Document Analysis and Classification : A robotic process Automation(RPA) and Machine Learning Approach

Abhishek Baidya

Why RPA? It's a Robotics Process Automation process that utilize user interface to capture data. It is frequently used to perform repetitive tasks. It also excludes human from the decision-making process, which excludes any judgement or use of external data.

This research is mainly focused on text classification from data extraction of different types of documents. Utilizing unstructured and semi-structured documents to identify, extract and structure data for analysis and classification

RPA and ML was conjuctionally used for Document classification, extraction and OCR (Character Recognition)

Major challenges that led this research to happen was RPA's limitation of using context data. It fails to recognized sections that add context to the words that we are looking for. Also, parsing the inconsistencies of natural language is hard to define. Especially from social media.

RPA faills to understand the nuance between different words.

> I leveraged RPA with Machine Learning to analyze structured, unstructured and semi-structured documents to identify, extract and structure data within them for further analysis and classification. As RPA and Machine Learning (ML) are evolving towards hyper-automation, I used the software robots in conjunction with ML to handle complex tasks such as Document classification, extraction and Optical Character Recognition. But there are few challenges using ML in RPA. One major challenge is in accounting for context data, e.g., the documents in RPA have structured sections that add context to the words within. Another challenge lies in parsing the inconsistencies of natural language in the social media like Tweeter, Google reviews, etc. In order to understand the nuance like these, different modes of ML are combined with software rules. Several works have emphasized the capability of gathering the process intelligence and using this intelligence for

RPA also requires well-defined, clear rules and parameters for decision making which isn't very effective when dealing with all sorts of data. Therefore, to overcome these limitations, combination of ML with RPA has shown to be highly effective.

This research consists of data extraction and classification which will eventually lead to categorization of the documents to wait for appropriate action to be taken.

The analysis and classification part are text-mining category which also overlaps the area of NLP

1. Document extraction is a feature extraction process referred as tokenization: removing any noise like special characters are removed from the text data.

2. Feature selection: keeping relevant data and removing unnecessary words or characters

   This feature selection used BAG OF WORDS (BOW), aka term frequency. And TF-IDF(Term frequency-inverse document frequency) , TF is a weight of each word in a document which depends on the distribution of each word in a document. It indicates the importance of the word in the text. DF indicates number of documents that contains the term in the total number of documents.

frequency (IDF). TF is a weight of each word in a document which depends on the distribution of each word in a document. It indicates the importance of the word in the text. IDF of each term in the documents' collection is a weight which depends on the distribution of each word in the documents' collection. It indicates the importance of each term in the entire collection. TF/IDF utilizes both TF and IDF to determine the weight of a term [12]. Mathematically TF/IDF can be defined as the product of the TF and IDF:

$$TF/IDF(t,d) = TF(t,d) * IDF(t,d) \quad (1)$$

TF(t,d) in the equation (1) means the number of times a term t occurs in a document d. IDF(t,d) is the inverse document frequency which is calculated according to the formula below:

$$IDF(t,d) = \log \frac{n_d}{1+DF(d,t)} \quad (2)$$

DF(d,t) in the equation (2) represents the number of documents d that contains the term t and n_d is the total number of documents. The result of TF/IDF is a vector with various terms along with the term weight.

Performance is measured with its' effectiveness, precision, recall and accuracy.

Problem statement: categorizing documents for the finance department. Purchase Orders, Contracts and invoices and then extract specific fields from each of the documents which is then categorized to trigger different actions.

Before RPA

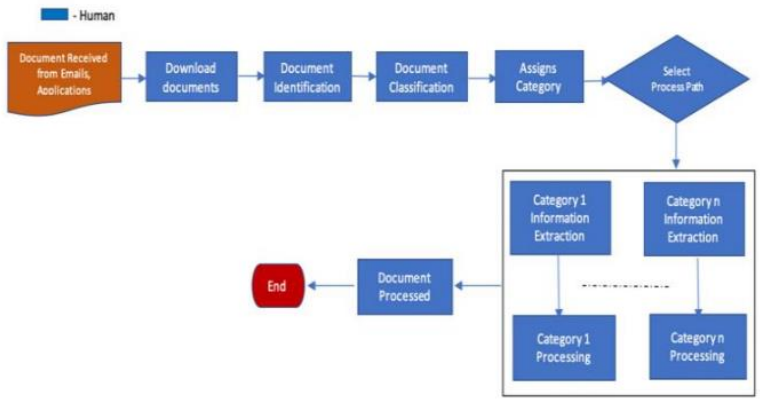
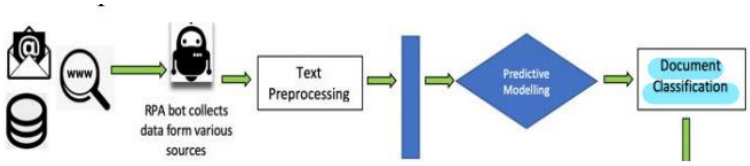


Fig. 1. Current Business Process

After RPA



RPA+ML

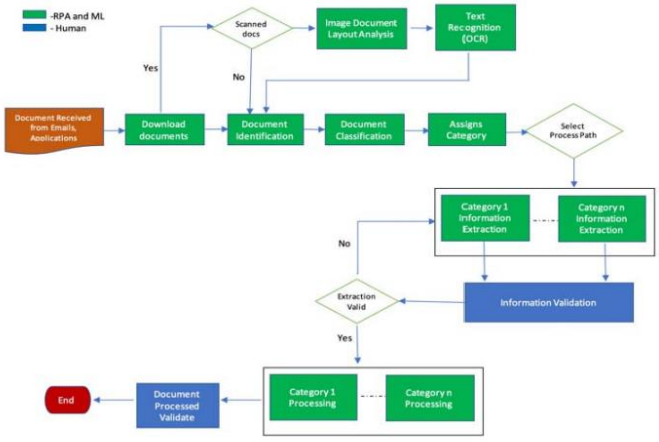


Fig. 3. Automated Business Process with Machine Learning and RPA

1. Automation starts with downloading the docs and checking for scanned images. -> text recognition software digitizes the image into readable format.

2. Using NLP techniques, it automatically classified text much faster. Which also allows for clear classifications based on the past observations.
3. By using pre-labeled examples as training data, the algorithm learns different association between documents.

The SVM classifier follows these steps.

1. Feature extraction where text is transformed into numerical representation.
2. Fed with training dataset that has features and tags.
3. Once the categories are assigned to the docs. It gets sent to specific paths to process.
4. During this process, human validation also occurs to check if there aren't any mistakes during the extraction process. If it wrongly classified, it gets sent back to the extraction process where it diagnoses the errors and correct the predictions.
5. Afterwards, robots send out emails or notification to responsible parties to process.

The research used 80:20 ratio for training set and test set to perform.
Overall, it resulted in 3860 hrs of reduced time. Compared to when only done by human.

the count of the documents that were used for our training and testing with 80%-20% train-test split and the amount of information extracted from different document categories during training and testing.

TABLE I. DOCUMENT COUNTS FOR DIFFERENT CATEGORIES

Categories	# Test documents	# Information extracted	# Train documents	# Information extracted
Purchase Orders	8180	32510	40900	161350
Invoices	5960	52457	29800	267240
Contracts	9860	87934	49300	339568

Figure 4 presents the performance of the SVM classifier on the different documents. The table describes the detailed classification metrics for the execution of the classification process. It represents the test accuracy, train accuracy, precision, recall, F1-score which is the weighted average of the precision and recall.

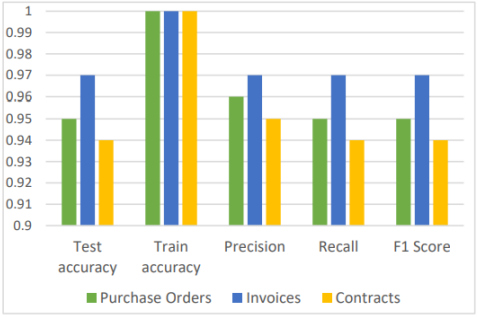


Fig. 4. Performance metrics for different document categories

Limitation of the paper was it didn't use actual live or production data.