Automatic Validation of Medical Practitioner’s Prescription

Dr. Suma Dawn, Vishant Chaudhary, Sajal Agarwal, Shristi Agarwal

Department of Computer Science And Engineering, Jaypee Institute Of Information Technology, Noida, India

[vishantchaudhary007@gmail.com](mailto:vishantchaudhary007@gmail.com)

[wrk.sajal@gmail.com](mailto:wrk.sajal@gmail.com)

[shristi200102@gmail.com](mailto:shristi200102@gmail.com)

Dr Suma Dawn- Mentor, Guided the whole project.

Vishant Chaudhary- Text extraction, Labelling the data (BIO), Training the Model using spacy and extracting the entities

Sajal Agarwal- Creating the data set of the prescriptions, fuzzy matching of the extracted entities and calculating the ratio based on weightage given to the entities and deciding the validity.

Shristi Agarwal- Creating the front-end of the project, a user application wherein a user uploads the image and the corresponding results are displayed after processing.

Automatic Validation of Medical Practitioner’s Prescription

Often medical stores and online phama companies struggle to validate a prescription written by a doctor. They do not know if the prescription is original or fake as it takes manual verification. We want to develop an application where the prescription can be verified. This will help in screening the prescription.

# Introduction

The most critical step in the prescription filling procedure may be verifying prescription correctness.

Validating a legal prescription is a vital step in assuring proper medicine supply and distribution through the proper channels. Tamper-resistant and carefully designed prescription pads, as well as computerised prescribing, have been developed to reduce prescription counterfeiting. As a result, pharmacists must train all personnel on what information to look for when determining the legitimacy of prescriptions, particularly those for restricted medications.

Based on a few industry reports and analysis, the E-pharmacy market in India is expected to increase at a higher CAGR of around 40-45% as compared to Global E-Pharmacy markets that are expected to increase at a CAGR of around 15-20%.

Validating a medical prescription is a tedious task that the E-pharma companies face. There is no central database that maintain record of medical practitioners in India. Even if one exists, it is only limited to government access. The central/state government only issues a registration number if a practitioner is legitimate. Each prescription needs manual verification. If a system is developed that can screen the prescriptions, the industry can expect even faster growth. This will also enable swift order dispatch and deliver thereby benefitting the end user that is customer.

Through our project we are trying to develop a system that can predict the validity of a prescription. This may serve as an initial screening for these pharmacies to either accept or reject a prescription.

A user uploads a prescription on our app and then the system extracts various entities from it. Like Doctor name, registration number, email, Phone No, designation, hospital or clinic name etc.

After this is extracted, the system verifies this by scanning the database. Each entity has separate weightage and based on it we calculate the ratio matching with the one in the database and then predict its chance of being valid or invalid.

# Background Study /Current State of the Art

Tesseract module of python acts as the brain of the application. The Tesseract after processing, gives a printed text of the image taken as the output. An application was tested on different categories of images and the accuracy of output varied greatly. It was observed that the application accuracy depends upon the nature of the image, layout of page and mobile computing power.

It was observed that Tesseract works best with clean black text on solid white background. It also gives good accuracy if the test is horizontal with text height being at least 20 pixels.

Named entity recognition (NER) is a natural language processing tool for information extraction from unstructured text data such as e-mails, newspapers, blogs, etc. NER is the process of identifying nouns like people, place, organization, etc., that are mentioned in the string of the text, sentence, or paragraph. Five well-known NER software were selected namely Stanford NLP, NLTK, Open NLP, SpaCy, Gate.

Stanford NLP and SpaCy usually performs the best. We have used SpaCy in this project for entity recognition.

Flask is a small and lightweight Python web framework that provides useful tools and features that make creating web applications in Python easier. We created a small web app wherein a user can upload the image (prescription) and process it. After processing it the predicted entities are matched with the one in the database. If the two match the prescription’s validity can be assessed.

1. **Methodology**

(a)The most challenging part of this whole project was gathering the dataset. Singe there wasn’t any pre-existing dataset available, we created our own. We asked for prescriptions from the people we knew. We were able to collect 100+ prescriptions of distinct doctors.

The dataset varies widely. There is no set format of how a medical prescription should look like and what all details are mandatory.

Over 90% prescriptions contained name of the doctor

Almost 40% prescriptions did not contain either the clinic name or the hospital name

Over 80% prescriptions contained contact number

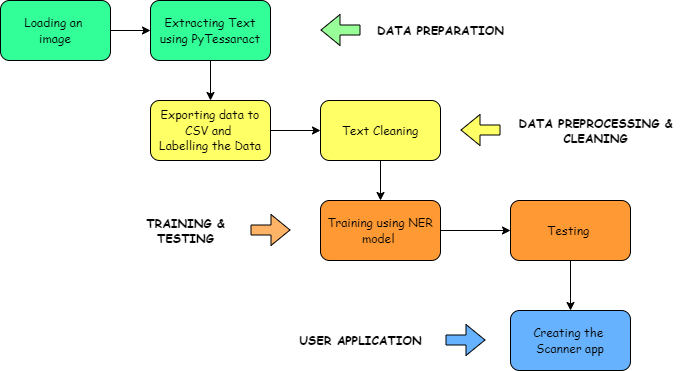
About 70% prescriptions contained doctor’s designation / speciality.

Over 30% prescriptions did not contain Registration number

**Train vs Test split:**

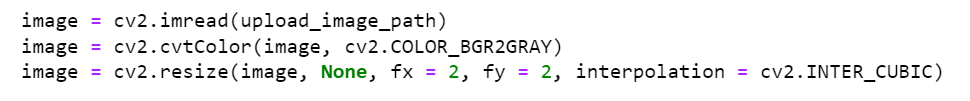
80% used to train the model 20% to test the model

**Workflow**

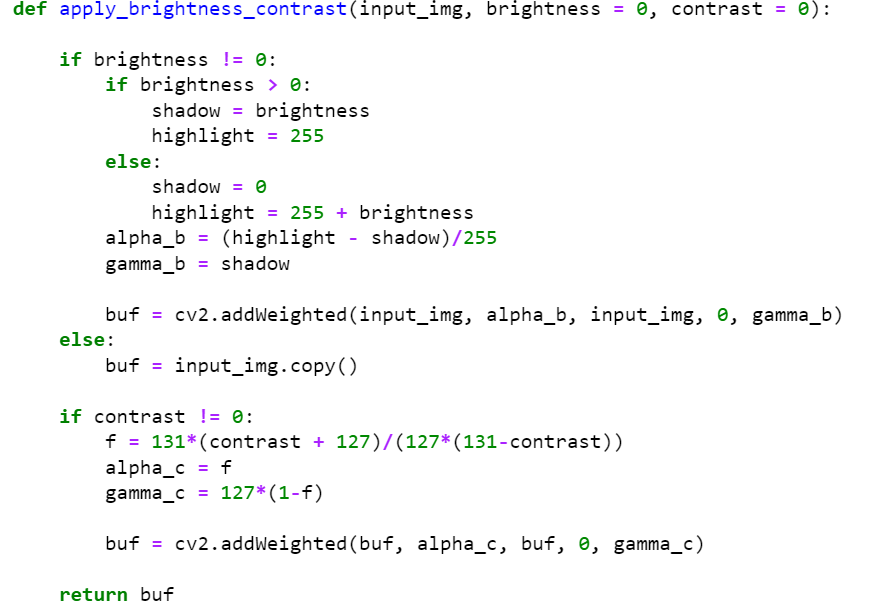


1. **Loading an image**

An image is first loaded using opencv2. Image is converted from RGB format to Grayscale for better text recognition. It is then resized on x and y axes.

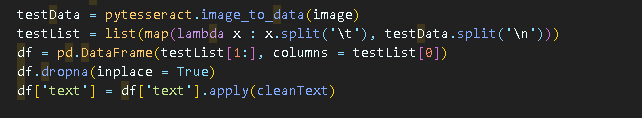


Applying Brightness contrast to the image

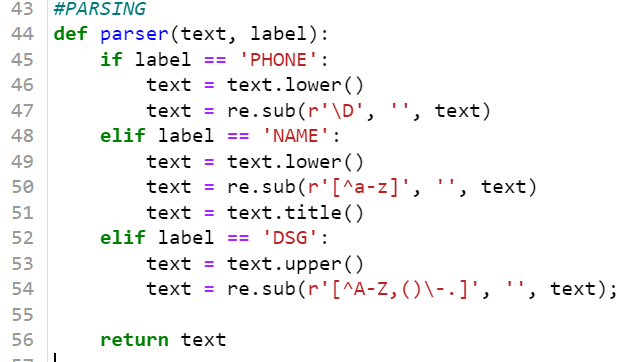


1. **Extracting Text Using Pytesseract**

Using pytesseract the raw text is extracted

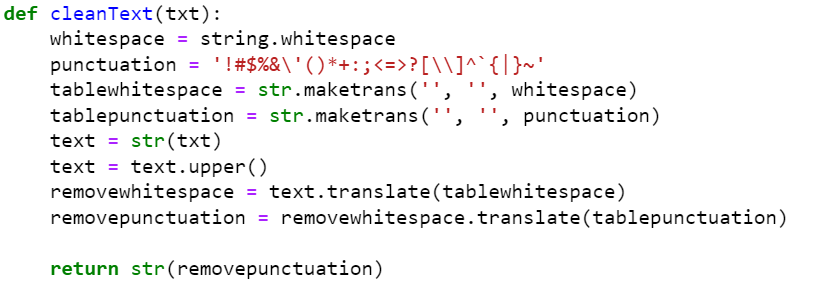


Then using Regular Expression, data is parsed.

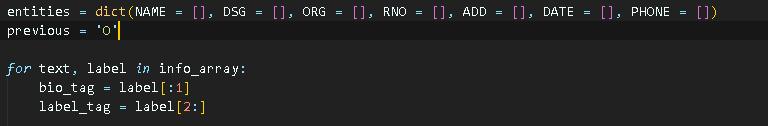


1. **Data Preprocessing and cleaning (labelling)**

Raw text is cleaned.



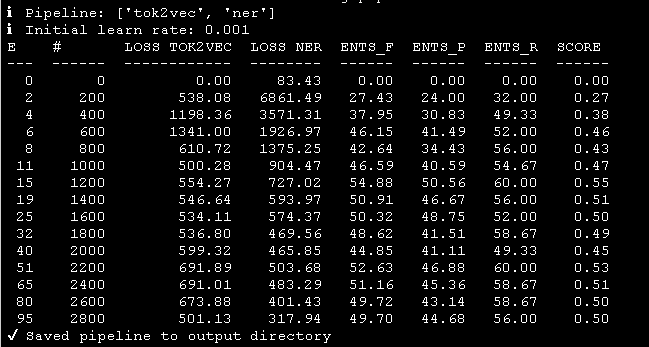
Data entities



Data is Labelled in CSV file.



1. **Training using NER model**

****

1. **Testing**

Ratio of various entities is calculated. Each entity is given weightage and based on the that a final ratio is calculated to classify a prescription as valid or invalid.

            name\_ratio=fuzz.ratio(str(entitylist[0]),str(i[0]))

            reg\_ratio=fuzz.ratio(str(entitylist[1]),str(i[1]))

            clinic\_ratio=fuzz.ratio(str(entitylist[2]),str(i[2]))

            phn\_ratio=fuzz.ratio(str(entitylist[3]),str(i[3]))

            if reg\_ratio>80:

                flagvalid=1

            p=[name\_ratio,reg\_ratio,clinic\_ratio,phn\_ratio]

            ratio=0.50\*name\_ratio+0.25\*reg\_ratio+0.20\*clinic\_ratio+0.05\*phn\_ratio

            if ratio>maxratio:

                maxratio=ratio

                finalresult=[i,maxratio]

        if maxratio >40 or flagvalid==1:

            # print (i,p,ratio,"Valid")

            finalresult.append("Valid")

        else:

            finalresult.append("Invalid")

1. **Creating the Scanner App**

Using Flask the a web app is created to

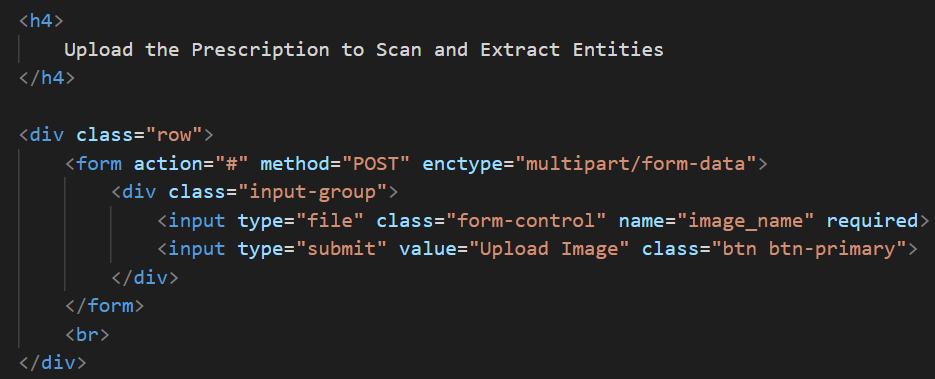
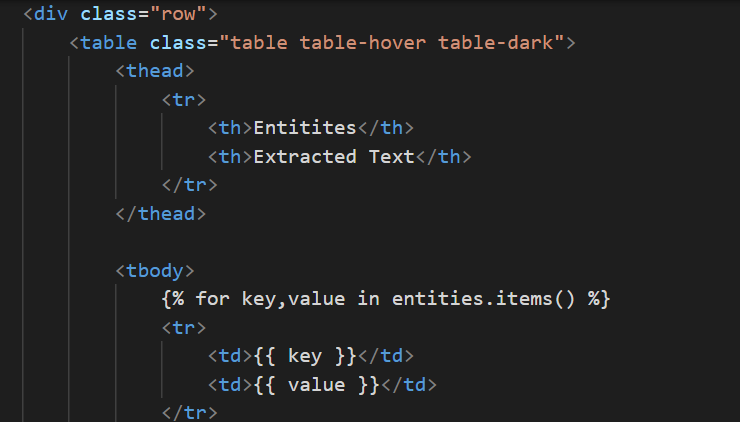
****

Table for the Extracted entities



1. **Results**

Considering the best fit model, the accuracy is 55%. The dataset is small and needs large scale training and testing.

1. **Discussion**

Sometimes the model confused registration number with pincode or std code for phone number. In case multiple doctor names were present in a single prescription, the accuracy had a hit. Also, if the image being uploaded wasn’t aligned straight or was not clear, the accuracy was poor. Overall, the results are quite impressive and promising considering the variation in the format of the prescriptions.

1. **Future works** - As the accuracy of our model can’t be considered near perfect, A new larger dataset could be used to re-train our data and develop a model with more accuracy.

This project can be extended to:

* check the validity of even the drugs prescribed by the doctor (if they are legible).
* Check if the prescribed medicine is within safe dosage range.
* Check if the drug is habit forming or narcotic

It can also be extended to prevent black marketing or hoarding of medicines if location is mapped with the prescription.

1. **References**

|  |  |
| --- | --- |
| [1] Remus Petresuca, Sergiu Manolache, Costin-Antin Boingiu, Giorgina Violetta, Cristian Avatavului, Marcel Prodan and Ion Bucur. “Combines Tesseract and aspire Results to improve OCR Text Detection Accuracy” Article in Journal of Information Systems Management · May 2019 | [8] Chaudhuri Arindam, Mandaviya Krupa, Badelia Pratixa, Ghosh Soumya K., Optical Character Recognition Systems, Springer International Publishing, 2017 |
| [2] Badawy, W. "Automatic License Plate Recognition (ALPR): A State of the Art Review." (2012) | [9] Chirag Patel, Atul Patel and Dharmendra Patel. “Optical Character Recognition By open source OCR Tool Tesseract”. International Journal of Computer Applications (0975 – 8887) ,October 2012 |
| [3] Ahmad T. Al-Taani, "Automatic Text Summarization Approaches", International Conference on Infocom Technologies and Unmanned Systems (ICTUS'2017). | [10] P. Koehn, F. J. Och, D. Marcu, “Statistical phrase-based translation,” in Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (Association for Computational Linguistics, Stroudsburg, PA, 2003), pp. 48–54. |
| [4] D. Bahdanau, K. Cho, Y. Bengio, “Neural machine translation by jointly learning to align and translate,” <http://arxiv.org/abs/1409.0473> (2015). | [11] Weizenbaum J., ELIZA—A computer program for the study of natural language communication between man and machine. Commun. ACM 9, 36–45 (1966). 10.1145/365153.365168 |
| [5] Dan ClaudiuCiresan and Ueli Meier and Luca Maria Gambardella and JurgenSchmidhuber, “Convolutional Neural Network Committees for Handwritten Character Classification”, 2011 International Conference on Document Analysis and Recognition, IEEE, 2011 | [12] <https://towardsdatascience.com/a-gentle-introduction-to-ocr-ee1469a201aa> |
| [6] <https://spacy.io/usage/spacy-101> | [13] <https://en.wikipedia.org/wiki/Optical_character_recognitio> |
| [7] <https://github.com/tesseract-ocr/> |  |