



Information Extraction in User's Complaints

by

T G Narayanan
MT20027

Under the Supervision of
Dr. Md. Shad Akhtar

Indraprastha Institute of Information Technology Delhi
August, 2022



Information Extraction in User's Complaints

by

T G Narayanan
MT20027

Submitted

in partial fulfillment of the requirements for the degree of
Master of Technology in Computer Science Engineering
(Specialization in AI)

to

Indraprastha Institute of Information Technology Delhi
August, 2022

Certificate

This is to certify that the thesis titled “**Information Extraction in User's Complaints**” being submitted by **T G Narayanan** to the Indraprastha Institute of Information Technology Delhi for the award of the Master of Technology in Computer Science and Engineering (Specialization in AI), an original research work carried out by him under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or whole to any other university or institute for the award of any degree/diploma.

August, 2022

Dr. Md. Shad Akhtar

Department of Computer Science & Engineering
Indraprastha Institute of Information Technology Delhi
New Delhi 110020

Acknowledgements

I would like to convey my sincere gratitude and thanks to my guide and advisor, Dr Md. Shad Akhtar. I admire his focus on the fields of Natural Language Processing and Deep Learning and the vast amount of knowledge he has in all associated fields.

He has always encouraged me not to limit myself to one domain but explore various aspects associated with the same. The covid times have been tough for everyone, and planning the workload according to his generosity to take into consideration my personal needs with providing constant guidance has helped me to be able to complete the work and acquire significant amount of knowledge which is way above what I could have achieved doing regular courses.

Thanks to the members of LCS2 for helping me at times with annotation and data analysis tasks.

I want to thank my family and close friends for being extremely patient with me and believing in me despite the testing times they have been through. Without their moral support, getting this work done would have been near impossible.

Abstract

The task of Named Entity Recognition is one of the most explored fields in the Natural Language Processing domain. Numerous existing works have tried to uncover different aspects of this common yet unique field. The NER task has been extended to a variety of domains (such as social media, judiciary, medical, or the general domain) and languages (English, European, Chinese, Hindi, etc.). However, each domain and language offer their core challenges particularly due to the syntactical complexities. In this research, we explore the named-entity recognition in the legal domain. Given a user’s complaint to report a crime, we intend to extract all relevant and necessary information considering the crime, victim, accused, etc. in an efficient manner. We collect publicly available user complaints and developed an entity and relationship annotated dataset, aka. Legal Document Processing (LDP) dataset. The instances of this dataset are densely annotated with more than fifty labels broadly revolving around victim and other crime details. Subsequently, we benchmark the dataset using multiple pre-trained language models and information extraction-based baselines. In particular, we finetune Multi-lingual BERT, HindiBERT, and HindiBERTa on the LDP dataset. Our evaluation shows Multi-lingual BERT reports the best performance among all baselines. The potential scope for the future work includes entity and relation-level knowledge graph creation as well as converting a user complaint to a technical and legal document.

Contents

| | | |
|----------|--|-----------|
| 1 | Introduction | 1 |
| 1.1 | Background | 2 |
| 1.2 | Motivation | 2 |
| 1.3 | Experiments | 3 |
| 1.4 | Contribution | 4 |
| 2 | Related Work | 5 |
| 2.1 | Studies on Named Entity Recognition | 5 |
| 2.2 | Question Answering Task | 5 |
| 2.3 | Comparison with existing works | 6 |
| 3 | Dataset and Tasks | 7 |
| 3.1 | Raw Data Source | 7 |
| 3.2 | Raw Data Collection Automation | 8 |
| 3.3 | Preprocessing | 8 |
| 3.4 | Labelling Task | 8 |
| 3.5 | LDP_NER and LDP_NER_combined Datasets | 9 |
| 3.6 | LDPQuAD Dataset | 11 |
| 3.7 | Challenges | 12 |
| 3.7.1 | Text Structure | 12 |
| 3.7.2 | NER Task using LDP_NER | 12 |
| 3.7.3 | Question Answering Task using LDPQuAD | 12 |
| 4 | Methodology | 14 |
| 4.1 | Inter Annotator Agreement | 14 |
| 4.2 | Importance of label distribution | 14 |
| 4.3 | Stratification vs cross fold validation | 14 |
| 4.4 | Training | 14 |
| 5 | Experiments and Results | 16 |
| 5.1 | Models Experimented | 16 |
| 5.1.1 | BERT | 16 |
| 5.1.2 | Multilingual BERT | 17 |
| 5.1.3 | DistilBERT | 17 |
| 5.2 | Experimental Setup for NER using LDP_NER and LDP_NER_combined datasets | 18 |
| 5.3 | Experimental Setup for Question Answering Task using LDPQuAD dataset | 18 |
| 5.4 | Evaluation Metrics | 18 |
| 5.4.1 | Cohen’s Kappa | 18 |
| 5.4.2 | F1-score | 19 |
| 5.4.3 | Exact Match | 19 |
| 5.5 | Experimental Results | 20 |

| | | |
|----------|-----------------------------------|-----------|
| 5.5.1 | NER_LDP Results | 20 |
| 5.5.2 | LDPQuAD Results | 21 |
| 6 | Conclusion and Future Work | 23 |

List of Tables

| | | |
|-----|---|----|
| 3.1 | Label Distribution | 13 |
| 5.1 | LDP_NER Dataset | 20 |
| 5.2 | LDP_NER_combined Dataset | 20 |
| 5.3 | LDPQuAD Dataset | 21 |
| 5.4 | Entities with best class wise F1-scores for Multilingual BERT | 21 |

List of Figures

| | | |
|-----|---|----|
| 2.1 | RoleFiller | 6 |
| 3.1 | A screenshot from Haryana FIR Portal | 7 |
| 3.2 | A screenshot from Delhi FIR Portal | 8 |
| 3.3 | Annotated Entities | 9 |
| 3.4 | LDP Dataset Creation | 9 |
| 3.5 | LDP_NER Dataset | 10 |
| 3.6 | An instance from the LDP Dataset | 10 |
| 3.7 | LDP_NER and LDPQuAD Datasets Creation | 11 |
| 3.8 | Question Generation for LDPQuAD Dataset | 11 |
| 5.1 | BERT Architecture | 16 |
| 5.2 | Tokenized instance of LDP dataset using AutoTokenizer from Huggingface and IndicBERT pretrained model | 17 |
| 5.3 | stride | 18 |
| 5.4 | Traning Loss and Validation Loss Curves for Multilingual BERT over 5 folds | 22 |

Chapter 1

Introduction

The use cases of Named Entity Recognition and Question Answering models for the purpose of handling user complaints are beyond one can imagine. From early crime detection to dynamic criminal record updation there are numerous ways in which we can greatly benefit from such deep learning models. In India for instance which is rising in terms of technologically getting equipped faster than in any nation, the digitization of physical records have started about a decade ago. This facilitates greatly with early and dynamic access to necessary records required to conduct following process such as banking related process which require lot of documentation or legal cases which are left pending due to lack of certain proofs.

We are given various provisions in the constitution, such as, we can file a complaint against an individual or a firm or express our concerns on the way systems function. People can file Public Interest Litigation for instance when they need to get information regarding the way a particular system works in any firm, or can take the matter to court incase of no judgement served. Now, certain complaints or reports are considered to be time sensitive as any delay in its addressal could prove to be harmful. The most complaints or reports are made in the legal field as with boom in the economy, there is a significant boom in the number of criminal offenses arising. Moreover, with every kind of advancement comes a new set of rules and restrictions attached to it. Cyber laws for instance was an unknown concept until past few decades.

We have studied the user complaints in the form of First Information Reports filed in police stations either in offline mode by filling up a huge form or online digitally. These reports are a great source of information as they contain the details of the crime reported by the victim. The demographics of the victim, the type of crime, and the way in which the crime took place can give great insights for developing early crime detection systems. Certain cities like Delhi and Haryana have all their FIR records maintained in digital form and available to public, unless held in disclosure due to the sensitivity of the crime. Bihar police website so far contains the uploaded copies of hand written FIR reports but they can also prove to be a great source of information.

The reports retrieved from these websites were preprocessed to get necessary portions of the records which give information exhaustively. The crimes were under various sections of the Indian Penal Code and studying all of them together would have made the task complex. We decided to limit our focus to the most frequent type of crime which is reported the most in police station i.e. theft related crimes. The theft related crimes comes under section 379 IPC.

Once the preprocessing was done, we proceeded with the annotation task. We used Doccano which is an open-source annotation tool with features to label the entities. It gives the option to create our own set of custom labels. Once the annotation task is done, we get our novel Legal Document Processing (LDP) Dataset by exporting the annotations in the form of .jsonl file.

1.1 Background

The rise of the economy in the past decade has peaked with the increase in criminal activities. Criminal cases have grown exponentially, making it nearly impossible to determine a verdict for most of them. Populous countries such as India, China, and the USA have numerous pending cases. The countries adopting the technologies associated with legal data logging sooner could benefit greatly. With almost every machine learning and deep learning model being trained on English datasets, the problem was not as severe as in other languages. Even China, for that matter, mainly having Chinese speakers, could get a massive number of resources enabling them to extend their research and development considerably several models in no time. The scenario differs in a multi-linguistic country like India, where all languages have fair representation. Almost every state has a different language. Thus, the databases also have been maintained in multiple languages making it troublesome to find the singular conventions to maintain legal databases.

1.2 Motivation

The lack of data resources in the legal domain motivated us to explore various legal websites that have maintained records to some extent and make a dataset that can help extend the study of legal information. More focus was laid on the fact that the dataset is created in the Hindi language as it has fewer resources comparatively despite being one of the most spoken languages in the world. We extracted FIR reports from Delhi, Haryana and Bihar Police websites using a simple code automated to get the necessary details from the portals. These reports were studied well to identify required and essential labels. The mix of English and Hindi languages along with certain Urdu words makes it a bit complex than existing works predominantly done with just English datasets. The study of language representations in code-mixed[3] setting could benefit greatly for analysing these reports. With increase in the rate of crimes and advancement of technology, every year or so new kind of laws needs to be incorporated in the constitution to comply with the needs of every person.

Due to the vast diversity in the type of criminal activities, it is crucial to not limit ourselves to traditional role filler entities such as nouns like person, place, and thing and proper nouns like names of businesses, places, people and organizations. We labelled each of the 1000 instances of FIR from Haryana Police with more than 50 labels, enabling the scope of the study for deeper analysis into the multi-label domain of named entity recognition. Although we had to limit ourselves to studying only a particular kind of criminal offense. The section 379 IPC is applicable for crimes related to theft. Our study was limited to this section but due to the nature of crime, the maximum number of crimes were found to be under this section. This motivated us to explore only this particular section, keeping open future scope for exploring other sections of IPC.

1.3 Experiments

The Legal Document Processing Dataset was prepared by extracting point number 12 from the FIR reports. This point contains the entire excerpt of the incident that occurred in a detailed manner. To explore the intensity of crimes, a particular focus was laid on labelling specific labels such as `item_desc`(which describes the lost item) and `location_reason` (which indicates the motive behind the crime). Labelling was done separately for vehicle categories like bike and car and mobile phone details such as number, colour, model, and many more, leaving no stone unturned. This allowed us to combine the labels as per need and decide the level of named entity recognition task we wanted to indulge with. We created multiple variations of this dataset namely LDP_NER, LDP_NER_combined and LDPQuAD. The LDP_NER dataset is made by converting the annotated entities to chunks and labelling them by prefixing I,O and B tags for marking the Inside, Outside and Beginning parts of the entity. The LDP_NER_combined dataset is created by merging few labels in the LDP_NER dataset under a parent label. For instance, `victim#name` and `victim#address` were combined under one label `victim` as both of them represent details of the victim. The third dataset was LDPQuAD (LDP Question Answering Dataset) which was inspired by the SQuAD dataset. The entity labels from the LDP_NER were used to create custom questions. Because of a large number of labels, there was a scope of binding certain labels under one question as well.

1.4 Contribution

The lack of hand-annotated resources in Hindi with many labels required understanding of the individual incidents reported in the FIR. The Legal Document Processing corpus thus opens the way to extend the work analysing different types of crimes, which will have maybe less but not more than the labels used in this dataset. The titles are self-indicators of the relation they hold with one another, such as victimname, victimrelation, victimrelative, crime_detailslocation, criminalweapon, etc, making them suitable for relation recognition. This approach to building dataset opens the gate for exploring further in the field of legal domain. Crime prevention and suspect detection programs can benefit in future with development of more and more models on such densely annotated datasets. Our brief exploration into question answering domain also opens the possibilities to develop automated e-fir-filing systems. Automated report filing systems based on information extraction in user complaints could help in speeding up the judicial system of the nation.

Chapter 2

Related Work

2.1 Studies on Named Entity Recognition

CONLL datasets have been released in the past with labelled entities by the Conference on Natural Language Learning. The dataset consists of columns namely Sentence number, entity, part-of-speech tag and label. The CONLL-2003 [16] raw dataset comprises of eight files. The files are in the form of training, testing and development files for training purpose. It also consists of raw dataset which is not annotated. The CONLL-2003 dataset was released as a part of the Shared Task for conducting language-independent NER. It consists of instances from both English and German languages. CONLL++ dataset is another dataset with about 5% of the test labels corrected in the CONLL-2003 dataset.

LDP raw dataset was converted to the same format as that of CONLL-2003 dataset after various preprocessing steps. The LDP dataset consists of novel set of labels and the FIR instances are in a mix of Hindi and English language, thus English POS was not incorporated in the final dataset.

2.2 Question Answering Task

SQuAD [15] or Stanford Question Answering Dataset is a accumulation of various paragraphs from the Wikipedia articles with questions provided by the crowdworkers and the answer to those questions is a span of from those passages. Not all questions are answerable. As the models trained on this dataset were not capable enough to identify the questions that cannot be or must not be answered, another dataset called SQuAD 2.0 was released with additional 50,000 unanswered questions on top of the 100,000 question from the original SQuAD dataset. This abstainence from not answering all questions was the objective behind this new dataset.

We experimented with a SQuAD like variation of our dataset LDPQuAD. In this dataset, we used the annotation labels to generate standard questions. These questions in a generic sense ask about the span of the entity denoted by the question. We incorporated the BERT model for this purpose. It was finetuned for the Question Answering task in the LDPQuAD variation of our LDP dataset.

The task of question answering in a language other than English is a challenge on its own. Future works could be around developing question answering models [5] solely based on various hindi corpus with analysis of differences in framing of questions in Hindi from English.

¹Source: <https://rajpurkar.github.io/SQuAD-explorer/>

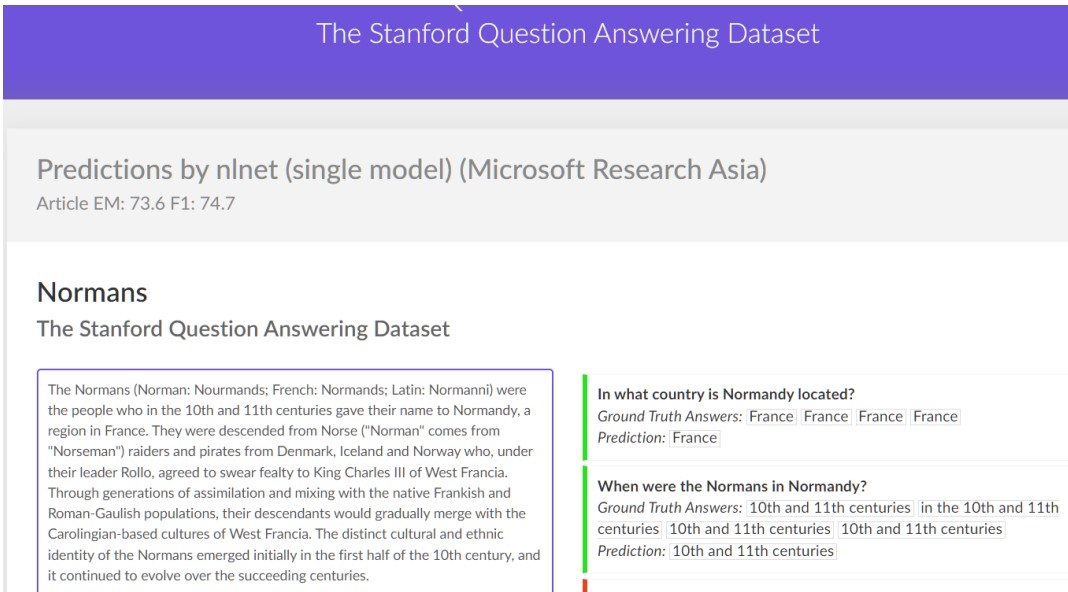


Figure 2.1: Screenshot of SQuAD Dataset Webpage¹.

2.3 Comparison with existing works

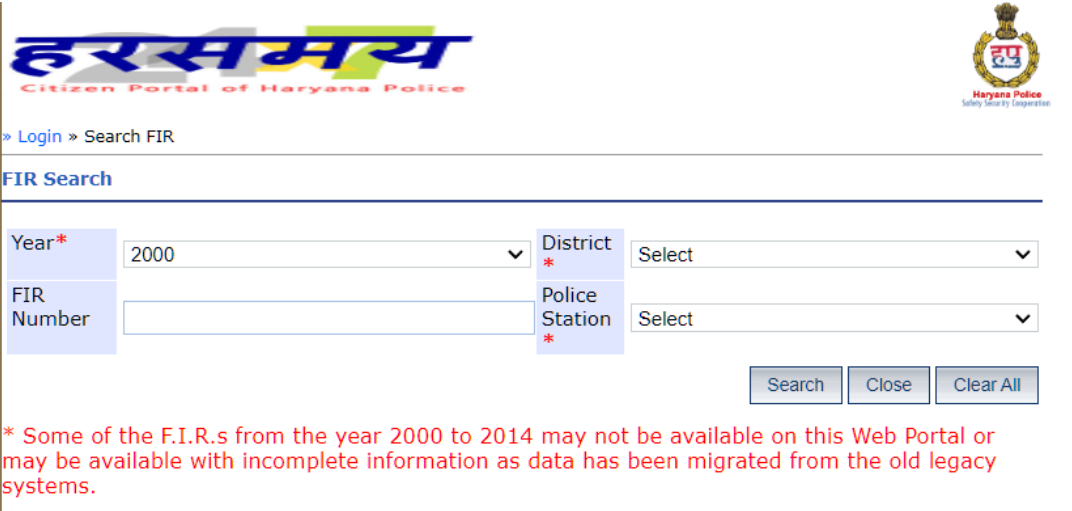
The existing works, such as ILDC [4], HLDC [1], and many more, predominantly focus on the court verdicts with a single binary class indicating whether the victim was sent to jail or was set free. Such datasets also contain fewer labels on a vast dataset. Moreover, not all are hand-labelled, which makes them inferior in terms of specificity. Various models are built on these datasets like LegalBERT [9]. With a considerable number of instances, these datasets are ideal for training models well and avoiding overfitting. In our dataset, with many labels, the training instances per label are less, affecting the resulting F1 scores. Despite that, with cross-validation tweaking of the parameters, Multilingual BERT performs significantly with our dataset giving an F1 score of over 70. The class-wise score is not very consistent, but for certain classes, the model can predict the entities quite accurately.

Chapter 3

Dataset and Tasks

3.1 Raw Data Source

Fetching data from the official website of Haryana Police and Delhi Police required manually entering details such as random FIR serial number, date and district. To simplify this task, we used Selenium which is an open-source automation tool. Once the raw repository of sources for our dataset was ready, we analysed the most occurring type of crime to keep the study focused on one section of law. Our dataset consists mainly of cases under Section 379 IPC, punishment for theft. It is a non-bailable offence. Section 379 has further subsections, such as Section 379A, which indicates theft by snatching.



The screenshot shows the 'Haryana Police Citizen Portal' with a search interface. The header includes the logo and the text 'हरसमय Citizen Portal of Haryana Police'. Below the header, there are navigation links for 'Login' and 'Search FIR'. The main section is titled 'FIR Search' and contains a form with the following fields:

| | | | |
|------------|------|-----------------|--------|
| Year* | 2000 | District* | Select |
| FIR Number | | Police Station* | Select |

At the bottom of the form, there are three buttons: 'Search', 'Close', and 'Clear All'. A red asterisk note below the form states: '* Some of the F.I.R.s from the year 2000 to 2014 may not be available on this Web Portal or may be available with incomplete information as data has been migrated from the old legacy systems.'

Figure 3.1: A screenshot from Haryana FIR Portal¹.

¹Source:<https://haryanapolice.gov.in/ViewFIR/FIRStatusSearch?From=LFh1ihlx/W49VSlBvdGc4w==>

²Source:<https://delhipolice.gov.in/viewfir>

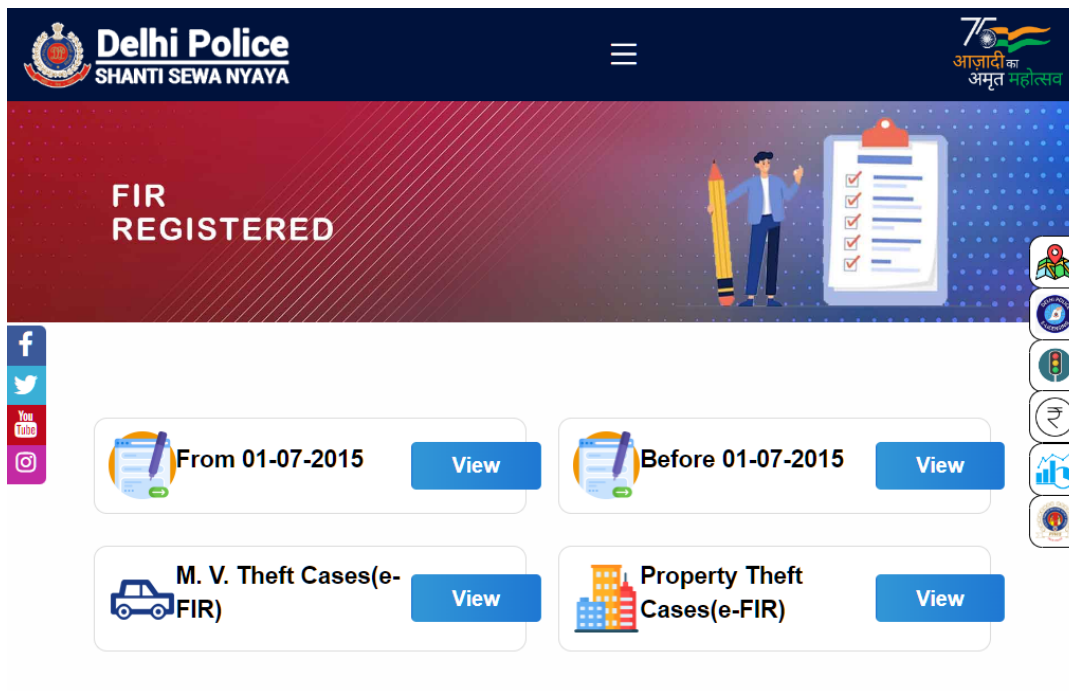


Figure 3.2: A screenshot from Delhi FIR Portal².

3.2 Raw Data Collection Automation

Selenium is a framework built for facilitating certain manual tasks with custom automation. It provides this compact interface that is used to easily create scripts in various languages such as Java, C, Python and many more.

3.3 Preprocessing

The dataset consisted of instances with unnecessary information such as the instance description preceding the main content and the use of certain delimiters which are not required and complicate the model training process. As the instances were annotated, the indices of the label span had to be adjusted as per the changes made in the instance text. Some delimiters such as hyphens, commas, colons and brackets were to be checked before removal as they held significance of their own. For instance, turning “Section 379 23-C, Ambala” to “Section 379 23 C, Ambala” by removing the hyphen from the address can lead to erroneous entity span detection.

3.4 Labelling Task

For labelling the entities in the instances of FIR, an open source tool, Doccano, was used. This application consists of a UI which enables the users to annotate the labels and choose from the custom labels made by user in the form of label_config.json file. As the relationship labelling feature was under development, we decided to label the entities extensively with titles such as victimname, victimrelation, victimrelative and victimresidence such that the victim’s connection with other entities such

as relation, relative and home can be captured. The tool lets the user download the annotations as a .jsonl file consisting of the entities' character span in each FIR instance.

```

|||Haryana_FIR_2015_AMBALA_AMBALA CANTT_0_8.txt::श्री मान जी, नकल लेख इस प्रकार से है - सेवा मे एस.एच.ओ.

अम्बाला सदर छावनी कैन्ट 07.01.15 । श्री मान जी, निवेदन है कि, मे अमित चावला पुत्र श्री वेद
•police_station#name •crime_details#date •victim#name •victim#relative
•victim#relation

प्रकाश चावला पता 6323 निकलसन रोड अम्बाला कैन्ट का रहने वाला हूँ । गाडी जीप महेन्द्रा नई जो
•victim#residence •lost_item#type
•car_details#name

मेरे नाम पर है जो गीता गोपाल स्संथा अम्बाला कैन्ट मे चलती है इस महेन्द्रा जीप का engine
•misc_info#victim •car_details#name
•lost_item#type

no. GHE1K59073 OR CHASSIS NO. MA1ZN2GHKE1K769800RM0DEL2014 है यह गाडी कल रात 9.30 बजे के करीब अनाज
•car_details#engine •car_details#chassis •car_details#model •crime_details#location
•crime_details#time

मन्डी अम्बाला कैन्ट गीता गोपाल भवन के पास खडी की थी तथा गाडी को LOCK कर दिया था । जो कल
•misc_info#lost_item

रात को कोई अज्ञात व्यक्ति गाडी का LOCK खोल कर गाडी को चोरी करके ले गया जो हमारी नई गाडी
•criminal#looks •crime_details#type
•criminal#type
•criminal#action

है जिसका हमने NO APPLY किया हुआ है जो हमने गाडी की बहुत तलाशा जो अब तक नहीं मिली है कृपा
करके हमारी गाडी की तलाश की जाए तथा मुजीरमान का पता लगाकर हमारी गाडी बरामद करी जाये । SD

```

Figure 3.3: Annotated Entities

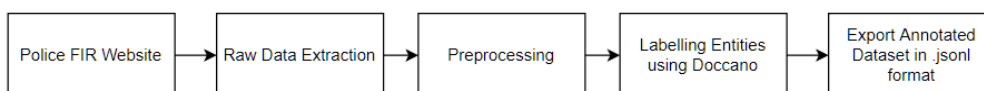


Figure 3.4: LDP Dataset Creation

3.5 LDP_NER and LDP_NER_combined Datasets

LDP dataset comprises of id, text and labels which we obtained after exporting the ldp_dataset.jsonl file using doccano tool. We required to convert this dataset to a format suitable for NER task. NER tasks are mostly done with the use of IOB tagging scheme (there are other tagging schemes too such as IOB2, BILOU, etc.)

³Source: <https://haryanapolice.gov.in/ViewFIR/FIRStatusSearch?From=LFlh1h1x/W49VSlBvdGc4w==>

| Sentence # | Tag | Word |
|------------|-----------------------|---------|
| 11 | B_police_station#name | हाऊसिंग |
| 12 | I_police_station#name | बोर्ड |
| 13 | I_police_station#name | कालोनी |
| 14 | I_police_station#name | अम्बाला |
| 15 | I_police_station#name | कैन्ट |
| 16 | O | श्रीमान |
| 17 | O | जी |
| 18 | O | निवेदन |
| 19 | O | है |
| 20 | O | कि |
| 21 | O | मैं |
| 22 | B_victim#name | कमल |
| 23 | I_victim#name | कान्त |
| 24 | B_victim#relation | S/O |
| 25 | B_victim#relative | श्री |
| 26 | I_victim#relative | पवन |
| 27 | I_victim#relative | प्रकाश |

Figure 3.5: LDP_NER Dataset

```
{
  "id": 416,
  "data": "|||Haryana_FIR_2015_AMBALA_AMBALA_CANTT_0_3.txt::सेवा मे प्रबन्धक
अफसर थाना अ0कैन्ट जय हिन्द श्री मान जी, निवेदन हैकि, आज मन ASI .....",
  "label": [[72, 84, "police_station#name"], [133, 154, "police_assigned#on_duty"], [157, 178,
"police_assigned#on_duty"],.....]
}
```

Figure 3.6: An instance from the LDP Dataset³.

and the most popular dataset was CONLL2003 Shared Task Dataset. This motivated us to convert our LDP dataset in the IOB tags format. LDP_NER Dataset was formed by splitting the tagged entities about the spaces and labelling the beginning word of the entity by prefixing the same label with "B_". Similarly, the following words were labelled with "I_" prefixed labels. The end of sentences or unlabelled entities were given "O" tags. The LDP_NER_combined dataset was formed simply by merging certain annotation labels in the LDP_NER dataset for conducting experiment with lesser number of distinct labels.

3.6 LDPQuAD Dataset

LDPQuAD Dataset is another variation of The Stanford Question Answering Dataset (SQuAD), which is one of the most popular datasets available for handling question answering tasks. SQuAD came into popularity especially after another version of this dataset (SQuAD v2.0), which was released with extra questions without answers to help the model abstain from answering non-answerable questions. Taking inspiration from the same, we created LDPQuAD (LDP Question Answering Dataset). The creation of this dataset was done by using the various labels available in the LDP dataset. A format was set-up to convert certain labels into questions. To generalise few answers, multiple labels were merged together under a common question. The dataset was made in a format similar to that of the SQuAD. The LDPQuAD has same structure with columns namely answers, context, id, question and title.

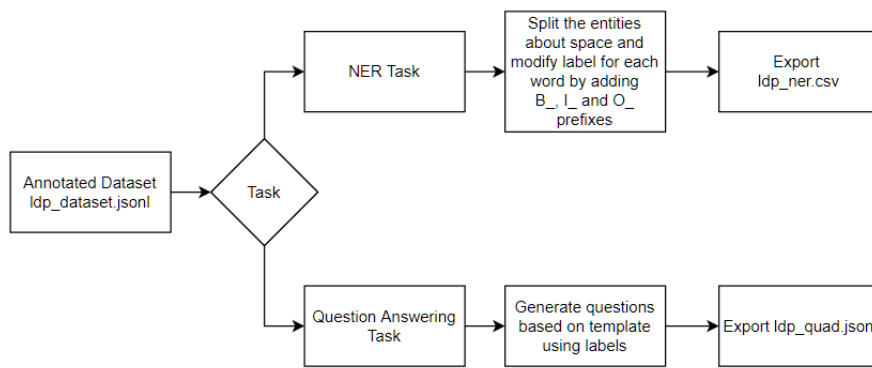


Figure 3.7: LDP_NER and LDPQuAD Datasets Creation

| Label | Question |
|------------------------|--|
| crime_details#date | अपराध कब हुआ है |
| crime_details#location | अपराध कहाँ हुआ है |
| crime_details#type | अपराध किस प्रकार का है |
| criminal#looks | अपराधी केसा दिखता था |
| criminal#qty | कितने अपराधी थे |
| criminal#transport | अपराधी किस वाहन का उपयोग कर रहे थे |
| police_assigned#name | रिपोर्ट दर्ज करने वाला पुलिसकर्मी कौन है |
| police_station#name | अपराध कहाँ दर्ज किया गया है |
| victim#contact | पीड़ित का संपर्क क्या है |
| victim#name | अपराध की रिपोर्ट करने वाला व्यक्ति कौन है |
| victim#relation | पीड़ित के साथ आने वाले व्यक्ति का पीड़िता से क्या संबंध है |
| victim#relative | क्या पीड़िता के साथ कोई व्यक्ति है |
| victim#residence | पीड़ित का निवास क्या है |

Figure 3.8: Question Generation for LDPQuAD Dataset

3.7 Challenges

3.7.1 Text Structure

One of the challenges we faced was with the dataset instances mostly being in Hindi with few words in English, and Hinglish (an informal way of recognising the language having English and Hindi words along with loosely transliterated Hindi words into English and vice versa). We initially proceeded with translating the instances entirely into English so that it is compliant with existing models. This could have been beneficial in model training but not aligned with the motivation behind creating this dataset. Moreover, translations such as "Aakash" are not achievable easily as the existing translation tools can very easily confuse it for "Sky" instead of a certain name or translation of a certain phrase in hindi to Police Thana in English instead of Police Station is like added responsibility that requires more work than simply heading forth with the raw instances for the task of labelling.

3.7.2 NER Task using LDP_NER

There were certain challenges which was realised post training. Calculation of F1-scores for huge number of entity labels is a challenge. Taking care of the occurrences per label was crucial to get class-wise f1-scores. Performing stratification also becomes tough as we cannot achieve a shuffle in which justice can be done to all the labels equally. Another issue was with a lower agreement score for certain labels which ultimately led to weak F1-scores.

3.7.3 Question Answering Task using LDPQuAD

The Question Answering task is where the models are capable of giving answers from the context. The FIR instances are huge in length for performing Question Answering task, especially using advanced models like BERT which gave breakthrough results with SQuAD dataset provided that the context length is within limits. Trimming them to the standard 512 tokens (-2 for CLS and SEP tags) [12] could work but then it leads to loss of information.

| Label | Frequency | Label | Frequency |
|--------------------------|-----------|----------------------------------|-----------|
| lost_item#type | 4324 | occupation_details#workplace | 268 |
| victim#name | 2778 | victim#cast | 224 |
| police_station#name | 1959 | crime_details#location_reason | 177 |
| police_assigned#on_duty | 1772 | lost_item#name | 163 |
| section | 1744 | criminal#transport | 162 |
| FIR_date | 1699 | occupation_details#work_location | 153 |
| crime_details#type | 1679 | criminal#qty | 140 |
| victim#relative | 1666 | criminal#action | 129 |
| victim#relation | 1657 | subject | 97 |
| victim#residence | 1653 | car_details#regno | 70 |
| criminal#looks | 1123 | affected#relative | 62 |
| crime_details#location | 1038 | car_details#name | 60 |
| victim#contact | 960 | crime_details#suspicion | 55 |
| crime_details#date | 911 | affected#relation | 51 |
| bike_details#regno | 906 | lost_item#desc | 46 |
| police_assigned#case | 820 | mobile_details#brand | 41 |
| police_assigned#name | 755 | car_details#engine | 39 |
| criminal#type | 754 | car_details#chassis | 39 |
| bike_details#name | 684 | misc_info#victim | 38 |
| crime_details#time | 629 | misc_info#lost_item | 36 |
| bike_details#chassis | 588 | car_details#color | 36 |
| bike_details#engine | 582 | car_details#model | 30 |
| bike_details#color | 577 | mobile_details#no | 29 |
| lost_item#contents | 575 | misc_info#criminal | 27 |
| police_assigned#location | 480 | mobile_details#imei | 16 |
| FIR_time | 459 | mobile_details#model | 15 |
| bike_details#model | 433 | mobile_details#color | 13 |
| occupation_details#type | 301 | criminal#weapon | 8 |
| lost_item#last_seen | 291 | victim#age | 7 |

Table 3.1: Label Distribution

Chapter 4

Methodology

4.1 Inter Annotator Agreement

For a fraction of the instances from the LDP Dataset, we annotated the cases in our ways. We agreed upon the definition of each label, but there was still an Cohen's Kappa score of 0.63 which is considered to be a good extent of convergence. Exclusion of certain ambiguous labels such as `lost_itemtype` and `lost_itemname` can sometimes be confused. If the gold chain is the lost item, then gold can be the name with the chain type, or gold can also be the type of item. As the disparity was noticed for only a few labels, we decided not to modify them and to keep the scope open for the future study into difference in perception among the annotators.

4.2 Importance of label distribution

We fine-tuned existing BERT models such as BERT, Multilingual BERT, HindiBERT and HindiBERTa on the full training set of LDP_NER dataset. Before training the model, we combined the bike and car labels under one category vehicle. This is still classified as retained labels as the `item_type` is an indicator of whether the lost item is a vehicle such as car or bike or say a gold chain. Now, with a vast number of labels, we tried to experiment by merging various labels which can be broadly classified as single label. For instance, `victim#name`, `victim#residence`, and few others can be clubbed as simply `victim`. This reduced the distinct number of labels significantly and gave better per instance count of instances. To avoid overfitting, this proved to be a good step. For question answering task using LDPQuAD dataset, the label distribution again was critical in identifying the kind of questions to be framed using them. The questions were made in Hindi and were assigned for corresponding labels.

4.3 Stratification vs cross fold validation

Due to the large number of labels, stratification on the basis of certain labels would compromise the distribution of the rest other. A better approach was to shuffle the dataset (`random_state 2018` for instance was used to get the results) and then perform five fold cross validation to determine the best division of the dataset for testing purpose.

4.4 Training

For the entity recognition task using LDP_NER dataset, upon various trials, the training and validation loss curves would converge within ten epochs. Based on this observation, we chose fine-tuned our models with ten epochs. In each epoch, we en-

sured that the previous gradients are cleared before beginning the backward pass. As we provide the labels to the model, we return the loss. To avoid exploding gradient problems, we clip the norm of the gradient. Average training loss is calculated as total loss divided by the length of train data and it is appended with each iteration. After completion of each epoch, we calculate the accuracies and f1-scores on the validation set. For the question answering task using LDPQuAD dataset, the `doc_stride` variable was used which determines the extent of overlap allowed when the instances in the dataset are split, enabling us to preserve the context.

Chapter 5

Experiments and Results

5.1 Models Experimented

5.1.1 BERT

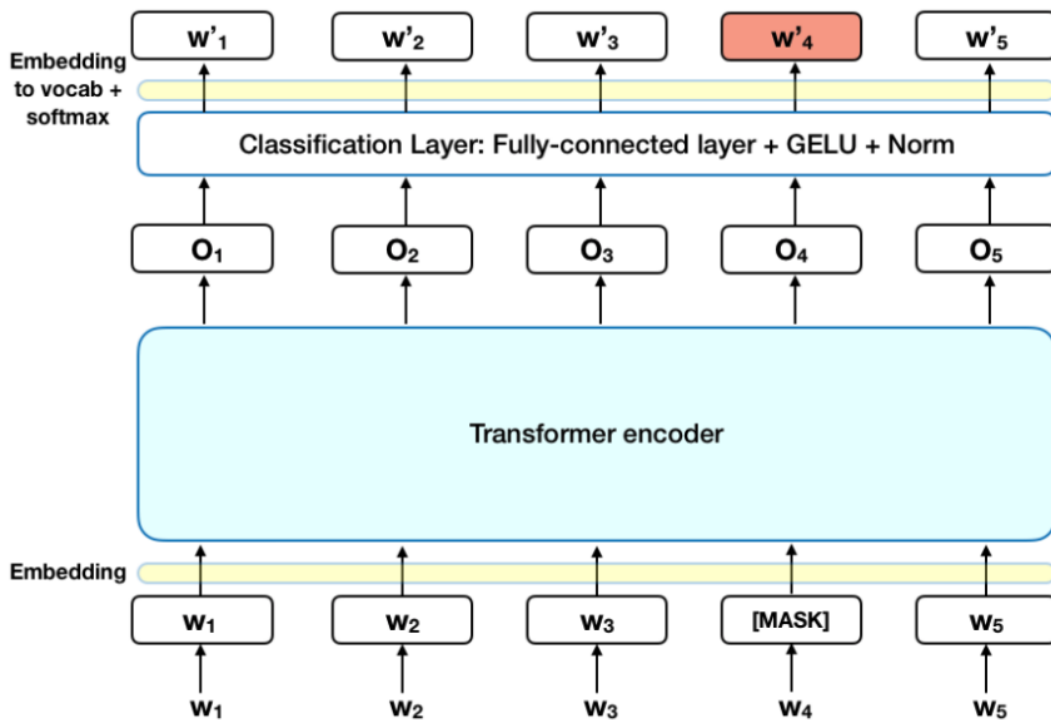


Figure 5.1: BERT Architecture¹.

BERT (Bidirectional Encoder Representations from Transformers) got instantly recognised worldwide for producing state of the art results on various NLP tasks including the CONLL-2003 Shared Task and SQuAD Question Answering Task. This model used the ability of the transformer of bi-directional training which was unlike any other previous models that did not perform tasks more than scanning text from left to right or vice versa or combining both. With bidirectional training, the model proved to capture more context that too not just within a sentence but across multiple sentences. The BERT was trained with the intention to achieve results with Masked Language Model and Next Sentence Prediction. Incorporating the usefulness and vision of BERT with our LDP Dataset by performing finetuning as per our needs greatly helps us in achieving greater than expected results in shorter amount of time.

¹Source: <https://aclanthology.org/N19-1423.pdf>

| | | | |
|----------|--------|---------|--------|
| [CLS] | 2 | . | |
| _सेवा | 1,078 | . | |
| _मे | 368 | _इन्द्र | 13,014 |
| _प्रबन्ध | 19,856 | ाज | 1,379 |
| क | 20 | _रि | 835 |
| _अफसर | 10,314 | कोर् | 27,807 |
| _थाना | 1,238 | ड़ | 764 |
| _अ | 56 | _थाना | 1,238 |
| 0 | 1,936 | _विधि | 3,692 |
| कै | 776 | नुसार | 6,916 |
| न्ट | 3,111 | _किया | 245 |
| . | | _गया | 257 |
| . | | _। | 52 |
| . | | [SEP] | 3 |

Figure 5.2: Tokenized instance of LDP dataset using AutoTokenizer from Huggingface and IndicBERT pretrained model

5.1.2 Multilingual BERT

Multilingual BERT (M-BERT) is also a breakthrough model developed with datasets in 104 languages. This model was trained with the intention of conducting cross-lingual model transfer with zero-shot setting. This deeply facilitates in fine-tuning one language model into another while model training. This model was trained on Hindi language as well, which motivated us to fine-tune this for our LDP dataset. The results achieved for NER task was about F1-score of 80 with fine-tuned version of this model.

5.1.3 DistilBERT

DistilBERT is a concise version of BERT. It is about 40% the size of BERT model but results in considerable accuracies. The DistilBERT used something called triple loss which combines distillation, losses from cosine distance and language modelling. For our LDPQuAD dataset, where BERT was taking huge time due to the construction of heavy models, DistilBERT could provide us the results in way less time.

¹Source: https://huggingface.co/docs/transformers/main_classes/tokenizer#transformers.PreTrainedTokenizer.push_to_hub.example

5.2 Experimental Setup for NER using LDP_NER and LDP_NER combined datasets

The experiments with baseline models of different variations of BERT were conducted with hidden_dropout set to 0.1 and attention dropout set to 0.1. Due to a large number of labels, performing stratification was a challenge and hence simple cross-validation over 5 folds was performed. Each fold gave better F1-scores for different entities because of varying supports.

5.3 Experimental Setup for Question Answering Task using LDPQuAD dataset

The experiments with baseline models of different variations of BERT were conducted with learning_rate = 2e-5, num_train_epochs = 2 and weight_decay = 0.01, with maximum answerlength being set to 30. The doc_stride, which determines the extent of overlap between two sentences of an instance in order to capture context was set to 128.

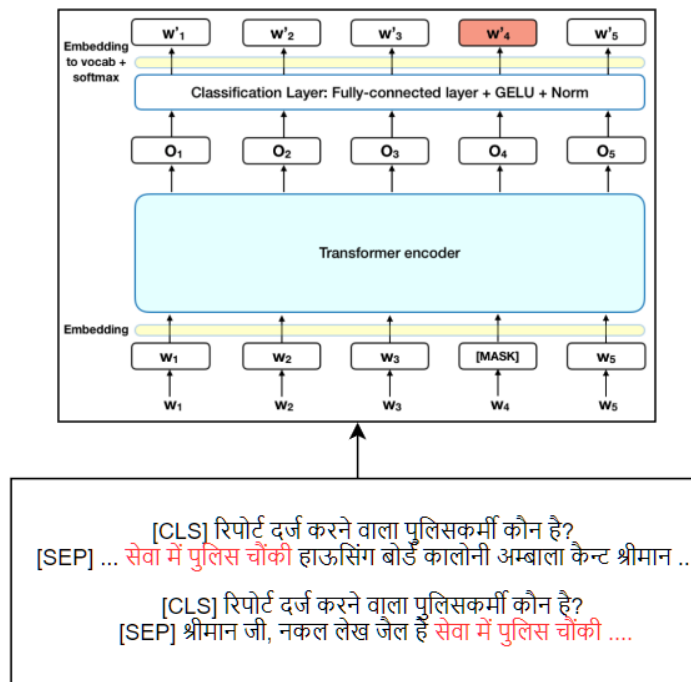


Figure 5.3: LDPQuAD instance : Context preserved using stride

5.4 Evaluation Metrics

5.4.1 Cohen's Kappa

Cohen's Kappa is a measure of reliability between two annotators. The two annotators are required to annotate the same set of instances for same set of

classes and cohen kappa score tells the extent of agreement/ similarity between their annotations. This score generally lies between 0 to 1, 0 being total disagreement and 1 being total agreement. It can also be negative, indicating the agreement being even worse than the situation where dataset is annotated randomly. Cohen's Kappa is measured as

$$K = \frac{P_o - P_e}{1 - P_e} \text{ where}$$

$$P_o = \text{Number in agreement} / \text{Total}$$

$$P_e = P_{correct} + P_{incorrect}$$

5.4.2 F1-score

For question answering task, F1-score is a measure of the extent to which the prediction was correct (it is not necessary that each and every word in the predicted result must match the target value). It is measured as

$$F1 = 2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall})$$

Where precision determines the ratio of number of words shared between target and prediction to the total number of words in the target, whereas recall determines the ratio of number of words shared between target and prediction to the total number of words in the ground truth.

The formula of F1-score is the same for NER task as well. The perception of precision and recall varies as

precision = true positive / (false positive + true positive) recall = true positive / (false negative + true positive) where true positive is the amount of chunks (not tokens) predicted correctly and false positive is the amount of chunks (not tokens) predicted incorrectly. True negative is the amount of chunks that are guessed incorrectly and false negative is the amount of chunks not recognised.

5.4.3 Exact Match

For question answering task, the exact match metric is to evaluate the spans which were predicted accurately. It is a strict measure where if the prediction is correct from starting till ending character then score is given 1 else 0. The metric is used in Question Answering task to determine whether the span of the answer from the context is predicted correctly or not.

For NER task, the exact match would be the measure of the occurrences where each chunk of an entity is predicted correctly.

5.5 Experimental Results

5.5.1 NER_LDP Results

| Dataset | F1-score |
|--------------------------|-------------|
| Multilingual BERT | 77.8 |
| HindiBERT | 18.2 |
| HindiBERTa | 29.4 |
| BERT | 68 |

Table 5.1: LDP_NER Dataset

| Dataset | F1-score |
|--------------------------|-------------|
| Multilingual BERT | 80.2 |
| BERT | 67.5 |
| HindiBERT | 20.08 |
| HindiBERTa | 30.08 |

Table 5.2: LDP_NER_combined Dataset

Fine tuning the Multilingual BERT model on our LDP_NER dataset gives the best F1-score of 77.8 whereas the LDP_NER_combined dataset which has certain high level entity labels merged to form one label, gives a F1-score of 80.2. The NER_LDP dataset uses all labels, which seemingly caused a drop in the overall F1-score with not just multilingual BERT but all the other models as well. Even the class-wise scores saw a significant improvement upon merging of the labels. The model finetuned on Indic-BERT did not perform that well. Interestingly, BERT performed slightly better on the LDP_NER dataset than the one with labels merged i.e. LDP_NER_combined. This could be taken as a positive sign that except for few labels, the large number of labels is not causing a drastic drop in F1-score, thus making this work capable of being extending this study to datasets with huge number of labels.

5.5.2 LDPQuAD Results

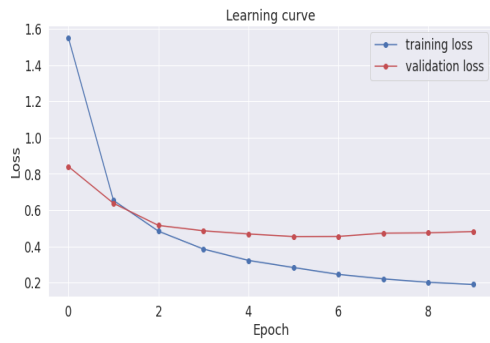
| Dataset | F1-score |
|-----------------------------|-------------|
| Distilled RoBERTa | 69.9 |
| Distilled BERT | 69.0 |
| Distilled Multilingual BERT | 37.25 |

Table 5.3: LDPQuAD Dataset

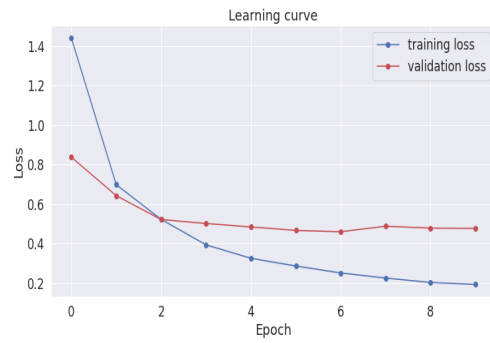
Fine tuning the distilled version of RoBERTa model performed the best with exact_match score of 68.51 and F1-score of 69.9. The LDPQuAD dataset was prepared using dataset with all labels but limited labels converted to questions for the same. The use of doc_stride variable to determine the extent of overlap between two consecutive span of instances have helped significantly in getting commendable result for larger size of instance.

| Entity | Class F1-score | Support |
|---------------------|----------------|---------|
| victim#name | 94 | 462 |
| victim#relation | 86 | 272 |
| victim#residence | 77 | 332 |
| victim#cast | 76 | 68 |
| police_station#name | 70 | 268 |

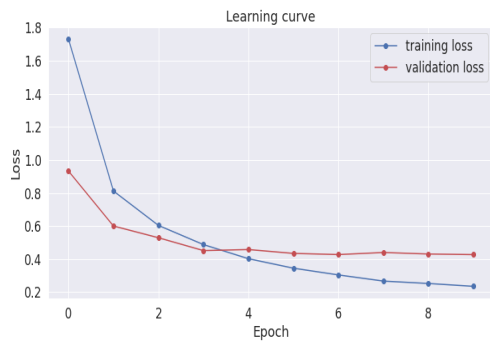
Table 5.4: Entities with best class wise F1-scores for Multilingual BERT



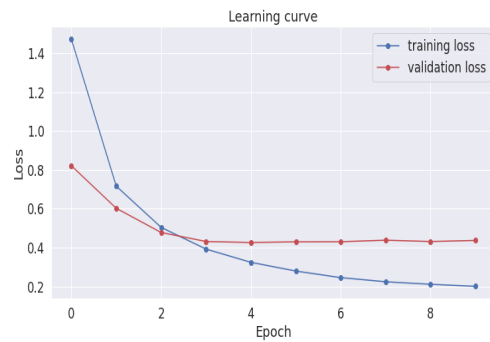
(a) first fold



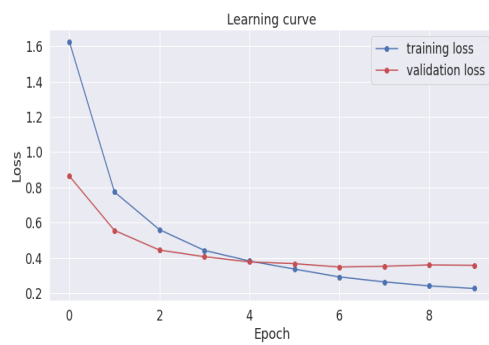
(b) second fold



(c) third fold



(d) fourth fold



(e) fifth fold

Figure 5.4: Training Loss and Validation Loss Curves for Multilingual BERT over 5 folds

Chapter 6

Conclusion and Future Work

The work has explored this potential of custom-built LDP dataset in the legal domain. Earlier works have tried to achieve F1 scores for a limited set of labels or identifying roles. In our work, with a vast collection of hand-labelled entities, future work can try various ways of incorporating these labels to extract deeper relations between the two entities. As the work solely focused on First Information Reports by victims and police officers, the present work is limited to identifying entities related to a particular crime scene under study. Although with a massive number of existing corpora related to judgment prediction in courts, this novel dataset can analyse how many cases have been to trial and have found the victim guilty. The dataset is open to extension using FIR reports available on the Haryana Police website to keep the dataset updated with time. In future we can look into better relation analysis between the entities with better feature [2] integration.

Bibliography

- [1] **May 2022** Arnav Kapoor et al. “HLDC: Hindi Legal Documents Corpus”. In: *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland: Association for Computational Linguistics, pp. 3521–3536. DOI: 10.18653/v1/2022.findings-acl.278. URL: <https://aclanthology.org/2022.findings-acl.278> (visited on 08/16/2022).
- [2] **June 2021** Lu Xu et al. “Better Feature Integration for Named Entity Recognition”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 3457–3469. DOI: 10.18653/v1/2021.naacl-main.271. URL: <https://aclanthology.org/2021.naacl-main.271> (visited on 08/16/2022).
- [3] **Aug. 2021** Ayan Sengupta et al. “HIT - A Hierarchically Fused Deep Attention Network for Robust Code-mixed Language Representation”. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, pp. 4625–4639. DOI: 10.18653/v1/2021.findings-acl.407. URL: <https://aclanthology.org/2021.findings-acl.407> (visited on 08/16/2022).
- [4] **Aug. 2021** Vijit Malik et al. “ILDC for CJPE: Indian Legal Documents Corpus for Court Judgment Prediction and Explanation”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 4046–4062. DOI: 10.18653/v1/2021.acl-long.313. URL: <https://aclanthology.org/2021.acl-long.313> (visited on 08/11/2022).
- [5] **Apr. 2021** Bineet Kumar Jha, Chandra Mouli Venkata Srinivas Akana, and R Anand. “Question Answering System with Indic multilingual-BERT”. In: *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 1631–1638. DOI: 10.1109/ICCMC51019.2021.9418387.
- [6] *Validating Label Consistency in NER Data Annotation* (Jan. 2021e). URL: <https://deepai.org/publication/validating-label-consistency-in-ner-data-annotation> (visited on 08/11/2022).
- [7] **Nov. 2020** Macarious Abadeer. “Assessment of DistilBERT performance on Named Entity Recognition task for the detection of Protected Health Information and medical concepts”. In: *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. Online: Association for Computational Linguistics, pp. 158–167. DOI: 10.18653/v1/2020.clinicalnlp-1.18. URL: <https://aclanthology.org/2020.clinicalnlp-1.18> (visited on 08/12/2022).

- [8] **July 2020** Xinya Du and Claire Cardie. “Document-Level Event Role Filler Extraction using Multi-Granularity Contextualized Encoding”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 8010–8020. DOI: 10.18653/v1/2020.acl-main.714. URL: <https://aclanthology.org/2020.acl-main.714> (visited on 08/11/2022).
- [9] **2020c** Ilias Chalkidis et al. “LEGAL-BERT: The Muppets straight out of Law School”. en. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 2898–2904. DOI: 10.18653/v1/2020.findings-emnlp.261. URL: <https://www.aclweb.org/anthology/2020.findings-emnlp.261> (visited on 08/11/2022).
- [10] **July 2020** Juntao Yu, Bernd Bohnet, and Massimo Poesio. “Named Entity Recognition as Dependency Parsing”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 6470–6476. DOI: 10.18653/v1/2020.acl-main.577. URL: <https://aclanthology.org/2020.acl-main.577> (visited on 08/16/2022).
- [11] **June 2019** Yi Luan et al. “A general framework for information extraction using dynamic span graphs”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 3036–3046. DOI: 10.18653/v1/N19-1308. URL: <https://aclanthology.org/N19-1308> (visited on 08/16/2022).
- [12] **June 2019** Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: <https://aclanthology.org/N19-1423> (visited on 08/11/2022).
- [13] **July 2019** Telmo Pires, Eva Schlinger, and Dan Garrette. “How Multilingual is Multilingual BERT?” In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 4996–5001. DOI: 10.18653/v1/P19-1493. URL: <https://aclanthology.org/P19-1493> (visited on 08/11/2022).
- [14] **Feb. 2017** Feifei Zhai et al. “Neural models for sequence chunking”. en. In: AAAI press. URL: <https://research.ibm.com/publications/neural-models-for-sequence-chunking> (visited on 08/12/2022).
- [15] **Nov. 2016** Pranav Rajpurkar et al. “SQuAD: 100,000+ Questions for Machine Comprehension of Text”. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 2383–2392. DOI: 10.18653/v1/D16-

1264. URL: <https://aclanthology.org/D16-1264> (visited on 08/16/2022).
- [16] **2003** Erik F. Tjong Kim Sang and Fien De Meulder. “Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition”. In: *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pp. 142–147. URL: <https://aclanthology.org/W03-0419> (visited on 08/16/2022).