IMPLEMENTATION OF LEGAL DOCUMENTS TEXT SUMMARIZATION AND CLASSIFICATION BY APPLYING NEURAL NETWORK TECHNIQUES

Siddhartha Rusiya (18UCS106)
Aditya Sharma (18UCS050)
Debajyoti Debbarma (18UCS052)
Samarjit Debbarma (18UCS059)

COMPUTER SCIENCE & ENGINEERING DEPARTMENT
NATIONAL INSTITUTE OF TECHNOLOGY,AGARTALA
INDIA-799046
December, 2021

IMPLEMENTATION OF LEGAL DOCUMENTS TEXT SUMMARIZATION AND CLASSIFICATION BY APPLYING NEURAL NETWORK TECHNIQUES

Report submitted to
National Institute of Technology, Agartala
for the award of the degree
of
Bachelor of Technology

by

Siddhartha Rusiya (18UCS106)
Aditya Sharma (18UCS050)
Debajyoti Debbarma (18UCS052)
Samarjit Debbarma (18UCS059)

Under the Guidance of

Dr. Anupam Jamatia
Assistant Professor, CSE Department, NIT,Agartala, India

COMPUTER SCIENCE & ENGINEERING DEPARTMENT
NATIONAL INSTITUTE OF TECHNOLOGY,AGARTALA
December, 2021

Dedicated To

To our Project Supervisor Dr. Anupam Jamatia, Assistant Professor, CSED, NIT, Agartala for sharing his valuable knowledge, encouragement and showing confidence on us all the time. Each of the faculties of the department to contribute in our development as a professional and help us to achieve this goal.
To all those people who have somehow contributed to the creation of this project and who have supported us.

❝You can't teach people everything they need to know. The best you can do is position them where they can find what they need to know when they need to know it.❞

-Seymour Papert (MIT Mathematician)

REPORT APPROVAL FOR B.TECH

This report entitled "*Implementation of legal documents text summarization and classification by applying Neural Network techniques*", by Siddhartha Rusiya (18UCS106), Aditya Sharma (18UCS050), Debajyoti Debbarma (18UCS052) & Samarjit Debbarma (18UCS059), is approved for the award of ***Bachelor of Technology*** in ***Computer Science and Engineering***.

Dr. Anupam Jamatia	Dr. Mrinal Kanti Debbarma
(Project Supervisor)	(Head of the Department)
Assistant Professor	Associate Professor
Computer Science and Engineering Department	Computer Science and Engineering Department
NIT, Agartala	NIT, Agartala

Date:_____

Place:NIT, Agartala

DECLARATION

We declare that the work presented in this report proposal titled "*Implementation of legal documents text summarization and classification by applying Neural Network techniques*", submitted to the Computer Science and Engineering Department, National Institute of Technology, Agartala, for the award of the ***Bachelor of Technology*** degree in ***Computer Science and Engineering***, represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

December, 2021
Agartala

Candidate's Name	Candidate's Name
Candidate's Name	Candidate's Name

CERTIFICATE

It is certified that the work contained in the report titled "*Implementation of legal documents text summarization and classification by applying Neural Network techniques*", by Siddhartha Rusiya (18UCS106), Aditya Sharma (18UCS050), Debajyoti Debbarma (18UCS052) & Samarjit Debbarma (18UCS059), has been carried out under my supervision and this work has not been submitted elsewhere for a degree.

Dr. Anupam Jamatia	Dr. Mrinal Kanti Debbarma
(Project Supervisor)	(Head of the Department)
Assistant Professor	Associate Professor
Computer Science and Engineering Department	Computer Science and Engineering Department
NIT, Agartala	NIT, Agartala

Acknowledgement

We would like to take this opportunity to express our deep sense of gratitude to all who helped us directly or indirectly during this thesis work.
Firstly, we would like to thank our supervisor, **Dr. Anupam Jamatia**, for being a great mentor and the best advisor we could ever have. His advise, encouragement and critics are source of innovative ideas, inspiration and causes behind the successful completion of this report. The confidence shown on us by him was the biggest source of inspiration for us. It has been a privilege working with her from last one year.
We are highly obliged to all the faculty members of Computer Science and Engineering Department for their support and encouragement. We also thank **Prof. H. K. Sharma**, Director, NIT Agartala and **Dr. Mrinal Kanti Debbarma**, H.O.D, CSED for providing excellent computing and other facilities without which this work could not achieve its quality goal.

- Siddhartha Rusiya **- Aditya Sharma**

- Debajyoti Debbarma **- Samarjit Debbarma**

List of Figures

List of Tables

Abstract

Legal documents are usually not well structured especially in case of judiciary it is a very difficult task for lawyers or legal advisors to proceed ahead with the case smoothly. Rhetorical role classification and summarizing of legal case documents enables legal advisors to analyse legal documents easily. Challenges are unstructured data and difference in significance of words at different sentences of a legal document. For Rhetorical role labelling, we did various experiments by trying different sampling and pre-processing methods on different models like BERT and Roberta. For finding relevant sentences in a legal document, we apply various models on our corpus. After doing various experiments in both rhetorical role labelling and finding relevant sentences in a legal document, we would conclude that selection sampling or pre-processing methods would be dependent on nature of corpus to get good results and also selection of deep learning methods like BERT, Roberta etc will be good choice where difference in significance of words exists at different places.

Contents

CHAPTER 1

Introduction

Text classification is one of the most common problems of NLP, which targets to assign labels or tags to textual data such as sentences, queries, paragraphs, and documents. It has a wide range of applications including question answering, spam detection, sentiment analysis, news categorization, user intent classification, content moderation, and so on. Rhetorical role labelling of sentences in a legal document refers to understanding what semantic function a sentence is associated with, such as facts of the case, arguments of the parties, the final judgement of the court, and so on. Identifying the rhetorical roles of sentences in a legal case document can help in a variety of downstream tasks like semantic search, summarization, case law analysis, and so on. However, legal case documents are usually not well structured, and various themes often interleave with each other. For instance, the reason behind the judgment (Ratio of the decision) often interleaves with Precedents and Statutes. Hence it sometimes becomes difficult even for human experts to understand the intricate differences between the rhetorical roles.Hence, automating the identification of these rhetorical roles is a challenging task. Prior attempts to automate the identification of rhetorical roles of sentences in legal documents rely on hand-crafted features such as linguistic cue phrases indicative of a particular rhetorical role, the sequential arrangement of labels, and so on. Some of these features, e.g., indicator cue phrases, are largely

dependent on legal-expert knowledge which is expensive to obtain. Also, the hand-crafted features developed in the prior works are often specific to one or a few domains categories. It has not been explored whether one can devise a set of features that works for documents across domains. Recently developed deep learning, neural network models do not require hand engineering features, but are able to automatically learn the features, given sufficient amounts of training data. Additionally, such models perform better in tasks like classification than methods using hand-crafted features.

1.1 Motivation

In today's world there are no lack of legal problems in every part of the world. Be it property disputes,theft,violence,etc,which cannot be solved without proper execution of the rules and hence leads to the filing of legal cases. With increasing legal problems the difficulty for the courts to solve the cases smoothly has been increasing day by day.The time taken by the legal courts to solve cases are also proportionally increasing with increasing legal problems.Even for lawyers and other legal consultants reading the legal documents are often very time consuming which eventually leads to slower progress of the cases.so,to counter this problem we have eventually thought of developing an automatic rhetorical role classifier and summarizer with the help of modern deep learning models and Natural Language Processing techniques.

In India alone, justice is often indefinitely kept pending. The result: there are 4.5 crores pending cases across all courts in India, as of September 15. In fact, in 2019, there were 3.3 crores pending cases — which means that in the last two years, India has added 23 cases every minute to its pendency list. Due to this fact, it will be a great help in solving the cases if we make proper classification and proper summarization. Since there is lack of research in this field,it motivates us to keep working in that field. One of the thing that also motivates us is the development of the country.

1.2 Goal

The purpose of this thesis is to do rhetorical role labelling and legal document summarization on Indian legal documents using modern deep learning methods.

By Rhetorical role classification it is meant that given multiple legal documents we are to classify the sentences in the documents according to the roles that usually is meaningful and provides a faster overview and understanding of the legal case documents.

We have to classify each sentence in the document in one of the rhetorical roles given below:-

- `Facts`: This refers to the occurrences of events that led to filing of the case.
- `Ruling by Lower Court`:Here,the documents given were from Indian courts.This refers to the judgements given by the previous courts before being presented in the current court.
- `Arguments`: This refers to the sentences that denote the arguments of the contending parties.
- `statute`:This refers to the relevant statute cited in the documents. A statute is a formal written enactment of a legislative authority that governs the legal entities of a city, state, or country by way of consent.
- `Precedent`: This refers to a statement of law found in decision of the superior court. Such decisions are binding to that court and the inferior courts have to follow. The cases based on similar set of facts decided by a court may arise in any future case
- `Ratio of the decision`: This refers to the sentences that denote the rationale/reasoning given by the Supreme Court for the final judgement
- `Ruling by Present Court`: This refers to the sentences that denote the final decision given by the Supreme Court for that case document.

The second goal of the thesis was to create a summary of the given judgements. It is divided into two parts. First part is finding of significance of a sentence in a legal document. Second part is to make a summary of legal documents by considering significant sentences.

1.3 Contribution

This thesis helps our judiciary system in giving judgement of legal legal cases in less time which ultimately reduces the pending cases and give justice to victim quickly or release the

innocent one quickly. It also helps lawyers and legal advisors in understanding the case easily and quickly by which they can manage more clients. People of country have to pay less to lawyers or legal advisors for their case. So that's way this thesis helps so many people from judiciary to a common victim.

1.4 Thesis Overview

This thesis is all about rhetorical role labelling and summarization of legal documents. The main purpose of this thesis is to simplify and automate the way of classification of legal documents. We use many models for classification of legal documents namely BERT with lemmatized text, BERT with basic preprocessed text, BERT with stemmed text, Roberta with lemmatized text, Roberta with basic preprocessed text, Roberta with stemmed text, XLNet with lemmatized text, XLNet with basic preprocessed text and XLNet with stemmed text. The best results of these models during training are categorical accuracy(0.9474), f1 score(0.9474), precision(0.9532) recall(0.9432).The results during validation are categorical accuracy(0.9167), f1 score(0.9167), precision (0.9203) recall (0.9140). The results during testing on macro average are f1 score(0.58), precision 1(0.57) recall 1(0.65).
The concluded version of this thesis is that the classification and summarization gives a good way to understand things quickly and smartly.

CHAPTER 2

Related Work

In this section we discuss prior works related to rhetorical role labelling of sentences and applications of deep learning in the legal domain.

Automatic labelling of rhetorical role of sentences relies heavily on manual annotation.while papers that aim to automate the task of semantic labelling also perform annotation analysis[1][2],other works focus on the process of annotation- developing a manual/set of rules for annotation, inter-annotator studies,curation of a gold standard corpus,and so on.Temis, a corpus of 504 sentences,that were annotated both syntatically or semantically,was developed in [3].An in depth annotation study and curation of a gold standard corpus for the task of sentence labelling can be found in [4],where assessor agreement was low for labels like facts and reasoning outcomes. Towards automating the annotation task,[5] discusses an initial methodoly using NLP tools on 47 criminal cases drawn from california Supreme court and state court of appeals.

There have been several prior attempts towards automatically identifying rhetorical roles of sentences in legal documents. Initial experiments for understanding the rhetorical/ thematic roles in court case documents/judgements/case laws were developed as apart of achieving the broader goal of summarizing these documents[6][7][8].For instance,Saravan et al[6] used conditional

random fields(CRF) for the task on 7 rhetorical roles.segmenting a document into functional and issue specific parts was looked into by [1] on U.S.court documents using CRF handcrafted features.

Most of the works related to rhetorical role labelling of sentences were done using handcrafted features.In comparison to those in these thesis we used modern deep learning models to perform the task of rhetorical role labelling of sentences.Deep Learning (DL) methods are increasingly being applied in the legal domain, e.g., classification of factual and non-factual sentences in a legal document[9],crime classification[10][11] and many other tasks. Some domain general approaches to segmenting texts into multi-paragraph passages by topic are based on statistical similarity and lexical cohesion, the repetition of similar words in coherent segments and the tendency for vocabulary to change across segment boundaries.Segmenting legal texts into topics or, as in our project, into functional sections or parts, has required the application of more legal domain-specific knowledge[12].For instance, one must first settle on the types of functional sections that are present in the legal texts of interest such as courts' legal decisions.One approach to segmentation has focused on automatically identifying the rhetorical roles of sentences. For instance, a case document to be summarized has been divided into parts for purposes of selecting the important sentences and organizing them into a summary based on a standard model of case structure[13]. The authors of [14] employed verb tense and aspect in sentences stating legal background knowledge, case description, or a judge's opinions. [15] employed other linguistic markers with contextual dependencies to construct a thematic structuring rule base for contextual exploration. There are many more works related to the segmentation of legal documents which can be found on the internet.

CHAPTER 3

Dataset

3.1 Understanding Dataset

The provided dataset contains 11,285 legal judgements along with its corresponding labels.There is total seven rhetorical labels in which all legal judgements is classified. The numbers of legal judgements of differs labels differs from as low as 341(3%) to as high as 4211(37%).By careful observation with percentage distribution we noticed about the dataset is unbalanced. As legal judgements contains major legal terms, we have to counter the significance of every word of legal judgement also as in legal significance. We also observe that some of the words have very little or significance in process of classification. One more thing that is being noticed that during classification, dates and numbers have no role to play.There is a need to filter the legal judgement for modelng a proper classifier.

The seven different rhetorical labels are Facts, Ratio of the Decision, Precedent, Argument, Statue, Ruling by Lower Court, Ruling by Present Court. In this dataset, majority of legal judgements related to Facts while minority of legal judgements related to Ruling by Present Court.

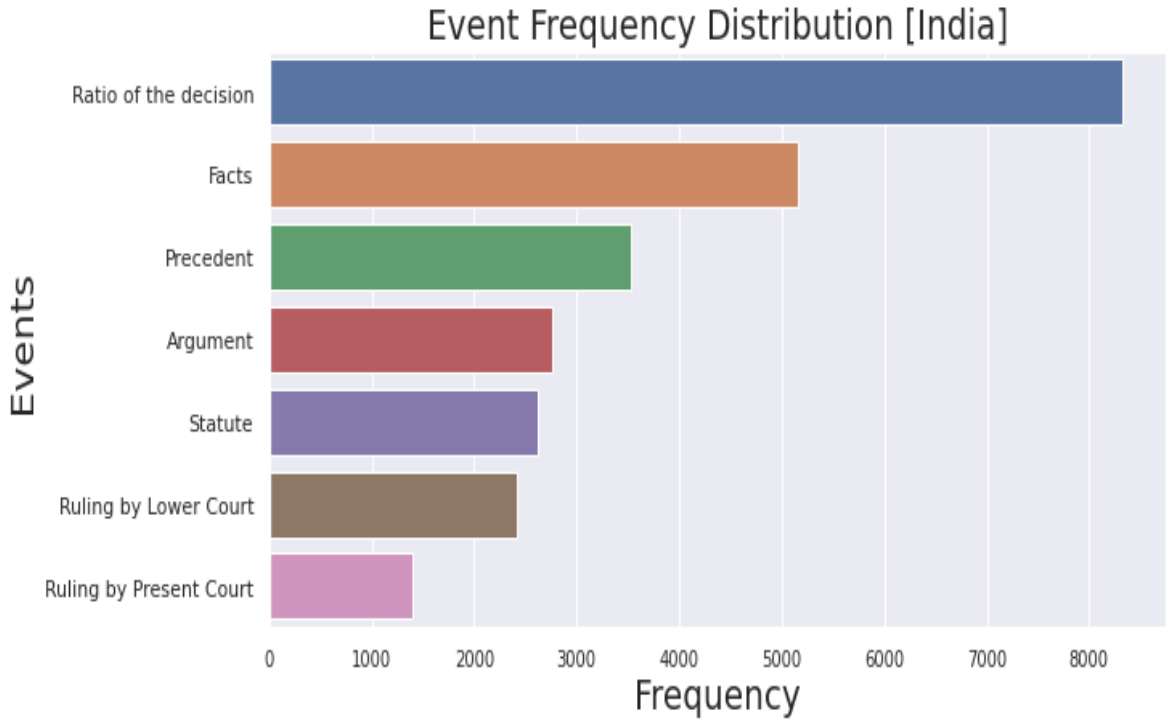

Fig-1:Number of statements against each label in original dataset

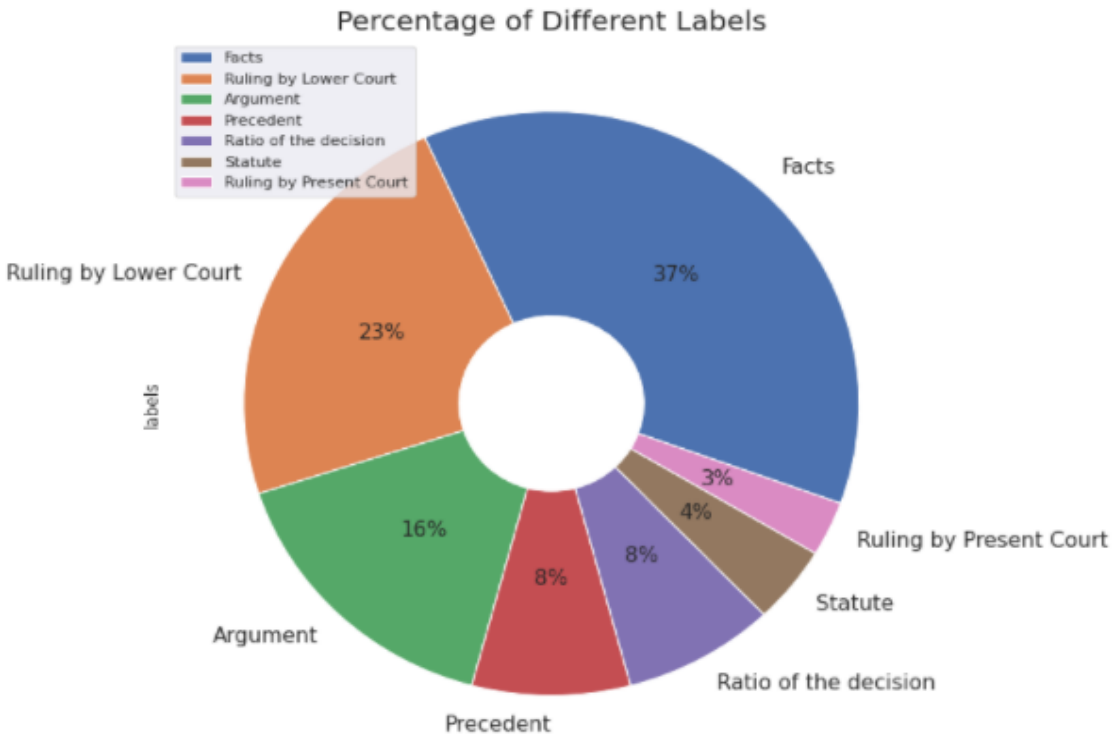

Fig-2:Percentage of statememnts against each label in original dataset

3.2 Preprocessing of Dataset

As per our observation regarding legal judgement, we remove commas, punctuations & special characters. Removing all this helps to get rid of unhelpful parts of the data and comes in handy when you want to do text analysis on important pieces of data. We also remove stopwords using nltk library. Sometimes use of these techniques also downgrade the model but after careful observation we noticed that need of these techniques exists. Although we also did experiment without applying these techniques also to observe the effect of applying them.

We also did experiments using legal judgement text that is either lemmatized or stemmed.Stemming and Lemmatization are Text Normalization (or sometimes called Word Normalization) techniques in the field of Natural Language Processing that are used to prepare text, words, and documents for further processing. Stemming and Lemmatization have been studied, and algorithms have been developed in Computer Science since the 1960's. In this tutorial you will learn about Stemming and Lemmatization in a practical approach covering the background, some famous algorithms, applications of Stemming and Lemmatization, and how to stem and lemmatize words, sentences and documents using the Python nltk package which is the Natural Language Tool Kit package provided by Python for Natural Language Processing tasks.

3.3 Balancing of Dataset

We noticed that difference between majority judgements containing labels and minority containing label is high. So, we use oversampling technique to reduce this difference. During oversampling, we take some predefined threshold values, based on range in which they lies we decide the degree of oversampling. After oversampling, numbers of legal judgements of differs labels differs from as low as 1398(5%) to as high as 8328(32%). Now the total judgements is 26,198. The percentage distribution among different labels after balancing of dataset are Facts(32%), Ratio of the Decision(10%), Precedent(11%), Argument(15%), Statue(9%), Ruling by Lower Court(20%), Ruling by Present Court(5%). As we can see now by numbers as well as by percentage composition that now our dataset is more balanced than the dataset before oversampling.

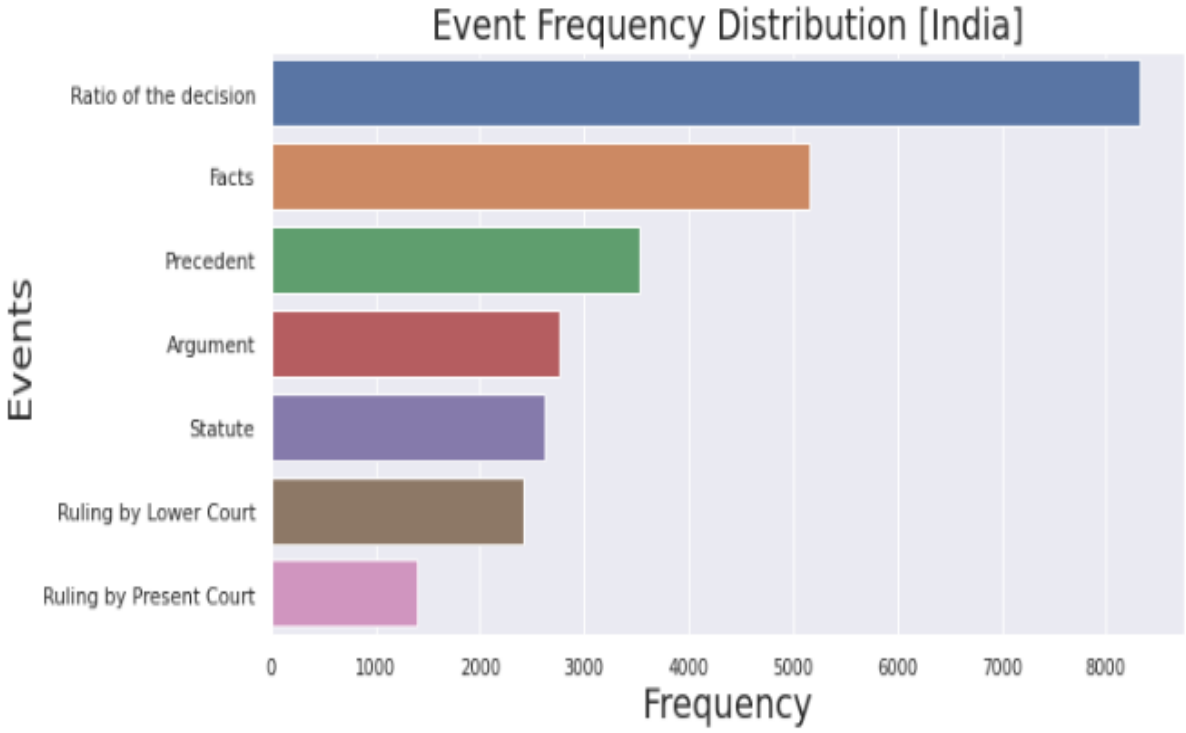

Fig-3:Number of statements against each label in balanced dataset

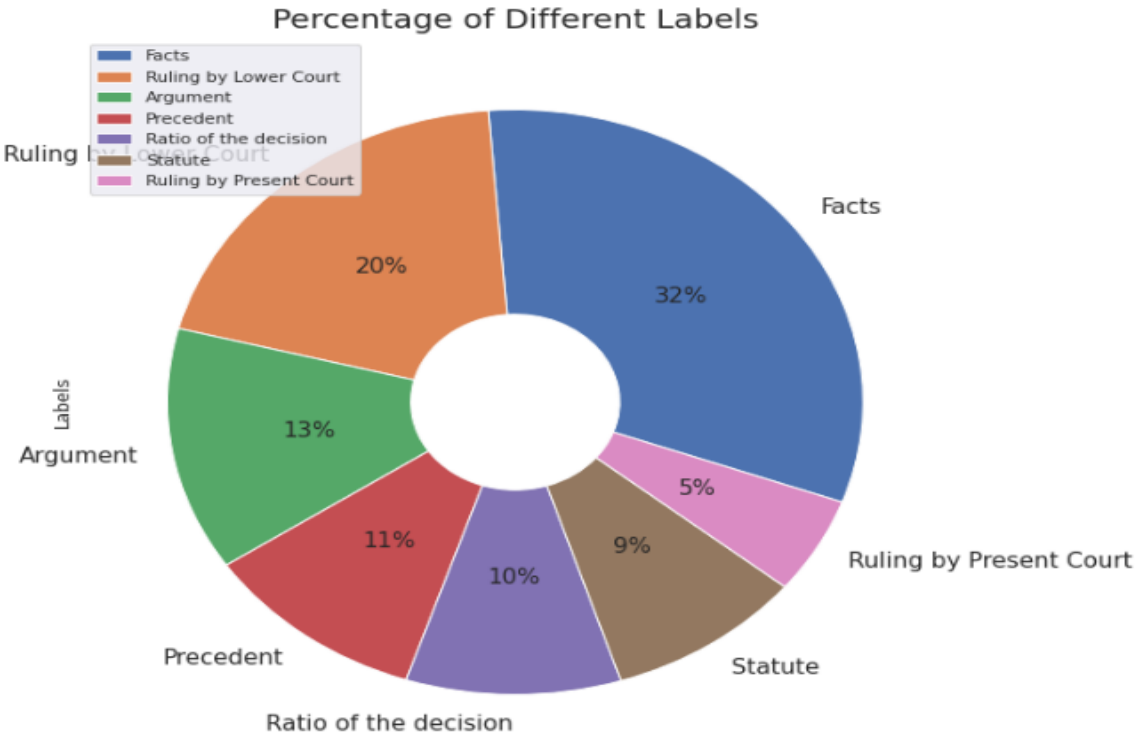

Fig-4:Percentage of statements against each label in balanced dataset

CHAPTER 4

Model Study

4.1 BERT model

BERT is an open source machine learning framework for natural language processing (NLP).It is designed to help computers understand the meaning of ambiguous language in text by using surrounding text to establish context.The model was pre-trained using text from wikipedia and can be fined-tuned with question and answer datasets.

BERT,which stands for Bidirectional Encoder Representations from Transformers,is based on Transformers, a deep learning model in which every output element is connected to every input element, and the weightings between them are dynamically calculated based upon their connection.In NLP this process is called attention. Historically,language models could only read text input sequentially either left to right or right to left but couldn't do both at the same time. BERT is different because it is designed to read in both directions at once. This capability, enabled by the introduction of transformers is known as bidirectionality. Using this bidirectional capability, BERT is pre-trained on two different, but related,NLP tasks.

The goal of any given NLP technique is to understand human language as it is spoken naturally. In BERT's case, this typically means predicting a word in a blank. To do this, models typically need to train using a large repository of specialized, labeled training data. This necessitates laborious manual data labeling by teams of linguists.BERT, however, was pre-trained using only an unlabeled, plain text corpus (namely the entirety of the English Wikipedia, and the Brown Corpus). It continues to learn unsupervised from the unlabeled text and improve even as its being used in practical applications (ie Google search). Its pre-training serves as a base layer of "knowledge" to build from. From there, BERT can adapt to the ever-growing body of searchable content and queries and be fine-tuned to a user's specifications. This process is known as transfer learning.

BERT is also the first NLP technique to rely solely on self-attention mechanism, which is made possible by the bidirectional Transformers at the center of BERT's design. This is significant because often, a word may change meaning as a sentence develops. Each word added augments the overall meaning of the word being focused on by the NLP algorithm. The more words that are present in total in each sentence or phrase, the more ambiguous the word in focus becomes. BERT accounts for the augmented meaning by reading bidirectionally, accounting for the effect of all other words in a sentence on the focus word and eliminating the left-to-right momentum that biases words towards a certain meaning as a sentence progresses. BERT is open source, meaning anyone can use it. Google claims that users can train a state-of-the-art question and answer system in just 30 minutes on a cloud tensor processing unit (TPU), and in a few hours using a graphic processing unit (GPU). Many other organizations, research groups and separate factions of Google are fine-tuning the BERT model architecture with supervised training to either optimize it for efficiency (modifying the learning rate, for example) or specialize it for certain tasks by pre-training it with certain contextual representations

4.2 ROBERTA model

RoBERTa stands for Robustly Optimized BERT Pre-training Approach. It was presented by researchers at Facebook and Washington University. The goal of this paper was to optimize the training of BERT architecture in order to take lesser time during pre-training. RoBERTa has almost similar architecture as compare to BERT, but in order to improve the results on BERT architecture, the authors made some simple design changes in its architecture and training procedure. These changes are:

- `Removing the Next Sentence Prediction (NSP) objective`: In the next sentence prediction, the model is trained to predict whether the observed document segments come from the same or distinct documents via an auxiliary Next Sentence Prediction (NSP) loss. The authors experimented with removing/adding of NSP loss to different versions and concluded that removing the NSP loss matches or slightly improves downstream task performance

- `Training with bigger batch sizes & longer sequences`: Originally BERT is trained for 1M steps with a batch size of 256 sequences. In this paper, the authors trained the model with 125 steps of 2K sequences and 31K steps with 8k sequences of batch size. This has two advantages, the large batches improve perplexity on masked language modelling objective and well as end-task accuracy. Large batches are also easier to parallelize via distributed parallel training.

- `Dynamically changing the masking pattern`: In BERT architecture, the masking is performed once during data preprocessing, resulting in a single static mask. To avoid using the single static mask, training data is duplicated and masked 10 times, each time with a different mask strategy over 40 epochs thus having 4 epochs with the same mask. This strategy is compared with dynamic masking in which the different masking for every time we passed data into the model.

The datasets used to train ROBERTA model were BOOK corpus, english wikipedia dataset,CC-NEWS,OPENWEBTEXT and STORIES. ROBERTA proved to be one of the best performing model in many tasks.Some of the results of ROBERTA are listed as follows:-

- On the GLUE benchmark NLP tasks, the model achieves a score of 88.5 on the public leaderboard and achieve the state-of-the-art score on 4 of GLUE tasks: Multi Natural Language Inference (MNLI), QuestionNLI, Semantic Textual Similarity Benchmark (STS-B), and Recognizing Textual Entailments (RTE) at the time of its release.

- At the time of its release, On the SQuAD 1.1 and SQuAD 2.0 datasets, it is able to match the previous state-of-the-Art results by XLNet.

- It also achieves better results than BERT(LARGE) model and XLNet on RACE benchmark datasets.

4.3 XLNet

XLNet is the latest and greatest model to emerge from the booming field of Natural Language Processing (NLP). The XLNet paper combines recent advances in NLP with innovative choices in how the language modelling problem is approached. When trained on a very large NLP corpus, the model achieves state-of-the-art performance for the standard NLP tasks that comprise the GLUE benchmark.XLNet is a BERT-like model instead of a totally different one. But it is a very promising and potential one. In one word, XLNet is a generalized autoregressive pretraining method.XLNet is an autoregressive Transformer that leverages the best of both autoregressive language modeling and autoencoding while attempting to avoid their limitations. Instead of using a fixed forward or backward factorization order as in conventional autoregressive models, XLNet maximizes the expected log likelihood of a sequence w.r.t. all possible permutations of the factorization order. Thanks to the permutation operation, the context for each position can consist of tokens from both left and right. In expectation, each position learns to utilize contextual information from all positions, i.e., capturing bidirectional context.Additionally, inspired by the latest advancements in autogressive language modeling, XLNet integrates the segment recurrence mechanism and relative encoding scheme of Transformer-XL into pretraining, which empirically improves the performance especially for tasks involving a longer text sequence. Autoregressive language model is a kind of model that using the context word to predict the next word. But here the context word is constrained to two directions, either forward or backward.he advantages of autoregressive language model are good at generative NLP tasks. Because when generating context, usually is the forward direction. autoregressive language model naturally works well on such NLP tasks.But autoregressive language model has some disadvantages, it only can use forward context or backward context, which means it can't use forward and backward context at the same time. Language model consists of two phases, the pre-train phase, and fine-tune phase. XLNet focus on pre-train phase. In the pre-train phase, it proposed a new objective called Permutation Language Modeling. We can know the basic idea from this name, it uses permutation.XLNet also outperformed RoBERTa substantially on large supervised classification tasks. XLNet's performance gain was significantly more significant for explicit reasoning tasks like SQuAD and RACE that involve more extended context requirements. XLNet beats BERT across twenty tasks, including:Text classification,Question answering,docs ranking,Natural language inference,Duplicate sentence detection.The model achieves state-of-the-art performance on eighteen out of the twenty tasks. XLNet was compared to more recent adaptions of BERT, including RoBERTa. To compare fairly, the same number of layers and hyper-parameters were used.

4.4 Challenges

The deep learning models comes with a huge advantages as it becomes many difficult and time consuming work a lot easier than usual.But it is not wrong to say it comes with few disadvantages or challenges as well.In this section we will be discussing about the challenges of the models mentioned above.

The main drawbacks of using BERT and other big neural language models is the computational resources needed to train/fine-tune and make inferences.BERT is a technology to generate contextualized word embeddings/vectors, which is its biggest advantage but also it's biggest disadvantage as it is very compute-intensive at inference time, meaning that if you want to use it in production at scale, it can become costly.However, more recent research have proposed different methods to overcome this issue. Knowledge distillation, quantization, pruning and other techniques make BERT models reasonable for production environments. Recently Google announced its production usage of BERT-based models on English searches.One of the limitation of BERT is lack of ability to handle long text sequence. By default, BERT supports up to 512 token.Also BERT is an AE language model.The AE language model also has its disadvantages. It uses the [MASK] in the pretraining, but this kind of artificial symbols are absent from the real data at finetuning time, resulting in a pretrain-finetune discrepancy. Another disadvantage of [MASK] is that it assumes the predicted (masked) tokens are independent of each other given the unmasked tokens. Now coming on models like XLNet,it combines the bidirectional capability of BERT with the autoregressive language modeling of Transformer-XL. The model outperforms BERT on various NLP tasks, often by a large margin. The model has been heralded by many as the new standard for language understanding. There are potential applications for XLNet in computer vision and reinforcement learning.But it also has some disadvantages as it is pre-trained to capture long-term dependencies, and combined with masking during permutation, the model can underperform on short sequences. XLNet is generally more resource-intensive and takes longer to train and to infer compared to BERT. Some NLP tasks can only be run on TPUs with sufficient memory to reach the reported performance. This additional resource requirement is needed to perform the sets of permutations across the input sequences.

CHAPTER 5

Experiment Setup and Result Analysis

In this part of thesis, We have discussed about the process of doing various experiments, challenges faced while doing those experiments, how to overcome those challenges, result analysis of experiments and error analysis of those challenges. One of the interesting things of our experiments is that we also observe the effect of lemmatization or stemmation, preprocessing on our results for different models.

5.1 Experiment Setup

We set up model environment using Google colab and use python for code implementation. We import libraries that is needed building our dataframe and other purposes. We also install various dependencies that is different for different models that we used. We used the datasets provided by the team of Artificial Intelligence for Legal Assistance(AILA) for training and testing of our models. We used our google drive for importing data to google colab. We also used Google colab GPU for running our code. We implement various models by importing BERT, RoberTa & XL-Net and applying different preprocessing techniques on it.

We use methods utilizes Transformers based models for the task of document classification. The proposed model uses a modified pretrained RoBERTa(a Robust and optimized BERT pretraining approach) encoder with an extra linear layer added to the pretrained RoBERTa base model, which was designed as an improvement of BERT by providing advanced masked language modeling and significantly increasing the magnitude of training data.

5.2 Feature Engineering

The parameters of the final model are selected according to the performance of each classifier on the development set. The fine-tuning parameters of the BERT model are selected according to the classification accuracy of the BERT on the development set. Since the maximum sentence length of 128 can cover all of the sentence length, so sequence length is directly set to 128, no parameter adjustment is required. The final BERT fine-tuning parameters, Batch_size=32, Learning_rate=5e-5, verbose = 1 and epoch=4.

The fine-tuning parameters of the Roberta model are selected according to the classification accuracy of the Roberta on the development set. Since the maximum sentence length of 128 can cover all of the sentence length, so sequence length is directly set to 128,no parameter adjustment is required. The final Roberta fine-tuning parameters, Batch_size=32, Learning_rate=5e-5, verbose = 1 and epoch=5.

The fine-tuning parameters of the XLNet model are selected according to the classification accuracy of the XLNet on the development set. Since the maximum sentence length of 128 can cover all of the sentence length, so sequence length is directly set to 128, no parameter adjustment is required. The final Roberta fine-tuning parameters, Batch_size=32, Learning_rate=5e-5, verbose = 1 and epoch=3.

5.3 Experiments using different models

For rhetorical role labelling, we did various experiments on these deep learning models namely BERT, Roberta & XLNet. Different experiments are basically variation of preprocessing and sampling techniques on different models. The experiments along with their results are as follows:

Experiment using BERT model with base oversampled and cleaned

In this experiment, we trained our model with cleaned dataset which is preprocessed as well as oversampled(due to the fact that so less judgement for some categories). In this experiment, we finetune our model by applying different values on learning parameters of model. The category wise f1 score of this model is Argument(0.62), Facts(0.63), Precedent(0.48), Ratio of the decision(0.49), Ruling by Lower Court(0.33), Ruling by Present Court(0.85) Statute(0.68). The overall f1 score, precision and recall of this experiment is 0.58, 0.57 & 0.65.

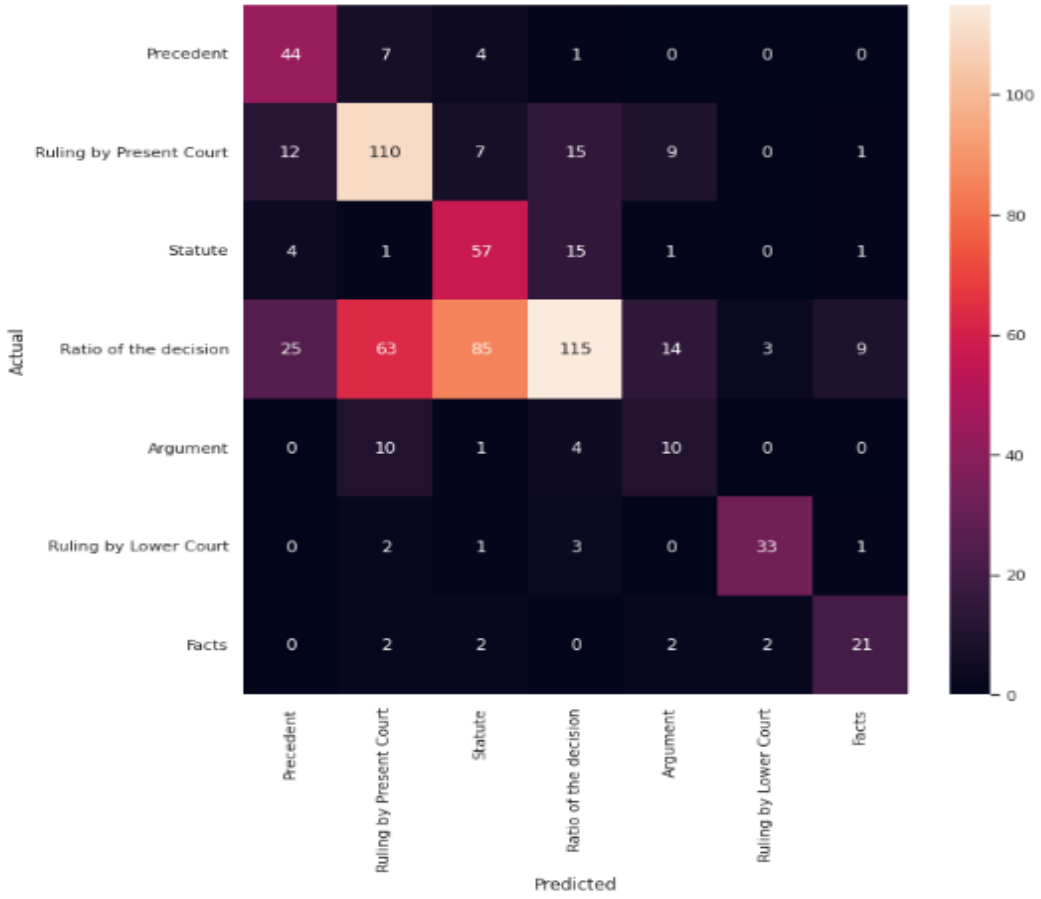

Fig-5:Confusion matrix for BERT model with base oversampled and cleaned

Experiment using BERT model with base lemmatized

In this experiment, we trained our model with cleaned dataset which is preprocessed, oversampled(due to the fact that so less judgement for some categories) and then lemmatized the legal

judgement. In this experiment, we finetune our model by applying different values on learning parameters of model. The category wise f1 score of this model is Argument(0.71), Facts(0.62), Precedent(0.46), Ratio of the decision(0.49), Ruling by Lower Court(0.30), Ruling by Present Court(0.74) Statute(0.64). The overall f1 score, precision and recall of this experiment is 0.57, 0.53 & 0.65.

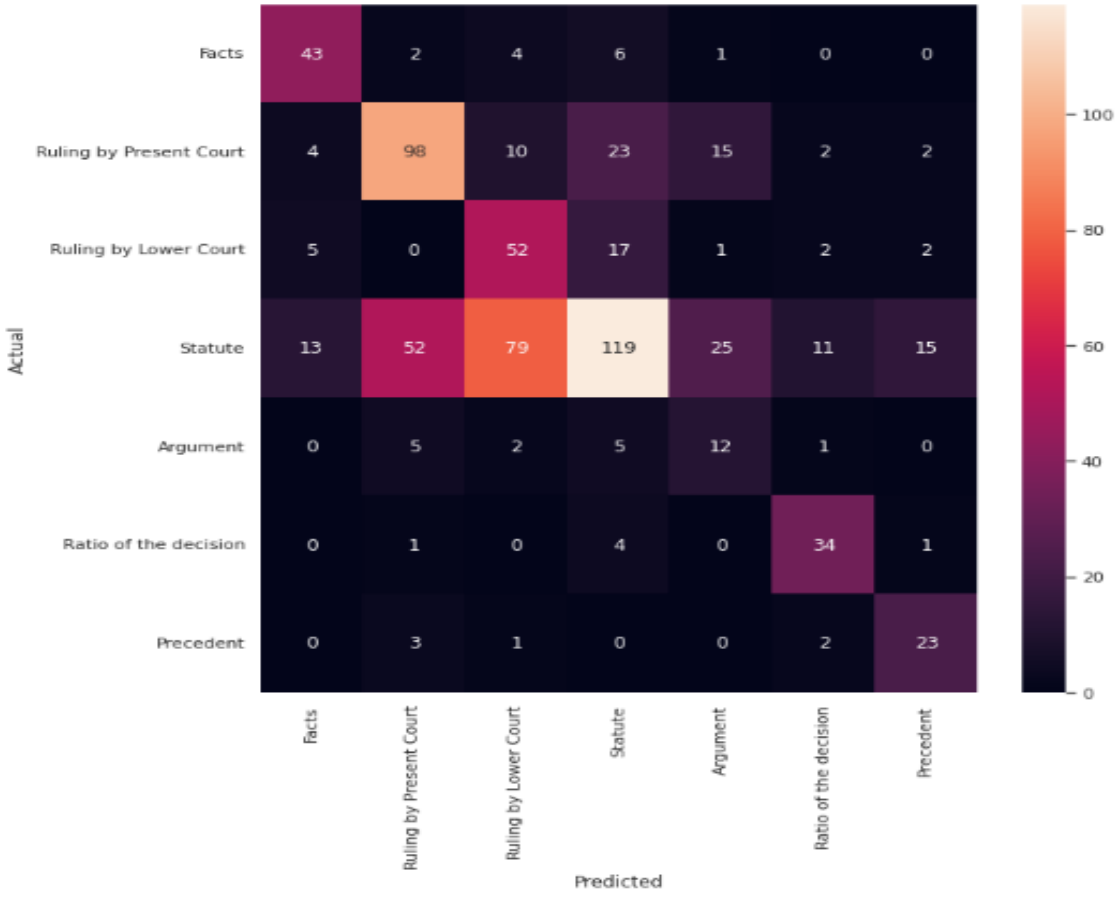

Fig-6:Confusion matrix for BERT model with base lemmatized

Experiment using BERT model with base stemmed

In this experiment, we trained our model with cleaned dataset which is preprocessed, oversampled(due to the fact that so less judgement for some categories) and then stemmed the legal judgement. In this experiment, we finetune our model by applying different values on learning parameters of model. The category wise f1 score of this model is Argument(0.54), Facts(0.51), Precedent(0.46), Ratio of the decision(0.44), Ruling by Lower Court(0.22), Ruling by Present Court(0.70) Statute(0.65). The overall f1 score, precision and recall of this experiment is 0.52, 0.50 & 0.60.

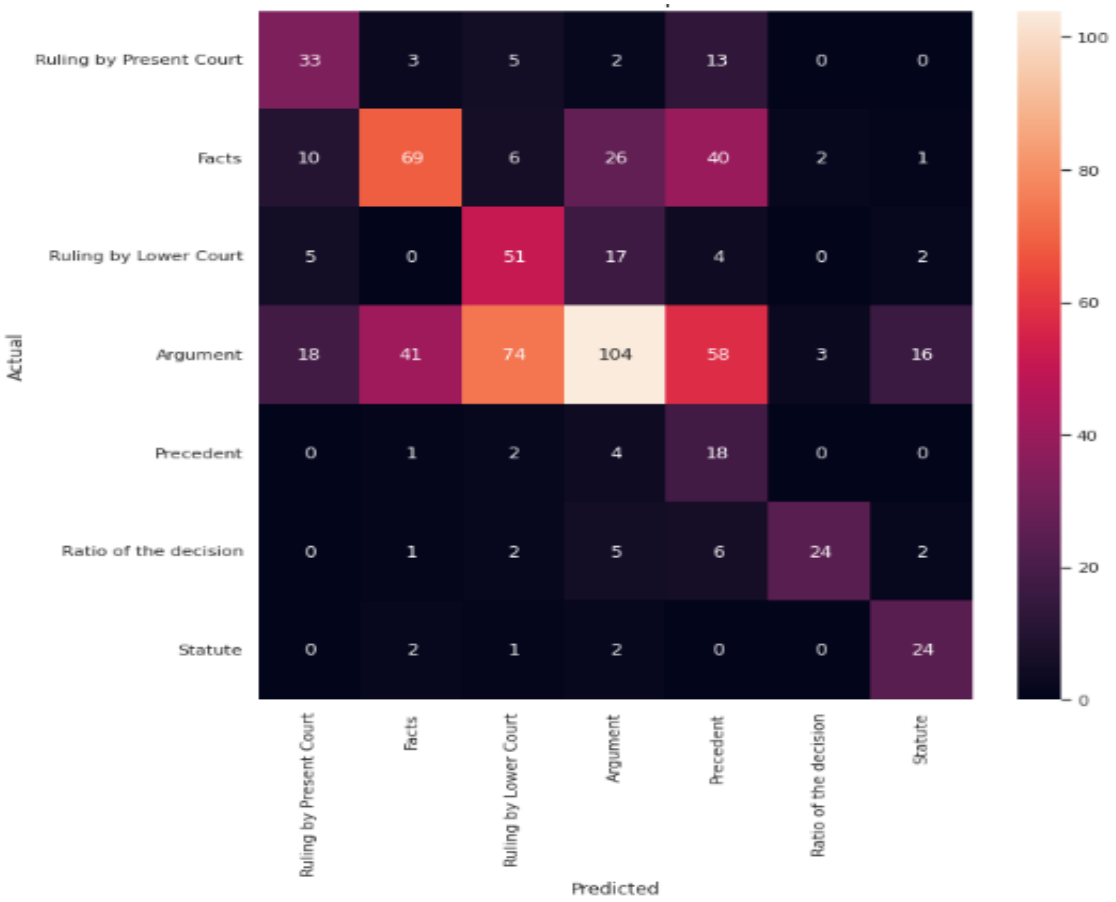

Fig-7:Confusion matrix for BERT model with base stemmed

Experiment using Roberta model with base oversampled and cleaned

In this experiment, we trained our model with cleaned dataset which is preprocessed as well as oversampled(due to the fact that so less judgement for some categories). In this experiment, we finetune our model by applying different values on learning parameters of model. The category wise f1 score of this model is Argument(0.69), Facts(0.57), Precedent(0.40), Ratio of the decision(0.54), Ruling by Lower Court(0.33), Ruling by Present Court(0.78) Statute(0.49). The overall f1 score, precision and recall of this experiment is 0.54, 0.55 & 0.58.

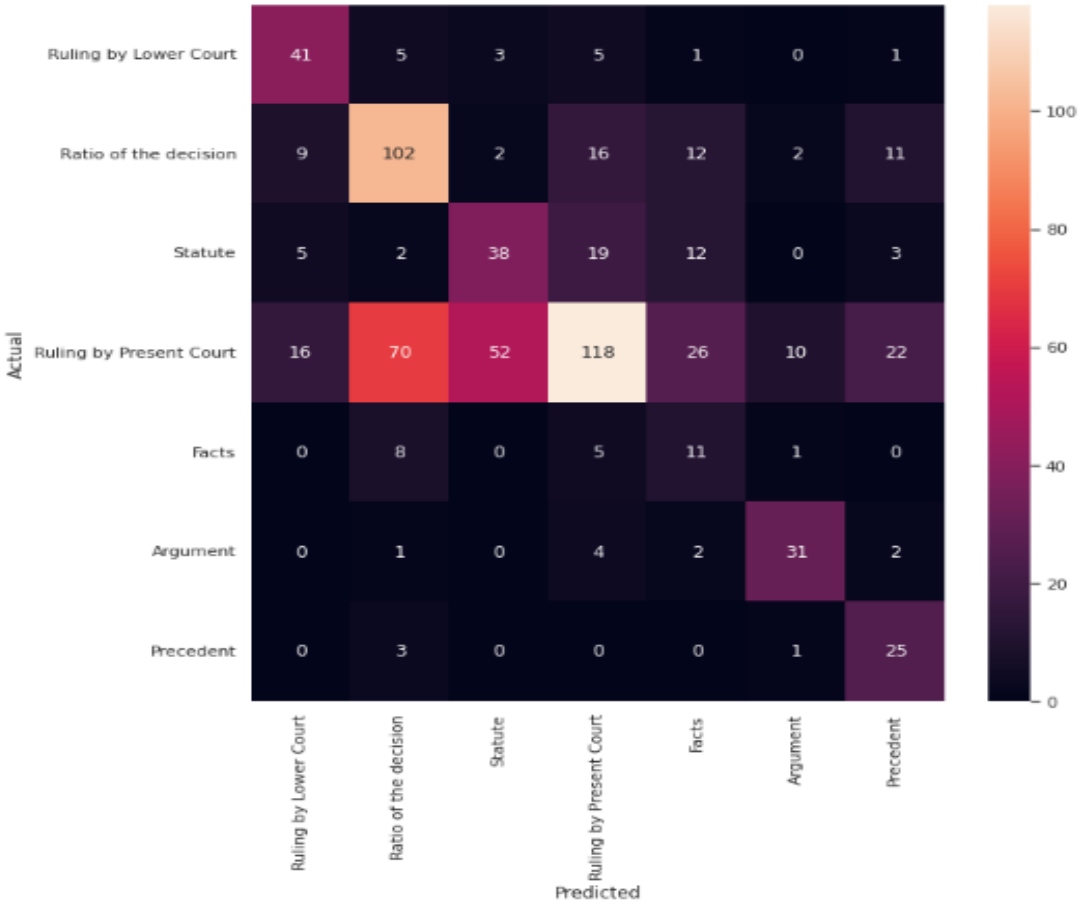

Fig-8:Confusion matrix for RoBERTa model with base oversampled and cleaned

Experiment using Roberta model with base lemmatized

In this experiment, we trained our model with cleaned dataset which is preprocessed, oversampled(due to the fact that so less judgement for some categories) and then lemmatized the legal judgement. In this experiment, we finetune our model by applying different values on learning parameters of model. The category wise f1 score of this model is Argument(0.67), Facts(0.60), Precedent(0.47), Ratio of the decision(0.50), Ruling by Lower Court(0.23), Ruling by Present Court(0.74) Statute(0.47). The overall f1 score, precision and recall of this experiment is 0.57, 0.57 & 0.59.

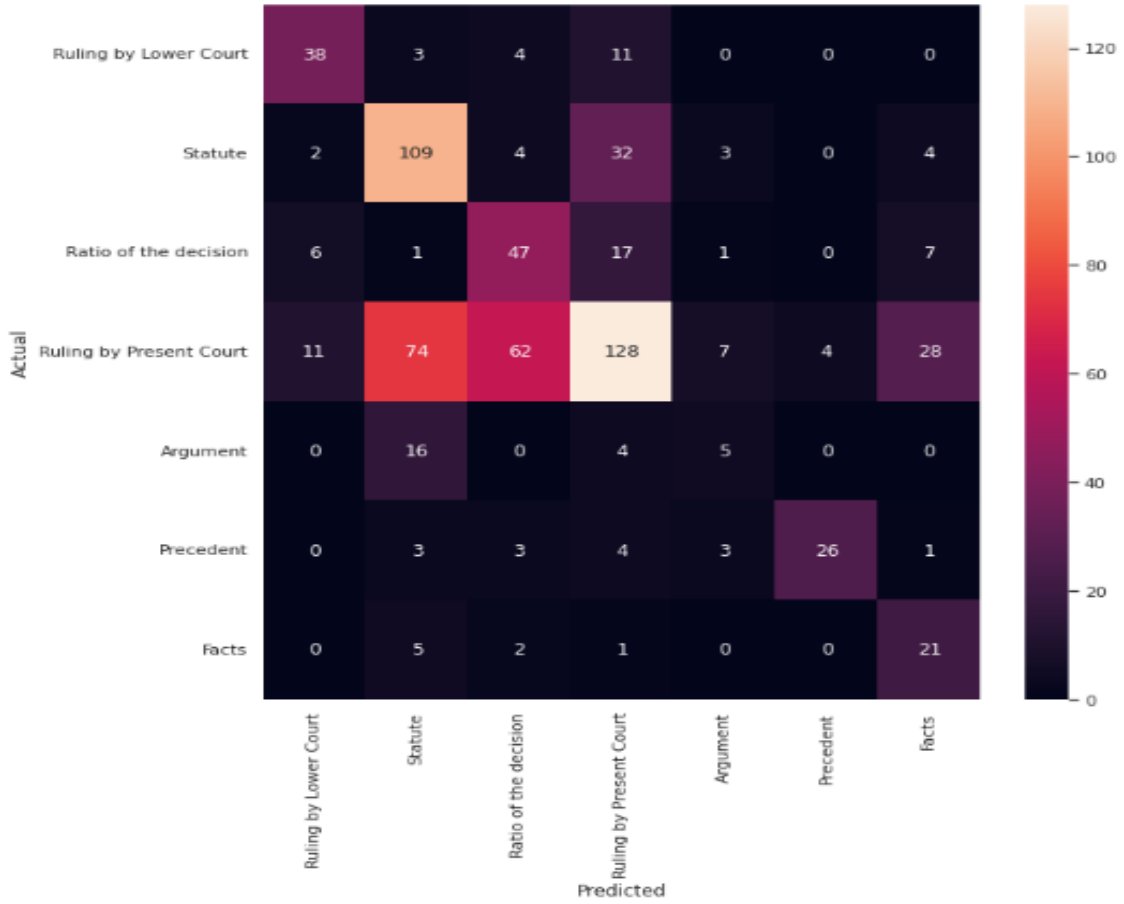

Fig-9:Confusion matrix for RoBERTa model with base lemmatized

Experiment using Roberta model with base stemmed

In this experiment, we trained our model with cleaned dataset which is preprocessed, oversampled(due to the fact that so less judgement for some categories) and then stemmed the legal judgement. In this experiment, we finetune our model by applying different values on learning parameters of model. The category wise f1 score of this model is Argument(0.65), Facts(0.60), Precedent(0.39), Ratio of the decision(0.51), Ruling by Lower Court(0.24), Ruling by Present Court(0.73) Statute(0.50). The overall f1 score, precision and recall of this experiment is 0.52, 0.50 & 0.60.

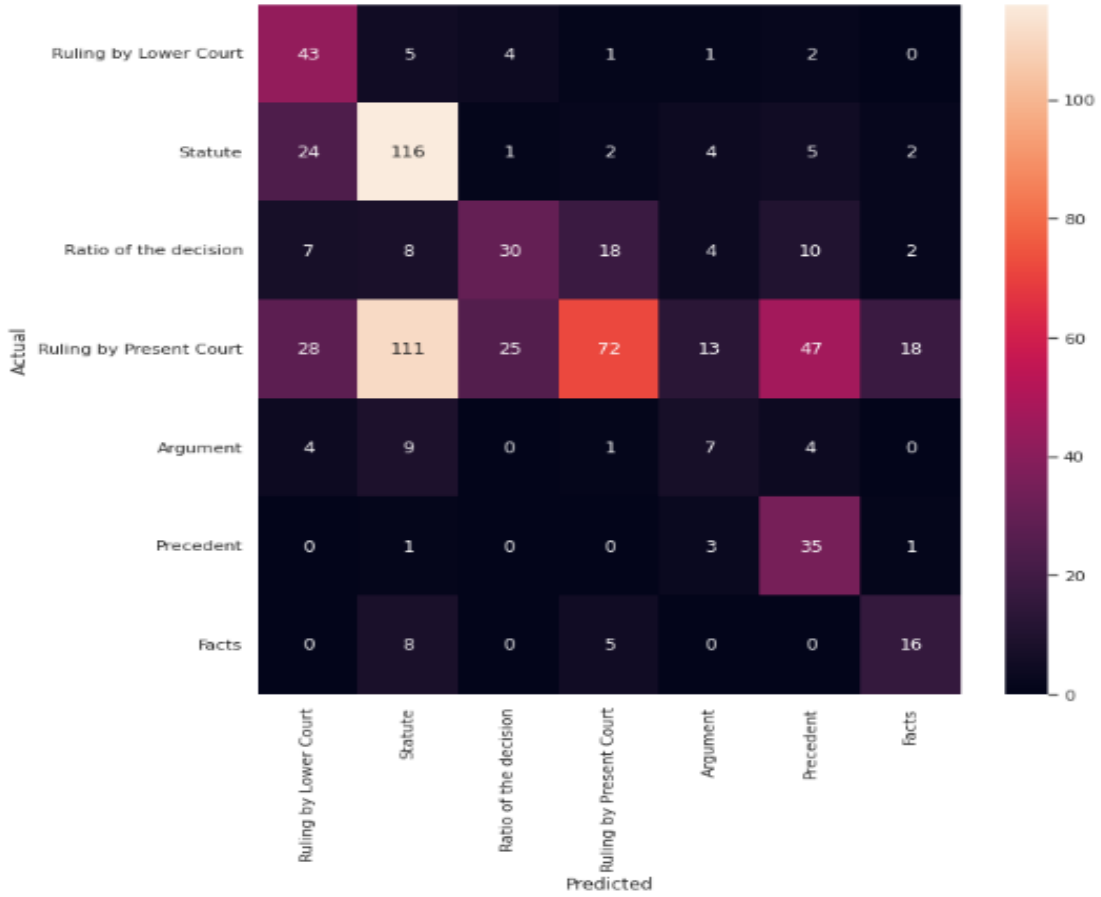

Fig-10:Confusion matrix for RoBERTa model with base stemmed

Experiment using XLNet model with base oversampled and cleaned

In this experiment, we trained our model with cleaned dataset which is preprocessed as well as oversampled(due to the fact that so less judgement for some categories). In this experiment, we finetune our model by applying different values on learning parameters of model. The category wise f1 score of this model is Argument(0.73), Facts(0.59), Precedent(0.42), Ratio of the decision(0.43), Ruling by Lower Court(0.27), Ruling by Present Court(0.80) Statute(0.56). The overall f1 score, precision and recall of this experiment is 0.54, 0.53 & 0.64.

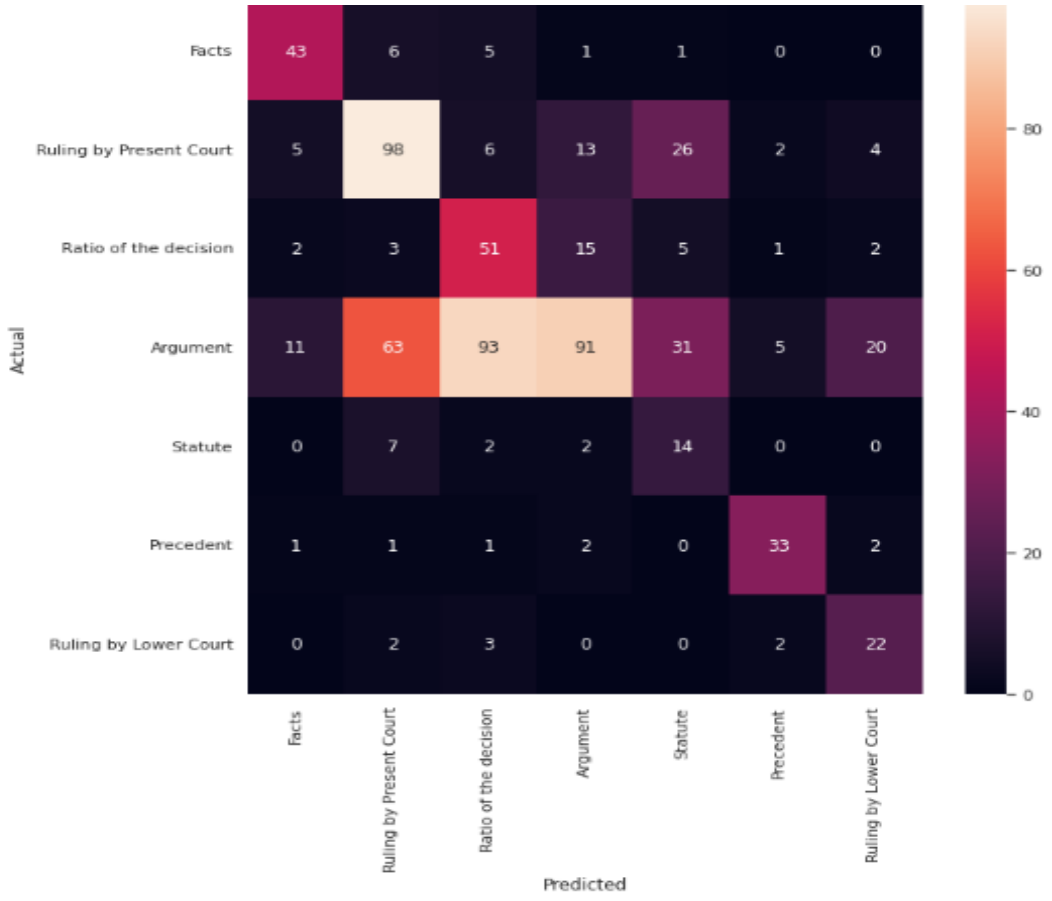

Fig-11:Confusion matrix for XLNet model with base oversampled and cleaned

Experiment using XLNet model with base lemmatized

In this experiment, we trained our model with cleaned dataset which is preprocessed, oversampled(due to the fact that so less judgement for some categories) and then lemmatized the legal judgement. In this experiment, we finetune our model by applying different values on learning parameters of model. The category wise f1 score of this model is Argument(0.72), Facts(0.61), Precedent(0.38), Ratio of the decision(0.54), Ruling by Lower Court(0.22), Ruling by Present Court(0.71) Statute(0.66). The overall f1 score, precision and recall of this experiment is 0.55, 0.54 & 0.60.

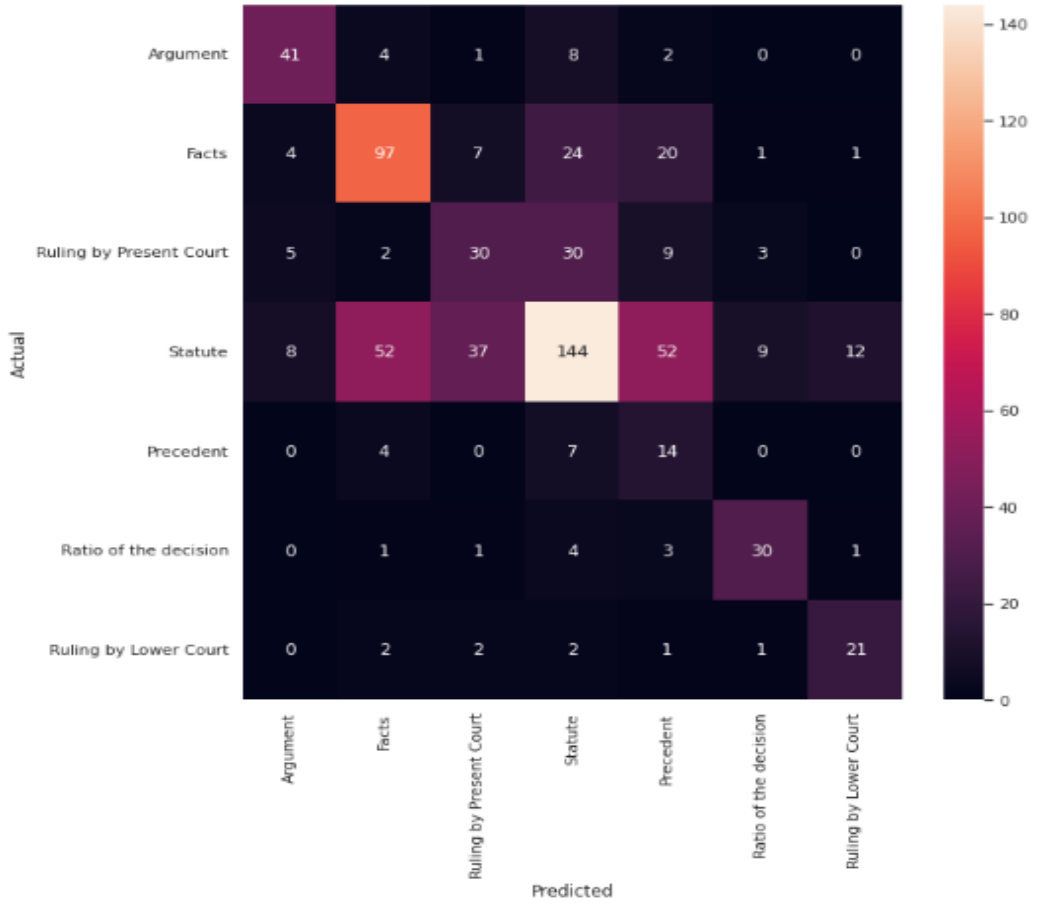

Fig-12:Confusion matrix for XLNet model with base lemmatized

Experiment using XLNet model with base stemmed

In this experiment, we trained our model with cleaned dataset which is preprocessed, oversampled(due to the fact that so less judgement for some categories) and then stemmed the legal judgement. In this experiment, we finetune our model by applying different values on learning parameters of model. The category wise f1 score of this model is Argument(0.69), Facts(0.61), Precedent(0.38), Ratio of the decision(0.35), Ruling by Lower Court(0.19), Ruling by Present Court(0.64) Statute(0.64). The overall f1 score, precision and recall of this experiment is 0.50, 0.53 & 0.59.

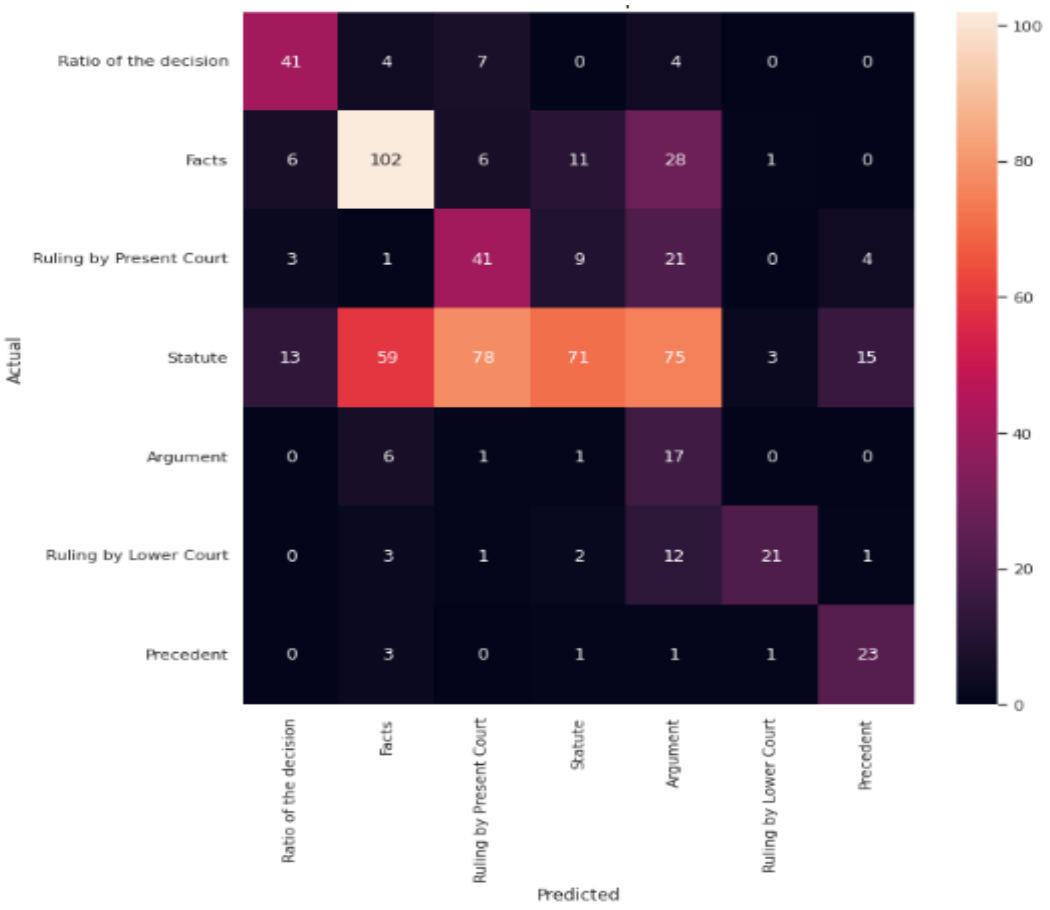

Fig-13:Confusion matrix for XLNet model with base stemmed

5.4 Results Analysis

The model was evaluated on the basis of classic classification metrics - macro averaged recall, precision and f1-score. The metrics were calculated for each label & for every document and then averaged over all the documents to get the overall results of the best model: Precision - 0.57, Recall - 0.65, Accuracy - 0.56 and F1-score - 0.58. The document-wise metrics of best performing model can be seen in Table 2. From document wise metrics, it can be observed that certain category scores much lower than the rest of the categories, these categories have a lower number of data for which the train data was less.

Model	Training	Validation	Testing
BERT	F1-score 0.9474 Accuracy 0.9474 Precision 0.9532 Recall 0.9432	F1-score 0.9167 Accuracy 0.9167 Precision 0.9203 Recall 0.9140	F1-score 0.58 Accuracy 0.56 Precision 0.57 Recall 0.65
BERT Lemmatized	F1-score 0.9317 Accuracy 0.9317 Precision 0.9389 Recall 0.9262	F1-score 0.9105 Accuracy 0.9105 Precision 0.9205 Recall 0.9057	F1-score 0.57 Accuracy 0.55 Precision 0.53 Recall 0.65
BERT Stemmed	F1-score 0.9181 Accuracy 0.9181 Precision 0.9260 Recall 0.9111	F1-score 0.8962 Accuracy 0.8962 Precision 0.9023 Recall 0.8920	F1-score 0.50 Accuracy 0.46 Precision 0.50 Recall 0.59
Roberta	F1-score 0.8679 Accuracy 0.8660 Precision 0.8780 Recall 0.8559	F1-score 0.8922 Accuracy 0.8924 Precision 0.9089 Recall 0.8899	F1-score 0.57 Accuracy 0.57 Precision 0.57 Recall 0.59

Roberta Lemmatized	F1-score 0.8386 Accuracy 0.8360 Precision 0.8448 Recall 0.8180	F1-score 0.8370 Accuracy 0.8380 Precision 0.8409 Recall 0.8280	F1-score 0.53 Accuracy 0.54 Precision 0.53 Recall 0.57
Roberta Stemmed	F1-score 0.8169 Accuracy 0.8097 Precision 0.8168 Recall 0.8080	F1-score 0.8221 Accuracy 0.8247 Precision 0.8376 Recall 0.8111	F1-score 0.52 Accuracy 0.52 Precision 0.50 Recall 0.60
XLNet	F1-score 0.8368 Accuracy 0.8374 Precision 0.8448 Recall 0.8222	F1-score 0.8443 Accuracy 0.8464 Precision 0.8556 Recall 0.8378	F1-score 0.54 Accuracy 0.51 Precision 0.53 Recall 0.64
XLNet Lemmatized	F1-score 0.7113 Accuracy 0.7213 Precision 0.7218 Recall 0.7009	F1-score 0.8038 Accuracy 0.8039 Precision 0.8110 Recall 0.8001	F1-score 0.55 Accuracy 0.54 Precision 0.54 Recall 0.60
XLNet Stemmed	F1-score 0.7627 Accuracy 0.7698 Precision 0.7778 Recall 0.7477	F1-score 0.8084 Accuracy 0.8061 Precision 0.8115 Recall 0.7990	F1-score 0.50 Accuracy 0.45 Precision 0.53 Recall 0.59

Table 1: Scores of Different Experiments

The best results of these experiments as well as their category wise results are as follows:

Label	Precision	Recall	F1-Score
Arguments	0.52	0.79	0.62
Facts	0.56	0.71	0.63
Precedent	0.36	0.72	0.48
Ratio of the decision	0.75	0.37	0.49
Ruling by Lower Court	0.28	0.40	0.33
Ruling by Present Court	0.87	0.82	0.85
Statue	0.64	0.72	0.68
Overall	0.57	0.65	0.58

Table 2: Category wise scores of best model

CHAPTER 6

Conclusion & Future Direction of Work

6.1 Conclusion

In this thesis we discussed about the automatic role labelling of sentences.The documents were provided by the Artificial Intelligence for Legal Assistance(AILA) team which were from Supreme Court of India.The main goal of the first task was to classify the sentences into seven rhetorical roles which are mainly facts(sentences that denote the chronology of events that led to filing of the case),Ruling by lower court,Arguments,Statute,Precedence,Ratio of decision,Ruling by the present court. Automatic classification of sentences is a very difficult tasks as Indian legal documents are not very well structured.The main goal of our task is to make it easier for the legally engaged individual to understand the court documents for the cases the person is handling in a short amount of time. For the task we firstly divide the documents into multiple sentences and perform mutiple basic preprocessing operations such as stemming,lemmatization,etc.we used different models such as BERT model,ROBERTa model and XLNET to train and test our datasets.Through training and testing of different models we used we found that BERT is the best performing model for the task of rhetorical role labelling of sentences.with the training and testing of different models we have achieved the required result

and has successfully completed our task.

6.2 Challenges of the work

Rhetorical role labelling of sentences of legal case documents is a very difficult task as mostly because legal documents are not very well structured.This is the same for Indian legal case documents. One of the major challenge we face is the difference in significance of legal words at different sentences. So because of the reason that different significance at different places we use deep learning methods for good performance of classifier model. Along with choosing deep learning models, we also finetune those models to get better results. For getting better values of parameters for our corpus, we use hit and trial method.

6.3 Future Direction of work

We show that deep learning models can much better identify rhetorical roles of sentences in legal documents which can in turn save a lot of time for both the court and legal person taking the cases.The principal advantage of neural models is that no hand-crafting of features is needed, hence expensive legal expertise is not essential. However, this property also poses difficulties in understanding why exactly a sentence is more likely to be assigned to one rhetorical role than the others.It also shows us the possibilty of being able to shorten the legal documents by summarizing the contents in it.

In future we are aiming on the tasks of summarization of legal documents using deep learning models.The tasks will consists of subtasks where one part will be summarizng the documents using only specific important information/sentences/words in the documents.The second subtasks will be a simple summarization of documents where a normal summary of the documents will be formed to get a basic understanding of the documents.

References

[1] J. Savelka and K. D. Ashley. Segmenting U.S. court decisions into functional and issue specific parts. In *Proc. JURIX, 2018.*

[2] O. Shulayeva, A. Siddharthan, and A. Z. Wyner. Recognizing cited facts and principles in legal judgements. In Artificial Intelligence and Law, vol. 25, no. 1, pp. 107–126, 2017.

[3] G. Venturi. Design and development of temis: a syntactically and semantically annotated corpus of italian legislative texts. In *Proc. Workshop on Semantic Processing of Legal Texts (SPLeT), 2012.*

[4] A. Z. Wyner, W. Peters, and D. Katz. A case study on legal case annotation. In Proc. JURIX, 2013.

[5] A. Wyner. Towards annotating and extracting textual legal case elements. *CEUR Workshop Proceedings, vol. 605, pp. 9–18, 01 2010.*

[6] M. Saravanan, B. Ravindran, and S. Raman. Automatic Identification of Rhetorical Roles using Conditional Random Fields for Legal Document Summarization. In Proc. International Joint Conference on Natural Language Processing: Volume-I, 2008.

[7] A. Farzindar and G. Lapalme. Letsum, an automatic legal text summarizing system. 2004.

[8] B. Hachey and C. Grover. Extractive summarisation of legal texts. In *Artificial Intelligence and Law, vol. 14, no. 4, pp. 305–345, 2006.*

[9] I. Nejadghoii, R. Bougueng, and S. Witherspoon. A semi-supervised training method for semantic search of legal facts in canadian immigration cases. In Proc. JURIX, 2017.

[10] P. Wang, Z. Yang, S. Niu, Y. Zhang, L. Zhang, and S. Niu. Modeling dynamic pairwise attention for crime classification over legal articles. In *Proc. ACM SIGIR, 2018.*

[11] P. Wang, Y. Fan, S. Niu, Z. Yang, Y. Zhang, and J. Guo. Hierarchical matching network for crime classification. In Proc. ACM SIGIR, 2019.

[12] Vanderbeck, S., Bockhorst, J., and Oldfather, C. Approach to Identifying Sections in Legal Briefs. In *MAICS 16-22. 2011.*

[13] Harasta, J., F. Kasl, J. Misek, and J. Savelka. Segmentation of Czech Court Decisions into Subtopic Passages. In CEILI Workshop on Legal Data Analysis JURIX, 2017.

[14] Grover, C., Hachey, B., Hughson, I., and Korycinski, C. Automatic summarisation of legal documents. In *Proc. 9th Int'l Conf. on Artificial intelligence and law 243-251. ACM. 2003.*

APPENDIX A

Biographical Sketch

Siddhartha Rusiya

Tulsi Nagar, Orai, Distt. Jalaun, Uttar Pradesh PIN-285001, e-Mail: siddhartharusiya84@gmail.com, Contact. No. +91-8423541688

- Pursuing B.Tech. in Computer Sc. & Engg. branch from N.I.T,Agartala with CGPA of 9.24/10.
- Intermediate from SVA International School, Orai under (C. B.S.E), Uttar Pradesh with 92% in 2017.
- High School from B.K.D. Aldrich Public School, Orai under (C.B.S.E), Uttar Pradesh with 9.6/10 in 2015.

Aditya Sharma

Bareth Road, Ganj Basoda, Distt Vidisha, Madhya Pradesh PIN-464221, E-Mail: 464221, Contact. No. +91-9829242482

- Pursuing B.Tech. in Computer Sc. & Engg. branch from N.I.T,Agartala with CPI of 8.34/10/00.
- Intermediate from Bharat Mata Convent sr. Sec. School, Ganj Basoda under (C.B.S.E), Madhya Pradesh with 91.2% in 2017.
- High School from Bharat Mata Convent sr. Sec. School, Ganj Basoda under (C.B.S.E), Madhya Pradesh with 10/10 in 2015.

Debajyoti Debbarma

Bishramganj,sepahijala,Tripura,PIN-799103, E-Mail: debajyotidebbarma55@gmail.com,
Contact. No. +91-7085519190

- Pursuing B.Tech. in Computer Sc. & Engg. branch from N.I.T,Agartala with CGPA of 7.71/10.
- Holy Cross School,Agartala (I.C.S.E), Tripura with 83.75% in 2018.
- St.Xavier's Higher Secondary School,sepahijala,Tripura under (C.B.S.E), with 10/10 in 2016.

Samarjit Debbarma

Ramhari para,Near Dayaram Para Hospital,PO-Golaghati,Bishalgarh,West Tripura,
PIN-799102, E-Mail: samarjitdebbarma816@gmail.com, Contact. No. +91-8787423644

- Pursuing B.Tech. in Computer Sc. & Engg. branch from N.I.T,Agartala with CGPA of 6.93/10.
- Intermediate from Modern HS School under (C.B.S.E), Tripura with 60% in 2018.
- High School from Kendriya Vidyalaya No.1, Kunjaban,Agartala under (C.B.S.E), Tripura with 8.4/10 % in 2015.