

SPAM MAIL FILTERING SYSTEM

Project ID:2020MP115 Review -II

Group Members

RA1611003030473 ASHAY MISHRA

RA1611003030490 GAURAV BAKHRU

RA1611003030505 KUSHAGRA SAXENA

RA1611003030446 AGAMYA MEHROTRA

Supervised By:

MS MEGHA AGARWAL

SRM Institute of Science & Technology



Table Of Contents

- 1 [Abstract](#)
- 2 [Literature Survey](#)
- 3 [Identification of Research Gap and Problem](#)
- 4 [Expected Impact on Academics/ Industry](#)
- 5 [Methodology of the Project Work](#)
- 6 [Detailed Design](#)
- 7 [Results Obtained](#)
- 8 [References](#)

Abstract

- Spam is a universal problem with which everyone is familiar. A number of approaches are used for Spam filtering .
- The most common filtering technique is content -based filtering which uses the actual text of message to determine whether it is a spam or not.
- The content is very dynamic and it is very challenging to represent all information in a mathematical model of classification. For instance, in content based spam filtering, the characteristics used by the filter to identify spam messages are constantly changing over time.
- Naive Bayes method represent the changing nature of message using probability theory and support vector machine represent those using different features.
- These two method of classification are efficient in different domain.
- These two methods do not consider the issue and it is interesting to find out the performance of both the methods in the problem of email spam filtering.
- In this project Naive Bayes and SVM classification techniques and other ML techniques will be implemented to classify

1. Paper: An Efficient Spam Filtering Techniques for Email Account
Author and Year: S. Roy, A. Patra, S.Sau, K.Mandal, S. Kunar 2013

Conclusion: The email client system that has capability to send email and receive email and project mainly concerned about an efficient email spam filtering techniques for an email account. For this system, we collected statistical data by which we create a training set. This dataset is updated time by time. The filtering techniques based on Naive bayes Theorem, which is a good one machine learning algorithm.

2. Paper : MACHINE LEARNING METHODS FOR SPAM E-MAIL CLASSIFICATION

Author and Date: W.A. Awad and S.M. Elseuofi, 2011

Conclusion: In this paper we review some of the most popular machine learning methods and of their applicability to the problem of spam e-mail classification. Descriptions of the algorithms are presented, and the comparison of their performance on the Spam Assassin spam corpus is presented, the experiment showing a very promising results specially in the algorithms that is not popular in the commercial e-mail filtering packages

Identification of Research Gap and Problem

- Many Email spam filtering systems are there but the cost is high.
- Spam filters based on deep neural network are effective but training costs are high and need lot of data.
- This system ,instead of using one system or algorithm combines several different algorithm to classify mail as spam or non spam

Expected Impact on Academics/ Industry

- If this system is used it will help to classify mail more effectively and with a much lesser cost.
- This idea of mail filtering maintain balance between cost effectiveness and classifying accuracy.
- The senders can be punished, if tracked, under the cyber crime law.

■ SVM(SUPPORT VECTOR MACHINE)

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space with the value of each feature being the value of a particular coordinate.



Decision Trees

Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter.

■ Vectorization

A vectorized function takes a nested sequence of objects or numpy arrays as inputs and returns a single numpy array or a tuple of numpy arrays. The vectorized function evaluates pyfunc over successive tuples of the input arrays like the python map function, except it uses the broadcasting rules of numpy.

■ Naive Bayes Algorithm

Naive Bayes is a classification algorithm for binary (two-class) and multiclass classification problems. It is called Naive Bayes or idiot Bayes because the calculations of the probabilities for each class are simplified to make their calculations tractable.

■ Data Preprocessing

Text Cleaning is a very important step in machine learning because your data may contains a lot of noise and unwanted character such as punctuation, white space, numbers, hyperlink and etc
Some standard procedures that people generally use are:

- convert all letters to lower/upper case
- removing numbers
- removing punctuation
- removing white spaces
- removing hyperlink
- removing stop words such as a, about, above, down, doing etc.

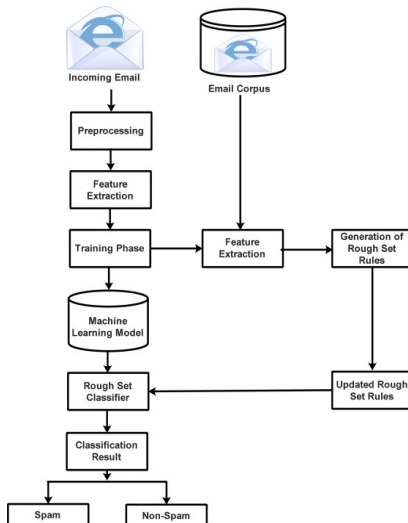
■ Stemming

Stemming is the process of reducing inflection in words to their root forms such as mapping a group of words to the same stem even if the stem itself is not a valid word in the Language. It is a technique in the field of Natural Language Processing that are used to prepare text, words, and documents for further processing

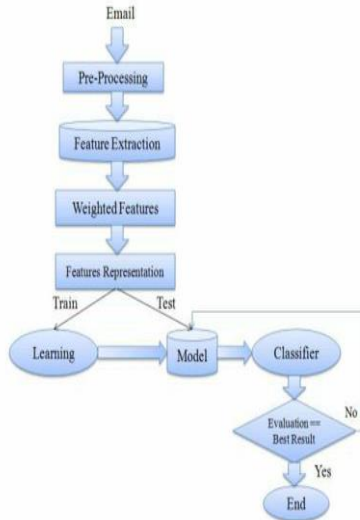
■ Word Cloud

Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud. Word clouds are widely used for analyzing data from social network website

ARCHITECTURAL DESIGN



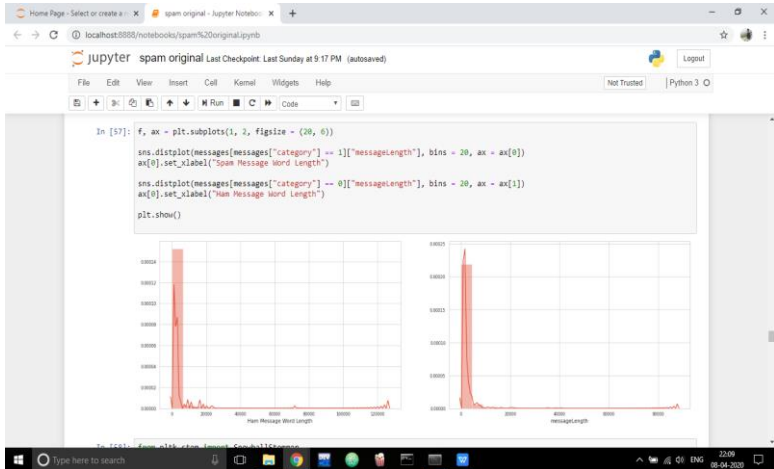
DATA FLOW DIAGRAM



[illegible]

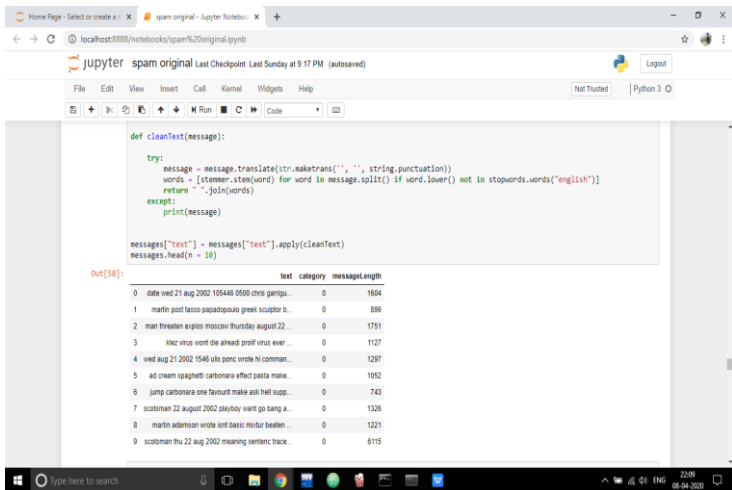
SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
Deemed to be University u/s 3 of UGC Act, 1956

Results Obtained



SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
Deemed to be University u/s 3 of UGC Act, 1956

Results Obtained



The screenshot shows a Jupyter Notebook running in a web browser at localhost:8888. The notebook is titled "spam original" and shows a code cell with a function `cleanText` and a data manipulation operation. The output of the code is a table of 10 rows, each representing a message and its category and length.

```
def cleanText(message):  
    try:  
        message = message.translate(str.maketrans('', '', string.punctuation))  
        words = [stemmer.stem(word) for word in message.split() if word.lower() not in stopwords.words("english")]  
        return " ".join(words)  
    except:  
        print(message)  
  
messages["text"] = messages["text"].apply(cleanText)  
messages.head(n = 10)
```

Out[58]:

	text	category	messageLength
0	date wed 21 aug 2002 105446 0500 chris garri...	0	1604
1	martin post tasso papadopoulos greek sculptor b...	0	899
2	man threaten explos moscow thursday august 22 ...	0	1751
3	klez virus wont die ahead profit virus ever ...	0	1127
4	wed aug 21 2002 1546 uis ponc vrole hi comman...	0	1297
5	ad cream spaghetti carbonara effect pasta make...	0	1052
6	jump carbonara one favourit make ask hell supp...	0	743
7	scotsman 22 august 2002 playboy want go bang a...	0	1326
8	martin adamson wrote iont basic mistur beaten ...	0	1221
9	scotsman thu 22 aug 2002 meaning sentenc trace...	0	6115



SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
Deemed to be University u/s 3 of UGC Act, 1956

References

- Christina, V., S. Karpagavalli, and G. Suganya. 2010. “Email Spam Filtering using Supervised Machine Learning Techniques”. International Journal on Computer Science and Engineering. 2(09):3126-3129
- Bahgat, E.M., S. Rady and W/ Gad. 2016. “An Email Filtering Approach using Classification Techniques”. In: The 1st International Conference on Advanced Intelligent System and Informatics (AISi2015), November 28-30, 2015. Springer International Publishing: BeniSuef, Egypt. 321331.
- <https://www.kaggle.com/ozlerhakan/spam-or-not-spam-dataset>
- https://www.researchgate.net/publication/320703241_E-Mail_Spam_Filtering_A_Review_of_Techniques_and_Trends

THANK YOU