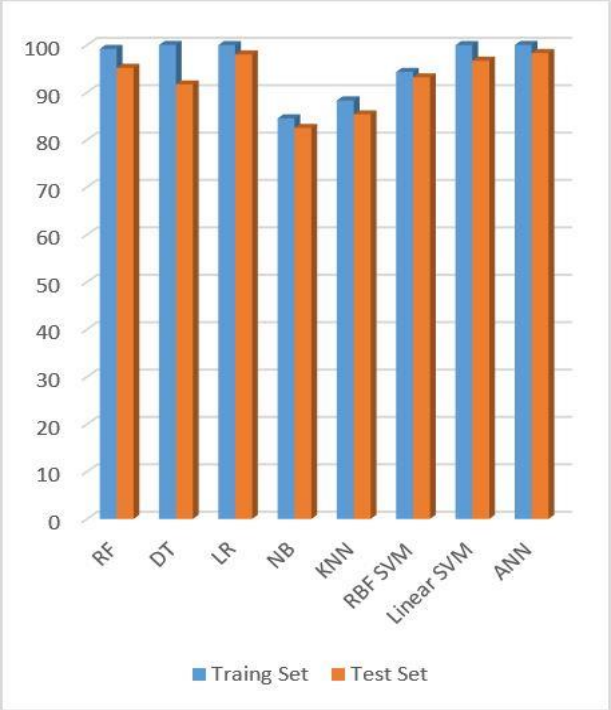

Figure 7: Accuracy of each classifier

V. Conclusion

In this paper, we have analyzed different Machine learning algorithm for the classification of email into Legitimate or spam class. Here we have normalized email addresses, URLs and numbers in an email to improve the performance of our classifier. We have observed that ANN classifier has the highest accuracy. In addition, we have seen Logistic Regression, Random Forest and SVM are having accuracy more than 90% for test set.

As a future work, we plan to get more dataset set to check how performance of these algorithms are being changed. We are also planning to improve performance of Decision Tree, Naïve Bayes classifier, as they are suffering from over-fitting and under-fitting problem. In future, we are planning to improve the performance of classifier based on the attached image with emails.

REFERENCES

[1] J. Brutlag and C. Meek, “Challenges of the email domain for text classification,” Proceedings of ICML, pp. 103-110, 2000.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[2] “Spam in Q1 2014: US Once Again the Prime Target for Malicious Emails,” Online, May, 2014. Available:http://www.kaspersky.com/about/news/spam/2014/Spam-in-Q1-2014-US-Once-Again-the-Prime-Target-for-Malicious-Emails

[3] Harpreet Kaur,Ajay Sharma, "Improved Email Spam Classification Method Using Integrated Particle Swarm Optimization and Decision Tree",2nd International Conference on NGCT, pp. 516 - 521, 2016

[4] M. K. Chae,Sasikumaran Sreedharan, "Spam Filtering Email Classification (SFECM) using Gain and Graph Mining Algorithm",2nd International Conference on Anti-Cyber Crimes (ICACC), pp. 217-222, 2017

[5] Suma T, Kumara swamy Y S,"Email classification using adaptive ontologies Learning",IEEE International Conference On Recent Trends In Electronics Information Communication Technology,pp. 2102 – 2106, 2016

[6] B. Ruttenbur, G. Spickler, and S. Lurie, “eLearning – The Engine ofthe Knowledge Economy”, Morgan Keegan &Co. Inc. eLearning Industry Report, 200

[7] C.-J. L. Chih-Wei Hsu, Chih-Chung Chang, “A Practical Guide to Support Vector Classification,” BJU Int., vol. 101, no. 1, 2008

[8] T. Kavzoglu and I. Colkesen, “A kernel functions analysis for support vector machines for land cover classification,” Int. J. Appl. Earth Obs. Geoinf., vol. 11, no. 5, pp. 352–359, 2009

[9] I. H. Witten, E. Frank, and M. a. Hall, Data Mining Practical Machine Learning Tools and Techniques Third Edition, vol. 277, no. Tentang Data Mining. 2011