International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0063
ISSN (Online): 2279-0071

International Journal of Software and Web Sciences (IJSWS)
(Open Access, Double Blind Peer-reviewed, Refereed and Indexed Journal)
www.iasir.net

Automatic Email Classifiers Using Supervised Learning

Deepansh Pandey
Amity School of Engineering, Amity University
Noida, Uttar Pradesh, India

Abstract: *In the digital era, on average a user receives hundreds of mail per day of which many of them are irrelevant mails. There is lot of work going on in the field on classifiers but there is no definite algorithm or method available for email classification. This paper researches an approach to classify such mails so that user response time can be decreased with increase in efficiency using supervised learning techniques. We have experimented on the efficiency of prominent Naïve Bayes, Maximum Entropy and Support vector machine techniques by creating own dataset. The highest efficiency technique is used for developing the system. The well-known application of email classifiers are spam filters. We have taken this one step forward by developing a system that can automatically classify the mails and assign them to different folders labelled as "Social", "Primary" and "Promotion" on the basis of specific keywords mapping rules that are described in the machine learning algorithm during the training period.*

Index Terms*: Automatic, Email, Classifiers, Supervised Learning, Machine Learning*

I. INTRODUCTION[1]

Emails are highly popular now a days because of their vast integrity in the field of communication for personal to professional use in day to day life. Emails are older than the internet. Emails provide much more easy accessibility and durability than the obsolete way of postcards that we are used to use few years back. Mails are used in every field irrespective of the expanse of the field. They provide significant information both user relevant or irrelevant. In the beginning of the email era we used to have simple end to end message sending devices that put the message in a prescribed user directory.

In the modern era we have devices that need only one button effort to send a mail to other user. With many software available in the market, they share a common flaw that all mails whether relevant or irrelevant fill the mailbox of the user.

The problem of automatic email classification is there from embark of the digital era when all of our society traditional ways are adapting to the new ways of digitalization. The Radicati Group report on email statistics depicts there is an average of 7% increase in the Worldwide Email Accounts every year with every average accounts per user is 1.9 [1]. With the increase in numbers of user there is lot of irrelevant data and Emails. There has been lot of work done on it, but there is no categorical solution to this problem have been found.

Emails classification can be content based or request based. In content based the classification is based on the more weightage of content in which category whereas in the request based classification the user feedback is vital factor for the classification of Emails. Major automatic email classification can be divided into three major categories. These are supervised, unsupervised and semi-supervised classification.

Supervised learning requires prerequisite set of user defined training data or rules which can be further used to train the classifier [2]. We have thoroughly develop a set of rules that will classify the emails into various categories which will provide more better user accessibility to the relevant content. The application of email classifiers that we have seen is spam filtering that can be found in every mailbox based on the user feedback.

II. MOTIVATION

Users want to save their time from searching an important mail in the stack of irrelevant that a user receive daily. The transactional email report from Experian marketing services depicts that less than 16% of the total bulk mails are open or clicked this shows that for 84% of the user those emails are irrelevant for users [3]. For example a user wants to check for a mail that he/she received few hours back but need to search few minutes to find that mail as on average a user receive 216 mails in a day as per the Email Statistics Report of 2016-2020 [4]. This has lead for need of automatic email classifiers that can deduce the number of irrelevant mails for the user. Until now we have email classifiers that can classify the mail whether they are Spam or Ham.

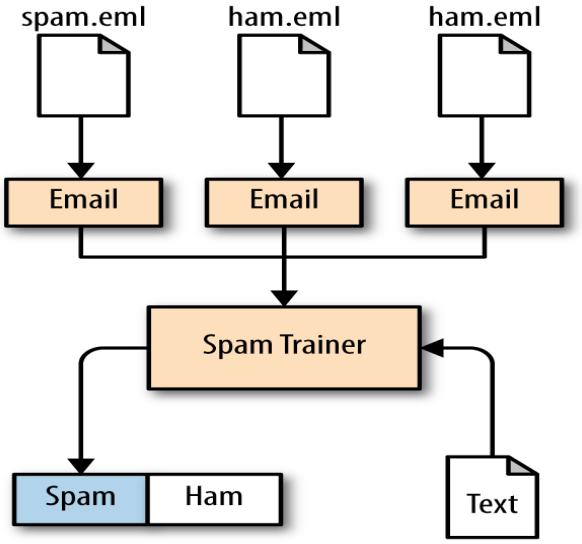

Fig. et al. [12]

We have used supervised learning technique through which the prescribed keywords will be used to classify the mail and assign them to different folders which will be labelled as “Social”, “Promotion” and “Primary”. These will help the user to save time and increase efficiency by reducing the interaction of user with the irrelevant mails. The flow chart of this will be as shown below.

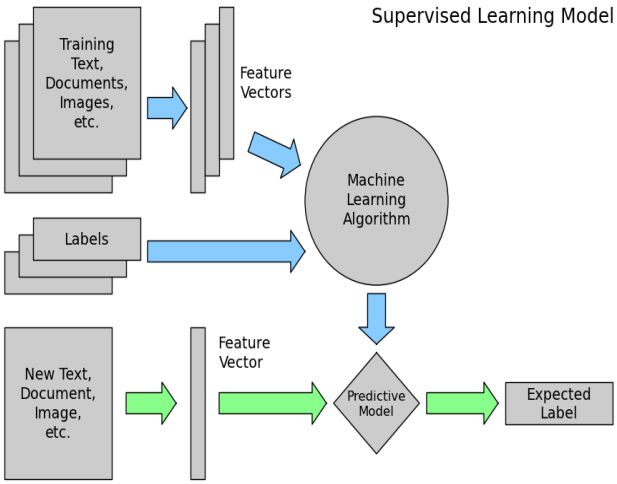

Fig. et. al. [13]

We have considered 30 mails from social, promotion sites altogether 90 mails in the training data set. Then we have created feature vectors for them which is the set of keywords that is for different categories. The labels are the name that are assigned to various categories. These labels are mapped with the feature vector using the machine learning algorithm. This whole process comes under the training phase. After that we have actual test of new data, machine learning algorithm extract the feature from new data and match it with the prescribed label during the training phase. The application of automatic emails classifier can be understood by this example: as per the Email Statistics Report of 2014-2018 [1] a user on average receives 88 consumer mails per day, we can deduce this number of mails by assigning them to different folder which will save lot of user time. Now the user have to search from the 124 mails which will increase user efficiency by 58.4 percent. We have used various data sets for classification on the basis of union of these datasets we have defined our ultimate datasets that can classify all mails with an efficiency of 90% (on running the datasets we are able to classify 9 mails successfully out of 10). Using supervised learning the user have one more benefit that can add his own keywords to the prescribed datasets.

III. EMAIL CLASSIFIER

The email classifier has become necessity and part of most of the stature mailboxes that users use now a days. However, there is no prescribed algorithm that is available on the internet that assures 100% efficiency. There is five procedural steps that depicts the working of an email classifier.

Fig. et al. [14]

The figure shows the gathering of information and analyzing the data for extracting point which is further used for identifying patterns that will be used on discovering answers for unknown inputs for which we have set rules called learning algorithms.

1. Gathering Information

The phase consist of gathering information about various user demands and mails that user gets per day that are both relevant and irrelevant with the prospect of user. The user feedback and response to various emails, for what purpose the user is using email accounts etc. The number of mails may vary from user to user, disorganized and are disintegrated on multiple accounts that one user have. The context of email include many different type of languages, amount of content and context of the email.

2. Analyzing Data

The data received from the first phase is then changed into statistically data that would be helpful for future inference. For example, according to 2015 Email Marketing Metrics Benchmark Study [5] that the best industry performing in click through rate is computer hardware where the worst industry is automobiles which help advertiser to run ad campaign more preferably on computer industry as they has higher click through rate.

3. Extracting Points

In this phase the noisy data which is irrelevant to the process is removed. The main keywords that is relevant for the classification process, these keywords are termed as extraction point. The noisy data may contain stop words, special symbols, hyperlinks etc.

4. Identifying Patterns

The keyword that are extracted in the above phase are analyzed and on the basis of them email will be classified in the prescribed labels. These keywords are stored in a temporary location for further matching with the prescribed rules that are described in the classifier during the training phase of the classifier. These keywords have high impact on the deciding factor to which the email will be assigned to the label.

5. Discovering Answers

During this phase we match the extracted keywords with the ones that is prescribed in the rule that we have defined during the training set. The maximum number of keywords that come under one of label will be assigned to that label and that email will be moved to that label folder. The supervised algorithm helps us to define the set of rule during the training phase which will be helpful during the unknown set of data input.

IV. FEATURE SELECTION

The aim of feature-selection methods is the reduction of the dimensionality of the dataset by removing features that are considered irrelevant for the classification [6]. This transformation helps us to interact with small dataset size which lead to increase the efficiency and response time of the time. Feature selection process should decrease the curse of dimensionality. During our research we have built various datasets for the classification process and using this we have calculated the efficiency of the system. For example we have constructed a dataset in which he have shortlisted all the possible email addresses of the promotion websites and set a rule which maps those email ids to the promotion label. Another dataset that have been constructed is searching for

keyword like Sale, Offer etc. This data set will compare the count of keywords with the other label keywords and the highest number of keyword in the respective label, the email is transferred to that label. The process of feature selection needs to choose unique feature that can be easily differentiate from other features of different labels. Feature selection leads to reduction is search space. The union of feature vector with following attributes that have been selected for three labels are as follows:

A. Social : "Friend Request", "Facebook", "Likes", "Accepted your Request", "Social", "Dating Service", "Added you in circle", "Updated Profile Picture", "Added", "Twitted", "Google+", "Instagram", "Hangouts", "Tweet", "Popular", "LinkedIn", "like", "Twitter".

B. Promotion : "Buy", "Join Us", "Follow Us", "Sign Up", "Offers", "Connect with us", "Get Discount", "Deal", "Grab Deal", "Below", "EBay", "eBay", "eBay:", "eBay!", "Starting at", "Shopping", "Cash on Delivery", "Off", "Return Policy", "Amazon", "EBay", "Flipkart", "Alibaba", "EBay", "Freecharge"

C. Primary: All the other emails which don't have any keyword matching these will be labelled as primary.

The following dataset that have been created for labelling the emails any keyword matching this dataset, will be awarded to their respective labels.

Feature Reduction

In this process of feature reduction we deduce the words or content that is not useful for classification part. For this we have few categories that can be easily deduced for the consideration of the feature vector. These are Stop words or filler words such as "is", "a", "the" used in the emails does not have any effect and therefore need to be filtered out. Another are slang words that is user own writing style and it varies from user to user, it is very difficult to remove slang words from the email. However in business communication we don't use slang words so is easy to label that email as primary email if there is slang words present in it. Other keywords can be extracted from the email.

1. Feature Transformation

Feature Transformation varies significantly from Feature Selection approaches, but like them its purpose is to reduce the feature set size [7]. Feature transformation is used to record potential non-linearity in the dataset. The various techniques used are linear discriminant analysis (LDA) that separates the inter-class groups or clusters. It is a method that is widely used in pattern recognition and machine learning. On the other hand we have sparse representation which uses a sparse dictionary combination. Principal Component Analysis is a well-known method for feature transformation [8]. Its aim is to learn a discriminative transformation matrix in order to reduce the initial feature space into a lower dimensional feature space in order to reduce the complexity of the classification task without any trade-off in accuracy [9]. The feature transformation plays a vital role in feature extraction process. It tends to decrease the size of search space area with selecting only those unique features that distinguishes from other features that in different class.

The main application of feature selection is to shorten the training time with easy interpretation of models by simplifying them. The main goal of feature selection is that the data contains many feature that irrelevant or redundant and thus can be removed without loss of important information. It is different from feature extraction in which we create new features from the function original values whereas in feature selection we choose a subset of the feature domain. Feature selection is used when there are many features and few number of samples available.

V. SUPERVISED LEARNING

Supervised learning entails learning a mapping between a set of input variables X and an output variables Y and applying this mapping to predict the outputs of the unseen data [2]. The process of supervised learning is guided by a trainer who set the rules for mapping between various sets i.e. from input to output basis on the training data set. These mapping rules are generally termed as machine learning algorithm. During the training phase the trainer teaches the classifier on the test data what should be the output for the following input. After the training phase unknown data is input and the trainer checks the output of the classifier. The difference between the expected value and the output defines the efficiency of the classifier. The various techniques that have been associated with supervised learning:

1. Naïve Bayes Classification

The naive Bayes classifier greatly simplify learning by assuming that features are independent given class [10]. Although independence is generally a poor assumption, in practice naive Bayes often competes well with more sophisticated classifiers. It is most commonly used in supervised learning. The basis of Naïve Bayes is probability theory. It depicts an approach to dealt different number of attributes.

$$c^* = \text{argmac}_c P_{NB}(c/d)$$

$$P_{NB}(c/d) = (P(c) \sum_{i=1} m \frac{P\left(\frac{f}{(c)^{ni(d)}}\right)}{P(d)})$$

In this above equation f represents the feature and ni(d) shows the count of fi that are found in the d which is set as 1 mail. The total number of feature vector is 3 with variable number of feature attributes in them. The three feature vectors are used for classifying them in different primary, social and promotion classes.

2. Maximum Entropy Classification

Kamal Nigam, John Lafferty and Andrew McCallum [11] maximum entropy is a probability distribution estimation technique used for classification. The underlying principle of maximum entropy is that without external knowledge, one should prefer distributions that are uniform.

$$P(c|d, \lambda) = \frac{\exp \sum_i \lambda_i f_i (c,d)}{\sum c\prime \exp \sum_i \lambda_i f_i (c',d)}$$

In the equation of Maximum Entropy, c-class, d-message, y-weight i.e. d is the message content of the email, c is class and λ is weight vector. The importance of features in classification is represented by weight vectors.

3. Support Vector Machines

The purpose of support vector machines is to find the decision boundary between the classes where the vectors are of size n.

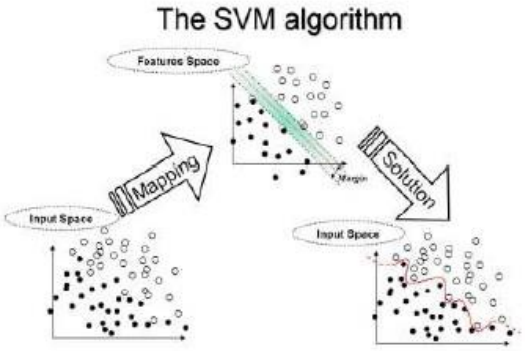

Fig. et. al. [15]

The figure above represents the algorithm for support vector machines in which the mapping of input space and feature space is done and in the end the solution is extracted that distinguishes the input space in different categories based on their features.

The above mentioned techniques are used for classification in the email classifier to check which provide the best efficiency. The classification process should have high response time as well as efficiency to label emails to various categories so the user's time can be saved without displaying irrelevant mails in the primary inbox. These supervised learning techniques comes under the machine learning algorithm in the flowchart [13]. Their efficiency varies with the size of data set and the parameters that are passed through them.

VI. RESULT

The result using the above 3 classifier i.e. Naïve Bayes, Maximum Entropy and Support Vector Machine we have created an email automatic classifier that will classify the mail in social, promotion and primary. These three classifier has provided different result or output every time for different set of data. The classifiers Naïve Bayes and Maximum Entropy have almost same result but support vector machines provide low efficiency results in many cases. The most efficient result in all the observations recorded is provided by the Naïve Bayes classifier. The reason of similar result for Naïve Bayes and maximum entropy classifier is that they have been provided with finite amount of data so they provided result but Naïve Bayes classifier can work on large amount with high efficiency while on the other hand Maximum Entropy classifier can affect the result if there is large amount of data to a high extent. The major differences that have been observed during the process are stated as:

1. Naïve Bayes and Support Vector Machine

Naïve Bayes and support vector machine have varied performance under the influence of variant, datasets and features. During the observation we have classified under which circumstance, the classifiers perform efficiently or not. Both of the classifier are used as baseline for other methods in text classification but we have used them to classify the email on the basis of the content of the emails. In practice Naïve Bayes classifier are more

efficient than support vector machine but in some case support vector machine have performed well. Both of the classifier provides different options, this include choice of kernel function, parameter, data sets etc. They are quite effected by the parameter that are used, quite simple change in parameters can change their output and efficiency significantly. This is termed as parameter optimization. This lead to the fact that for one parameter you may find Naïve Bayes classifier performing more efficiently. Contrary on the other parameter input you may find support vector machine performing more effectively. It is found that if there is negligible amount of overlapping between the classes them Naïve Bayes will be more effective whereas support vector machines work well for high amount of overlapping between the classes. The reason why Naïve Bayes classifier are used so widely nowadays in the field of text classification is that they provide high speed even on large amount of data.

2. Naïve Bayes and Maximum Entropy

There is very finite line that differentiates Naïve Bayes classifier and Maximum Entropy classifier. Naïve Bayes classifiers are probabilistic classifier that uses Bayesian theorem. On the other hand, maximum entropy uses stated prescribed data. So, the main difference between Naïve Bayes and Maximum Entropy is that Naïve Bayes performs more effectively with independent features while maximum entropy perform well with dependent features. We need to have large amount of datasets to train the maximum entropy classifier as training on small dataset can affect the result significantly whereas on the other hand Naïve Bayes cane be trained on a small dataset. Naïve Bayes doesn't uses any algorithm, it is a probabilistic classifier whereas maximum entropy uses algorithm like generalized interactive scaling [16]. The applicability of maximum entropy is vast as compared to Naïve Bayes. Moreover, the feature selection in not an issue in case of Naïve Bayes whereas in maximum entropy classifier it is a complex process as even one feature can affect the result significantly.

VII. CONCLUSION

We have developed own application that can classify the mail in different labels. The need for this application is that nowadays user mailboxes are filled with irrelevant amount of emails. There is no direct algorithm available in books and on web that can be used directly for classification of emails. We have used three classifiers i.e. Naïve Bayes, Maximum Entropy and Support Vector Machine. This paper compares the efficiency of these classifier. The most efficient result is provided by Naïve Bayes whereas Support Vector Machine provided the worst result. There is imperceptible line between Naïve Bayes and Maximum Entropy. During this process we have depicted:

1. We have created a data set of 30 mail of every category set i.e. promotion, social and primary which is used for training the classifier.
2. We have developed a process to delete the irrelevant data such as stop words, special symbols and hyperlinks
3. We have created own set of keywords for classification of mails to the social, promotion and primary label.
4. We have developed an algorithm that will map the input to output using the keyword for matching and assigning to label primary, social and promotion.

The researched system can increase the user efficiency by 58.4 percent by transferring the number of 88 mails consumer mails out of 212 average mails user receive per day.

VIII. REFERENCES

[1] The Radicati Group, Inc. Email Statistics Report, 2014-2018 Editor: Sara Radicati, PhD

[2] **S**upervised Learning P´adraig Cunningham, Matthieu Cord, and Sarah Jane Delany

[3] The transactional email report, Experian Marketing Services

[4] The Radicati Group, Inc. A Technology Market Research Firm, Email Statistics Report, 2016-2020

[5] 2015 Email Marketing Metrics Benchmark Study, IBM Marketing Cloud

[6] Forman, G., An Experimental Study of Feature Selection Metrics for Text Categorization. Journal of Machine Learning Research, 3 2003, pp. 1289-1305

[7] Han X., Zu G., Ohyama W., Wakabayashi T., Kimura F., Accuracy Improvement of Automatic Text Classification Based on Feature Transformation and Multi-classifier Combination, LNCS, Volume 3309, Jan 2004, pp. 463-468

[8] Zu G., Ohyama W., Wakabayashi T., Kimura F., "Accuracy improvement of automatic text classification based on feature transformation": Proc: the 2003 ACM Symposium on Document Engineering, November 20-22, 2003, pp.118-120

[9] M. Ikonomakis, S. Kotsiantis, V. Tampakas, "Text Classification Using Machine Learning Techniques", Wseas Transactions on Computers, Issue 8, Volume 4, August 2005, pp. 966-974

[10] I. Rish T.J. Watson Research Center, "An empirical study of the naive Bayes classifier"

[11] Kamal Nigam, John Lafferty and Andrew McCallum "Using Maximum Entropy for Text Classification"

[12] https://www.safaribooksonline.com/library/view/thoughtful-machine-learning/9781449374075/assets/thml_0402.png

[13] http://radimrehurek.com/data_science_python/plot_ML_flow_chart_11.png

[14] https://www.dtreg.com/uploaded/pageimg/SvmFlow.jpg

[15] http://rainbowresearch.org/wp-content/uploads/data_extraction-300x276.png

IX. ACKNOWLEDGEMENT

I would like to show my gratitude to Dr. Seema Verma, she is associate professor at Amity School of Engineering, Amity University, Noida for sharing her pearls of wisdom with me during the course of this research and I thank her for her so-called insights. I am immensely grateful for her comments on an earlier version of the manuscript, although any errors are my own and should not tarnish her reputation.