Journal of Current Research in Engineering and Science
JCRES
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 4- Issue 2, Paper 11, July 2021

SPAM EMAIL CLASSIFICATION USING NLP

Er. Farhana Siddiqui[1], Khan Suhail Nisar[2], Talha Atique Ansari[3], Kazi Zaki Haseeb[4]
[1]Assistant Professor, Dept. Of Comp. Engg
M. H. Saboo Siddik College of Engineering, Byculla, Mumbai 400 008

Abstract— *As title of the project suggests spam email Classification using NLP (i.e. Natural Language Processing) it is one of the major issue faced by every working professionals , organizations as well common to every people using electronic mail . As we receive bulging amount of spam mails daily it gets really hard for us to differentiate them, as well it is time consuming. In this project we will classify mail as spam and ham(i.e. not spam) by supervised training of the model using Naive Baye's classifier method .Naive Baye's classification is based on Baye's Theorem.*

I. INTRODUCTION

Email is one of the most commonly used modes of communication in this modern era for Education , Banking , Advertisement etc . As the technology is advancing cyber crime has also increased . Spam mail means unwanted mail or unused mail, that means there are no use in present and not in future . Ham is opposite of spam means wanted and useful mail . There are mails containing advertisement of some commercial websites for purchasing their products . As they are from unknown sources our inbox get filled with very huge amount of spam messages .

So to overcome with we will prepare a model that will categorize the messages received by our devices as spam or ham . In order to achieve this , data from the messages is to be collected first and natural language processing techniques are to be applied on it . This spam filtering technique will help the mobile user to have better visualization of the inbox . Unnecessary mails will be marked as spam so mobile user need not to waste their time reading them.

II. LITERATURE SURVEY

Identifying spam messages has been done by various methodologies . Below are some of the approaches .

1. Sethi, P., Bhandari, V., &Kohli, B. (2017). "SMS spam detection and comparison of various machine learning algorithms." 2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN).
2. DelviaArifin, D., Shaufiah, &Bijaksana, M. A. (2016)."Enhancing spam detection on mobile phone Short Message Service (SMS) performance using FP-growth and Naive Bayes Classifier."2016 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob).
3. Gupta, M., Bakliwal, A., Agarwal, S., &Mehndiratta, P. (2018). "A Comparative Study of Spam SMS Detection Using Machine Learning Classifiers." 2018 Eleventh International Conference on Contemporary Computing (IC3).
4. A. Malge and S. M. Chaware, ''An efficient framework for spam mail detection in attachments using NLP,'' Int. J. Sci. Res., vol. 5, no. 6, pp. 1121–1125, May 2016.

Journal of Current Research in Engineering and Science
JCRES
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 4- Issue 2, Paper 11, July 2021

III. Existing system

There are three classification technique that can be used for Classification of email Naive Bayes Classifier (NB) ,

1. Naive Bayes classification (NB) :-

Naive Bayes Classifier was proposed for spam recognition. Naive Bayes Classifier works best with Natural language Processing (NLP) problems . Naive Bayes uses probability theory and Bayes theorem to predict the tag of text . It is probabilistic classifier , that means it calculates the probability of each tag for the given text and then output the tag with highest one . The way this probabilities are calculated is by Bayes theorem that describes probability of a feature based on prior knowledge of conditions that may be related to that feature . Formula for Naive Bayes is as given below . Bayes Theorem:

Prob (B given A) = Prob (A and B) / Prob (A)

2. K – Nearest Neighbor (KNN) :-

KNN is a simple supervised machine learning algorithm that can be used for Classification as well as regression. KNN works by finding the distance between the query and all the examples in data and selects the specified number of examples closest to the query after that it votes for the most frequent label . But KNN's main disadvantage is that it becomes slower in practical aspects when size of the dataset increases .

3. Support Vector Machine (SVM) :-

Cortes and Vapnik (1995) implemented support vector machines to reduce classification error while increasing the margin between two classes (Vahid et al., 2018). Decision planes, which describe decision boundaries, are the foundation of support vector machines. A decision plane divides a group of objects that belong to various classes (Kishore et al., 2012).

IV. System implementation

For our Project we have used Naive Bayes Classifier.

Algorithm:

Step 1: Select the email

Step 2: Extract features with help of tokenization and word count algorithm.

Step 3: Training the dataset with the help of Naive Bayesian Classifier.

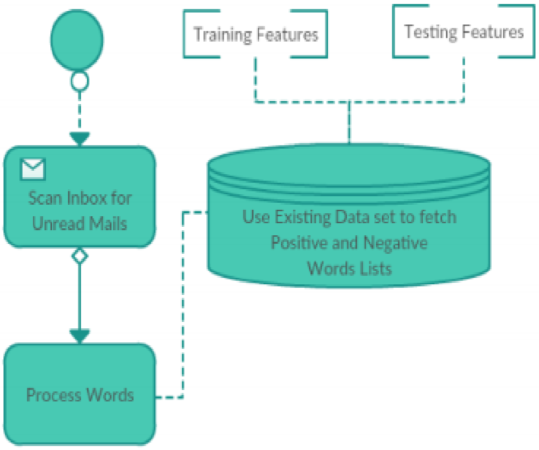
Training Features
Testing Features
Scan Inbox for Unread Mails
Use Existing Data set to fetch Positive and Negative Words Lists
Process Words

Figure 4: Word processing and classification for training using existing dataset

Step4: Find the probability of spam and non-spam mails.

Prob_spam = (sum(train_matrix (spam_indices,)) + 1) ./ (spam_wc + numtokens)

Prob_nonspam=(sum(train_matrix(nonspam_indices,)) + 1)./(nonspam_wc+numtokens)

Step 5: Testing the dataset

log_a = test_matrix*(log(prob_tokens_spam))' + log(prob_spam)

log_b = test_matrix*(log(prob_tokens_nonspam))'+ log(1 - prob_spam)

if

output = log_a > log_b then document are spam

else the document are non-spam

Step 6: Classify the spam and non-spam mails.

Step 7: compute the error of the text data and calculate the word which is wrongly classified.

Numdocs_wrong = sum(xor(output, text_lables))

Step 8: display the error rate of text data and calculate the fraction of wrongly classified word

Fraction_wrong = numdocs_wrong/numtest_docs

V. METHODOLOGY

We have proposed following methodology to implement our project

1. Preprocessing

The messages have to be pre-processed for the removal of unwanted punctuation, grammar, stop words etc.

2. Label Encoding

Label Encoder encode labels with values between „0‟ and „n-1‟ where n represents the number of distinct labels for the classes. Same value is as assigned to the labels which are repeated earlier. We converted the class labels to binary values in our experiment, with 0 indicating ham and 1 indicating spam.

3. StopWord Removal

When using Natural Language Processing(NLP), our goal is to perform some analysis or processing so that a computer can respond to text appropriately.

A machine cannot understand the human readable form. So, data has to be pre-processed in order to make it machine-readable. This is "pre-processing" of which one of the major forms is to filter out useless data. This useless data (words) is generally referred as „stop words‟ in Natural Language Processing(NLP).

4. Lemmitization

It is the method of combining a word's various inflected forms so that they can be analysed as a single item. For example, "include", "includes," and "included" would all be represented as "include". In contrast to stemming, lemmatization preserves the context of the sentence (another buzz word in text mining which does not consider meaning of the sentence).

5. Vectorizing the text

What does it mean by vectorize our data . The aim of vectorization is to break down our features into distinct values that can be used later. One simpe way to vectorize our data is to look at an email and count every word that is used . So for this purpose we will be going to use CountVectorizer.

5. Feature Generation

Feature engineering is the process of constructing features for machine learning algorithms by using the knowledge of that specific domain. The words in each text message are the features on which the algorithm will predict the output. Tokenizing each term will be needed for this reason. The most common 1500 words that are generated in feature generation will be used as our features. Then the data is split in training and testing datasets with a test size of 25%.

5. Implementation of Algorithm

We need to import each algorithm from scikit-learn library along with performance metrics. We require accuracy score and classification report metrics to predict the accuracy and give a classified report on the output.

VI. Conclusion

The previously collected mails are taken as dataset and for each input in the set, a class is predicted and given as output. The messages are first tagged correctly to apply algorithms on them. Applying various classifiers helps us to know the best and the worst algorithms for a problem.

The Naive Classifiers has given us very good accuracy of about 98%. This model need to be improved to understand sarcasm, context on the whole which could be essential while detecting spam.

Acknowledgement

I would like to thank Mrs. Farhana Siddiqiui, Professor, CSE department, for guiding us throughout the project.

References

[1] Masurah Mohammad, Ali Selman "An Evaluation on the Efficiency of Hybrid Feature Selection in Spam Email Classification", IEEE,2015.

[2] C. Bala Kumar, D. Ganesh Kumar "A Data Mining Approach on Various Classifiers in Email Spam Filtering", IJRASET, May 2015.

[3] Vinod Patidar, Divakar Singh, Anju Singh "A Novel Technique of Email Classification for Spam Detection ", International Journal of Applied Information Systems (IJAIS), Volume 5 – No. 10, August 2013.

[4] Archit Mehta ,Raunakraj Patel "Email Classification using data Mining", IJARCCE, 2011.

[5] Rachana Mishra and R.S. Thakur, "Analysis of Random Forest and Naïve Bayes for Spam Mail using Feature Selection Categorization", *International Journal of Computer Applications*, vol. 80, no. 3, pp. 43-48, 2013.

[6] Priyanka Sao and Kare Prashanthi, "E-mail Spam Classification Using Naïve Bayesian Classifier", *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, vol. 4, no. 6, pp. 2792-2797, 2015.

[7] Payal Prajapati, Tarjani Vyas and &Somil Gadhwal, A Survey and Evaluation of Supervised Machine Learning Techniques for Spam E-Mail Filtering, IEEE.