CUSTOMIZED MAIL SPAM CLASSIFIER USING MACHINE LEARNING ALGORITHMS

[*]R. Deepak & [#]M. Suresh Kumar

[*]*PG Student, MCA, GIS, GITAM, Visakhapatnam, deepakrella9@gmail.com*

[#]*Assistant Professor, GIS, GITAM, Visakhapatnam, smaddila@gitam.edu*

Abstract: *An email-spam is a major problem for users in the present generation, preventing them to rapidly process the important emails in shortest time. Recent methods of machine learning are being used to successfully detect and filter spam emails. It presents a systematic review of some of the popular machine learning based email spam filtering approaches. In this paper, an email recommended system is proposed using user actions and statistical approaches. Instead of a two-class classification with Ham and Spam, it treats the problem as a multi-class classification in which each class is a recommended action from user to an email. The most common actions are: Inbox, Archived, Trash and Sent. In this paper it reviews some of the most popular machine learning methods (i.e.; k-NN and Logistic Regression) and of their applicability to the problems of spam Email classification.*

Keywords: Mail Spam Classifier, Machine Learning Algorithms, KNN, Logistic Regression.

1. INTRODUCTION

An email is one of the finest applications which are utmost commonly used services over internet, allowing people to send messages to one or more recipients. Lately, undesired bulk emails otherwise known as spam mails have been a major disappointment for the email users, causing a major trouble for them. These emails are purposely been send by unidentified scammers, who collects such email id's of general public logged by them on various sites on internet. Spams stop the individuals from the absolute usage of time, fuller storage space and network bandwidth. It acts as a carrier of malware which creates the fraud schemes, phishes messages and extracts explicit contents of the individual. Spam mails have been increased over the year which has significantly increased the cost of internet users. To wipe out the inconvenience faced by individuals, various email providers have employed techniques such as machine learning for spam filters. Such techniques have the capability to catch hold of spam mails and filter them over vast number of computers. Nowadays machine learning technique is being widely used because of its efficiency and been studied with various algorithms, can be well used for spam filtering.

Bag of words: Bag of Words (BOW) is a technique to excerpt features from text documents. These features care used for training text data sets in machine learning. It creates a terminology of all the inimitable words occurring in the documents in the trained data sets. The BOW model is a method for representing text data when demonstrating text with machine learning algorithms. This model is simple to understand and implement and has seen great accomplishment in problems such as document classification and language modelling.

TF-DIF: TF-IDF mass is composed in two terms, they are- the normalized Term Frequency (TF), or the number of times a word appears in a document divided by the total number of words in that document and the second is the Inverse Document Frequency (IDF), which is computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the definite term appears. TF-IDF is simply the TF multiplied by IDF.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

Term Frequency: Term Frequency means which measures how often a term occurs in a document. Since every document is dissimilar in length, it is possible that a term would appear frequently in larger than shorter ones. It is the number of times a word appears in a document divided by the total number of words in the document. Each and every document has its own term frequency.
TF (t) = (Number of times‘t’ terms occurs in a document) / (Total number of terms in the document).

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}}$$

Inverse Document Frequency: Inverse Document Frequency measures how significant a term is. While computing TF, all terms are considered similarly important. However, it can be seen that certain terms like- "is", "of", and "that", may appeared multiple times but have little importance. The log of the number of documents divided by the number of documents that comprises the word ‘***w***’. This frequency determines the mass of rare words across all documents in the body.
IDF (t) = log_e (Total number of documents / Number of documents having ‘t’ term in it).

$$idf(w) = log(\frac{N}{df_t})$$

2. MACHINE LEARNING ALGORITHMS

2.1. K-nearest neighbors Algorithm

A K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique. KNN can be used for both regression predictive problems and classification. However, it is mainly used for classification predictive problems in machine learning. K-nearest neighbors (KNN) algorithm uses ‘feature similarity’ to predict the values of new data points which further means that the new data point will be assigned a value based on how exactly it matches the points in the training set.

It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

The following two properties would define KNN well −

a) **Lazy learning algorithm-** KNN is a lazy learning algorithm because it does not have a specialized training phase and uses all the data for training while classification.

b) **Non-Parametric learning algorithm-** KNN is also a non-parametric learning algorithm because it doesn’t assume anything about the underlying data.

2.2. Logistic Regression Algorithm

Logistic Regression is categorized in ‘Supervised’ Machine Learning (ML) a method which is also called as a ‘Statistical Learning’ technique. Logistic regression is a classification algorithm used to assign observations to a separate set of classes. Logistic regression converts its output using the logistic sigmoid function to return a probability value.

Logistic regression uses an equation as the representation, which is similar to linear regression. Input values (x) are merged linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to forecast an output value (y). A key difference from linear regression is that the output value being imitated is a binary value (0 or 1) rather than a numeric value.

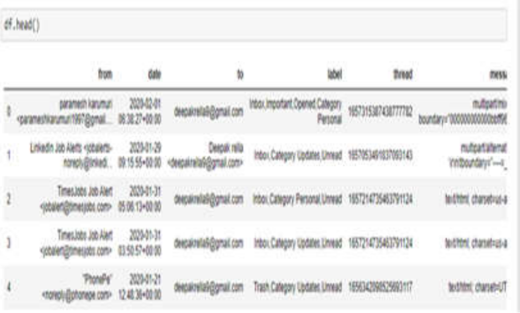
An example for logistic regression equation is: y = (b0 + b1*x)^e / ((b0 + b1*x)^e+1)
Here, the predicted output is y.

- The bias or intercept term is b0.
- The coefficient for the single input value (x) is b1.

Each column in input data has a relation with the coefficient b (a constant real value) that must be understood from the training data.

3. ABOUT DATA-SET AND ATTRIBUTES

The examination was supervised on Real Time data set and it is in the format of Mbox. It has been downloaded from the Google data for assessing the precision and presentation of machine learning techniques. This dataset contain 24 attributes namely thread, message, label and etc. To identify the best classifier the data set is divided into trained data and test data. Preparing data includes cleansing the data, altering the data and splitting the data. The data set is splitted into trained data and test data: 70% of training data and 30% of testing data. And Again the Testing data is splitted into trained data and test data, i.e. 80% of trained data and 20% of testing data.



```
df.head()
```

	from	date	to	label	thread	mess
0	paramesh karumuri <parameshkarumuri1997@gmail...	2020-02-01 06:38:27+00:00	deepakrella9@gmail.com	Inbox,Important,Opened,Category Personal	1657315387438777782	multipart/mi boundary="000000000000bbff96
1	LinkedIn Job Alerts <jobalerts-noreply@linked...	2020-01-29 09:15:55+00:00	Deepak rella <deepakrella9@gmail.com>	Inbox,Category Updates,Unread	1657053491637093143	multipart/alternat \r\nboundary="----=_
2	TimesJobs Job Alert <jobalert@timesjobs.com>	2020-01-31 05:06:13+00:00	deepakrella9@gmail.com	Inbox,Category Personal,Unread	1657214735463791124	text/html; charset=us-a
3	TimesJobs Job Alert <jobalert@timesjobs.com>	2020-01-31 03:50:57+00:00	deepakrella9@gmail.com	Inbox,Category Updates,Unread	1657214735463791124	text/html; charset=us-a
4	"PhonePe" <noreply@phonepe.com>	2020-01-21 12:48:36+00:00	deepakrella9@gmail.com	Trash,Category Updates,Unread	1656342098525693117	text/html; charset=UT

Fig.1. Mail Spam Classifier Dataset

4. RESULTS AND ANALYSIS

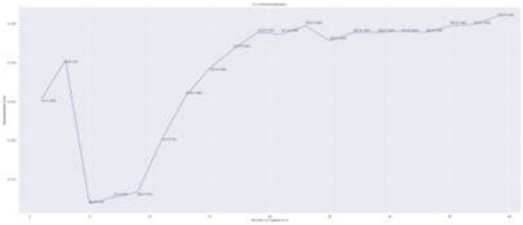


Fig.2. KNN using bag of words

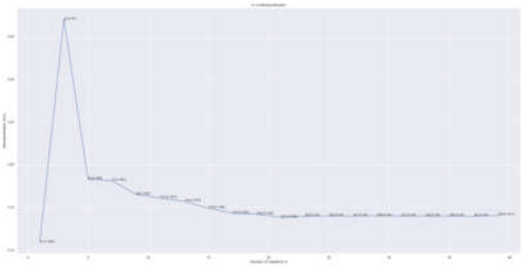


Fig.3. KNN using TF-DIF

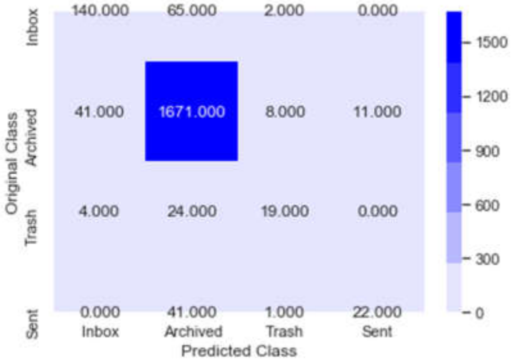


Fig.4.Confusion matrix of logistic regression using bag of words

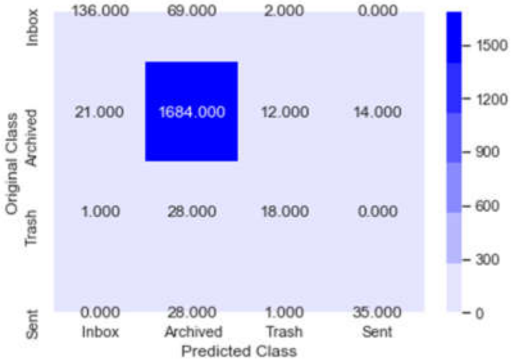


Fig.5.Confusion matrix of logistic regression using TF-DIF

A comparison of different algorithms is performed on Mail Spam Classifier data set. To select the best out of the entire model created can be done by comparing the accuracy of the entire model and selecting the one which gives the maximum accuracy in both training data and test data. The results of the algorithms are shown below:

Table- I: Comparison values

Algorithm: KNN	**Accuracy Score**	**F1 Score**
Bag of Words	87.6	-
TF-DIF	88.1	-

Table- II: Comparison values

Algorithm: Logistic Regression	**Accuracy Score**	**Precision Score**	**Recall Score**	**F1 Score**
Bag of Words	90.0	89.5	90.3	89.7
TF-DIF	91.0	90.8	91.4	90.8

5. CONCLUSION

In this paper we review a number of the foremost popular machine learning methods and of their applicability to the matter of spam e-mail classification. A review of the state of the art algorithms been applied for classification of messages as either spam or hamis provided. The attempts made by different researchers to solving the Problem of spam through the use of machine learning classifiers were discussed. From Table -1 & 2, we can easily Logistic Regression is the best algorithm when compared with the other algorithm. Having discussed the open problems in spam filtering, further research to enhance the effectiveness of spam filters need to be done. This will make the development of spam filters to continue to be an active research field for academician and industry practitioners researching machine learning techniques for effective spam filtering. Our hope is that research students will use this paper as a spring board for doing qualitative research in spam filtering using machine learning, deep learning and deep adversarial learning algorithms.

REFERENCES

Journal Articles

[1] *D. K. Renuka, T. Hamsapriya, M. R. Chakkaravarthi and P. L. Surya, "Spam Classification Based on Supervised Learning Using Machine Learning Techniques", in proc. IEEE- International Conference on Process Automation, Control and Computing, 2011, pp. 1–7.*

[2] *R. Shams and R. E. Mercer, "Classifying spam emails using text and readability features", in proc. IEEE International Conference on Data Mining (ICDM), 2013, pp. 657–666.*

[3] *J. Clark, I. Koprinska and J.Poon, "A Neural Network-Based approach to automated email classification", in proc. IEEE- WIC International Conference on Web Intelligence, 2003.*

[4] *A. Harisinghaney, A. Dixit, S. Gupta, and Anuja Arora, "Text and image based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN Algorithm", in proc. IEEE-International Conference on Reliability, Optimization and Information Technology (ICROIT), 2014, pp.153-155.*

[5] *L. Firte, C. Lemnaru, R. Potolea, "Spam Detection Filter using KNN Algorithm and Resampling", in proc. IEEE- 6th International Conference on Intelligent Computer Communication and Processing, 2010, pp.27- 33.*

[6] *M.Prilepok and P. Berek, "Spam Detection Using Data Compression and Signatures and Signatures," in Cybernetics and Systems: An International Journal, Vol. 44, pp. 533–549, 2014.*

[7] *T. A. Almeida and A. Yamakami, "Content-Based Spam Filtering", in proc. IEEE- International Joint Conference of Neural Networks (IJCNN), pp. 1-7, 2010.*