International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor (2018): 7.426

Intelligent Email Extraction and Classification with NLP & Deep Learning

Madhu Sudhan H V

M.B.A. - Finance, Christ University, Bangalore, India

Abstract: ***The information extraction of data from unstructured sources has opened new opportunities for querying, organizing and analyzing data with the help of Artificial Intelligence and Deep Learning techniques. The field of information extraction has its root to natural language processing by using widely used techniques like named entity recognition, parts of speech tagging etc. Recent reports confirm that email is still number one online activity. Corporate users send and receive an average of 110 messages per day, out of which about one third are messages sent. IT Helpdesk normally receives queries through emails, it is required for a human to understand the customer problem and extract certain information from the mail and then create a ticket to take further steps. Manual Intervention can be avoided with the help of Natural Language Processing and Deep Learning.***

Keywords: Artificial Intelligence, Natural Language Processing, Deep Learning, Python

1. Introduction

Intelligent Email Extraction (IEE) is a combination of IT Systems, tools and methodologies that enables machines to interact naturally with their environment, people and data. These systems create more intuitive interactions and extend the capabilities of what either human or machine can do on their own. IEE is developed using NLP & Deep learning algorithms. It is integrated with other advanced automation components like Email ingestion, Document integration, etc. to provide a seamless end to end 'Email management solution'. This solution has been programmed to suit 60+ business rules with an ability to detect 55 languages for appropriate processing, making it completely customized for the client's processing.

2. Process & Architecture

Below figure depicts the process flow for the AI engine, it has been broken into various modules. Individual modules are as below.

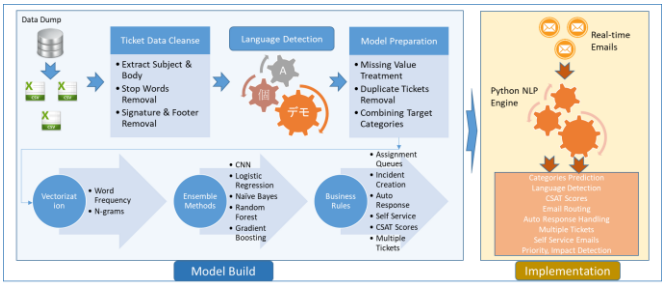

Figure 1: Architecture Design

2.1 Data Dump

Historical ticket's data are received periodically from data source team. This data can be residing in a ticketing tool like Service Now, SalesNow etc. This data is further used to train the model at a stipulated time interval. The data is received as flat files, or it will be dumped in a certain location on SFTP to pick it up.

2.2 Ticket Data Cleanse

This module is used to cleanse the raw data dump received, subject and body will be extracted from the module, stop words as well signature will be removed from the raw data. Data preprocessing will be performed using NLP for stemming, lemmatization etc. To extract relevant information from the subject and body of the mail, named entity recognition and POS tagging are used. Regex is also implemented to extract data with certain patterns.

2.3 Language Detection

This module is used to detect the language of the email by using the Body and Subject content from the mail. It supports up to 55 languages. Python module "langdetect" is used to achieve the task.

2.4 Model Preparation

This module is run during the model refresh, Module is capable of handling missing values, duplicate tickets removal from raw email dump and combine some of the categories to improve the model accuracy.

2.5 Vectorization

Features extraction is performed with word frequency and TFIDF from the Subject and Body of the mail. CountVectorizer module from Scikit Learn is used for feature extraction. Ngrams are used while feature extraction to boost the accuracy.

2.6 Ensemble Methods

Ensemble of algorithms were tried to achieve the best accuracy with the voting classifier. We have used CNN, Logistic Regression, Naïve Bayes, Random Forest and Gradient Boosting.

2.7 Business Rules

IEE has various modules for Business Rules implementation, below is the list of business rules implemented for IEE

- Self Service Mails
- Priority, Impact and Urgency of Tickets
- Ticket Assignment Process
- Auto Resolution
- HR Tickets
- External Resources Requests
- Follow-up Tickets
- Spam Filtering

2.8 CSAT Prediction

Customer Satisfaction Scores are predicted whether he would be satisfied even before solving the ticket so that agents can address the tickets with more care to improve the customer satisfaction. Model was trained based on historical customer satisfaction survey.

3. Implementation

3.1 POST Request Details

Raw email will be received through a web service API. The fields that we receive through the API would be from, to, subject, body, date and other details. Below is sample for request.json that will be fed to Intelligent Email Extraction service.


```
{
    "request_id": "5143",
    "CustomerEmailAddress": "madhu.sudhan.h.v@test.com",
    "CreationDate": "2016/03/21 12:12",
    "Subject": "Laptop Heating Problem",
    "Body": "Dear Sir,\n My laptop is getting heated up always and
    restart automatically. Can you please advise what would be the
    problem and rectify as soon possible. It is currently causing
    delay for my deliverables. Thanks,\n Madhu Sudhan H V",
    "attachment": "attachment1.txt",
    "Region": "IN"
}
```


Figure 2: POST request.json details

3.2 AI Engine

AI Engine is deployed on a Python Django Framework where it will receive the POST request.json for further processing. Email subject and body will be combined and fed through data preprocessing with NLP modules and Information Extraction modules. Later email classification ensemble model will identify the category based on the subject and body content. Various business rules are applied for assigning the queue, language identification and for the priority of the tickets. Finally the response.json is generated that will communicate with RPA to create tickets on ticketing tool.

3.3 Generate Response Json

Once request.json goes through the AI engine, it generates the response.json that will be consumed by the RPA to communicate with ticketing tool. Below is sample response.json with classified and extracted content.


```
{
    "PredictedCategory1": "Laptop",
    "AssignedGroup": "Bangalore Team",
    "AutoAssigned": "No",
    "AutoResponseEmail": "No",
    "Body": "Dear Sir,\n My laptop is getting heated up always
    and restart automatically. Can you please advise what would
    be the problem and rectify as soon possible. It is currently
    causing delay for my deliverables. Thanks,\n Madhu Sudhan H V",
    "CSATProbability": 0.5750961773386505,
    "CSATResult": "Satisfied",
    "From": "madhu.sudhan.h.v@test.com",
    "PredictedCategory2": "Heat",
    "Company": "Test",
    "ContactType": "Email",
    "FirstName": "Madhu Sudhan",
    "FollowupMailFlag": 0,
    "Impact": "4 - Low",
    "IncidentCreated": "Yes",
    "Language": "English",
    "LastName": "H V",
    "Location": "Bangalore",
    "Priority": "4 - Low",
    "RequestID": "5143",
    "Response": "Yes",
    "SelfServiceFlag": "No",
    "State": "New",
    "Subject": "Laptop Heating Problem",
    "Urgency": "4 - Low"
}
```


Figure 3: Web Service Response Json

4. Other recommendations

The text classification model accuracy can be improved by using artificial neural networks. Distance measures like cosine similarity for information extraction around a keyword from the body of the mail can be implemented for improved data extraction.

References

[1] Madden, M.—Jones, S.: Networked Workers. PewInternet report, Pew Research Center; September 24, 2008.
[2] HP, The Radicati Group, Inc.: Taming the Growth of Email – An ROI Analysis (White Paper), 2005.
[3] Jones, J.: Gallup: Almost All E-Mail Users Say Internet, E-Mail Have Made Lives Better. 2001.
[4] META Group Inc.: 80 % of Users Prefer E-Mail as Business Communication Tool. 2003.
[5] Whittaker, S.—Sidner, C.: Email Overload: Exploring Personal Information Management of Email. In Proceedings of ACM CHI96, pp. 276–283, 1996.
[6] Fisher, D.—Brush, A. J.—Gleave, E.—Smith, M. A.: Revisiting Whittaker & Sidner's "Email Overload" Ten Years Later. In CSCW2006, New York ACM Press 2006.
[7] Laclav´ık, M.—Maynard, D.: Motivating Intelligent Email in Business: An Investigation into Current Trends for Email Processing and Communication Research.

Author Profile

Madhu Sudhan H V received the MBA Degree in Finance from Christ University, Bangalore. He has done Btech in Electronics and Communication from VTU, Belgaum. Currently he is working as an Artificial Intelligence Associate Principal at Accenture and delivering values to clients in Artificial Intelligence, NLG, NLP and Big data.