# Information Extraction from Court Room Records

By
**Nida Ahmed**
**2018-NUST-MS-CS-08-275470**

Supervisor
**Dr. Faisal Shafait**
**Department of Computing**

A thesis submitted in partial fulfillment of the requirements for the degree of Masters in Computer Science (MS CS)

In
School of Electrical Engineering and Computer Science,
National University of Sciences and Technology (NUST),
Islamabad, Pakistan.

(November 2021)

# Dedication

I dedicate this thesis to my parents, brothers and family.

# Acknowledgment

First and foremost, praises and thanks to the God, the Almighty, for His showers of blessings throughout my research work to complete the research successfully.

I would like to express my gratitude to my research supervisor, Dr. Faisal Shafait for his guidance and support throughout this journey. Alongwith my GEC members, Dr. Adnan Ul-Hasan, Dr. Seemab Latif and Dr. Rabia Irfan.

I am extremely grateful to my parents for their love, prayers, sacrifices and endless strength for me to carry on. Their unwavering confidence in me was what led me to finish this journey.

I would like to say special thanks to my friends turned family, Ayesha Sadia Hashmi, Ayesha Sarwar, Ayesha Ashfaq, Hiba Faisal, Maha Rasheed, Maria Rasheed and Yumna Qamar for their patience and moral-boosts during this journey.

Finally, my gratitude goes to all the people who have supported me to complete the research work directly or indirectly.

# Contents

# List of Abbreviations

| | |
|---|---|
| ACE | Automatic Content Extraction |
| Bert | Bidirectional Encoder Representations from Transformers |
| CoNLL | Computational Natural Language Learning |
| CRF | Conditional Random Fields |
| IBO | Inside Beginning Outside |
| IREX | Information Retrieval and Extraction Exercise |
| LHC | Lahore High Court |
| SCP | Supreme Court of Pakistan |
| LSTM | Long Short Term Memory |
| NLP | Natural Language Processing |
| NE | Named Entity |
| NER | Named Entity Recognition |
| QA | Question Answering |
| TREC | Text Retrieval Conference |

# List of Tables

# List of Figures

# Abstract

The legal domain remains among various areas that have many opportunities when it comes to improvement and innovation through computational advancements. In Pakistan, in the recent past, the courts have made reported judgments available to the public. As this data continues to grow at a rapid rate, it has become essential to process this massive chunk of data to better meet the requirements of the respective stakeholders. However, extracting the required information from this unstructured legal text is the main issue. Therefore, our goal is to have a machine learning system that can automatically extract information out of these publicly available judgments of the Supreme Court. Once this information has been extracted, it can then be used by the lawyers, judges as well as civilians and also for policy making in Pakistan. For the purpose of our work, a total of thirteen entities are being extracted including dates, case-numbers, respondent names, reference cases, FIR no., person names, references etc. A labeled dataset is created using the publicly available legal judgments from the Supreme Court of Pakistan by using annotation guidelines. A pre-trained BERT model is then further trained and fine-tuned on the created dataset for Named Entity Recognition to extract the desired information. Our model also improved the results of the similar dataset available consisting of judgments from Lahore High Court which has smaller number of labels.

# Chapter 1

# Introduction and Motivation

Documents, digital or printed, are one of the most effective and time tested means for retaining knowledge and experience. With the advent of computing, a rise in the digitization of public documents/records has been observed which has resulted in the ease of access of such records in many areas of the society. As the volume of the documents is increasing, it has become strenuous to maintain a structured data/information that can be traceable. Organisations often require specific information they need from the document, to visualize relationships between different set of entities or build information and prediction systems.

There is a requirement of information extraction in the legal domain as well. As now judgments have been digitized, different types of systems can be helpful to lawyers and judges to speed up their work processes and save their valuable time. Information extraction can be one of the core systems on which other tasks could be dependent. Since in legal domain, the decisions are made based on laws. If laws have been retrieved from the associated judgments, it can be used to identify the case similarities to help lawyers, and judges look at a more concrete dataset.

This chapter includes an introduction to our work along with the description of the data being used, and challenges etc.

## 1.1 Introduction

In Pakistan, in the recent past, the major courts have made their reported judgments available to the public. These judgments are available on the official websites of the courts. These include the High Courts as well as the Supreme Court of Pakistan. The Supreme Court of Pakistan, alone, has a total of 1,694 judgments publicly available on their website. With an increase in the number of cases being filed each year, this data continues to grow at a rapid rate. It has now become essential to process this massive chunk of data to automate the judicial process in some way.

Traditionally, when preparing for a case, the lawyers have to not only use laws to make their case but also refer to similar cases from the past. These cases are searched manually which takes a considerable amount of time. This process can be made straightforward by extracting useful information such as case details, laws and case references in the judgment etc. Information extraction can be an important part of judicial automation and it can provide a great deal of benefits to all the stakeholders associated in the legal context.

## 1.2 What is a Judgment?

A court of law is a body of persons that have the judicial authority to hear and resolve disputes in different cases, be criminal, civil, military, human rights issues, etc. When a case is presented to a court, the court then decides what really happened and what needs to be done about it. In a criminal case, the court gets to decide whether the person committed the crime and how they should be punished and what the punishment should be. In civil cases the court can present a peaceful way to resolve any disputes that may be present between the parties involved.

The decision the courts make regarding any given case is released as a judgment. A judgment, in law, is the decision of a court regarding a legal action or proceeding considering the rights and liabilities of the parties involved. It settles the dispute between the parties by determining their rights and obligations.

A judgment contains all the important information pertaining to the proceeding of a given case. This information includes the names and details of the people and organizations involved, people that are present during the hearings, the judges hearing the case, date of the final hearing when the judgment is being announced, some information regarding cause of the case/suit, as well as any laws or previous cases being referred to for the given case.

This judgment, once announced, is handed over to the parties involved.

The judgment from a given court is either acted upon or the parties involved can appeal against the court's decision to a higher court. In Pakistan, the highest court a judgment can be appealed against at is the Supreme Court of Pakistan. The judgment from the Supreme Court itself can also be reviewed.

Depending on the conditions pertaining a case that is brought to a court, the court can assign it a case type. This is done to categorize all the cases that a court is hearing into different types. In the subsections 1.2.1, 1.2.2 and 1.2.3 we shall describe the different case types, the type we chose to work with for the purposes of this work as well as the format of the judgments.

### 1.2.1 Case Types

According to the Constitution of Pakistan, the Supreme Court of Pakistan exercises original, appellate and review jurisdiction. The Supreme Court has original jurisdiction when it comes to inter-governmental disputes and can issue declaratory judgments, and can also enforce fundamental rights involving an issue of public importance. The Supreme Court also holds appellate jurisdiction for appeals against the judgments passed by High Courts for different criminal and civil cases as well as for the interpretation of the Constitution. Other than this,as the Supreme Court is the highest court in Pakistan, the court can also review its own judgments.

Depending on the circumstances pertaining a case, it is filed under one of the different categories at a given court. The Supreme Court Of Pakistan has a total of 28 case types. All the cases filed at the Supreme Court Of Pakistan are divided into these 28 case types. These categories are assigned depending on whether a case is an appeal or a petition, and whether it is of the civil or criminal nature, and if it is regarding Shariat law or human rights issue etc. These include Civil and Criminal appeals, petitions, miscellaneous appeals, Shariat review petitions, Human Rights, etc. The names of all the case types in Supreme Court of Pakistan are as follows:

1. Civil Appeal (C.A.)
2. Criminal Appeal (Crl.A)
3. Criminal Shariat Appeal (Crl.Sh.A.)
4. Civil Shariat Appeal (C.Sh.A.)
5. Civil Petition (C.P.)
6. Criminal Petition (Crl.P.)

7. Civil Review Petition (C.R.P.)
8. Criminal Review Petition (Crl.R.P.)
9. Criminal Shariat Petition (Crl.Sh.P.)
10. Civil Shariat Review Petition (C.Sh.R.P.)
11. Jail Shariat Petition (J.Sh.P.)
12. Jail Petition (J.P.)
13. Constitution Petition (Const.P.)
14. Criminal Shariat Review Petition (Crl.Sh.R.P.)
15. Human Rights Case (H.R.C.)
16. Criminal Original Petition (Crl.O.P.)
17. Suo moto Review Petition (S.M.R.P.)
18. Criminal Miscellaneous Application (Crl.M.A.)
19. Civil Misc. Appeal (C.M.Appeal)
20. Civil Miscellaneous Application (C.M.A)
21. Criminal Miscellaneous Appeal (Crl.M.Appeal)
22. Criminal Suo Moto review Petition (Crl.S.M.R.P.)
23. Suo Moto Case (S.M.C.)
24. Reference (Reference)
25. Intra Court Appeal (I.C.A.)
26. Criminal Suo Moto Shariat Review Appeal (Crl.S.M.Sh.R.P.)
27. Human Rights Miscellaneous Appeal (H.R.M.A.)
28. Civil Shariat Petition (C.Sh.P)

### 1.2.2 Civil Appeal

All cases regarding civil issues are filed under a civil category, be it civil appeal, civil petition, civil review petition, etc.

The Supreme Court of Pakistan has the highest number of available judgments for cases filed under the category of Civil Appeal. Due to the highest number of available judgments, we shall be using the judgments for Civil Appeal cases for the purposes of our work. These are all appeals against the judgments of High Courts for that fall under civil case types in the Supreme Court.

According to the Supreme Court of Pakistan, a case can be filed as Civil Appeal if it falls under the following conditions as per the Constitution of Islamic Republic of Pakistan, within 30 days from date of order/judgment of High Court.

- Article 185 (2) (d): if the amount or value of the subject-matter of the dispute in the court of first instance was, and also in dispute in appeal is, not less than fifty thousand rupees or such other sum as may be specified in that behalf by Act of the parliament and the judgment, decree or final order appealed from has varied or set aside the judgment, decree or final order of the court immediately below; or

- Article 185 (2) (e): if the judgment, decree or final order involves directly or indirectly some claim or value and the judgment, decree or final order appealed from has varied or set aside the judgment, decree or final order of the court immediately below;

### 1.2.3 Judgment Format

The Supreme Court of Pakistan uses several different formats for its written judgments. Formats are not only different between different case types and categories of the Supreme Court but judgments from the same case category can also have differing formats from each other. This means that the order in which the information is mentioned in the beginning of these judgments can be slightly different.

Regardless of these differences, the main information mentioned in the judgments tends to be similar. Every judgment has the name of the judges present at the final hearing, date of judgment, case number of the case the judgment is for and names of parties involved, etc.

Figure 1.1 shows one of the Civil Appeal judgments from the Supreme Court of Pakistan.

**IN THE SUPREME COURT OF PAKISTAN**

(Appellate Jurisdiction)

**PRESENT:**

MR. JUSTICE UMAR ATA BANDIAL
MR. JUSTICE MUNIB AKHTAR
MR. JUSTICE YAHYA AFRIDI

**CIVIL APPEALS NO. 1660 AND 1661 OF 2014**
*(On appeal against the judgment dated 16.05.2012 of the Islamabad High Court, Islamabad passed in Tax Appeal No.7 and 8 of 2005)*

**M/s Al-Khair Gadoon Ltd.** ...Appellant(s)

**versus**

**The Appellate Tribunal etc.** ...Respondent(s)

For the Appellant(s): Mr. Saood Nasrullah Cheema, ASC (in both cases)

For the Respondent(s): Dr. Farhat Zafar, ASC (in both cases)

Date of Hearing: 21.01.2019

**JUDGMENT**

**Yahya Afridi, J.—** Leave to appeal was sought by M/s Al-Khair Gadoon Limited, appellant in both the appeals, challenging the decision of the Islamabad High Court, Islamabad in Tax Appeals No.7 and 8 of 2005 both dated 16.05.2012. This Court allowed leave vide order dated 25.11.2014 in terms that:

> "In order to consider whether the impugned judgment dated 16.5.2012 is erroneous as the learned High Court failed to take into account the fact that the learned Tribunal gave the clear cut findings that provisions of Section 4(2) of the Central Excise Act, 1944 (the Act) were not attracted and that the show cause notice was issued on the basis that the petitioner was evading central

Figure 1.1: Judgment for a Civil Appeal to the Supreme Court of Pakistan

## 1.3 Named Entities

The term ‘Named Entity’ was coined for the first first time ever at the 6th Message Understanding Conference [18] When it comes to entity extraction task, Named Entities (NE) were defined as proper names and quantities that are of interest. Person name, location names, and organizations, as well as dates, times, percentages and monetary amounts can be considered named entities.

Named Entities can either be generic or domain specific. Generic entities include person names, organizations, date, location, etc. Domain specific entities can differ greatly depending on the domains. For example, for biomedical texts, the cycle days and cycle lengths, drugs, treatments and symptoms along with diagnosis can be the important entities.

### 1.3.1 Named Entities in a Judgment

Judgments are long and complex documents and it is, understandably, difficult for a human being to understand and acquire information from. As multiple older judgments and cases are needed to prepare for a new case, the process of acquiring important information gets very tiresome and error-prone.

Following are some of the most common named entities that are available in a given Civil Appeal judgment of the Supreme Court of Pakistan.

- Judges Name –These are the names of any judges that might be present at the final hearing of the case including the judge that announces the judgment.
- Parties Involved –These include the appellants and the respondents of the case.
- Case Number –These are the case numbers of the cases whose judgment is being announced. These also include case numbers of cases whose decision had been appealed against at the Supreme Court and will be decided in the judgment.
- Laws –These are different laws that the lawyers referred to uphold the case that were mentioned in the judgment. These also include the laws that the judges refer to to justify their final decision.
- Date –These are the important dates mentioned in the judgment, including the date of the final hearing/announcement of the judgment,

dates the judgment for the cases being appeal against was released, and other important dates pertaining the case.

- Other Cases –These are older cases that have already been decided and the judgment has been announced. These are often cases that the lawyers for the parties may use to strengthen their case or that the judges may use to support their decision.

The Figure 1.2 shows some of the entities that are mentioned in the beginning of a given judgment released by the Supreme Court of Pakistan.

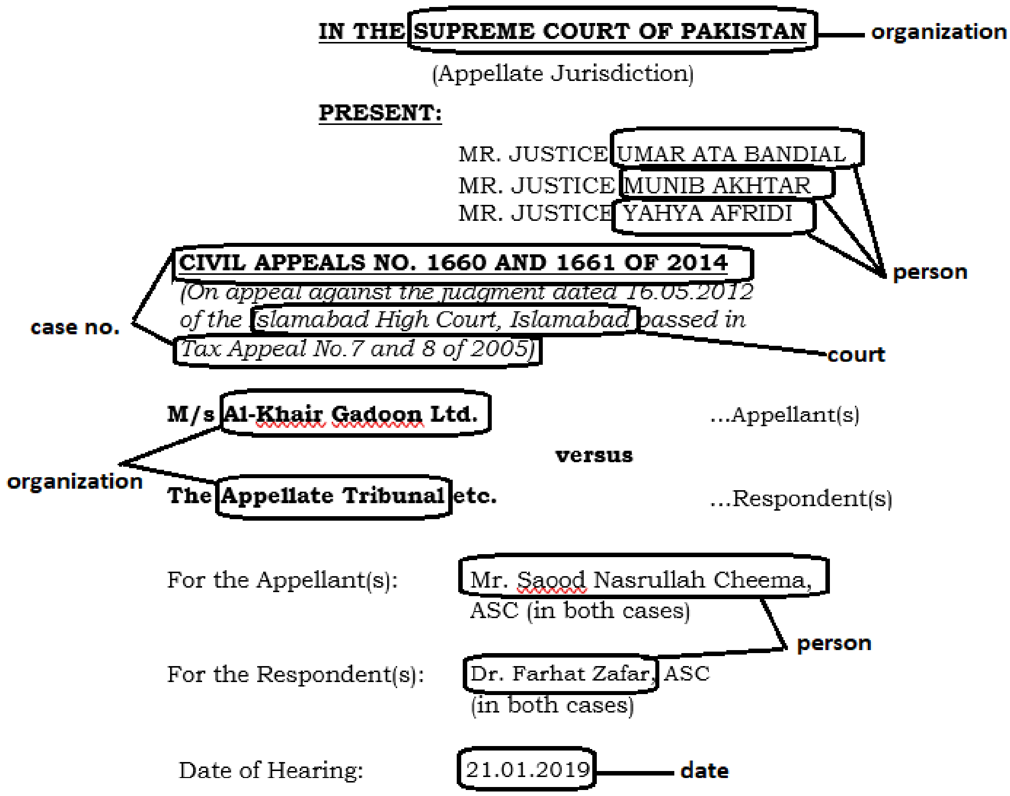


Figure 1.2: Some Named Entities in a Judgment from the Supreme Court of Pakistan

## 1.4 Named Entity Recognition

Named Entity Recognition (NER) is a Natural Language Processing (NLP) technique for Information Extraction (IE) that pinpoints named entities in a given text and then classifies the entities into certain predefined categories. This technique is also often referred to as Entity Identification, Entity Extraction, and in some cases, Named Entity Recognition and Classification (NERC). NER serves as the basis of many other crucial areas in Information Management, such as Ontology Population, Semantic Annotation, Opinion Mining and Question Answering. These named entities entities can be names of people, organizations, locations, monetary values, referred documents, quantities, and more.

Different methods can be used for the purposes of Named Entity Recognition, including Traditional approaches, unsupervised methods, feature-based supervised methods, as well as deep learning methods. These would be explained in Chapter 2.

In the judicial domain, named entity recognition includes the recognition of domain specific as well as generic named entities. Named entity recognition in judicial context can help us extract important information from judicial documents, in our case these documents are judgments. Using NER, we can extract the named entities mentioned in 1.3.1 and any other necessary information that might be needed.

NER systems can act as a basis for automation systems as these extracted named entities can then be used to build knowledge graphs and ontologies. When it comes to the judicial domain, NER can extract important information that might be needed by lawyers or judges to prepare for a case. Traditionally the way these cases are prepared is by the lawyers manually studying older cases and looking up similar cases to support theirs.

Using named entity recognition, we can extract the laws that were referred to in a given case along with the referred cases. This makes the search for similar cases much easier and can help save a lot of time. NER can also extract the original case that might have been appealed against in a given judgment which can give us the important information to look up the original case information and judgment from the lower court that passed the judgment.

## 1.5 Challenges of NER in Judgments

Following are some of the challenges in the field of NER in the context of legal judgments.

**Types of Entities to Identify**

The types of named entities that may need to be extracted from court judgments are not usually considered in international evaluation. Some examples of this include the different case and law references that are mentioned in judgments. Different judgments need to be studied in order to decide the information that needs to be extracted and the different named entities need to be identified. Another challenge was the same names often fell under different named entity labels depending on the context. For example, an organization name can be labelled as an organization or a respondent depending on the context in which the name of said organization is being mentioned. Following are some of the challenges due to the entity types in these judgments:

- A given court can be the one whose judgment is being appealed against or whose past judgment is being used to strengthen the argument in the present judgment.
- An organization can be a respondent for a given case as well.
- A case number mentioned in a judgment can be the number for the case that is being decided or the judgment that is being appealed against.
- The judgment refers to different laws depending on the backgrounds of the involved parties, e.g., the parties involved can claim for a different division of inheritance depending on the religious background of the parties.

**Annotation Guidelines**

As the entities to be extracted from judgments are domain specific, and are unique to the judicial documents, the annotation guidelines for said entities need to be decided. A well defined guide for the labeling of these entities is not present. As mentioned in the previous section these entity labels will also be dependent on the context in which an entity is mentioned.

## 1.6 Problem Statement

Irrespective of the decades of research in information extraction, a system is not publicly available to extract information from the judgments from the courts of Pakistan. The methods already proposed only work under certain conditions, and the available dataset does not cover all the important information that may be present in a judgment.

## 1.7 Solution Statement

Developing a robust system to solve the information extraction problem that makes it invariant to changing document layouts and create a dataset that has all important information in a judgment labelled with guidelines that can be implemented to all court judgments regardless of the level of court as long as the case category remains the same.

## 1.8 Key Contributions

We propose a robust information extraction approach for court judgments using Bidirectional Encoder Representations from Transformers (BERT) and provide a labelled dataset based on the judgments of the Supreme Court of Pakistan containing entities that can help extract important information from the judgment text.

## 1.9 Upcoming Chapters

Later part of the thesis document is organized in the following chapters.

#### Literature Review

This chapter serves as a window into the notable work that has been done on named entity extraction in the past. This section sets a research direction in this dissertation.

#### Design and Methodology

This chapter discusses our proposed named entity labeled dataset and the NER method in detail. It breaks down our approach into different modules and provides an insight into their technical details.

### Experiments, Results and Analysis

This chapter presents the experiments and their results. It also provides the analysis of the results in detail with handpicked examples.

### Conclusion and Future Direction

This chapter provides the final conclusive remarks and sheds light upon the future direction for the research community.

# Chapter 2

# Literature Review

Named Entity Recognition can not only be considered an independent means for Information Extraction (IE), but it also has a fundamental role to play in multiple other applications of natural language processing (NLP), including but not limited to information retrieval [19, 44], question answering [30], text translation [2], automatic document summarizing [26], knowledge base construction [16] and text understanding [62, 9], etc.

The first instance of the phrase "Named Entity" (NE) being used is at the sixth Message Understanding Conference (MUC-6) [18]. Since then, an overwhelming interest in different NER tasks can be observed. Multiple scientific events have also devoted much effort to this topic, e.g, ACE [14], ConLL03 [49], IREX [12], TREC Entity Track [4]. The Named Entity Recognition problem has captured the interest of the research community for decades. There are two different streams for NER techniques, traditional and deep learning approaches. The traditional techniques further include approaches include rule based approach, unsupervised learning and feature based supervised learning approaches. This literature discusses all these in detail.

## 2.1 Traditional Approaches to NER

When it comes to Named Entity Recognition, the traditional approaches can be essentially classified into three main streams, i.e., rule-based, unsupervised learning and feature-based supervised learning approaches [39, 59].

### 2.1.1 Rule-based Approaches

Rule-based NER systems do not require previously annotated data and rely on a set of handcrafted rules which can be based on gazetteers that can be domain specific [16, 50], as well as different syntactic-lexical patterns [61].

Kim [24] proposed a system that can automatically generate rules based on Brill's part-of-speech (POS) tagger [8]. To overcome the challenge of the existence of multiple synonyms for individual genes and proteins, Hanisch et al. [20] proposed a system called ProMiner that utilizes pre-processed synonym dictionary and identifies the name instances in biomedical tests and associate the protein and gene database identifiers. Quimbaya et al. [46] introduced a dictionary-based NER system focused on improving the entity recall while having limited decline in precision in electronic health records.

Rule-based NER systems perform impressively well when lexicon is comprehensive but because of rules that tend to be domain-specific and incomplete dictionaries, these systems often have result with high precision and low recall and cannot be used for a different domain. Some of the better known rule-bases NER systems include LaSIE-II [22], SAR [1] and LTG [38].

### 2.1.2 Unsupervised Learning Approaches

Clustering is a quintessential approach when it comes to unsupervised learning. NER systems based on clustering extract entities based on context similarity from clustered groups. Collins et al. [11] provided two un-supervised algorithms for named entity classification and presented their observation that using data that is not labeled massively lessens the need of supervision to mere 7 basic rules referred to as "seed" rules.

The system KnowItAll [16] took advantage of a set of declared names for input and bootstraps its process for recognition from a limited set of generic extraction patterns. Nadeau et al. [40] proposed a system that does not require manual labeling of training data. The system combines the tasks of entity extraction and disambiguation and can be used for gazetteer building. Zhang and Elhadad [61] proposed an unsupervised system for biomedical named entity recognition (BM-NER). The system leveraged terminologies, corpus statistics and syntactic knowledge. Experimental results on two different datasets, clinical notes and biomedical literature, with different entity types shows the proficiency and versatility of their unsupervised approach.

### 2.1.3 Feature-based Supervised Learning Approaches

As an application of supervised learning, NER often falls under the division of sequence labeling or multi-class classification tasks. Features are carefully designed from the given annotated data samples to represent individual training examples. Machine learning algorithms are then trained to recognize any similar patterns in unseen data in order to extract the named entities. When it comes to supervised NER systems, feature engineering is of critical importance. The representation of feature vector is an abstraction over text where a word is represented by one or multiple Boolean, nominal, or numeric values [39, 51]. Word-level features (e.g, part-of-speech tag, morphology, and case) [64, 52], document and corpus features (e.g., numerous occurrences and local syntax, etc), and list lookup features (e.g., DBpedia gazetteer and Wikipedia gazetteer) have been used widely in different supervised NER systems.

Many machine learning algorithms have been used in NER based on these features, including Decision Trees, Maximum Entropy Models (MEM), Hidden Markov Models (HMM), Support Vector Machines (SVM) and Conditional Random Fields (CRF). The first HMM-based NER system was proposed by Bikel et al. [6], named IdentiFinder, which can be used in the identification and classification of dates, names, numerical quantities, and time expressions. Szarvas et al. [57] created an NER system using AdaBoostM1 and C4.5 Decision Tree algorithms. The system was trained on newswire articles and was portable across multiple languages. Multiple independent decision tree classifiers were trained and their decisions were combined using a majority voting scheme.

Borthwick et al. [7] applied maximum entropy theory and proposed "Maximum Entropy Named Entity" (MENE), which made use of a diverse range of knowledge sources in making its tagging results. Other maximum entropy models were also proposed by Bender et al. [5] and Chieu and Ng [10]. McNamee and Mayfield [37] proposed an SVM (Support Vector Machines) based model called SNOOD that required minimal linguistic knowledge and could be applied to different target languages without any major adaptation. Their classifiers was trained on 258 orthography and punctuation, and 1000 language related features. Each classifier made binary decision regarding the current token belonging to one of each of the 8 classes or not. One downside of using SVMs is that they donot take neighboring words of the current token into account.

Conditional Random Fields (CRFs) can be used to overcome this con. McCallum and Li [36] introduced a feature induction model for CRFs in named entity recognition. Krishnan and Manning [25] proposed a two step approach using two coupled CRF classifiers. The output of the first CRF was used to obtain latent representations that can be used by the second CRF. CRF based NER has been used in multiple domains including chemical text [48], tweets [47, 34] and biomedical text [52, 32]. In the legal domain Sharafat et al. [53] used CRF model to extract named entities from civil court proceedings.

## 2.2 Deep Learning Techniques For NER

Deep learning models consist of multiple processing layers to learn representations of data with multiple levels of abstraction. Recently, deep learning models have become dominant in the field of named entity recognition and have achieved state-of-the-art results. The main advantage of deep learning systems is the representation learning and semantic composition accredited to both the vector representations and neural processing.

Yao et al. [60] proposed a Bio-medical Named Entity Recognition (Bio-NER) system based on deep neural network architecture where each layer abstracts features based on the features generated by lower layers. Zheng et al. [63] proposed a single model solution using a novel tagging scheme which converted the joint extraction of entities and relations task into a tagging problem. The model used word embeddings learned on NYT corpus by word2vec toolkit. Ma et al. [35] used a combination of bidirectional LSTM, CNN and CRF models for an end to end system that required no feature engineering and not data pre-processing. Oeters et al. [43] proposed ELMO word representations which are functions of the entire input section and computed on top of two-layers biLMs with character convolutions.

Baevski et al. [3] proposed a method which achieved state-of-the-art results on CoNLL03 by pre-training a bi-directional transformer model in a cloze-style manner. Lee et al. [27] applied transfer learning by training an ANN model trained on a large labelled dataset and transferring it to another dataset with limited number of labels. Shen et al. [54] proposed that the amount of training data can be drastically reduced by using a lightweight CNN-CNN-LSTM model incrementally training for the NER task with each batch of new labels.

Qu et al. [45] proposed that named entity types that might be related regularly share lexical and context features. This approach finds the correlation among source and target types of named entity employing a neural network consisting of two layers. Peng and Dredze [42] explored the idea of transfer learning in a multi-domain setting. The novelty in this task was not only domain mismatch but also label mismatch.

Huang et al. [21] proposed a multi-tasking deep structural model to integrate "partially annotated" dataset to jointly identify all entity types that might appear in a training corpora. Liu et al. [33] proposed that the chances of neural models can be reduced by utilizing external gazetteers. Zie and Lu [23] proposed a dependency-guided LSTM-CRF model that can encode the complete dependency trees and capture syntactic relations and infer the existence of names entities for the NER task.

Strubell et al. [56] proposed a faster alternative to Bi-LSTMs called Iterative Dilated Convolutional Neural Networks (ID-CNNs) which has improved capacity for large context and structure prediction than traditional CNNs. Ghaddar and Langlias [17] observed that lexical features are often discarted in neural network approaches to NER and proposed to embed words and entity types into low-dimensional vector spaces that is trained from annotated data. They computed a feature vector representing each word and used a vanilla recurrent neural network (RNN) for entity recognition.

Devlin et al. [13] introduced a new language representation model called BERT (Bidirectional Encoder Representations from Transformers). Li et al. [31] used BERT with Dice Loss to overcome data imbalance issue: negative examples significantly outnumbering positive examples. The Dice loss gives resembling importance to false positives and false negatives and is relatively immune to the data imbalance issue.

Lee et al. [28] introduced BioBERT (Bidirectional Encoder Representations from Transformations for Biomedical Text Mining). BioBERT is a domain-specific language representation model that has been pre-trained on large-scale biomedical corpora. The analysis show that pre-training BERT on biomedical corpora helps it in understanding complex biomedical texts and improves the entity and relation extraction scores.

In the legal domain, Dozier et al. [15] proposed three different methods for Named Entity Extraction and Resolution including look-up methods, context rule based methods and statistical models.

Leitner et al. [29] created a dataset consisting of German court decisions and used CRFs and Bi-LSTMs for Named Entity Recognition from said dataset. The proposed word was carried out under the European LYNX project that includes the development of a semantic platform for the creation of different document processing applications in the legal domain.

Skylaki [55] proposed the use of Pointer Generator Network for NER in noisy text obtained from PDF files of US court judgments by formulating the NER task as a text-to-text sequence generation task and then training a pointer generator network to generate the the entities in the document rather than label them. Wang et al. [58] proposed a Sequence Tagging Model (STM) that was created by combining an Inter-Dilated Convolution Neural Network (IDCNN) and a Bi-LSTM model. The model could be used for large scale data from Brazilian legal documents. The paper also compared the results of the model with IDCNN-CRF based model.

### Summary

In this chapter we mentioned different methods for Named Entity Recognition including Traditional Approaches, which include rule-based approaches, unsupervised learning approaches as well as feature based approaches, along with different deep learning techniques for NER. We also mentioned some works of named entity recognition in the legal domain. In the next chapter, we propose our methodology for NER for Judicial judgments from the Supreme Court of Pakistan.

# Chapter 3

# Design and Methodology

From our literature review we can see that transformer based Named Entity Recognition systems can perform better than other deep learning models. This chapter will include the details regarding the dataset creation from judicial judgments as well as our method for Named Entity Recognition.

Our work includes two important parts.

1. Dataset creation
   This includes the details regarding the creation of our dataset of Labelled Civil Appeal Judgments from the Supreme Court of Pakistan.

2. Named Entity Extraction
   This includes the details of the model we used for named entity recognition.

This chapter discusses both these sections one by one in detail.

## 3.1 Dataset Creation

One of our biggest challenges was the creation of a labelled dataset. For this we needed to download the required judgments, define the unique named entities that are present in a given judgment and describe labeling guidelines for the annotation process. We will discuss these steps in details in the subsections 3.1.1, 3.1.2 and 3.1.3.

### 3.1.1 Data Acquisition

In order to create a dataset of the appropriate size, we downloaded the Civil Appeal judgments available on the Supreme Court's website. These judgments were downloaded as pdf files from the Supreme Court of Pakistan

Website using BeautifulSoup python library. The files that were corrupted or were in Urdu language were then manually removed and a total of 214 Civil Appeal judgments were selected for annotation.

### 3.1.2 Named Entity Identification

Some of the named entities we needed to extract from judgments were straightforward. These included person, organization, location, date, and monetary amounts. Other than these, the judgments included the names of other courts of Pakistan. These courts were either mentioned in the beginning of the judgment, to signify the court the judgment is from, alongside mention of other cases as either the court that passed the judgment that is being appealed at the Supreme Court or as the court that passed the judgment that the Supreme Court used to justify their ruling.

Other named entities include law references mentioned in the reference text, FIR numbers and respondents for a case.

For this task we referred to [53] where judgments from Lahore High Court were used and added extra entities which are important for not only Supreme Court but also for High Courts.

After studying different judgments from the Civil Appeal category of the Supreme Court of Pakistan, we decided on a total of 14 different labels for Named Entities as given below.

1. Person
2. Organization
3. Location
4. Date
5. Case Number
6. Respondent
7. Money
8. FIR Number
9. Reference
10. Reference Case
11. Reference Court

12. Appeal Case number
13. Appeal Court
14. Approved

### 3.1.3 Annotation Guidelines

We decided on specific guidelines for the annotation of judgments using the named entities given in 3.1.2. The labels for the named entities and their annotation guidelines are as below:

1. **Per** –These are the names of people mentioned in any judgment, with an exception of respondents of the case). This includes the names of judges, lawyers, people involved in the case etc.

2. **Org** –These are the names of any organizations that might be mentioned in a judgment. These might be involved in the case that is currently being decided or organizations from cases that are being referenced in the judgment.

3. **Loc** –These are any locations/addresses mentioned in a judgment. These include cities, districts, provinces, etc.

4. **Date** –These are any dates mentioned in a judgment. These only include complete dates (day-month-year) mentioned in whatever format. Only months or years will not be labelled as a Date entity.

5. **CaseNo** –These are the case numbers of the case being decided. In the Supreme Court of Pakistan often more than one cases are being decided in the same judgment. All those case numbers are to be labelled as CaseNo.

6. **Resp** –These are the respondents of the case being decided. These can include people or organizations and a single case can have multiple respondents.

7. **Money** –These are mentions of any amount of money in the judgment. These are usually money a party owes the other or the amount a party has to pay, etc.

8. **FIRNo** –These are The FIR numbers mentioned in the judgment. This can include the FIR that might have been filed against any of the parties involved in the case that might be mentioned because it is important for the decision on said case.

9. **Ref** –These are law references mentioned in a judgment. These might include Pakistan's law, Shariah Law, etc. Law references are included in a judgment in the paragraphs where the lawyers are using them for their defense and also where the judges mention them to justify their decision.

10. **RefCase** –These are cases that are mentioned in a judgment to help decide a judgment. These can include older cases from Pakistan or any other country's courts that might involve a similar situation. These cases can also be referred to by the lawyers or the judges.

11. **RefCourt** –These are the courts that decided the cases that were referred to in the judgment; the courts that decided the cases being labelled as 'RefCase'.

12. **AppealCaseNo** –These are the cases that are being appealed against at the Supreme Court. The cases in the Supreme Court of Pakistan are often for appeals against judgments passed by any of the High Courts so an AppealCaseNo is essentially the case number of the case a High Court decided and that decision is now being challenged against in the Supreme Court.

13. **AppealCourt** –These are the courts that decided the case that is now being appealed against in the Supreme court; the courts that decided the cases being labelled as AppealCaseNo.

14. **Approved** –This will show whether or not the judgment has been approved for reporting. A judgment of the Supreme Court of Pakistan can be either approved or not approved for reporting, as mentioned at the end of a judgment. This label will only be present in the judgments being approved for reporting.

### 3.1.4 Data Annotation

For the annotation process, the selected judgments were converted to text files and annotated using the open source annotation tool Doccano [41] as shown in Figure 3.1. The judgments were annotated using the annotation guidelines given in Section 3.1.3.

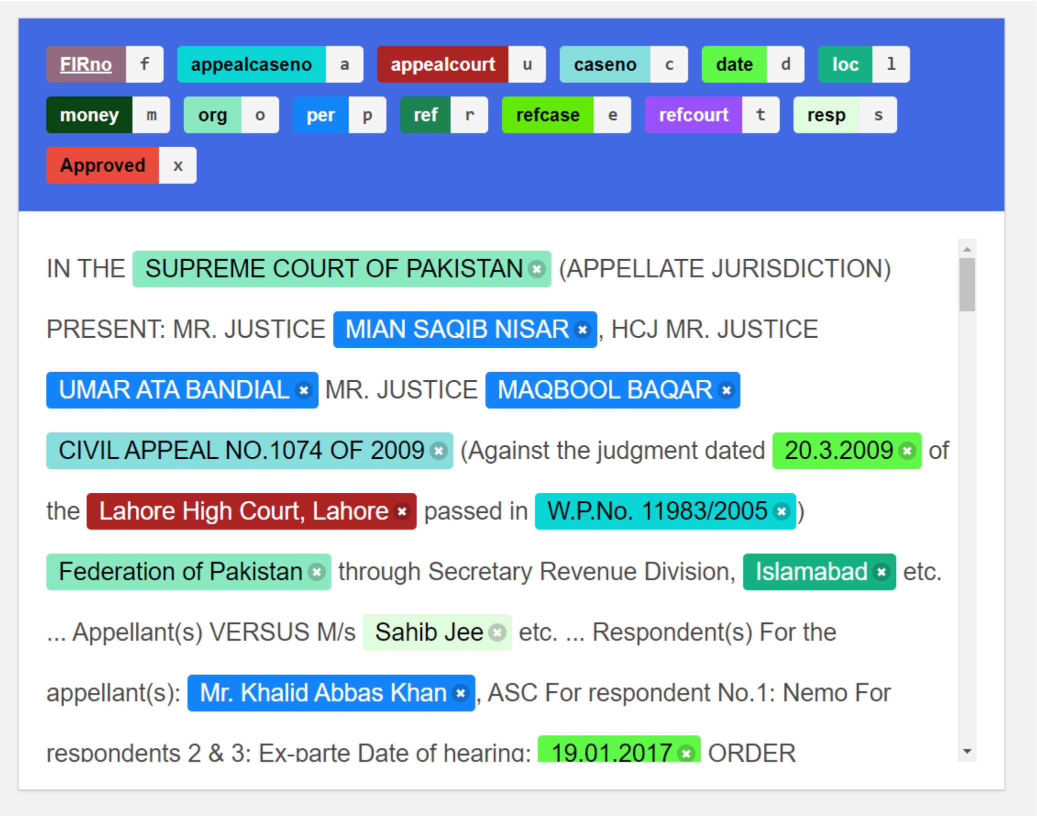


Figure 3.1: Labelling of Named Entities in judgments in Doccano

A dataset was then created from these annotated documents that followed the IOB (Inside, Outside and Beginning) format. The named entity chunks were labeled with the Beginning and Inside tags.

## 3.2 Named Entity Recognition

For the purposes of Named Entity Recognition, we use BERT, a state-of-the-art model that has achieved exceptional results in different NLP tasks.

### 3.2.1 BERT

BERT (Bidirectional Encoder Representations from Transformers) was published by Devlin et al. [13], a team of researchers at Google AI Language. The key innovation of BERT is its bi-directionality as this characteristic enables it to understand the context of words depending on its surrounding words.

BERT uses Transformers which are an attention mechanism that have the ability to learn contextual relations between words in a piece of text. In it most basic form, a transformer combines two separate mechanisms; an encoder that can read the input text, and a decoder that gives a prediction for the given task. As opposed to traditional models, transformers read the whole sequence of words concurrently and thus is considered bi-directional.BERT resolves the constraint of uni-directionality with the use of Masked Language Model (MLM) pre-training objective.

#### Masked Language Model

A Masked Language model randomly selects and masks some of the tokens from the input text. The model then predicts the original vocabulary id of the word that had been masked based purely on its context.

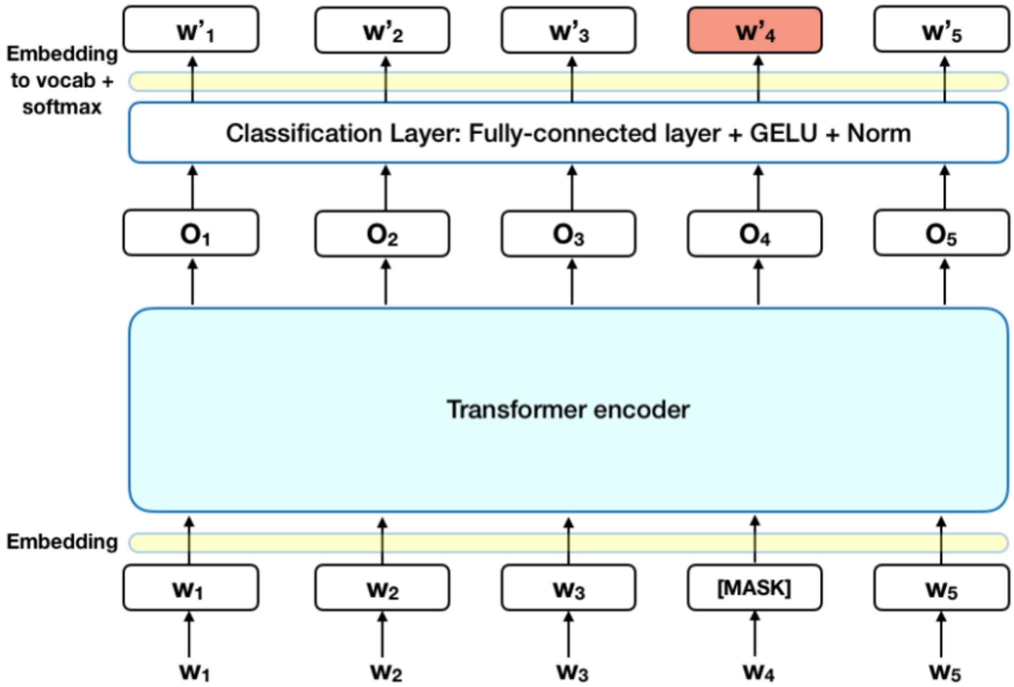


Figure 3.2: Prediction of masked words in a Masked Language Model

Unlike the traditional left-to-right pre-training for language, the masked language model facilitates the fusion of right and left context by the representation. This in turn enables the pre-training of a deep bi-directional Transformer.

For this part of the pre-training process, 15% of words in every sequence are changed to [MASK] token and then the word sequences are fed to the model. The model then tries to predict the word that was originally there based on the non-masked words present in the sequence, as shown in Fig3.2. The loss function BERT uses only focuses on the prediction of masked words and ignores the non-masked word predictions. This causes the model to converge much slower than some directional models.

### Next Sentence Prediction

The second step in the pre-training process of the BERT model is the Next Sentence Prediction (NSP). In this stage, the model is given pairs of sentences as inputs. The model then learns to predict whether the second sentence in the given pair is the subsequent one in the original document. For the training data, 50% of the sentence pairs contain a random sentence from the document as the second sentence.

The remaining pairs contain subsequent sentences as the second sentence in the pairs. The assumption for this task is that the randomly selected sentences will have a disconnect from the first sentence.

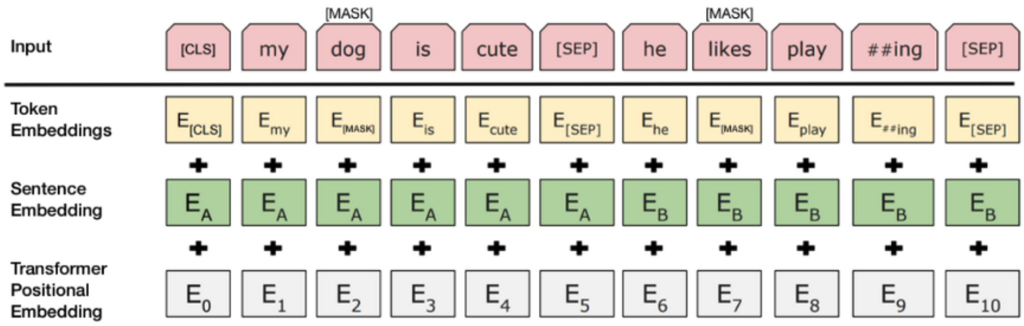


Figure 3.3: BERT input representation

When training the BERT model, Masked Language Model and Next Sentence Prediction are trained together. The goal of this training is the minimization of the combined loss function of the two strategies.

BERT is available pre-trained on Wikipedia and Book Corpus. The dataset contains more than 10,000 books of different genres, a total of 3.3 Billion words 2.5B from Wikipedia and 0.8B from BookCorpus [13].

### 3.2.2 NER using BERT

In order to use BERT for a specific NLP task, the pre-trained model can be further trained and fine-tuned on domain specific data.

For the purposes of legal NER we used BERT-base. BERT-base has a total of 12 layers/transformer blocks, Hidden size of 768 and 12 self attention heads (L=12, H=768, A=12).

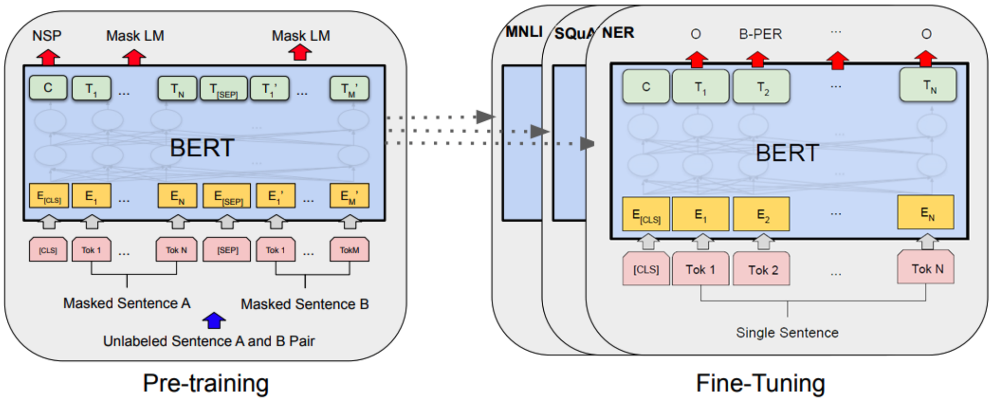


Figure 3.4: Finetuning Pre-trained BERT for NER

For finetuning purposes we used the labelled dataset created using the Civil Appeal judgments from the Supreme Court of Pakistan. This dataset included generic named entities, i.e., Person, Organization, Location, etc, as well as domain specific named entities, i.e., Law references, Case references, Case number, etc.

### Summary

This chapter included the detailed description of the dataset creation process as well as the model we use for the purposes of NER from said dataset. The model was then tested on different datasets. The details of the experiments and their results are mentioned in the next chapter.

# Chapter 4

# Experiments, Results & Analysis

This chapter describes the details of the performance evaluation, experiments that have been carried out and their results. The model was tested on three different dataset, the descriptions of which are mentioned in the Section 4.2. These is a detailed analysis based on the results as guidance for the future work.

## 4.1 Performance Evaluation

In order to evaluate the performance of a system, different performance measures are used. For NER systems, we include precision, recall, and F1 score for evaluation. The NER approach we used is also compared to the results of [53].

### 4.1.1 Evaluation Metrics

In order to evaluate the performance of our model for Named Entity recognition on legal judgments, we take into consideration the precision, recall and F1 score. The comparison with previously published results is using F1 scores for the individual entities.

#### Precision

Precision is the measure of the portion of positive identifications that were correct. The more higher percentage of named entities correctly recognised by the model, higher the precision.

$$precision = \frac{TP}{TP+FN} \tag{4.1}$$

#### Recall

Recall as a metric quantifies the number of correct positive predictions made out of all positive predictions that could have been made. For named entities, this means the higher number of Named Entities of a given label the model recognizes correctly as compared to the total number of named entities of that label, higher the recall.

$$recall = \frac{TP}{TP+FN} \tag{4.2}$$

#### F1 Score

F1 score conveys the balance between precision and recall. F1 score is the harmonic means of the precision and recall and is therefore widely used as higher value of precision along with higher recall is often needed for information retrieval as well as NER systems.

$$F = 2. \frac{precision \, . \, recall}{precision+recall} \tag{4.3}$$

## 4.2 Experiments

In order to evaluate the performance of BERT for named entity recognition for generic as well as legal named entities, we have run experiments on 3 different datasets which include CONLL2003 dataset, a dataset of Civil judgments from Lahore High Court, Pakistan that was used by [53], and the dataset of Civil Appeal judgments from the Supreme Court of Pakistan. The information of the datasets along with the results are in the following sections.

### 4.2.1 CoNLL-2003 Dataset

CoNLL-2003 is one of the most known datasets for named entity recognition. the dataset was released as a part of the shared CoNLL-2003 task of language independent named entity recognition. The dataset covers two languages, English and German. Both languages have their separate training and testing files along with development files, as well as a large files consisting of unannotated data.

For our purposes, we used the training and testing files of the English language dataset. The English data for CoNLL-2003 dataset was taken from the Reuters Corpus which contains news stories from August 1996 to August 1997. The entities labelled in the dataset are Organizations, locations, person names and miscellaneous. The data is labelled in the IBO (Inside-Beginning-Outside) tagging. The number tags in the dataset are given in Table 4.1

| Labels | Count | Labels | Count |
|---|---|---|---|
| B-PER | 10059 | I-PER | 6991 |
| B-LOC | 10645 | I-LOC | 1671 |
| B-ORG | 9323 | I-ORG | 5290 |
| B-MISC | 5062 | I-MISC | 1717 |

Table 4.1: Labels present in CoNLL2003 dataset and their count.

### 4.2.2 LHC Dataset

This is the dataset that was used by [53]. The dataset consists of 100 Civil proceeding judgments from Lahore High Court, Pakistan. The exact case type or category is not mentioned and the cases were chosen at random. The dataset contains a total of 10 named entities and is labelled using the IOB format as well. The names of the labels and their counts are given in Table 4.2.

| Labels | Count | Labels | Count |
|---|---|---|---|
| B-per | 1081 | I-per | 1602 |
| B-loc | 255 | I-loc | 217 |
| B-org | 289 | I-org | 918 |
| B-caseNo. | 147 | I-caseNo. | 485 |
| B-Misc.name | 297 | I-Misc.name | 573 |
| B-date | 879 | I-date | 66 |
| B-refCourt | 475 | I-refCourt | 576 |
| B-ref | 422 | I-ref | 2405 |
| B-refCase | 243 | I-refCase | 605 |
| B-money | 109 | I-money | 63 |

Table 4.2: Labels present in the LHC dataset and their count.

### 4.2.3 Supreme Court Dataset

This is the dataset we created using the Civil Appeal judgments from the Supreme Court of Pakistan. The datasets consists of 214 Civil Appeal judgments and has a total of 14 named entities. Detailed description regarding the creation of this dataset is mentioned in Sections 3.1.1 and 3.1.4. The entities where labeled using the guidelines mentioned in Section 3.1.3 in the IBO format. The names and counts of these labels are mentioned in the Table 4.3.

| Labels | Count | Labels | Count |
|---|---|---|---|
| B-per | 4961 | I-per | 9749 |
| B-loc | 1703 | I-loc | 1163 |
| B-org | 3050 | I-org | 4934 |
| B-caseno | 1497 | I-caseno | 5151 |
| B-resp | 487 | I-resp | 3048 |
| B-date | 3850 | I-date | 1854 |
| B-refCourt | 306 | I-refCourt | 888 |
| B-refCase | 2301 | I-refCase | 32576 |
| B-ref | 4775 | I-ref | 32099 |
| B-appealcourt | 422 | I-appealcourt | 1778 |
| B-appealcaseno | 770 | I-appealcaseno | 3990 |
| B-money | 446 | I-money | 208 |
| B-FIRno | 23 | I-FIRno | 52 |
| B-Approved | 160 | I-Approved | 1 |

Table 4.3: Labels present in the Supreme Court dataset and their count.

## 4.3 Results and Analysis

We trained BERT on the NER dataset for the datasets mentioned in Section 4.2. Here we present the results and analysis of said results for said datasets.

### 4.3.1 CoNLL-2003 Results

For CoNLL-2003 dataset, we achieved an F1 score of **96.37** which is similar to what was published by [13].

The highest F1 score was 98.42 for the label I-PER and the lowest F1 score was 83.15 for the label I-MISC. The variation in the individual F1 scores of all the named entity labels in the CoNLL-2003 English dataset can be seen in the graphs in Figures 4.1 and 4.2.

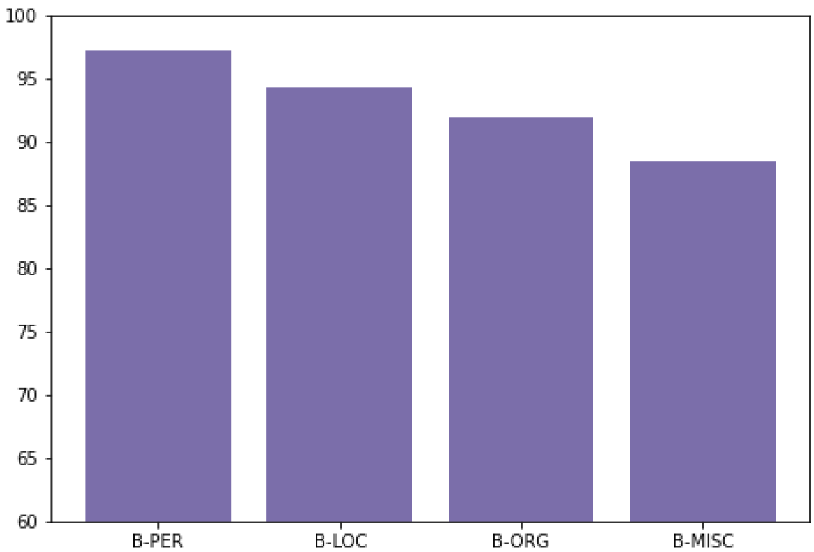


Figure 4.1: F1 Scores of BERT for the 'B' labels of CONLL-2003 Dataset

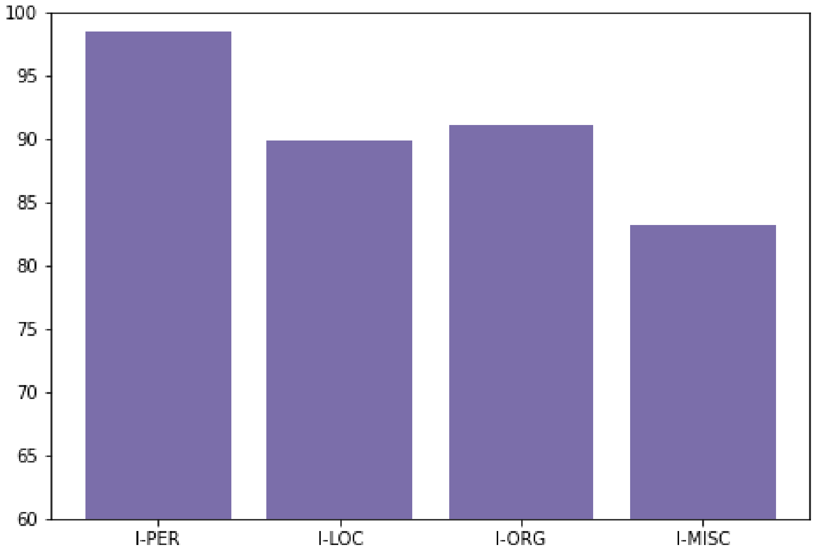


Figure 4.2: F1 Scores of BERT for the ‘I’ labels of CONLL-2003 Dataset

| Labels | F1-score |
|---|---|
| B-PER | 97.14 |
| I-PER | 98.42 |
| B-LOC | 94.23 |
| I-LOC | 89.82 |
| B-ORG | 91.90 |
| I-ORG | 91.12 |
| B-MISC | 88.46 |
| I-MISC | 83.15 |

Table 4.4: F1 scores for individual labels of the CoNLL2003 dataset.

The individual F1 scores for all the labels in the CoNLL-2003 English dataset are mentioned in the Table 4.4. The confusion matrix for the results is shown in Figure 4.3.

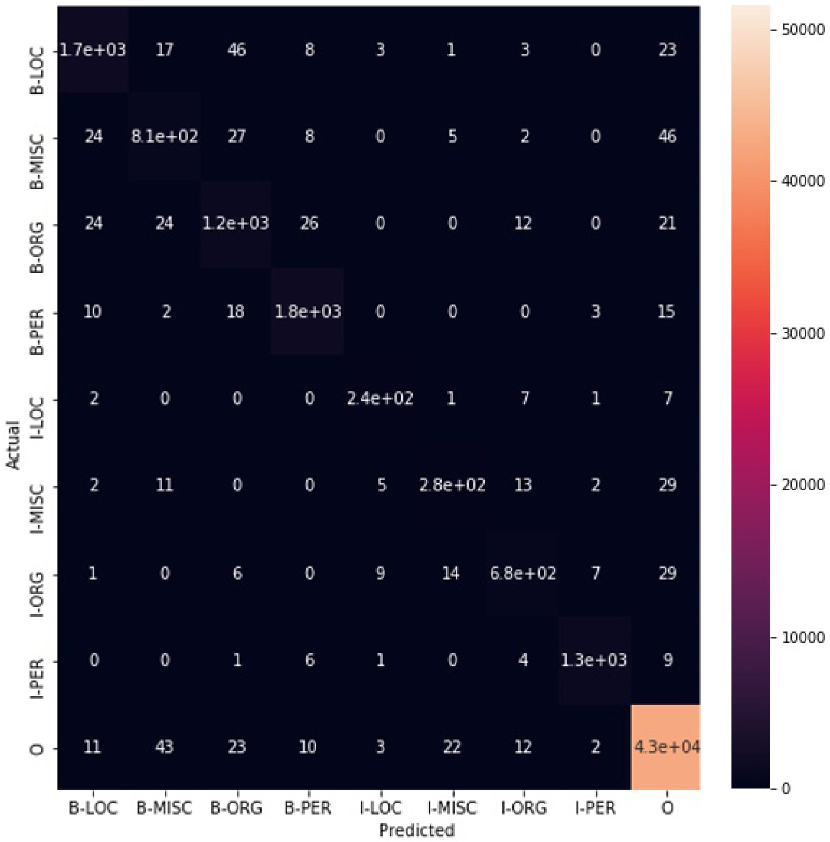


Figure 4.3: Confusion Matrix for BERT results for CoNLL-2003

### 4.3.2 LHC Results

For the Lahore High court dataset used by [53], with the highest published F1 score was **86.62** using a CRF model for Named Entity Recognition.

The highest F1 score achieved by the CRF model for this dataset is 98.72 for the label of B-refCase. The lowest F1 score achieved was 55.10 for the label I-money and the second lowest was 65.79 for the label I-loc.

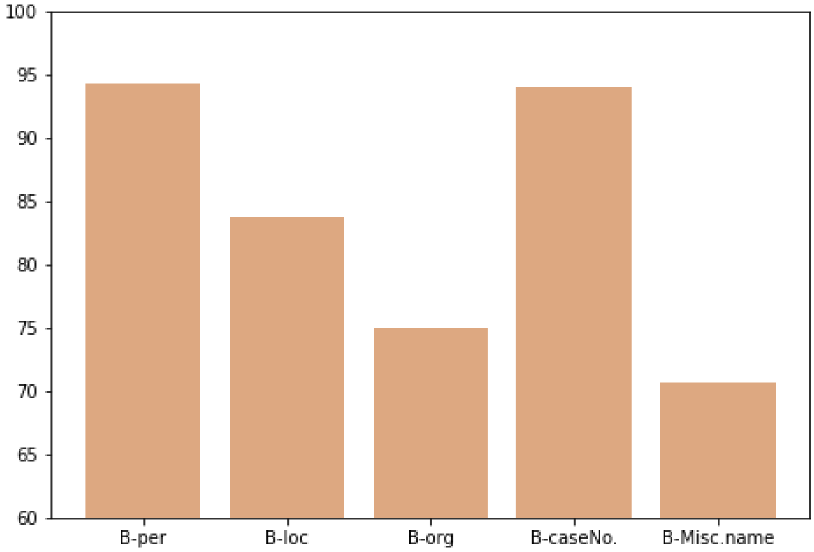


Figure 4.4: F1 Scores of CRF model for the first 5 ‘B’ labels of Lahore High Court Dataset

From the graphs in Figures 4.4, 4.5, 4.6 and 4.7, we can see how the F1 scores of the CRF model vary for different labels. We can see that the F1 scores for the labels of B-Misc.name, I-Misc.name, B-org, I-loc, B-ref and I-money are relatively lower than the F1 scores for other labels. The tabe

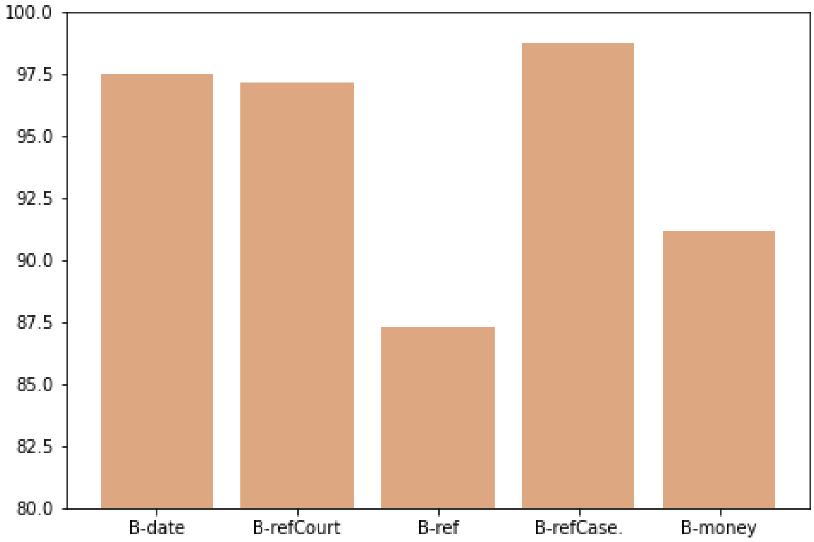


Figure 4.5: F1 Scores of CRF model for the last 5 ‘B’ labels of Lahore High Court Dataset

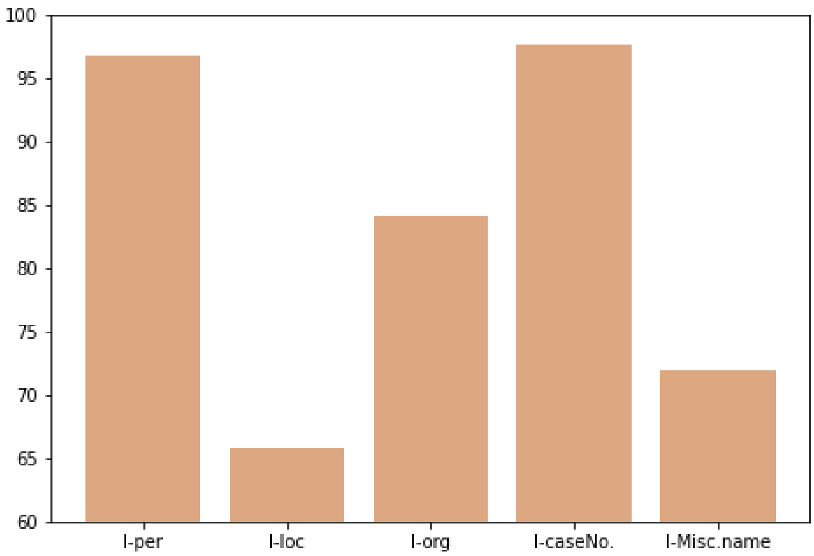


Figure 4.6: F1 Scores of CRF model for the first 5 ‘I’ labels of Lahore High Court Dataset

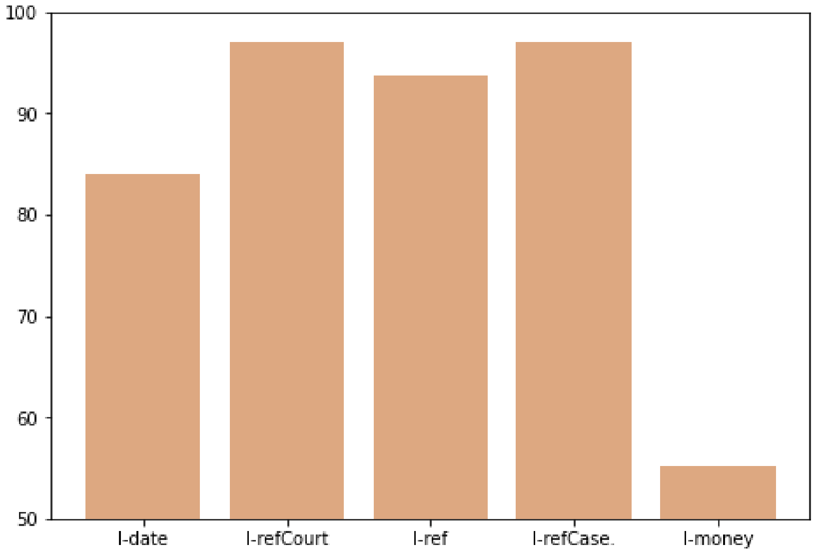


Figure 4.7: F1 Scores of CRF model for the last 5 ‘I’ labels of Lahore High Court Dataset

With BERT, we managed to achieve an F1 score of **93.21** which is considerably higher than the preciously published results.

Using BERT, we achieved the highest F1 score of 100 for the labels B-date, I-date, B-refCase, I-refCase, B-money, and I-money. The lowest F1 score observed was 70.43 for the label B-org. From a total of 20 labels, we achieved an F1 score greater than 90 for 15 of the labels.

We can see the variations in the F1 scores of different labels in the Figures 4.8, 4.9, 4.10 and 4.11 and the individual F1 scores are given in the Table 4.6.

| Labels | CRF model F1-score |
|---|---|
| B-per | 94.28 |
| I-per | 96.72 |
| B-loc | 83.68 |
| I-loc | 65.79 |
| B-org | 75.04 |
| I-org | 84.11 |
| B-caseNo. | 94.06 |
| I-caseNo. | 97.57 |
| B-Misc.name | 70.64 |
| I-Misc.name | 71.91 |
| B-date | 97.49 |
| I-date | 84.0 |
| B-refCourt | 97.16 |
| I-refCourt | 97.08 |
| B-ref | 87.26 |
| I-ref | 93.64 |
| B-refCase | 98.72 |
| I-refCase | 96.93 |
| B-money | 91.15 |
| I-money | 55.10 |

Table 4.5: F1 scores for individual labels of the LHC dataset using CRF model.

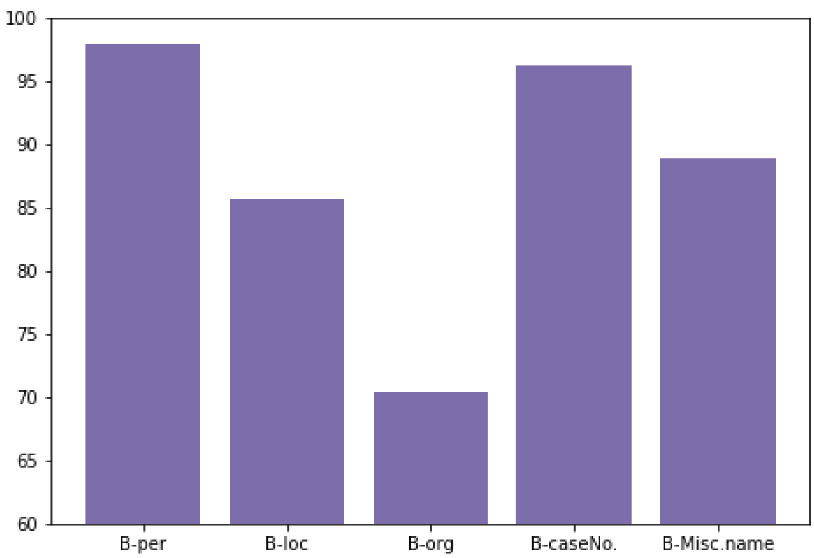


Figure 4.8: F1 Scores of BERT for the first 5 ‘B’ labels of Lahore High Court Dataset

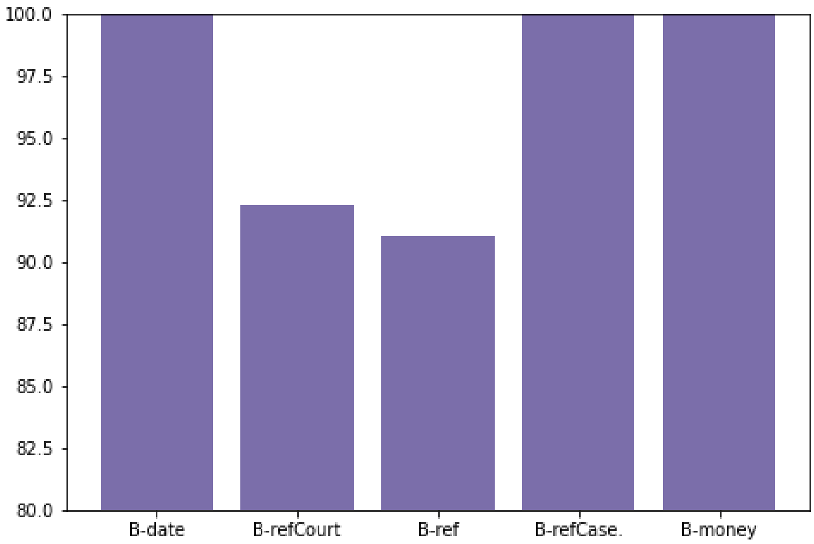


Figure 4.9: F1 Scores of BERT for the last 5 'B' labels of Lahore High Court Dataset

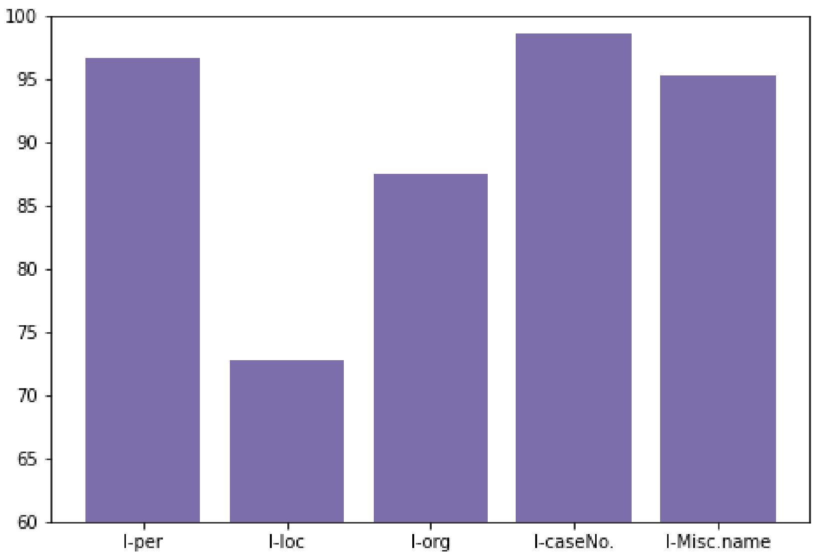


Figure 4.10: F1 Scores of BERT for the first 5 'I' labels of Lahore High Court Dataset

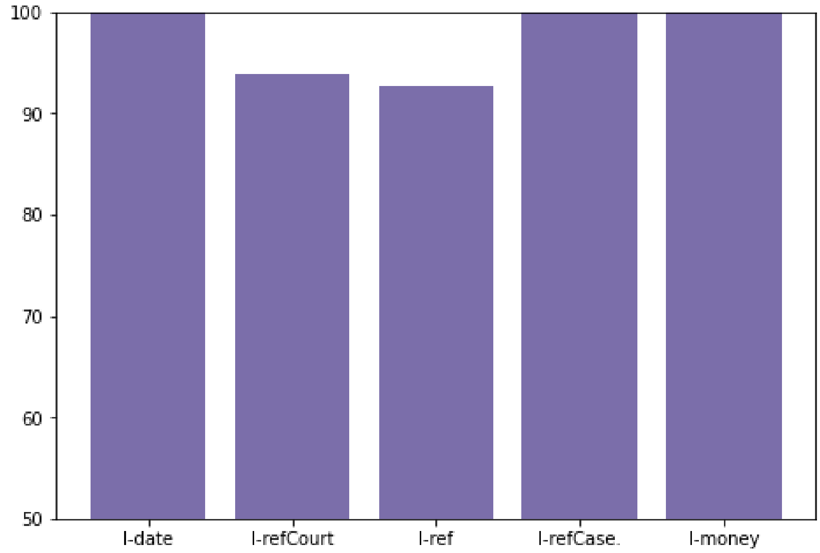


Figure 4.11: F1 Scores of BERT for the last 5 'I' labels of Lahore High Court Dataset

The comparisons of the F1 scores for all the individual entities in the dataset are as mentioned in Table 4.7

From these results, we can see that BERT is better at recognizing most of the named entities than CRf models. As BERT is pre-trained on a large corpus and is able to learn information from both left and right side of a given token, it gives better results when it comes to identifying named entities, especially when it comes to judgments where same words can often have different labels depending on their context, e.g., a person name can be labeled as an person (per) or a respondent (Misc.name) depending on the context in which the name of the persons is mentioned.

The confusion matrix for BERT for the results on the LHC dataset is given in Figure 4.12. The Figures 4.13, 4.14, 4.15,4.16 and 4.17 contain graphs comparing the F1 scores of BERT and CRF models on the LHC dataset.

| Labels | BERT F1-score |
|---|---|
| B-per | 97.94 |
| I-per | 96.58 |
| B-loc | 85.62 |
| I-loc | 72.73 |
| B-org | 70.43 |
| I-org | 87.50 |
| B-caseNo. | 96.29 |
| I-caseNo. | 98.63 |
| B-Misc.name | 88.89 |
| I-Misc.name | 95.24 |
| B-date | 100.0 |
| I-date | 100.0 |
| B-refCourt | 92.31 |
| I-refCourt | 93.94 |
| B-ref | 91.02 |
| I-ref | 92.59 |
| B-refCase | 100.0 |
| I-refCase | 100.0 |
| B-money | 100.0 |
| I-money | 100.0 |

Table 4.6: F1 scores for individual labels of the LHC dataset using BERT.

| Labels | CRF model F1-score | BERT F1-score |
|---|---|---|
| B-per | 94.28 | 97.94 |
| I-per | 96.72 | 96.58 |
| B-loc | 83.68 | 85.62 |
| I-loc | 65.79 | 72.73 |
| B-org | 75.04 | 70.43 |
| I-org | 84.11 | 87.50 |
| B-caseNo. | 94.06 | 96.29 |
| I-caseNo. | 97.57 | 98.63 |
| B-Misc.name | 70.64 | 88.89 |
| I-Misc.name | 71.91 | 95.24 |
| B-date | 97.49 | 100.0 |
| I-date | 84.0 | 100.0 |
| B-refCourt | 97.16 | 92.31 |
| I-refCourt | 97.08 | 93.94 |
| B-ref | 87.26 | 91.02 |
| I-ref | 93.64 | 92.59 |
| B-refCase | 98.72 | 100.0 |
| I-refCase | 96.93 | 100.0 |
| B-money | 91.15 | 100.0 |
| I-money | 55.10 | 100.0 |

Table 4.7: Comparison of F1 scores for individual labels of the LHC dataset.

We can see that BERT has a lower F1 score for the refCourt labels for the Lahore High Court dataset. From the confusion matrix in Figure 4.12, we can see that most of the confusion is because of the model mistaking them for organizations. One of the main reasons for this is that the Courts are labelled as organizations when a case from them is being appealed against and as refCourts when a case is being referenced.

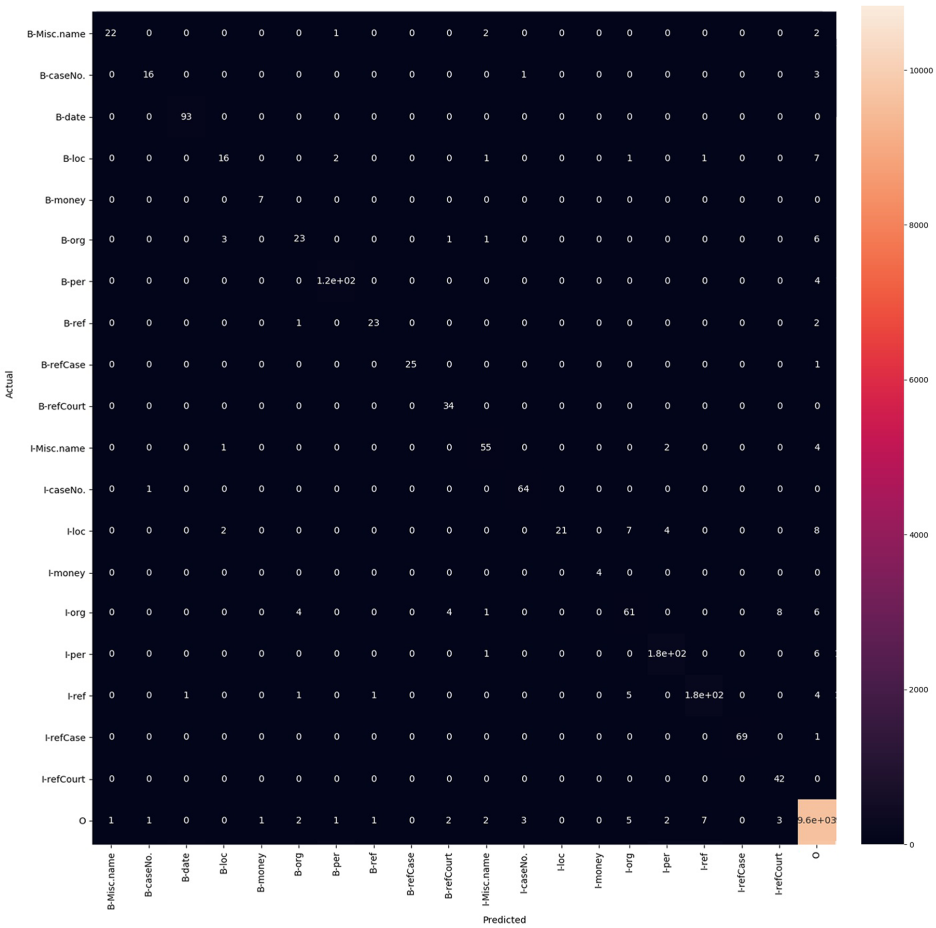


Figure 4.12: Confusion Matrix for BERT results for LHC Dataset

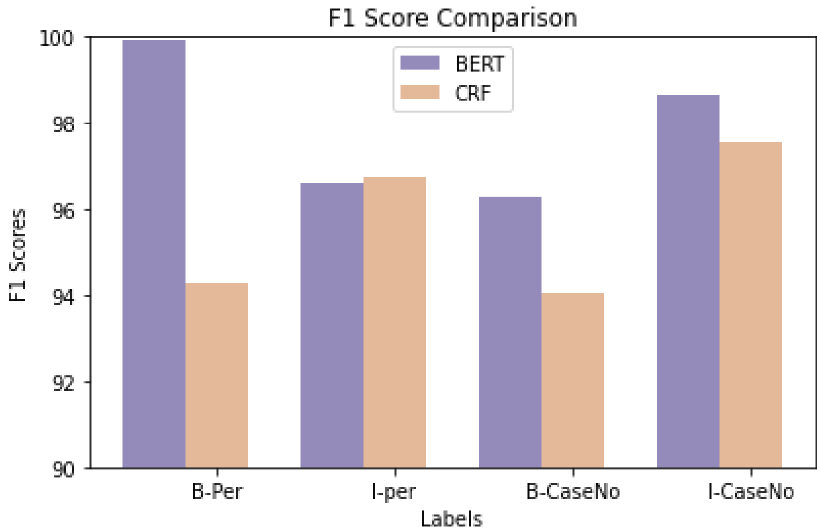


Figure 4.13: Comparing results of the performance of BERT and CRF models for ‘per’ and ‘CaseNo’ labels.

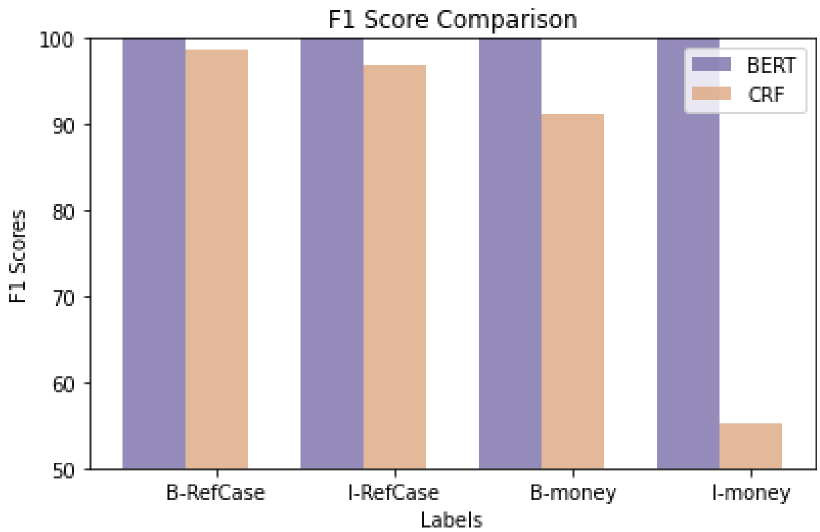


Figure 4.14: Comparing results of the performance of BERT and CRF models for ‘RefCase’ and ‘money’ labels.

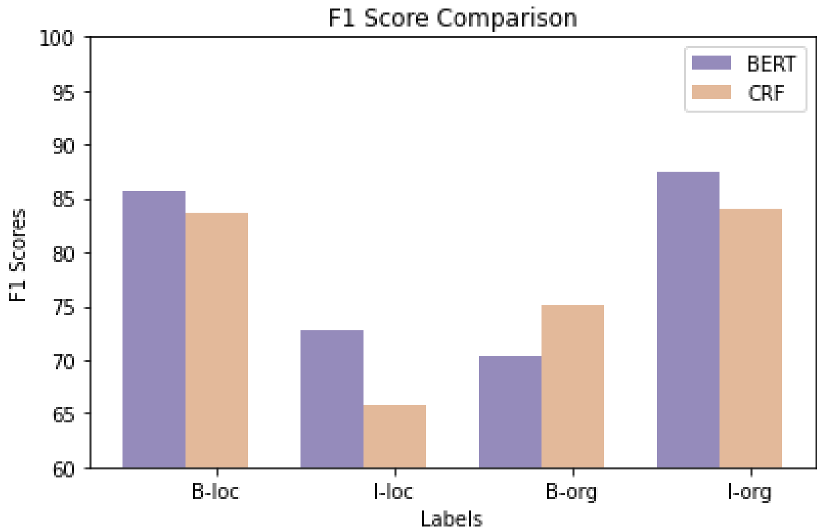


Figure 4.15: Comparing results of the performance of BERT and CRF models for 'loc' and 'org' labels.

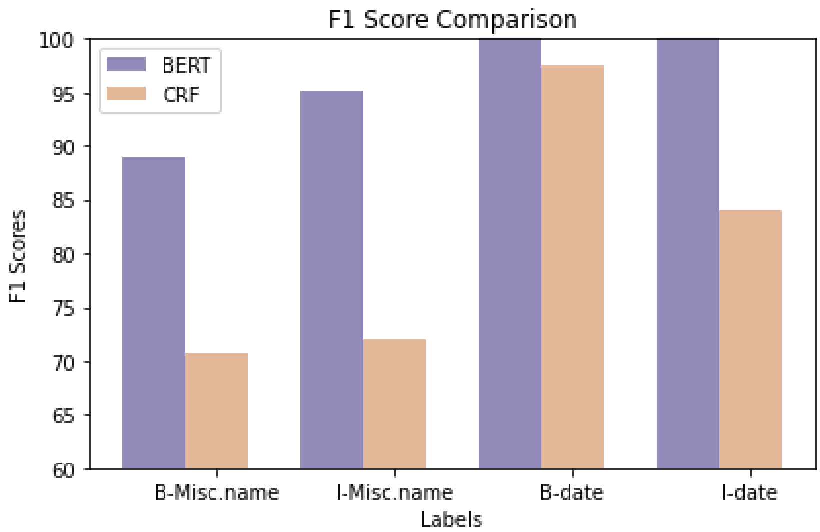


Figure 4.16: Comparing results of the performance of BERT and CRF models for 'Misc.name' and 'date' labels.

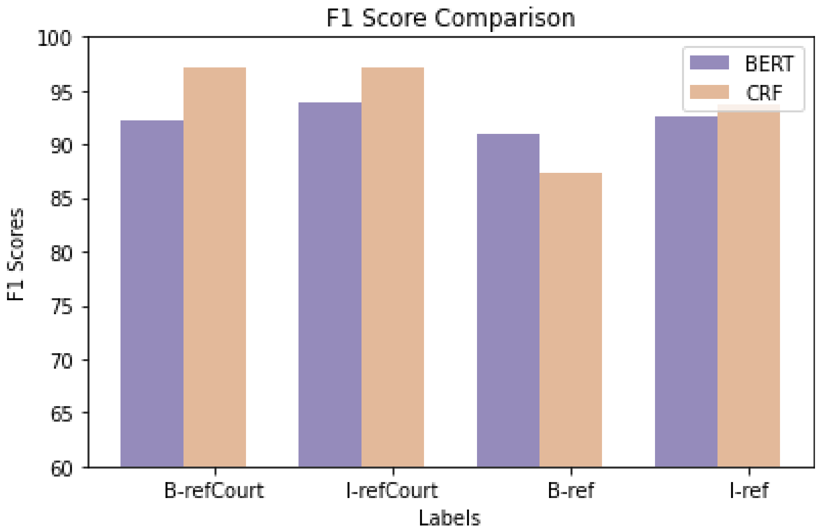


Figure 4.17: Comparing results of the performance of BERT and CRF models for ‘refCourt’ and ‘ref’ labels.

Other than these, we can see minor confusions between person names and locations. This is to be expected to some extent as some places are named after public figures.

Regardless of these, we can see an overall improvement in the average F1 score for the named entity recognition in this dataset with BERT. Especially for the labels of I-money, I-loc, B-Misc.name, and I-Misc.name.

### 4.3.3 Supreme Court Results

With BERT, we achieved an overall F1 score of **92.72** and an average F1 score of 91.08 for the Named Entities for the dataset consisting of judgments from the Supreme Court of Pakistan.

The highest F1 scores achieved for this dataset were 100 for the entities 'B-FIRno.' and 'I-FIRno.'. The lowest F1 score achieved was 87.54 for the entity 'I-refCourt'. Only two labels have the F1 score below 90 which are 'B-refCourt' and 'I-refCourt'.

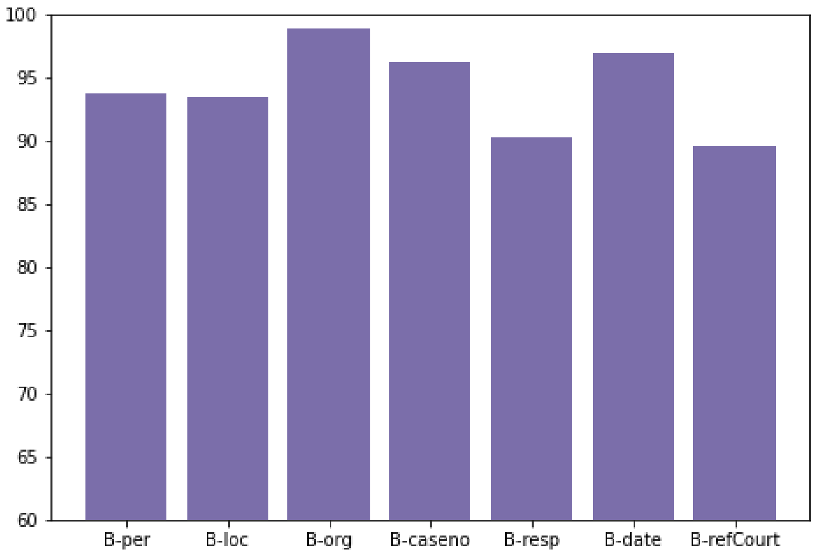


Figure 4.18: F1 Scores of BERT for the first 7 'B' labels ofthe Supreme Court Dataset.

From the Figures 4.18 and 4.19 we can see that almost all of the beginning ('B') labels have an F1 score of over 90 with the only exception being the 'B-refCourt' label.

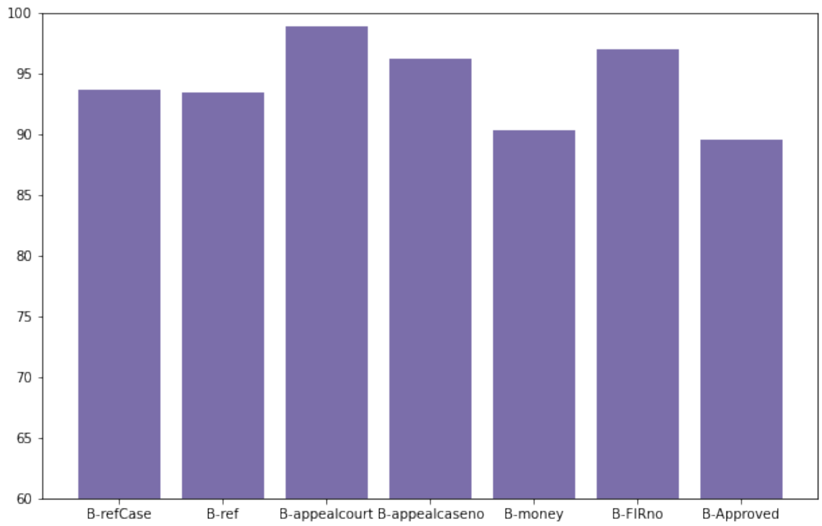


Figure 4.19: F1 Scores of BERT for the last 7 'B' labels ofthe Supreme Court Dataset.

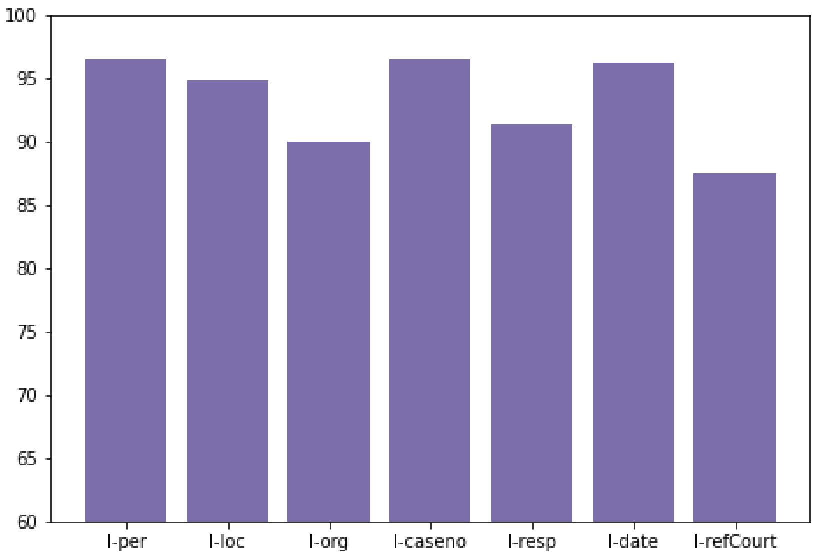


Figure 4.20: F1 Scores of BERT for the first 7 'I' labels ofthe Supreme Court Dataset.

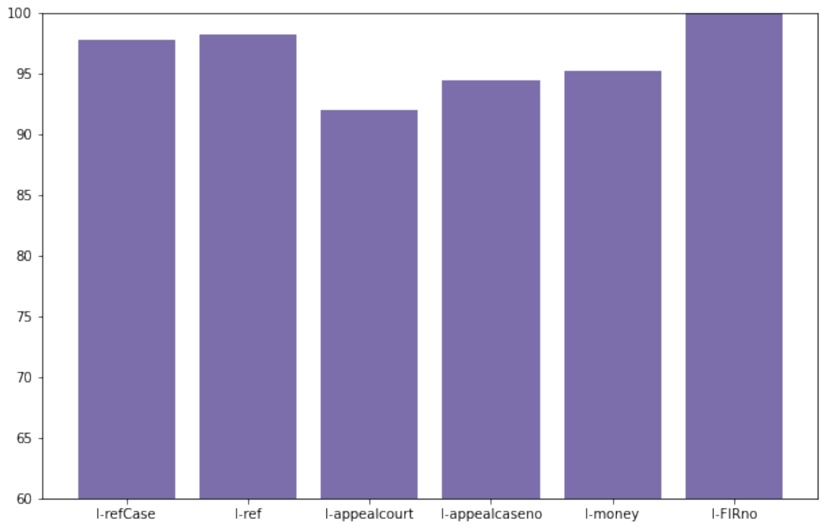


Figure 4.21: F1 Scores of BERT for the last 6 'I' labels ofthe Supreme Court Dataset.

From the Figures 4.20 and 4.21 we can see that almost all of the inside ('I') labels have an F1 score of over 90 with the only exception being the 'I-refCourt' label.

The individual F1 scores of all the named entity labels in the dataset are mentioned in the Table 4.8.

| Labels | F1-score | Labels | F1-score |
|---|---|---|---|
| B-per | 93.67 | I-per | 96.49 |
| B-loc | 93.46 | I-loc | 94.90 |
| B-org | 98.88 | I-org | 90.02 |
| B-caseno | 96.25 | I-caseno | 96.55 |
| B-resp | 90.27 | I-resp | 91.30 |
| B-date | 96.96 | I-date | 96.22 |
| B-refCourt | 89.5 | I-refCourt | 87.54 |
| B-refCase | 98.82 | I-refCase | 97.75 |
| B-ref | 92.17 | I-ref | 98.21 |
| B-appealcourt | 90.77 | I-appealcourt | 91.97 |
| B-appealcaseno | 90.32 | I-appealcaseno | 94.37 |
| B-money | 96.77 | I-money | 95.20 |
| B-FIRno | 100 | I-FIRno | 100 |
| B-Approved | 92.00 | I-Approved | - |

Table 4.8: F1 scores for individual labels of the Supreme Court dataset.

From the Table 4.8, we can see that BERT performs well on the Supreme Court Dataset that we created as well. The confusion matrix for the experiment on Supreme Court dataset is given in Figure 4.22.

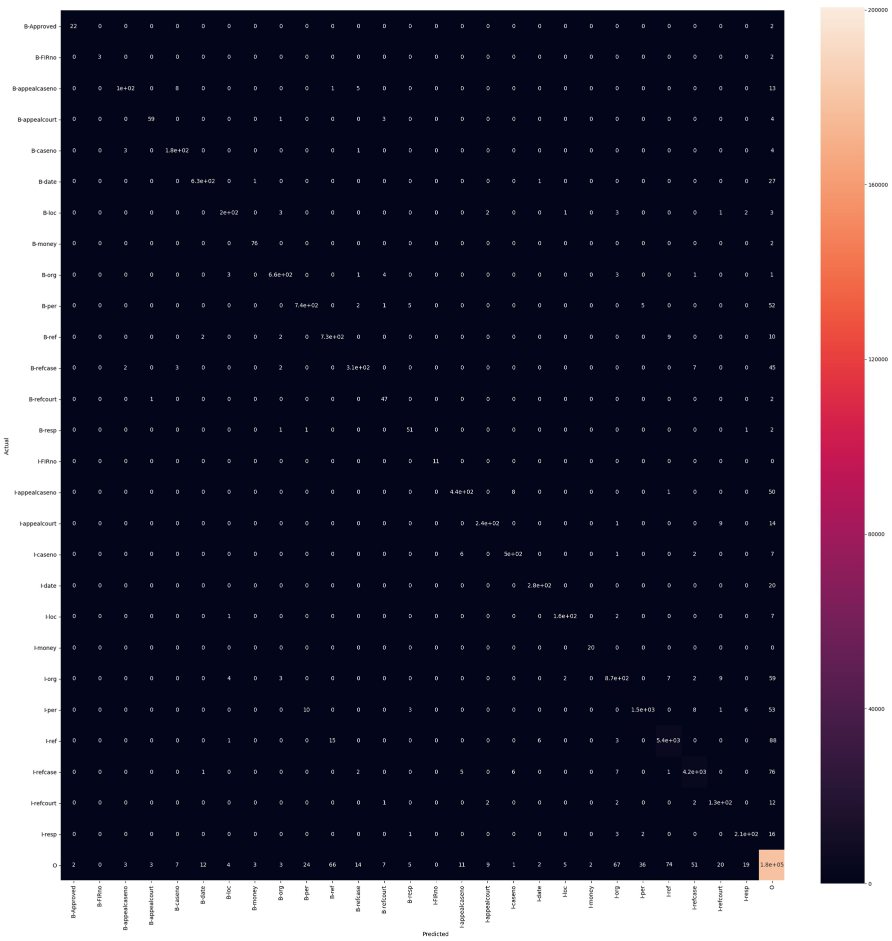

Figure 4.22: Confusion Matrix for BERT results for the Supreme Court Dataset

### 4.3.4 Analysis

Using the pre-trained BERT, we managed to produce an F1 score of 96.37 for the CoNLL-2003 dataset which is similar to the one that was published by Devlin et al. [13] for the said dataset.

For the Lahore High Court dataset, BERT achieved an F1 score of 93.21

which is a considerable improvement from the results published by Sharafat et al. [53] using the CRF model. From Table 4.7 we can see that the label 'B-org' has an F1 score of 70.43 which is lower than the previously published F1 score of 75.04. Looking at the confusion matrix in Figure 4.12 we can see the cause of this is that the model is predicting the entities as 'B-loc' or as 'O' (Outside label). These are words like 'the' mentioned in the judgments that are being labelled as 'B-org' in the dataset even when the word 'the' is not in the official name of the organization. Another label that has a relatively low F1 score even though it is an improvement in the previously published one is the label of 'I-loc' with the F1 score of 72.73. From the confusion matrix in Figure 4.12 we can see that this is because these words are being predicted as 'I-org', 'I-per' and 'O'. Part of this confusion is because of locations being named after people and organization named having names of places in their names, e.g. 'Lahore High Court', 'Punjab University' etc.

When it comes to the Supreme Court of Pakistan dataset, BERT achieved an F1 score of 92.72. From the Table 4.8 we can see that the label 'I-approved' does not have any results. This is because the label only exists due to the data being in the IBO (Inside-Beginning-Outside) format and the actual entity being labelled in the judgments is a single word so there are no words being labelled as I-approved when the labelled corpus is being converted to the IBO format. From the Table 4.8, we can also see that the labels 'B-refcourt' and 'I-refcourt' have slightly lower F1 scores. From the confusion matrix in Figure 4.22 we can see that this is because of some words being predicted ad 'B-appealcourt' and 'I-appealcourt' as most of the word in these labels are the same, i.e., the names of different courts of Pakistan.

Overall, for a model pre-trained on non-legal dataset, BERT has performed well after being finetuned on judicial data for NER.

**Summary**

This chapter included the details of the three different datasets; CoNLL-2003, Dataset consisting of Civil judgments from Lahore High Court and the dataset consisting of Civil Appeal judgments from the Supreme Court of Pakistan that we created. We achieved an F1 score of 96.37 for CoNLL-2003 dataset, 93.21 for the LHC dataset, which was a considerable improvement from the previously published f1 score of 86.62, and an F1 score of 92.72 for the Supreme Court of Pakistan dataset.

# Chapter 5

# Conclusions and Future Work

The courts of Pakistan are producing huge amounts of textual data in the form proceedings/judgments that is then being made publicly available for the sake of awareness and guidance. As the size of this available data grows, it becomes more and more difficult for it to be manually processed by a human being. There is an opportunity for the automated processing of this data by creating systems using machine learning and deep learning models. For the development of information extraction systems, labelled data is needed for training the algorithm to predict labels that can be used to extract important information from a given document.

In this study, we used a pre-trained BERT model and finetuned it for Named Entity Recognition on different datasets.We presented the results for our approach for the CoNLL-2003 dataset for generic NER as well as on datasets consisting of court judgments. For this, we used two different datasets, first of which was the dataset containing 100 civil judgments, as used by Sharafat et al. [53]. This dataset consisted of civil judgments from different categories and consisted of a total of 10 Named Entity labels including names of locations, person, organization, miscellaneous names, case number, court name being referred, case being referred in the judgment, dates, monetary amounts and law references. For this Dataset, we achieved an F1 score of 93.21, which was a considerable improvement from the previously published f1 score of 86.62.

Moreover, we labelled 214 judgments of Civil Appeal category from the Supreme Court of Pakistan using a total of 14 labels which included names of people, organizations, locations, respondents, dates, case numbers, appealed cases, appealed courts, referred cases, referred courts, law references, FIR numbers, monetary amounts and whether the judgment was approved for reporting. For this dataset we got an F1 score of 92.72. These results appear to be promising when compared to previously reported results.

## Future Work

A notable progression for the current work is the addition of judgments of other different categries and case types in the training dataset. These judgments would not only come from the other categories of Supreme Court of Pakistan but also the other courts of Pakistan including the High Courts. Alongside this, more Named Entities can be included which would be present in the newly added categories that were not available in the Civil Appeal judgments that we used in the current dataset and thus did not include in current work.

Other than the expansion the training dataset for NER, a separate dataset can also be produced that can be used to pre-train the BERT model on judicial data. This process would need the preparation for an much larger dataset but as a result we can possibly achieve better results for not only Named Entity Recognition for judgments but also for different NLP tasks like legal text summarization and the creation of question answering systems.

A variety of systems can also be built using the results from NER systems. For example, the extracted named entities can be used for certain question answering systems where named entities are required, a knowledge base can be populated by using the relationships between entities, any personal information of any individuals that might be mentioned in the judgment can also be anonymized using the results of NER, which can also be a future work.

# References


[1] Chinatsu Aone et al. “SRA: Description of the IE2 system used for MUC-7”. In: *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998*. 1998.

[2] Bogdan Babych and Anthony Hartley. “Improving machine translation quality with automatic named entity recognition”. In: *Proceedings of the 7th International EAMT workshop on MT and other language technology tools, Improving MT through other language technology tools, Resource and tools for building MT at EACL 2003*. 2003.

[3] Alexei Baevski et al. “Cloze-driven pretraining of self-attention networks”. In: *arXiv preprint arXiv:1903.07785* (2019).

[4] Krisztian Balog, Pavel Serdyukov, and Arjen P de Vries. *Overview of the TREC 2010 entity track*. Tech. rep. NORWEGIAN UNIV OF SCIENCE and TECHNOLOGY TRONDHEIM, 2010.

[5] Oliver Bender, Franz Josef Och, and Hermann Ney. “Maximum entropy models for named entity recognition”. In: *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*. 2003, pp. 148–151.

[6] Daniel M Bikel, Richard Schwartz, and Ralph M Weischedel. “An algorithm that learns what’s in a name”. In: *Machine learning* 34.1 (1999), pp. 211–231.

[7] Andrew Borthwick et al. “NYU: Description of the MENE named entity system as used in MUC-7”. In: *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998*. 1998.

[8] Eric Brill. *A simple rule-based part of speech tagger*. Tech. rep. PENNSYLVANIA UNIV PHILADELPHIA DEPT OF COMPUTER and INFORMATION SCIENCE, 1992.

[9] Pengxiang Cheng and Katrin Erk. "Attending to entities for better text understanding". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05. 2020, pp. 7554–7561.

[10] Hai Leong Chieu and Hwee Tou Ng. "Named entity recognition: a maximum entropy approach using global information". In: *COLING 2002: The 19th International Conference on Computational Linguistics*. 2002.

[11] Michael Collins and Yoram Singer. "Unsupervised models for named entity classification". In: *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. 1999.

[12] Gianluca Demartini, Tereza Iofciu, and Arjen P De Vries. "Overview of the INEX 2009 entity ranking track". In: *International Workshop of the Initiative for the Evaluation of XML Retrieval*. Springer. 2009, pp. 254–264.

[13] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[14] George R Doddington et al. "The automatic content extraction (ace) program-tasks, data, and evaluation." In: *Lrec*. Vol. 2. 1. Lisbon. 2004, pp. 837–840.

[15] Christopher Dozier et al. "Named entity recognition and resolution in legal text". In: *Semantic Processing of Legal Texts*. Springer, 2010, pp. 27–43.

[16] Oren Etzioni et al. "Unsupervised named-entity extraction from the web: An experimental study". In: *Artificial intelligence* 165.1 (2005), pp. 91–134.

[17] Abbas Ghaddar and Philippe Langlais. "Robust lexical features for improved neural network named-entity recognition". In: *arXiv preprint arXiv:1806.03489* (2018).

[18] Ralph Grishman and Beth M Sundheim. "Message understanding conference 6: A brief history". In: *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*. 1996.

[19] Jiafeng Guo et al. "Named entity recognition in query". In: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 2009, pp. 267–274.

[20] Daniel Hanisch et al. "ProMiner: rule-based protein and gene entity recognition". In: *BMC bioinformatics* 6.1 (2005), pp. 1–9.

[21] Xiao Huang et al. “Learning a unified named entity tagger from multiple partially annotated corpora for efficient adaptation”. In: *arXiv preprint arXiv:1909.11535* (2019).

[22] Kevin Humphreys et al. “University of Sheffield: Description of the LaSIE-II system as used for MUC-7”. In: *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998*. 1998.

[23] Zhanming Jie and Wei Lu. “Dependency-guided LSTM-CRF for named entity recognition”. In: *arXiv preprint arXiv:1909.10148* (2019).

[24] Ji-Hwan Kim and Philip C Woodland. “A rule-based named entity recognition system for speech input”. In: *Sixth International Conference on Spoken Language Processing*. 2000.

[25] Vijay Krishnan and Christopher D Manning. “An effective two-stage model for exploiting non-local dependencies in named entity recognition”. In: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. 2006, pp. 1121–1128.

[26] Bjornar Larsen. “A trainable summarizer with knowledge acquired from robust NLP techniques”. In: *Advances in automatic text summarization* 71 (1999).

[27] Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. “Transfer learning for named-entity recognition with neural networks”. In: *arXiv preprint arXiv:1705.06273* (2017).

[28] Jinhyuk Lee et al. “BioBERT: a pre-trained biomedical language representation model for biomedical text mining”. In: *Bioinformatics* 36.4 (2020), pp. 1234–1240.

[29] Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. “Fine-grained named entity recognition in legal documents”. In: *International Conference on Semantic Systems*. Springer. 2019, pp. 272–287.

[30] Jing Li et al. “A survey on deep learning for named entity recognition”. In: *IEEE Transactions on Knowledge and Data Engineering* (2020).

[31] Xiaoya Li et al. “Dice loss for data-imbalanced NLP tasks”. In: *arXiv preprint arXiv:1911.02855* (2019).

[32] Shifeng Liu et al. “Hamner: Headword amplified multi-span distantly supervised method for domain specific named entity recognition”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05. 2020, pp. 8401–8408.

[33] Tianyu Liu, Jin-Ge Yao, and Chin-Yew Lin. “Towards improving neural named entity recognition with gazetteers”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* 2019, pp. 5301–5307.

[34] Xiaohua Liu et al. “Recognizing named entities in tweets”. In: *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies.* 2011, pp. 359–367.

[35] Xuezhe Ma and Eduard Hovy. “End-to-end sequence labeling via bi-directional lstm-cnns-crf”. In: *arXiv preprint arXiv:1603.01354* (2016).

[36] Andrew McCallum and Wei Li. “Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons”. In: (2003).

[37] Paul McNamee and James Mayfield. “Entity extraction without language-specific resources”. In: *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002).* 2002.

[38] Andrei Mikheev, Marc Moens, and Claire Grover. “Named entity recognition without gazetteers”. In: *Ninth Conference of the European Chapter of the Association for Computational Linguistics.* 1999, pp. 1–8.

[39] David Nadeau and Satoshi Sekine. “A survey of named entity recognition and classification”. In: *Lingvisticae Investigationes* 30.1 (2007), pp. 3–26.

[40] David Nadeau, Peter D Turney, and Stan Matwin. “Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity”. In: *Conference of the Canadian society for computational studies of intelligence.* Springer. 2006, pp. 266–277.

[41] Hiroki Nakayama et al. *doccano: Text Annotation Tool for Human.* Software available from https://github.com/doccano/doccano. 2018. URL: `https://github.com/doccano/doccano`.

[42] Nanyun Peng and Mark Dredze. “Multi-task domain adaptation for sequence tagging”. In: *arXiv preprint arXiv:1608.02689* (2016).

[43] Matthew E Peters et al. “Deep contextualized word representations”. In: *arXiv preprint arXiv:1802.05365* (2018).

[44] Desislava Petkova and W Bruce Croft. “Proximity-based document representation for named entity retrieval”. In: *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management.* 2007, pp. 731–740.

[45] Lizhen Qu et al. "Named entity recognition for novel types by transfer learning". In: *arXiv preprint arXiv:1610.09914* (2016).

[46] Alexandra Pomares Quimbaya et al. "Named entity recognition over electronic health records through a combined dictionary-based approach". In: *Procedia Computer Science* 100 (2016), pp. 55–61.

[47] Alan Ritter, Sam Clark, Oren Etzioni, et al. "Named entity recognition in tweets: an experimental study". In: *Proceedings of the 2011 conference on empirical methods in natural language processing*. 2011, pp. 1524–1534.

[48] Tim Rocktäschel, Michael Weidlich, and Ulf Leser. "ChemSpot: a hybrid system for chemical named entity recognition". In: *Bioinformatics* 28.12 (2012), pp. 1633–1640.

[49] Erik F Sang and Fien De Meulder. "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition". In: *arXiv preprint cs/0306050* (2003).

[50] Satoshi Sekine and Chikashi Nobata. "Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy." In: *LREC*. Lisbon, Portugal. 2004, pp. 1977–1980.

[51] Satoshi Sekine and Elisabete Ranchhod. *Named entities: recognition, classification and use*. Vol. 19. John Benjamins Publishing, 2009.

[52] Burr Settles. "Biomedical named entity recognition using conditional random fields and rich feature sets". In: *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications (NLPBA/BioNLP)*. 2004, pp. 107–110.

[53] Shahmin Sharafat, Zara Nasar, and Syed Waqar Jaffry. "Legal data mining from civil judgments". In: *International Conference on Intelligent Technologies and Applications*. Springer. 2018, pp. 426–436.

[54] Yanyao Shen et al. "Deep active learning for named entity recognition". In: *arXiv preprint arXiv:1707.05928* (2017).

[55] Stavroula Skylaki et al. "Named Entity Recognition in the Legal Domain using a Pointer Generator Network". In: *arXiv preprint arXiv: 2012.09936* (2020).

[56] Emma Strubell et al. "Fast and accurate entity recognition with iterated dilated convolutions". In: *arXiv preprint arXiv:1702.02098* (2017).

[57] György Szarvas, Richárd Farkas, and András Kocsor. “A multilingual named entity recognition system using boosting and c4. 5 decision tree learning algorithms”. In: *International Conference on Discovery Science*. Springer. 2006, pp. 267–278.

[58] Zhili Wang et al. “Named Entity Recognition Method of Brazilian Legal Text based on pre-training model”. In: *Journal of Physics: Conference Series*. Vol. 1550. 3. IOP Publishing. 2020, p. 032149.

[59] Vikas Yadav and Steven Bethard. “A survey on recent advances in named entity recognition from deep learning models”. In: *arXiv preprint arXiv:1910.11470* (2019).

[60] Lin Yao et al. “Biomedical named entity recognition based on deep neutral network”. In: *Int. J. Hybrid Inf. Technol* 8.8 (2015), pp. 279–288.

[61] Shaodian Zhang and Noémie Elhadad. “Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts”. In: *Journal of biomedical informatics* 46.6 (2013), pp. 1088–1098.

[62] Zhengyan Zhang et al. “ERNIE: Enhanced language representation with informative entities”. In: *arXiv preprint arXiv:1905.07129* (2019).

[63] Suncong Zheng et al. “Joint extraction of entities and relations based on a novel tagging scheme”. In: *arXiv preprint arXiv:1706.05075* (2017).

[64] GuoDong Zhou and Jian Su. “Named entity recognition using an HMM-based chunk tagger”. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. 2002, pp. 473–480.