International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 09, September-2015

Spam Detection using KNN, Back Propagation and Recurrent Neural Network

Kulwinder Kaur
M.E, Computer Science & Engineering
University Institute of Engineering & Technology,
Panjab University, Chandigarh

Dr. Mukesh Kumar
Computer Science & Engineering
University Institute of Engineering & Technology,
Panjab University, Chandigarh

***Abstract*— This paper presents our research in Email classification using different classification methods. In today's world, people are so much inclined towards Social Networking due to which it has become easy to spread spam contents through them. One can have access to the details of any person easily through these social networking websites. No one is secure inside the social media sites. Social Networking sites are fast becoming popular in the recent years, among which Email is one of the fastest growing sites. It plays a double role of Online Social Networking and Micro Blogging. Spammers try to attack the Email trending topics to harm the useful content. Social spamming is more successful as compared to the email spamming by using social relationship between the users. Spam detection is very important because Emails is mainly used for the commercial advertisements and professional purposes. The spammers attack the private information of the user and also the reputation of the user is harmed. Spammers can be detected by using the content and user based attributes. This research deals with the detection of spam in Emails. We introduce a methodology for data preprocessing and a new methodology for classification of unsolicited related emails. We generate comprehensive lists of spam words for spam classification. We compare the classification results against manually labeled emails to estimate the effectiveness of the classifiers and the comparisons shown.**

Keywords –Email Classification, KNN (K-Nearest Neighbour), Back propagation Neural Network(BPNN),Recurrent Neural Network(RNN),Enron Data set.

I. INTRODUCTION

Email has become efficient and styled communication technique as a result of the variability of the web users can increase. Therefore, email management became a really necessary and growing downside for the folks and organizations as results of it's liable to misuse. The blind posting of unsought email messages, mentioned as a spam, is example of misuse. Spam is usually printed as causing of the unsought bulk email - that is, email that wasn't asked for by several recipients. A further common definition of a spam is restricted to the unsought business email a definition does not take under the consideration of non-commercial solicitations like the political or religious pitches, though unsought, as spam. Email has out and away the foremost common style of the spamming on net.

According to the information estimable by analysis, spam accounts for v-j day to twenty of email at U.S.-based company organizations. 0.5 users unit of the measurement receives 10 or extra spam emails per day whereas variety of unit of measurement receives up to various whole bunch unsought emails. International info cluster expected that world email traffic surges to the sixty billion messages daily by 2006. It includes identical or nearly identical unsought messages to oversized style of recipients. Not like legitimate business email, spam is usually sent whereas not the specific permission of the recipients, and sometimes contains several tricks to bypass email filters. Moderate computers sometimes some ability to send spam. The only real necessary ingredient is the list of addresses to specialize in.

Spammers can get email addresses by the style of means: harvest addresses from the Usenet postings, DNS listings, or web pages; estimation common names at noted domains; and "expending" or looking for email addresses comparable to particular persons, like residents in exceedingly half. Several spammers utilize programs are mentioned as web spiders to go seeking out the email addresses on web page, although it is possible to fool the net spider by work the "@" image with another image, as example "#", where the posting email address. As result, users can compelled to waste the valuable time to delete the spam emails. Moreover, spam emails can stock up the house for storing the data processor quickly; they may cause varied downside for several websites with thousands of users. Presently, the torrential work on the spam email filtering have done victimization the methods like decision trees, neural networks, Naive theorem classifiers etc. To agitate the matter of grow volumes of unsought emails, various different ways that for email filtering unit of measurement being used in multiple business product. This tendency is to make a framework for a cost-effective email filtering victimization philosophy. Metaphysics give machine-understandable linguistics of the data, therefore it can be utilised in system. It is important to share information with each other for lots of sensible spam filtering. Therefore, it is necessary to create metaphysics and framework for the economical email filtering. Victimization philosophy has been designed to filter spam with bunch of bulk email that may possibly be filtered out on system.

II. DEFINITION AND CHARACTERISTICS OF SPAM

There exist various definitions of what spam in addition mentioned as junk mail is and therefore manner it differs from the legitimate mail in addition mentioned as non-spam, real mail or ham. The shortest definitions characterize spam as "unsolicited bulk email". Generally the word business is extra, but extension is argued. The TREC Spam Track

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

depends on the equivalent definition: spam is "unsolicited, unwanted email was sent indiscriminately, directly or indirectly, by sender having no current relationship with user". Another wide accepted definition states that "Internet spam is one or extra unsought messages, sent or announce as locality of much bigger assortment of messages, all having the significant identical content". Promoting the association planned to use word "spam" only for the messages with the positive varieties of the content, but this concept met no the enthusiasm, being thought of endeavor to legitimate fully different types of the spam. There is a tendency to area unit able to purpose is that spam is unsought, per cited formula "spam is regarding consent, not content". It is important to mention the notion of being unsought is tough to capture. In fact, despite wide agreement on the type of definitions filters have to be compelled to deem content and ways within which of the delivery of messages to acknowledge spam from legitimate mail. Among latest work it is fascinating to mention, UN agency still value highly to deem content and a user's judgement personally to stipulate spam.

It used for the spam mail identification and therefore manner can be utilised in the conjunction with machine learning theme. Feature ranking techniques are like data gain, Chi-square, Symmetrical uncertainty, Gain ratio, Relief, One and Correlation area unit applied to a replica of the information. Once feature selection set with perfect advantage is utilized to cut the spatial property of initial information and therefore the testing information. Every reduced datasets can be possibly passed to machine learning theme for testing. Results area unit attained by pattern Random Forest and 0.5 classification techniques.

The problem of unwanted electronic messages is a significant issue, as spam constitutes up to 75–80% of quantity of email messages. Spam causes various issues, a number of leading to direct monetary losses. a lot of exactly, spam causes the misuse of traffic, space for storing and machine power; spam creates users inspect and kind out the extra email, not solely waste their time and inflicting loss of productivity, however additionally irritating them and, as several claim, violating the privacy rights ; finally, spam causes legal issues by advertising the pyramid schemes etc. The full worldwide monetary losses caused by the spam in 2005 that were calculable by Ferris analysis analyser info Service at $50 billion.

There is a growing scientific address characteristic of spam development. In general, spam is used to advertise the completely different types of merchandise and services, and therefore the share of advertisements dedicated to particular reasonably merchandise or services changes over time. Very often spam serves wants of on-line frauds. A special case of the spamming activity is phishing, specifically attempting to search sensitive info by imitating the requests from trustworthy authorities, like server administration, banks or service suppliers..

Learning-based strategies of spam filtering

Filtering could be a widespread resolution to matter of spam.. Existing filtering algorithms area unit is quite effective, typically shown accuracy of higher than ninetieth throughout the experimental analysis. It is potential to use spam filtering algorithms on different phases of an email transmission: At routers, at destination of mail server, or the destination mailbox. It can be mentioned the filtering on destination purpose that solves the issues caused by the spam partially: a filter prevents the end-users from the wasting the time on junk messages, however it doesn't forestall resources misuse, as result of all messages area unit delivered nonetheless.

In general, spam filter is application that implements a function:

$f(m, \theta) = \begin{cases} \text{cspam is taken into account spam} \\ \text{cleg, if the message m is taken into account legitimate email} \end{cases}$

Where m could be a message to be classified, θ could be a vector of parameters, and cspam area unit labels allotted to messages.

Most of the spam filters area unit supported the machine learning classification techniques. In exceedingly learning-based technique the vector of parameters θ is the results of coaching classifier on a pre-collected dataset:

Where m1,m2 …… mn area unit collected messages, y1, y2 . . . yn area unit the corresponding labels, and is a coaching perform.

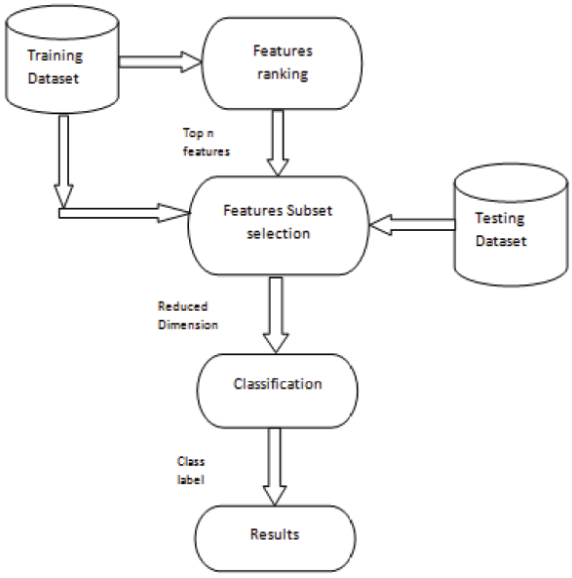


Fig1. Stages of Spam Email Classification

In following subsections tend to discuss essential ideas associated with the work. It involves short background on the feature ranking techniques, classification techniques and results.

III.DATA SET

The dataset can used for experiment is spam base. The last column of the 'spam base' denotes whether or not e-mail was thought-about spam or not. Most of attributes indicate frequency of spam connected term occurrences. The first 48 set of attributes provide the term frequency and inverse the document frequency values for the spam related words, whereas consequent half-dozen attributes provide term frequency and inverse the document frequency values for the spam connected terms. The run-length attributes used the length of sequences of consecutive capital letters, capital_ character long longest, capital_ character long average and

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

capital_ character long total. Therefore, dataset has fifty seven attributes serving as input options for spam detection and thus last attribute represent the category. We got additionally used one public dataset Enron. The "pre-processed" directory contains the messages within pre-processed format. Each message is exceedingly separate computer file. The body of email contains particular data. This data has to be extracted before the running of a filter method to suggest that of pre-processing. The aim for pre-processing is to rework messages in mail into an identical format that can be understood by training rule. Following square measure the steps concerned in the pre-processing:

1. *Feature extraction*: Extracting the options from e-mail in vector house.
2. *Stemming*: Stemming is a method to remove an individual morphological and in-flexional endings from the words in English.
3. *Stop word removal*: Removal of the non-informative words.
4. *Noise removal*: Remove obscure text or symbols from the options.
5. *Representation*: tf-idf can be applied mathematics won't to calculate vital a word is to document in exceedingly feature corpus. Word frequency has established by the term frequency, variety of the word seems within message yields the importance of word to document. The term frequency is increased with the inverse of document frequency that measures frequency of word occurring in all the messages.

IV.PROPOSED WORK

We analyze an impact of pre-processing of Email data for a spam detection task. More specifically the purpose is

- To evaluate the impact of using different attributes of Email data on classification problem.
- To pre-process the data using various steps such as tokenizing, stemming, stop word removal etc.
- To extract features from Email data for classification
- To classify Emails spam using different algorithms.

In this, we have to consider the email spam classification using some technique to remove it. Spam is irrelevant or malicious mail that comes in your personal or business which we have to remove.

In this we considered the conditions to evaluate the quality of classification or prediction. These prediction metrics can be used to evaluate the best quality of email prediction. The true positive indicates to spam detection tool that predicts the email called spam and truly it was a spam. The True Negative indicates the tool or email system to predict the email is normal or it is not spamming correctly it was so. Moreover, False Positive indicates by mistake this tool that predicts a good email is spam. At the end, False Negative indicates to another mistake which is predicted to spam email is normal. In such the perfect detection system has the values: TN 100%, TP 100%, FP 0%, or FN 0%. In the reality of perfect situation is impractical and impossible. FP and TP complement to each other for quality 100%. In this, same thing can be applied for TN and FN.

The main challenge of email detection system is restricted with various spam-detection roles, TP can be high, but the account of taking several false alarms. On the other side very restricted rules can get high TN but the account of FN.

Speed is other challenge in emails spam detection. Consider security, performance and speed which is always in trade off with the security where many roles go slow down the system.

In addition to the spam based classification, papers can conduct the research in emails to discuss the other aspects like: folder classification & Automatic subject, emails and contacts clustering, priority based filtering of email messages, etc.

When deal with large-scale datasets, often it is a practical necessity to look to decrease the size of dataset, acknowledge in several cases patterns that are in data would exist if representative subset of the instances were opted. Select the reduced dataset that can be less noisy than original dataset to produce superior generalization performance of the classifiers trained on reduced dataset.

Most of the cases, Content based spam filters are useless if they could not understand the meaning of words and phrases in email. Now days, spammers can change one or more characters of the offensive words in spam to foil the content based filters. But it is important to observe the spammers to change the words in that way which the human being can use and understand the meaning of words without any difficulty. Spammers do not make drastic change in words so it can easily recognize by humans. Based on observations, develop a rule based on word stemming technique to match words on those look alike and sound alike.

The main goal of this instance selection is to choose a representative subset of the instances to enable the size of newest dataset to reduce the Spam is described as unsolicited commercial email that becomes one of biggest worldwide problems to face the Internet today. This thesis proposed efficient and effective email classification methods based on the data filtering scheme into training model. The focus of this thesis is to decrease the instance of email corpora from the training model using the ISM that is less significant into relation of classification. Our empirical evidence displays the proposed technique provides better accuracy with the reduction of instances from email corpora.

This thesis proposed to identify the span in Email Spam classification. To overcome this problem, we proposed the three technologies, which is displayed at below:

- KNN
- Back Propagation Neural Network
- Recurrent neural network

Before application of any classifier, we need to convert the text data into a numerical form. We have utilised TF -IDF score based method and applied text mining to it.

The email data is pre-processed and Enron dataset is used.Basic steps of preprocessing using Text mining and Machine Learning basic concepts, approaches are also discussed. It also includes the introduction to KNN and Naïve Bayes Multinomial along with algorithms of them.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

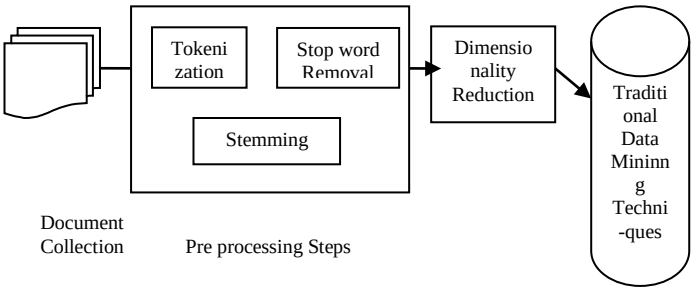


Fig2.The different steps performed in text mining are as follows:

Step 1: Preprocessing- It is used to distill unstructured data to structured format. There are different preprocessing steps performed in Text mining such as tokenization, stop word removal and stemming. These algorithms are discussed below.

i. Tokenization: The purpose of tokenization is to remove all the punctuation marks like commas, full stop, hyphen and brackets. It divides the whole text into separate tokens to explore the words in document.
ii. Stop word removal: The purpose of this process is used to eliminate conjunction, prepositions, articles and other frequent words such as adverbs, verbs and adjectives from textual data. Thus it reduces textual data and system performance is improved.
iii. Stemming: Stemming is used to reduce the words to their root words e.g. words like "computing" ,"computed" and "computerize" has it root word "compute". The purpose of stemming is to represent the words to only terms in their document. There are different algorithms to perform stemming such as Lovins Stemmer, Porters Stemmer, Paice/Husk Stemmer ,Dawson Stemmer, N-Gram Stemmer, YASS Stemmer and HMM Stemmer.
iv. Weighting Factor: - Features are extracted from overloaded large datasets.TF-IDF (Term frequency-Inverse document frequency) score is generally is used to give weight to each term. TF-IDF is multiplication of term frequency and inverse document frequency.

$$\mathrm{TF-IDE} = n_w^d \, log_2(\frac{N}{N_w}) \qquad (1)$$

Where n_w^d = frequency of word w in document d.

N= total document and N_w= document congaing word w.

v. Term - document matrix – After initial steps of preprocessing text in documents is converted into term-document matrix. Rows in matrix represents document in which word appears and columns represent the words that are extracted from documents. The cell of matrix is filled with TF-IDF score.

Step 2: Dimensionality Reduction – After preprocessing steps, dimensionality reduction is performed. Here original TDM (term document matrix) is replaced with smaller matrix by using a SVD (singular value decomposition technique). This technique discards unimportant word and relevant and important word are filtered out. The new matrix is generated of terms and documents.

Step 3: Mining the reduced data with traditional data mining techniques- Classification, clustering and predictive methods are applied to the reduced datasets using data mining techniques to analyze the pattern and trends within data.

Machine Learning Approaches For Classification

There are various approaches to design machine learning algorithms. The purpose of ML algorithms is to use observations as input and this observation can be a data, pattern and past experience. Thus Ml algorithms use to improve the performance of instances, which can be done by any classifier by trying to classify the input pattern into set of categories or to cluster unknown instances. As the nature of ML algorithms it enhances its performance from past experience or by receiving feedback. It can be divided into two categories supervised and unsupervised approach .

Supervised: In supervised learning, the instances are labeled with known or target classes labels. Here before classification the dataset knows the target class. Thus it is very helpful for the problems which have known inputs.

Unsupervised: In unsupervised learning, the algorithm groups the instances by their similarities in values of features and makes different clusters. In it no prior class or clusters are given, the algorithm itself defines their clusters automatically and statistically.

KNN Algorithm

KNN is type of instance based learning or lazy learning. In this learning the function is approximately locally and all computation is deferred until classification. It is simplest of all machine learning algorithms. In KNN classification, the output is class membership. An object is classified by majority votes of its neighbors by the object being assigned to class most common among its k nearest neighbor (k is positive small integer). The nearest neighbor is determined using similarity measure usually distance functions are user. Following are the distance function used by KNN [50].

Euclidean distance function $\sqrt{\sum_{i=1}^{N}(a_i - b_i)^2}$

Manhattan distance function $\sum_{i=}^{N}|a_i - b_i|$

Where {(a_1, b_1), (a_2,b_2) ,(a_3,b_3)…… (a_N, b_N)} is training datasets.

In KNN algorithm all the distance from testing point to training point are computed. Then these all testing points are sorted ascending order. Then class labels are added for each K nearest neighbors and sign of sum are used for prediction. The value of k in k-nearest neighbor is challenging task. As choosing smaller value of k. e.g. by choosing k=1 may lead to risk of over fitting and choosing larger value of k e.g. k=N may lead to under fitting. Therefore optimal value of k has been chosen between the values 3-10, which gives better result.

Back Propagation Neural Network

In the behavioral sciences we applied mathematics analyses on the victimization ancient algorithms that don't invariably

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

result in satisfactory answer, notably classification analysis. Current classification strategies make to suppose the constant quantity or non-parametric variable analyses following: discriminant analysis, cluster analyses, etc.

These strategies are square measure that usually rather inefficient information square measure to nonlinearly distributed, once the variable transformation. Therefore, the tendency to propose the classification technique that supports the principles of artificial neural networks. Throughout the eighties, the utilization of NN developed explosively within areas of word recognition

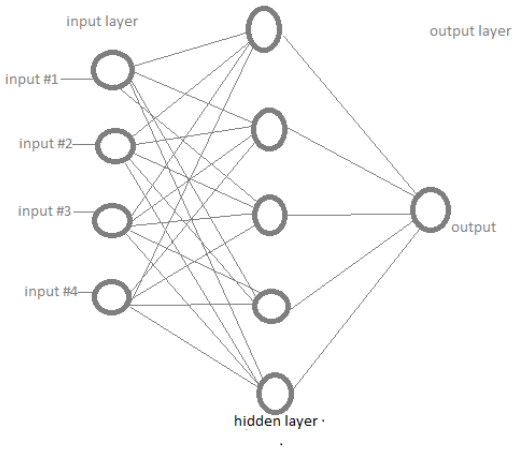


Fig. 3. Structure of a neural network used in the experiments

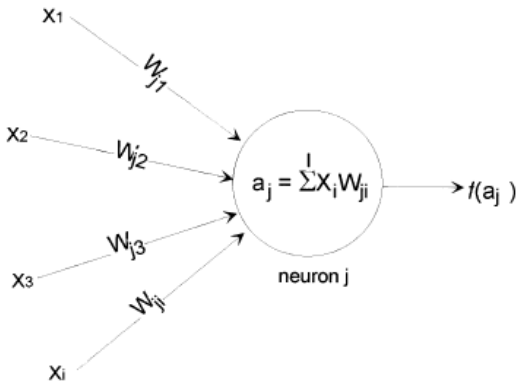


Fig. 4. Detail of one neuron

For the classification functions, NN can be used to analysis the macromolecule structure, the classification of seaweeds, and thus the recognition of impulsive noises in the marine mammals. With this paper, NN square measures the accustomed discriminate vocalizations of 4 male Dama dama, deer, throughout the rutting amount.

Theory of artificial neural networks starts and developed in the line with elementary principle of operation of neural system. Since, a awfully kind of networks have been created. All square measure are composed of units, and connections between them, that along confirm the behaviour of network. The selection of network depends on matter to be solved; rear propagation gradient network is most often used. This network includes 3 or lot of somatic cell layers: one is input layer, one output layer and a minimum hidden layer. In most of the cases, network with hidden layer is used the limit calculation time, specifically once results are attained the square measure satisfactory. All somatic cells of layer square measure connect by the nerve fiber to neuron of the successive layer a pair of 1. Signal propagation is the input layer that includes n neurons code for n items of signaling of network. The quantity of the neurons of hidden layer is selected with empirical observation by user. Finally, output layer consist k neurons for k categories. Every association between the 2 neurons is related to weight issue; this weight can be changed by the ordered iterations through the coaching of network in line with the input and output information. With the input layer, the state of somatic cell is set by input variable; the opposite neurons judge the state of signal from previous layer.

$$a_j = \sum_{i=1}^{I} X_i W_{ji}$$

Where aj is that the web input of somatic cell j; Xi is that output worth of somatic cell i of the previous layer; Wji is that the weight issue of association between somatic cell i and somatic cell j. The activity of neurons is typically determined via sigmoid function:

$$f(a_j) = \frac{1}{1 + \exp^{-a_j}}$$

Thus, weight factors can represent the response of NN to matter being Janus-faced.

Training the network

The back propagation method is supervised learning because the network has trained with expected replies. Each iteration modifies the association weights to decrease the error of the reply. Adjustment of weights, layer by layer can be calculated from output layer back to input layer. This correction is created by:

$$\Delta W_{ji} = \eta \delta_j f(a_i)$$

Where DWji is adjustment of the weight between somatic cell j and somatic cell i from previous layer; f(ai) is output of the somatic cell i, h is learning rate, and dj depends on layer. For output layer, dj is:

$$\delta_j = (Y_j - \hat{Y}_j) f_i'(a_j)$$

Where Yj is mean (‘observed value’) and Y. j is current output worth of somatic cell j. For hidden layer, dj is:

$$\delta_j = f_j'(a_j) \sum_{k=1}^{K} \delta_k W_{kj}$$

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

Where, K is variety of neurons within next layer. The learning rate plays a crucial role in training. Once the rate is low, the convergence of burden to optimum is extremely slow, once speed is too high, the network will oscillate, or a lot of seriously it will bog down during a native minimum. To cut back these issues, a momentum term a is employed and Wji becomes:

$$\Delta W_{ji} = \eta \delta_j f(a_i) + \alpha \Delta W_{ji}^{\text{Prev}}$$

Where, Wji Prev denotes the correction within the previous iteration. In our study, at the start a zero.7 and h 0.01, then they're changed in line with the importance of the error by subsequent algorithm:

If present–error\previous–error * one.04

$$\eta = \eta * 0.75,$$
$$\alpha = 0,$$
$$\eta = \eta * 1.05,$$
$$\alpha = 0.95,$$

The performed of representative information set, runs till total square of errors is minimized:

$$\text{SSE} = \frac{1}{2} \sum_{p=1}^{P} \sum_{j=1}^{N} (Y_{pj} - \hat{Y}_{pj})^2$$

The structure of network, the variety of records within information set and thus number of iterations confirms the coaching period. In our study, 100 of records are characterized by thirty two input variables and 4 output variables, and one hidden layer with ten neurons, three hundred iterations last regarding three minutes with AN Intel 486 DX2–66 processor.

Testing the network

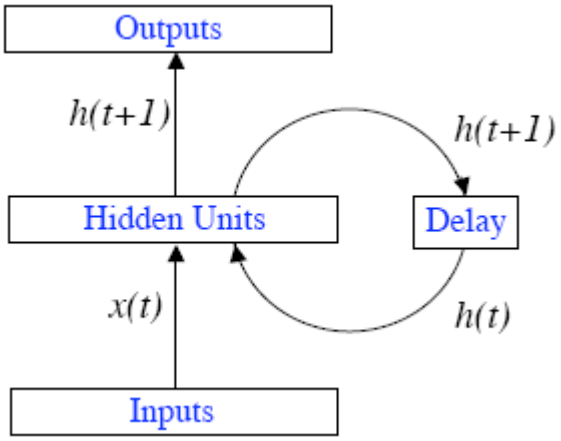
The network performance of the network should be tested. In discriminant analysis a main indication is given by proportion of the correct classifications of coaching set records. All same, the performance of network with take a look at set is a lot of relevant.

In this step, the computer file square measure fed into network and thus desired values square measure compared to network's output values. The agreement or disagreement of results offers sign of the performance of the trained network.

4.5 Recurrent neural network

The fundamental feature of RNN is the network contains one feed-back connection, so activations can flow round in loop, which enables to the networks to learn sequences and temporal processing, e.g., perform sequence recognition or temporal prediction. Recurrent neural network architectures have several different forms. One common type consist standard Multi-Layer Perceptron plus added loops. This exploit is a powerful non-linear mapping capability of MLP, and have form of memory. They have uniform structures, potentially with neuron connected to all others, and have stochastic activation functions. To simple architectures and deterministic functions, learn the achieved gradient descent procedures to leading to back-propagation algorithm for the feed-forward networks. When activations are stochastic, simulated annealing approaches can be more appropriate. The following will look at few of most important types and features of the recurrent networks.

The simple form of recurrent neural network is MLP with previous set of hidden activations feed back into network along with inputs:



Note the time t has to be discretized with activations updated at time step. The time scale may correspond to operation of the real neurons, or artificial systems time step size to the appropriate for given problem. A delay unit requires introducing the hold activations until processed at next time step.

V. RESULTS AND DISCUSSION

The chapter presents the results obtained after applying methodology discussed . The results are compared using different algorithms on four components of Eclipse using performance metrics.

Results:

In this research, component specific dictionaries are created of four component of Eclipse. These dictionaries are created using top 1250 terms using two feature selection methods; namely- info-gain and Chi square. The set of dictionary terms are then fed to two widely used ML algorithms named Naïve Bayes and KNN for classification task and performance is analyzed in terms of precision and accuracy. The result is documented below. The total number of correct classified emails as spam and non-spam out of a total data set of 8000

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

emails is given

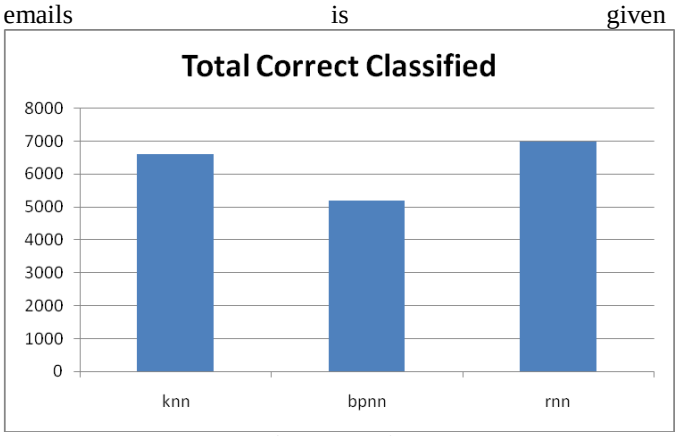


Fig 5 : Results of email classification algorithm

Overview of Performance Measures

We perform a binary classification as to whether a given tweets is spam or not. In case of a highly imbalanced data accuracy alone is not sufficient measure to estimate the performance of classifier. For example, if had only 1% of positive instances in data we can achieve an accuracy of 99% but simply classifying all instances as false. In such situations the following measure are more useful and informative for evaluating the performance of such binary classifiers.

Accuracy: Accuracy is calculated as fraction of sum of correct classification to total number of classification. It is defined as:

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

In simple terms this means the number of correctly classified positive instances out of the total instances which are positive. For example, if we had 10 fruits of which 5 are apples then recall will be how many of the 5 apples are correctly classified over number of apples. Recall is the same as sensitivity.

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

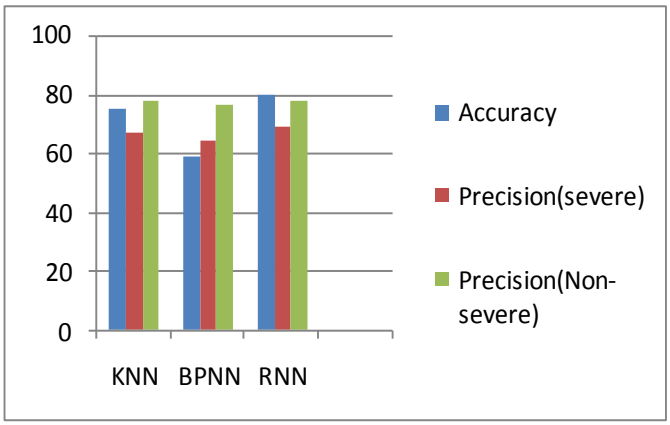
From the above fruits example, it would mean ratio of how many correctly classified apples over total number of fruits classified as apples.

These measures are computed for all the classification tests in this work. In general we can say that if the above measures have higher values the classifier performs well and accuracy is higher. However a high accuracy but lower values for above measures may indicate a poor classifier performance.

Table 1: Comparison of KNN, BPNN and RNN

Component	Accuracy	Precision (Severe)	Precision (Non-Severe)
KNN	75.00%	66.77%	78.03%
BPNN	59.09 %	64.18%	76.40%
RNN	79.54%	68.69%	78.03%

As observed the results of Recurrent Neural Network are found to be quite better than their other counterparts in terms of accuracy. The KNN is found to perform better over back propagation neural network but is still lagging behind recurrent neural network.



VI. CONCLUSION AND FUTURE SCOPE

The thesis dealt with the study of spam classification techniques in emails. The email dataset was obtained the document was pre-processed. Various special characters, stop words etc have been removed and finally the TF-IDF scores of each word are calculated. A novel Recurrent Neural Network based algorithm was utilized for classifying the email dataset as spam and non-spam. The results have been found to be quite satisfactory in terms of accuracy, precision and recall. The results have also been compared to two other algorithms which have been implemented namely: KNN and Back Propagation Neural Network.

It was found that while the KNN performs better than BPNN, the results of RNN are the best. In future, other algorithms can be implemented for comparison purpose. Also, hybrid of RNN and evolutionary algorithms can be utilised for the same. It is assumed that all data set has no noise and missing values because artificial neural network (RNN) have high tolerance to noisy data. So missing value handling along with continues and categorical values handling should be done as future enhancement. The study may be extended using data sources of other open source projects for the validation of proposed approach and findings. The approach used in this study takes KNN and Artificial Neural Network into account for classification. A comprehensive study could be conducted of other ML algorithms. Also, other feature selection methods could be used for crating dictionary of terms. The work in thesis only proposes to implement the proposed approach on offline database downloaded from Enron repository and is not designing any such automated system for online classification. Therefore a study could be conducted to make the system online for real time classification of spam reports.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

REFERENCES

[1] Alsmadi, Izzat, and Ikdam Alhami. "Clustering and classification of email contents." *Journal of King Saud University-Computer and Information Sciences*27.1 (2015): 46-57.

[2] Sculley, D., and Gordon V. Cormack. "Filtering Email Spam in the Presence of Noisy User Feedback." *CEAS*. 2008.

[3] Awad, W. A., and S. M. ELseuofi. "Machine Learning methods for E-mail Classification." *International Journal of Computer Applications (0975–8887)*16.1 (2011).

[4] Dasgupta, Anirban, Maxim Gurevich, and Kunal Punera. "Enhanced email spam filtering through combining similarity graphs." *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 2011.

[5] Zhou, Bing, Yiyu Yao, and Jigang Luo. "A three-way decision approach to email spam filtering." *Advances in Artificial Intelligence*. Springer Berlin Heidelberg, 2010. 28-39.

[6] Elssied, Nadir Omer Fadl, OTHMAN IBRAHIM, and Waheeb Abu-Ulbeh. "AN IMPROVED OF SPAM E-MAIL CLASSIFICATION MECHANISM USING K-MEANS CLUSTERING." *Journal of Theoretical & Applied Information Technology* 60.3 (2014).

[7] Blanzieri, Enrico, and Anton Bryl. "A survey of learning-based techniques of email spam filtering." *Artificial Intelligence Review* 29.1 (2008): 63-92.
Attenberg, Josh, et al. "Collaborative Email-Spam Filtering with the Hashing Trick." *Proceedings of the Sixth Conference on Email and Anti-Spam*. 2009.

[8] Idris, Ismaila, and Shafii Muhammad Abdulhamid. "An Improved AIS Based E-mail Classification Technique for Spam Detection." *arXiv preprint arXiv:1402.1242* (2014).

[9] Youn, S., and McLeod, D. Efficient Spam Email Filtering using an Adaptive Ontology. Proceedings of 4th International Conference on Information Technology: New Generations (ITNG07), Las Vegas, NV, April, 2007.

[10] My Chau Tu, Dongil Shin, Dongkyoo Shin, "A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms", dasc, pp.183-187, 2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, 2009.

[11] Liangxiao Jiang, Harry Zhang, Zhihua Cai, and Jiang Su, "One Dependence Augmented Naive Bayes", accessed online from citeseerx.ist. psu.edu/viewdoc/ download?doi= 10.1.1.90. 437.

[12] Patil BM, Joshi RC, Toshniwal D, Biradar S., "A New Approach: Role of Data Mining in Prediction of Survival of Burn Patients", accessed online from www.springerlink.com/index/8pnh75n137t99892.pdf.

[13] Hu H., Li J., Plank A., Wang H. and Daggard G., "A Comparative Study of Classification Methods for Microarray Data Analysis", In Proc. Fifth Australasian Data Mining Conference, Sydney, Australia (2006).
Martin, Steve, et al. "Analyzing Behavioral Features for Email Classification."*CEAS*. 2005.

[14] Gomez, Juan Carlos, and Marie-Francine Moens. "PCA document reconstruction for email classification." *Computational Statistics & Data Analysis* 56.3 (2012): 741-751.

[15] El-Alfy, El-Sayed M. "Discovering classification rules for email spam filtering with an ant colony optimization algorithm." *Evolutionary Computation, 2009. CEC'09. IEEE Congress on*. IEEE, 2009.

[16] Kumar, R. Kishore, G. Poonkuzhali, and P. Sudhakar. "Comparative study on email spam classifier using data mining techniques." *Proceedings of the International MultiConference of Engineers and Computer Scientists*. Vol. 1. 2012.

[17] Youn, Seongwook, and Dennis McLeod. "Spam Email Classification using an Adaptive Ontology." *JSW* 2.3 (2007): 43-55.

[18] Cui, B.,Mondal, A.,Shen, J.,Cong, G., and Tan, K. On Effective E-mail Classification via Neural Networks. Proceedings of the 16th International Conference on Database and Expert Systems Applications (DEXA05), Copenhagen, Denmark, 85-94, 2005.

[19] Crawford, E.,Koprinska, I., and Patrick, J. Phrases and Feature Selection in E-Mail Classification. Proceedings of the 9th Australasian Document Computing Symposium (ADCS04), Melbourne, Australia, 59-62, 2004.

[20] Diao, Y., Lu, H., and Wu, D. A comparative study of classification based personal e-mail filtering. Proceedings of the 4th Pacific-Asia Conference of Knowledge Discovery and Data Mining (PAKDD00), Kyoto, Japan, 2000.

[21] KIRITCHENKO, S., AND MATWIN, S. EMAIL CLASSIFICATION WITH CO-TRAINING. PROCEEDINGS OF WORKSHOP OF THE CENTER FOR ADVANCED STUDIES ON COLLABORATIVE RESEARCH (CASCON01), ONTARIO, CANADA, 2001.

[22] Matwin, S., Kiritchenko, S and Abu-Hakima, S. Email Classification with Temporal Features. Proceedings of the International Intelligent Information Systems (IIS04), Zakopane, Poland, 523-533, 2004.

[23] Youn, S., and McLeod, D. A Comparative Study for Email Classification. Proceedings of International Joint Conferences on Computer, Information, System Sciences, and Engineering (CISSE06), Bridgeport, CT, December, 2006

[24] Islam, Rafiqul, and Y. Xiang. "Email classification using data reduction method." *Communications and Networking in China (CHINACOM), 2010 5th International ICST Conference on*. IEEE, 2010.

[25] Renuka, D. Karthika, and T. Hamsapriya. "Email classification for spam detection using word stemming." *International journal of computer applications*1.5 (2010): 45-47.

[26] Tretyakov, Konstantin. "Machine learning techniques in spam filtering." *Data Mining Problem-oriented Seminar, MTAT*. Vol. 3. No. 177. 2004.

[27] Zhao, W.Q., Zhu, Y.L. An email classi¯cation scheme based on decision-theoretic rough set theory and analysis of email security, Proceeding of 2005 IEEE Region 10 TENCON, pp. 1-6, 2005.

[28] D. Vira, P. Raja, and S. Gada, ―An Approach to Email Classification using Bayesian Theorem,‖ GJCST. (USA), vol. 12, Issue 13, ver. 1.0, 2012.

[29] A Patra, K.Mandal, S.Roy, S.Sau and S. Kunar, ―An Efficient Spam Filtering Techniques for Email Account,‖ American Journal of Engineering Research, vol. 02, Issue 10, pp. 63-73, 2013

[30] F. Temitayo, O. Stephen, and A. Abimbola, ―Hybrid GA-SVM for Efficient Feature Selection in Email Classification,‖ IISTE, vol. 3, no. 3, 2012.

[31] M. Chang, C.K Poon, ―Using Phrases as Features In Email Classification,‖ ELSEVIER,vol.82, 2009, pp. 1036-1045.

[32] S. J. Delany, P. Cunningham, and B. Smyth, ―ECUE: A spam filter that uses machine learning to track concept drift,‖ Proceedings of the 17th European Conference on Artificial Intelligence (PAIS stream), pp.627-631, 2006.

[33] Youn, Seongwook, and D. McLeod. "A comparative study for email classification." *Advances and innovations in systems, computing sciences and software engineering*. Springer Netherlands, 2007. 387-391.

[34] Zhou, Bing, Yiyu Yao, and Jigang Luo. "Cost-sensitive three-way email spam filtering." *Journal of Intelligent Information Systems* 42.1 (2014): 19-45.

[35] Bergholz, André, et al.,"New filtering approaches for phishing email", Journal of computer security 18.1 (2010): 7-35.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)