# Semantics using NLP for handwritten Punjabi FIR Reports

## ***Overview***

**Natural Language Processing** is a subfield of Machine learning that is concerned with linguistics, computers and Artificial intelligence. NLP provides a great significance to the interaction of Human languages with Computers.

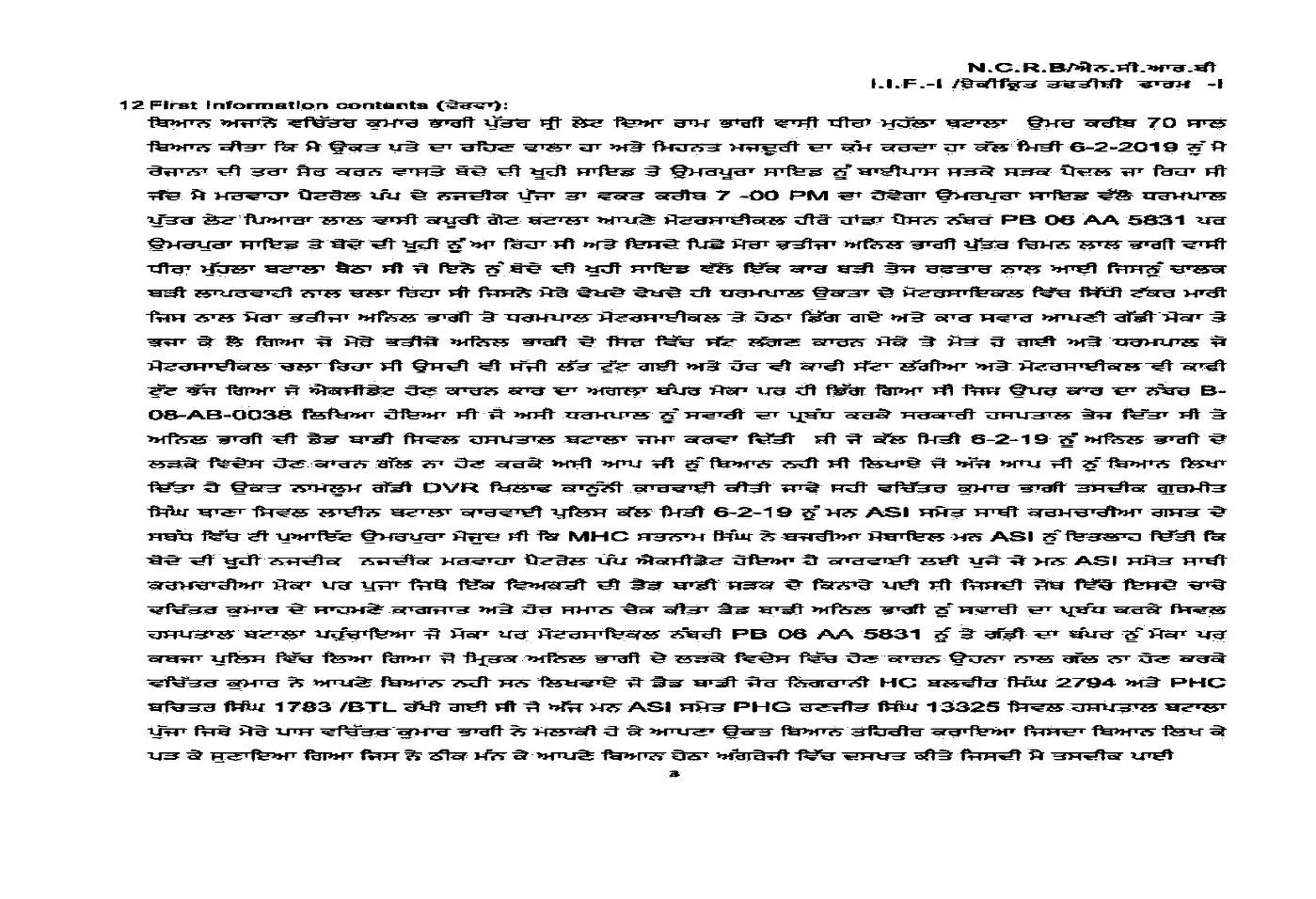
**Semantics using NLP for handwritten Punjabi FIR** is NLP project which takes into account the Punjabi language text for operations.

The project has been aimed to extract semantics from Punjabi handwritten FIR reports and provide Punjab police with the domain of FIR. A subdivision of Project also includes extracting Named entity recognition from Punjabi text and Using to fill various columns of FIR reports.

As it is a live project for Punjab police, an extension to this involves the creation of Chatbot that would ask Punjab police to provide further information of what is not provided in the FIR.

## ***Implementation***

FIR reports were provided in the form of PDF of generally 5 images. The general format of the FIR report contained Information content of the FIR written in the paragraph form starting from the third page.



### Pdf To Image

Firstly, we were given the pdfs of Punjab Police FIR which we converted to images using PDF2Image library of python. With the use of OS library, we iterated through the folders and subfolders and searched for the file that ends with "pdf" extension and stored them in the list. Further, the list is passed through the "convert\_from\_path" class of the pdf2image library the pdfs are converted to the images. The name of the image is kept with the filename\_index (Index is basically from 1 to 4 which are the number of pages of particular FIR report.) and then the image is stored at a specific folder ( naming convention used for the folder name is according to the name of the month of PDF)

### Extracting text from images

FIR report images contained handwritten texts. There are various Image to Text libraries for English language but the arduous task here was to convert the Image to Text for the Punjabi language as there is very less work done for this language. OCR tesseract was the one that solved the purpose. It was used to convert Punjabi FIR images by mentioning the language name in options of OCR tesseract. Multiple language text can also be recognized with the use of OCR tesseract.

### Converting text to Dataframe

The result after Image to text conversion was FIR reports in text form . Further processing is applied on text before changing it to data fram.

Then, we converted the Text to data frame for all the FIR texts present in a particular folder.

## ***Data Preprocessing***

The data frame received after Image text extraction is saved in the form of data frame of all FIR reports, which is further used for Preprocessing.

Below are some of the operations performed as part of Data preprocessing :

### Special symbols

In data preprocessing, the data is cleaned using various lambda functions which removed the special symbols from the data frame making it easy to work for existing libraries.

### Stopwords

Removing stop words for English has a predefined library and can easily remove stop words using it. But in case of Punjabi language, we made a list of Punjabi stop words and used it to remove those words from the FIR report data frame. Also, we used a cnltk library for removing some other Punjabi stop words

### Tokenization

Afterwards, the Punjabi words are tokenized using "inltk" library. This library is basically for NLP for Indian Languages. Tokenize is the function that is used to tokenize the Punjabi text when language\_code='pa' is passed.

## **Dictionary approach**

### Reading the dictionary

We were given a few pages of a bilingual dictionary and we were told that we had to map the dictionary words with the words of the data frame. The basic intuition was that any Punjabi word in data frame would be replaced by the English word of the dictionary. So firstly we read the image of dictionary pages by using the tesseract library of python. In it, the library reads the text from the images and then stored it in the CSV file.

### Cleaning

The text that was obtained by reading dictionary contains too many special symbols. Then the lambda function is defined that basically cleans the text and remove the special symbols and numeric values. Then the words are tokenized on spaces. Then we had to build the data frame that contains the English words in one column and its all 4-5 Punjabi meanings in another column.

### Word Replacement using Dictionary

The occurrence of any Punjabi word of dictionary in FIR report data frame is replaced with its similar English word to get some useful words in English to further understand the semantics of FIR

### Word Embeddings after Dictionary replacement

Further, the converted English words in the FIR data frame are used to get word embeddings and know the domain of FIR by understanding its semantics.

## **Fasttext approach**

Fasttext is a python library which contains different models for various languages used throughout the world.

We used fasttext to download model of Punjabi language in the form of vector and binary file.

These files contain large number of Punjabi words as keys and their embeddings as value.

We used this key-value pair for getting embeddings of all Punjabi words which are present in the FIR data frame.

We also created a document matrix by using Tf IDF on Punjabi words in FIR data frame.

The dot product of Tf IDF and Word embeddings generated using fasttext model gave us Document embeddings for FIR reports.

Further, topic modelling and K means clustering approach has been finalised to know the domain of FIR.

### Google Trans

We also tried using the google trans and then to convert the given Punjabi text to English text. The basic idea behind this was that we were not able to find the embeddings of the Punjabi language. So we thought of using the google trans and converting the given Punjabi text to English tex

## **References:**

<https://www.geeksforgeeks.org/python-reading-contents-of-pdf-using-ocr-optical-character-recognition/>

<https://medium.com/better-programming/beginners-guide-to-tesseract-ocr-using-python-10ecbb426c3d>

<https://www.analyticsvidhya.com/blog/2019/09/introduction-information-extraction-python-spacy/>

<http://docs.cltk.org/en/latest/punjabi.html>

<https://shirishkadam.com/2016/12/23/dependency-parsing-in-nlp/>

<https://medium.com/huggingface/state-of-the-art-neural-coreference-resolution-for-chatbots-3302365dcf30>

<https://pypi.org/project/googletrans/>

<https://stackoverflow.com/questions/48760628/which-the-efficient-python-library-that-convert-pdf-to-images-apart-from-wand-im>

<https://iq.opengenus.org/pdf_to_image_in_python/>

<https://machinelearningmastery.com/clean-text-machine-learning-python/>

<https://medium.com/@datamonsters/text-preprocessing-in-python-steps-tools-and-examples-bf025f872908>

<http://zetcode.com/python/googletrans/>