Email Classification using NLP & Machine Learning Techniques

Akash Kumar Singh[1] Akshay Nair[2] Krishnakant Mahto[3] Kundan Gadgil[4] Prashant G. Ahire[5]

[1,2,3,4,5]Department of Computer Engineering

[1,2,3,4,5]Dr. D. Y Patil Institute of Technology, Pune-18, India

Abstract— Due to Exponential Growth of online user's individual email accounts are been generally flooded with heap of emails. Most of the times it will be tedious job to individually read and classify these emails based on the desired classification protocol. Many methods are existing which are classifying the emails based on the host location and domain, this often results in unsatisfied outcomes, also there are several systems which use different AI techniques, algorithms like Statistical Bayesian , Naïve Bayes algorithm for classification of emails, but there is hardly any system developed so far to filter e-mails using the Fuzzy ANN. The drawback of existing classification systems is that they are not efficient enough, percentage of accuracy is low and are not capable of handling sarcasm. This paper proposes an approach towards building a model using NLP, Fuzzy ANN and machine learning techniques for classification of emails using pre-defined protocols. This model can be used in big data handling companies, news agencies and government offices.

Key words: Email Classification, Fuzzy ANN, NLP, Machine Learning

I. INTRODUCTION

E-mail is one of the most important and proficient communication methods in our professional and personal lives. It is a cost-effective method of communication commonly found in all areas of industries. Email data mining and analysis can be used for various purposes such as spam detection and classification. Email classification is addressed basically for two purposes: spam mails and general mails. Further, the general mails need to be classified into various folders as per requirements.

A. Fuzzy Logic (FL)

Itis a method of reasoning that resembles human reasoning. The approach of FL imitates the way of decision making in humans that involves all intermediate possibilities between digital values YES and NO. The conventional logic block that a computer can understand takes precise input and produces a definite output as TRUE or FALSE, which is equivalent to human's YES or NO. The inventor of fuzzy logic, LotfiZadeh, observed that unlike computers, the human decision making includes a range of possibilities between YES and NO, such as −

CERTAINLY YES
POSSIBLY YES
CANNOT SAY
POSSIBLY NO
CERTAINLY NO

The fuzzy logic works on the levels of possibilities of input to achieve the definite output.

Fuzzy logic is useful for commercial and practical purposes.

- It can control machines and consumer products.
- It may not give accurate reasoning, but acceptable reasoning.
- Fuzzy logic helps to deal with the uncertainty in engineering.

B. Machine Learning

In simple words, we can say that machine learning is the competency of the software to perform a single or series of tasks intelligently without being programmed for those activities. This is part of Artificial Intelligence. Normally, the software behaves the way the programmer programmed it; while machine learning is going one step further by making the software capable of accomplishing intended tasks by using statistical analysis and predictive analytics techniques. You may have noticed that whenever we like or comment a friend's pictures or videos on a social media site, the related images and videos are posted earlier and keeps on displaying. Same with the people you may know' suggestions, the system suggests us other user's profiles to add as a friend who is somehow related to our existing friend's list. Wondering! How does the system know that? That is called Machine learning. The software uses the statistical analysis to identify the pattern as a user you are performing, and using the predictive analytics it populates the related news feed on your social media site.

C. Natural Language Processing

Natural Language Processing (NLP) refers to AI method of communicating with an intelligent systems using a natural language such as English. The field of NLP involves making computers to perform useful tasks with the natural languages humans use. The input and output of an NLP system can be −

- Speech
- Written Text
- Steps in NLP

There are general five steps

1) Lexical Analysis

It involves identifying and analyzing the structure of words. Lexicon of a language means the collection of words and phrases in a language. Lexical analysis is dividing the whole chunk of txt into paragraphs, sentences, and words.

2) Syntactic Analysis (Parsing)

It involves analysis of words in the sentence for grammar and arranging words in a manner that shows the relationship among the words. The sentence such as "The school goes to boy" is rejected by English syntactic analyzer.

3) Semantic Analysis

It draws the exact meaning or the dictionary meaning from the text. The text is checked for meaningfulness. It is done by mapping syntactic structures and objects in the task domain. The semantic analyzer disregards sentence such as "hot ice-cream".

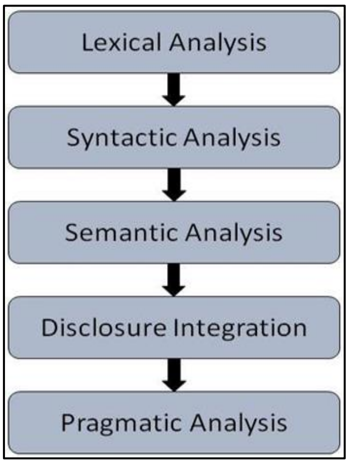

Fig. 1:

4) Discourse Integration

The meaning of any sentence depends upon the meaning of the sentence just before it. In addition, it also brings about the meaning of immediately succeeding sentence.

5) Pragmatic Analysis

During this, what was said is re-interpreted on what it actually meant. It involves deriving those aspects of language which require real world knowledge.

This paper proposes an approach towards building a model using NLP, Fuzzy ANN and machine learning techniques for classification of emails using pre-defined protocols. This model can be used in big data handling companies, news agencies and government offices.This research paper dedicates section 2 for analysis of past work as literature survey, section 3 deeply elaborates the proposed technique and whereas section 4 evaluates the performance of the system and finally section 5 concludes the paper with traces of future enhancement.

II. LITERATURE SURVEY

[2] Y. H. LI & A. K. JAIN, ”Classification of text document“ The author investigate four different methods for document classification, the naïve Bayes classifier, the nearest neighbor classifier, decision trees and a subspace method. These classifier where applied to the seven class yahoo news groups (business, entertainment, health, international, politics, sports and technology) individually and in combination. A study is carried out for three classifier combination approaches, simple voting, dynamic classifier selection and adaptive classifier combination. The author states that naïve Bayes classifier and the subspace method outperform the other two classifier on the given data sets. The author also states that combination of multiple classifier did not always improve the classification accuracy compared to the best individual classifier.

[3]KamalanathanKandasamy & PreethiKoroth, “An Integrated Approach to Spam Classification on Twitter Using URL Analysis, Natural Language Processing and Machine Learning Techniques” Presently people are so much habituated to social networks that it is very easy to spread spam contents through them and one can access the details of any person very easily through these sites. The author concentrate more on spammers in twitter and propose an application which classify a twitter user into spam or legitimate. To achieve this an integrated approach, which contains URL analysis, Natural language Processing and machine learning technique are used. The user of the application has to enter the username of the account to be checked into the interface provided. So, basically the input is the username and output is either spam or legitimate. The last 10 tweets of a user is used for the whole process. The entire work uses three techniques that are URL analysis, natural language processing and machine learning techniques. In URL analysis comparison with a set of blacklisted URLs and comparison with a set of already identified expression is done.

[4]Yilin Yang, Xinhai Huang, XuefeiHao, Zicong Liuand Zhenyu Chen“Natural Language Processing Based Test case Prioritization” The author states that in mobile application development, the frequent software release limits the testing time resource. To detect bugs in early phases, researchers proposed various test case prioritization (TCP) techniques in past years. This paper conducted an extensive empirical study to analyze the performance of three NLP based TCP technologies, which is based on 15059 test cases from 30 industrial projects. The result shows that all of these three strategies can help to improve the efficiency of software testing, and the Risk strategy achieved the best performance across the subject programs.

[5]I.Alsmadi and I. Alhami, Clustering and classification of email contents". The author states that a large set of personal emails is used for the purpose of folder and subject classifications. Algorithms are developed to perform clustering and classification for this large text collection. Classification based on NGram is shown to be the best for such large text collection especially as text is Bi-language.

[6] S.Sayed, Three-Phase Tournament-Based Method for Better Email Classification". The author proposes a tournament-based method to evolve email classification performance utilizing World Final Cup rules as a solution heuristics. The proposed classification method passes through three phases: 1) clustering (grouping) email folders (topics or classes) based on their token and field similarities, 2) training binary classifiers on each class pair and 3) applying 2-layer tournament method for the classifiers of the related classes in the resultant clusters. The first phase evolves K-mean algorithm to result in cluster sizes of 3, 4, or 5 email classes with the pairwise similarity function. The second phase uses two classifiers namely Maximum Entropy (MaxEnt) and Winnow. The third phase uses a 2-layer tournament method which applies round robin and elimination tournament methods sequentially to realize the winner class per cluster and the winner of all clusters respectively.

[7] M. Fuad, D. Deb and M. Hossain, A trainable fuzzy spam detection system".The author presents the design and implementation of a trainable fuzzy logic based e-mail classification system that learns the most effective fuzzyrules during the training phase and then applies the fuzzy control model to classify unseen messages. Their findings imply that automatically trainable fuzzy spam filters are practically viable and can have a significant effect on spam detection.

[8] XiangHui, Statistical-based Bayesian Algorithm for Effective Email Classifcation", IEEE 2015. The author proposes, a spam detection method is proposed upon

statistical based Bayesian algorithm. Firstly, the method use actual priori probability of spam instead of constant probability.Secondly, the selective range and rules of tokens is improved. Finally, our method add URLs and images into detection content. The experiment result shows that the improved statistical based Bayesian classification algorithm works well in practice.

III. PROPOSED METHODOLOGY

This section of the paper describes the implementation strategy of email Classification process with below described steps.

1) Step 1: In the first step of the proposed model a workbook is fed to the system which contains come attributes like to Email id, Subject and body of the Email. This workbook is read in the form of two dimension list using an external API called JXL.
2) Step 2: Once the Workbook data is read into a double dimension list, then it is subjected to email engine. This Email engine is made by using external API provided by the Google for Gmail hosts. Which eventually sends the email to all of the userid that are mentioned in the workbook with the respective subject and contents.

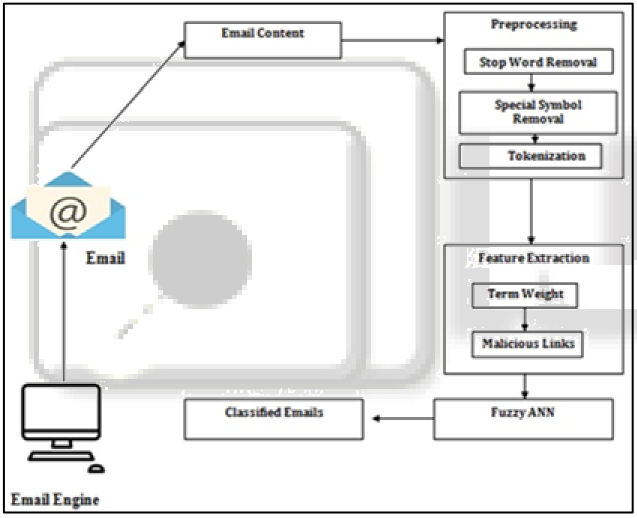

Fig. 1: System Overview

3) Step 3: Email Reader - This is the part of the proposed model where it reads all the unread Emails from the recipient email id using the passwords which is stored in database. To do this proposed model again uses the external API provided by the Google for the grail hosts. Then this all body and subjects of the respective Emails are stored in the two dimensional list that is forwarded to the preprocessing process.
4) Step 4: Preprocessing -From the Email list obtained through the prior step, each of these email string is being accessed to preprocess the same. Preprocessing of the string is the process of reducing the data by removing redundancy from it. This process contains the following divisions as mentioned below

– Special Symbol Removal

Here in this step all the words are being separated into a list. Then these words are being checked for the special symbols like,,.,:,;,?,{,},[,] etc…If a word contains any of these special symbols, then it is being replaced with the empty character to get the plane word.

– Stopword Removal

Here this process eventually drops off the conjunctive words in the English language. For example is, of, the, and, are we, etc... To do this, first special symbol freed word is iteratively checked in the list of stopwords for the presence of the same (that is collected through the different web respiratory). If the word contains any of the Stopword then it is being replaced with the empty character to get rid of it.

By doing this the meaning of the phrase remained unchanged. For example, if the sentence is there like “This is a good mouse". Then after Stopword removal process it becomes “good mouse". This shows the soul of the phrase never changed even though it shreds off the conjunctions form it. Rather than this it reduces the weight of the sentence to lessen the space complexity of the instance.

– Stemming

Stemming is the technique of trimming a word for any kind of extension regarding its tenses. For Example, if a word is there like "going", this word is the extension of the base word "go". So in stemming process word is being trimmed with the tense extension to become "go". So the base word "go" retains the soul of "going".

To attain this many stemming algorithms are existed like port stemmer, stans stemmer and many more. Each individual algorithm follow their own protocol and database to stem a word to bring down the same to its base form. Again, this process also reduces the space complexity of the process and thereby increase the efficiency of the same.So the proposed model of stemming uses the string replacement process to shred the extension tenses.

5) Step 4: Feature Extraction - This is the step where some important feaures are been extracted from email body string as mentioned below.

– Term Weight

Here term weight is being measured as the frequency of each words from the string. For this process email body string is split on the space character and then it is being added into a list. Then this list is subjected to hash set from where we can get unique words.

Then each of this unique word is been counted for its kind in the string to get the frequency of the same.

– Positive and negative words

Here in this step positive and negative words are stored in the bag of list. And then each of the words from the preprocessed list is being checked for its type to classify them.

These each words are further scrutinize for the sarcasm words based on the word tree protocol. All of these words are stored in seperatelise to process further to classify the email content.

– Malicious Links

Here in this part of the proposed model, email strings are being checked for special characters that are arriving in malicious URLs. It is searching for characters like %, / and many more and store the URL in a list.

6) Step 5: Fuzzy ANN - This is the final step of the classification process of the proposed model. Where each of the extracted features are been converted into numerical score based on the positive and negative

presence of the strings which are catalyzed by thier respective term weight or malicious links.

Then these values are subjected to Fuzzification process where each of the Email body is considered as the neuron representing with a score. These Email contents are converted into fuzzy crisp value ranges like Very Low, Low, Medium, High and Very High. Based on these crisp values IF- THEN rules are being applied to get the classified emails according to the given criteria. The process of Fuzzy ANN can be shown with the below mentioned alogorithm1.

A. Algorithm 1: Fuzzy ANN

```
//Input : Email Score Vector $E_v$
//Output: Classified list $C_L$
Start
Set small=0, big=0
    For i=0 to size of $E_v$
        $T_{Set} = E_{vi}$  [ $T_{set}$ = Temporary Set]
        Sc=$T_{set[1]}$
        IF ( Sc <small) [ sc= Score]
        small=Sc
        IF(Sc>big)
        big=Sc
End for
    d=( big-small)/5    [ d= Distance ]
      For i=1 to 5
        IF(i==0)
      Fc(min=small, max=d) [ Fc = Fuzzy Crisp Set ]
        else
        Fc(min=Fci-1(max),max= Fci-1(max)+d
      End For
        For i=0 to Size of Fc
          For j=0 to size of Ev
          $T_{Set}$ = Evi   [ Tset = Temporary Set]
          Sc=Tset[1]
            IF Sc ∈ Fci
            add $T_{Set}$ to $C_L$
            END IF
          End For
        End For
return $C_L$
```

IV. Result & Discussion

The proposed methodology of email classification model is deployed in windows based java machine using NetBeans as IDE. To prepare Email engine model uses the java mail API for Gmail host.

To measure the performance of the system for the accuracy of the classification, model considers the precision and recall as the measuring parameter.

Precision and recall are considered as the one of the best parameter to measure the performance of our system. Precision can be defined as the positive classified values that indicates the amount of relevant classification done through the system. Precision can be described as the ratio of number of relevant emails are classified for the input number of emails to the sum of number of relevant and irrelevant emails are classified for the input number of emails. Relative effectiveness of the system can be evaluated thoroughly by using precision parameters.

Recall indicates the relevant results extracted over the extracted relevant results. Recall can be described as the ratio of number relevant emails are classified to the sum of relevant emails are not classified. Absolute accuracy of the system can be properly denoted by this system.

Precision and recalls can be more effectively explained as below

- A = The number of relevant emails are classified for the given number of Emails
- B= The number of irrelevant emails are classified for the given number of Emails
- C = The number of relevant emails are not classified for the given number of Emails
  1) So precision can be given as
  2) Precision = (A / (A+ B)) *100
  3) Recall = (A / (A+ C)) *100

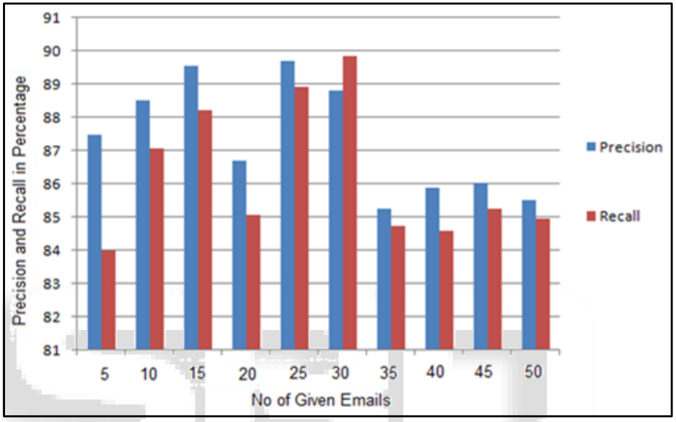

Fig. 2: Performance Evaluation through Precision & Recall

The graph in above figure 2 clearly indicates that proposed model for user behavior classification achieves 87.34 % of average precision and 86.27 % of average recall for the given number of Email. Which is high and indicates successful deployment of the proposed idea.

V. Conclusion

The proposed methodology of Email Classification is developed on real time emails from gmail host. All the emails are first collected from the host through java mail API supporting for gmail. Each email content are efficiently scrutunized using the preprocessing and feature extracting process. By using Efficient Classification theory like fuzzy ANN emails are classified for the products with the labels like Happy, Satisfactory, Just Okay, Disappointed and Worst. The precison and recall of our system shows that it yields better result by providing high values.

Proposed model can be design and deploy in real time email cloud servers using distributed computing.

References

[1] Akash kr singh, Akshay nair, Krishnakant mahto and Kunan gadgil "Survey paper on Email classification" Department of Computer Engineering, DIT, Pune

[2] Y. H. LI and A. K. JAIN,"Classification of text document"Department of Computer Science and

Engineering, Michigan State University, East Lansing, Michigan,USA.

[3] KamalanathanKandasamy and PreethiKoroth, “An Integrated Approach to Spam Classification on Twitter Using URL Analysis, Natural Language Processing and Machine Learning Techniques” 2014 IEEE Students’ Conference on Electrical, Electronics and Computer Science.

[4] Yilin Yang, Xinhai Huang, XuefeiHao,ZicongLiuandZhenyu ChenAn Industrial Study of Natural Language Processing Based Test case Prioritization, 10th IEEE International Conference on Software Testing, Verification and Validation

[5] Alsmadi and I.Alhami, "Clustering and classification of email contents", Journal of King Saud University - Computer and InformationSciences, vol. 27, no. 1, pp. 46-57, 2015.

[6] S. Sayed, "Three-Phase Tournament-Based Method for Better Email Classification", International Journal of Artificial Intelligence and Applications, vol. 3, no. 6, pp. 49-56, 2012.

[7] M. Fuad, D. Deb and M. Hossain, "A trainable fuzzy spam detection system", in 7th International Conference on Computer and Information Technology, 2004.

[8] XiangHui Statistical-based Bayesian Algorithm for Effective Email Classification 2016 3rd International Conference on Information Science and Control Engineering.