**A NOVEL APPROACH FOR CLASSIFYING EMAILS**

**Submitted by,**

Vishnu Sivan

Arya Rajan K

Hema Dileep

Hima M Nair

M A College of Engineering Kothamangalam

**SYSTEM STUDY**

**EXISTING SYSTEM**

The existing system is completely manual. All emails are reading and analyzing by a person for understanding information such as vendor of the business, description, price and tax of the product. Also the process of classifying the invoices according to vendors and determining which invoice is suitable for each product in terms of cost is completely done manually. When it comes to deal with a huge number of emails the task becomes more complex.

Maintaining such a system manually is a tedious and time-consuming process. It is very difficult to create a consistent and integrated data of all invoices. So an ultimate solution for handling a huge amount of emails that contains invoice details is needed.

**PROPOSED SYSTEM**

For overcoming the limitations of the existing system, we have planned to design an automated system, which uses the Robotic Process Automation (RPA) techniques to track e-mails and analyze the invoice. For implementing such a system, we uses the concept of Natural Language Processing (NLP) and the programming language used is Python which is one of the open source programming language that supports scripting as well as rich in built in modules.

**Architecture Diagram**

Fetch data from Email

(Using Stmplib, email modules)

Excel Sheet Creation / Updating

( Using xlsxViewer module )

NLP

Backend Submission

Noise Elimination

Feature Extraction on Textual data

Attachment Analysis

Classification

(Using Naive Bayes classifier)

Fetch data from PDF

(Using PDF2 module)

Fetch data from JPEG Image

(Using Tesseract OCR)

Feature extraction on Attachments

The procedure is given below,

1. Read the entire content of newly arrived emails from the dedicated email ID using stmplib, email modules.
2. Categorize the email based on the predefined features using NLP techniques:
   1. The emails may contain some unwanted data (data that are not relevant for classification). They are eliminated.
   2. The emails usually contain textual data and attachments. The sender id, subject and description are considered as textual data. Feature extraction on the textual data is carried out by the help of scikit-learn module. Using this module, we can find the features that are best for classifying mail into particular class.
   3. The attachments contains information regarding invoice details of the product. The attachments can be of any format. For reducing the complexity the format is restricted to JPEG or PDF. Attachment analysis phase is used to extract features from an attachment. To do this, read entire contents on the attachment into a textual format or as a string.
      1. Data from pdf attachments are converted into textual format by using xlsxViewer module.
      2. Data from jpeg attachments are converted into textual format by using Optical Character Recognition process. Tesseract OCR algorithm is used for this purpose.
   4. After performing feature extraction on the data, we need to classify it according to different products and corresponding vendors. For this we use Naïve Bayes Classifier. Naïve Bayes classification method is one of the most practical approach to classify data. It uses Bayes theorem for classification.
3. For further proceedings, store the results into an excel sheet using xlsxViewer module.
4. For more secured access, we backup the classified data into a database by the help of pymysql module and also filter the result.