Role of Rhetorical Labels in Legal Domain : A Chronological Survey

[1]Ramyashree P M, [2] Dr. B G Prasad

[1]Student, [2]Professor, Department of Computer Science, B.M.S. College of Engineering
[1]Department of Computer Science and Engineering,
[1]B.M.S. College of Engineering, Bengaluru, Karnataka, India

***Abstract*—Annotation, a primary task in legal domain, which was initially as vague as just identifying and labeling entities such as name, place, date of proceeding and few other basic labels has gradually leveled up to rhetorical labels with explosive growth in Natural Language Processing(NLP). These labels intend to understand the context of sentences. As an interwoven part of semantic theory, the rhetoric in Law delve into the words used in legal proceedings to persuade or impress particular audiences. Identifying rhetorical roles of sentences in legal case documents can help legal practitioners in many ways. They can serve as better annotation strategies to various downstream tasks like case similarity checks, summarization, providing recommendations on specific precedents and language that may appeal to a given judge, etc. This paper aims to provide a chronological survey of the research works done on rhetorical role labeling on texts in legal case documents. The sum and substance of this paper are : How did the concept of rhetoric labels ingress into legal field, How did it evolve with the invent of natural language processing and How it is being worked upon currently with advanced models like deep neural networks.**

***Index Terms*—Legal text analytics, Rhetorical role, Semantic analysis, Sentence labeling models.**

I. INTRODUCTION

In a large democratic country like India, high courts alone hear anywhere between 20 and 150 cases a day, averaging at 70 hearings per day. Lawyers are some of the most hardworking people out there.They work between 60 and 80 hours weekly on average. The routine of a trial lawyer is excruciatingly writing and researching oriented. Much of the job involves drafting summaries, records of law, and motions. Litigators spend tons of hours on document hoarding and review, deciding if each of it is to be ceded to the court or to the opposite party. Legal profession can be intellectually demanding, but much of a barrister's work is in reality, mundane and repetitive. New lawyers, particularly those in large law firms, are often bludgeoned with monotonous tasks of document inspection, cite checking, and customary research. This is where Artificial Intelligence(AI) can make its grand entry. Contrary to the popular notion that AI is replacing humans, the roles assigned to computers and AI technology still remain within the realms of the mundane or repetitive. Thus in the field of Law, lawyers, judges and other legal field workers can aim to use AI to automate manual processes to become more productive as currently AI is far more unlikely to be assigned to less set-in-stone elements of the legal process such as critical and creative thinking, empathy, compassion, interpretation of gray areas, conflict resolution and negotiation and various other skills that requires humans to perform well in their role which AI simply cannot currently emulate. Law is a profession of words. It rejoices linguistic supremacy, and persuasive argument. Thus legal speech and writings follow a rhetorical structure quite distinguishable from that of a common communicative language that can be exploited to one's own benefit. Each sentence in a legal document can tell a lot about the case from the person handling it to the judge judging it. Training a machine to autonomously tie rhetorical roles to sentences in a legal case document makes the document more readable, leads to easy understanding of the notable points and further can also help law students in their law practice without having to go through the vast text and manually label them which saves their valuable time. While humans might find mastering the rhetoric of a language an uphill task that requires years of practice, a machine can do it with a shorter span of time with good training in the field of NLP.

Artificial intelligence is not a new concept in the legal world. Many ideas were put forth stressing the importance of AI in law in the early 1950s by Loevinger [1] Allen[2] and Mehl[3] where they have attempted justifying the need of mechanizing information retrieval and discussed approaches of using symbolic logics in form of mathematical notations to construct rules for working on litigation data. A more appealing work was proposed by Buchanan and Headrick [4] in 1970 clearly unraveling the problem behind meager progress in involving computer science in legal works. Suspecting misconception between computer scientists and lawyers about the potential in each other's discipline, they have pointed out the key for resolving this as indulging in interdisciplinary research that can align the design of machine methods in doing some of legal reasoning processes.

A few other eminent works followed in 1980s namely Carole Hafner's work on conceptual retrieval[5], Anne Gardner's work on contract law[6] and Rissland's work on legal hypotheticals[7]. In 1987, the International Conference on AI and Law (ICAIL), was instituted [8] which since then has been serving as a venue for publishing and developing ideas connecting AI and Law.

In this paper, the next section will be about the research methodology followed to gather the presented information, section 3 will discuss various works done particularly on rhetorical labeling and section 4 will be a conclusion discussing the advantages and disadvantages of works mentioned in the literature review section and future scope.

II. Research Methodology

1. Literature selection

An organized search has been carried out to identify various works done in legal domain on annotation based on rhetorical roles. The search includes Journal of Artificial Intelligence and Law (Springer), Journal of Machine Learning (Springer) and the International Conference on Artificial Intelligence and Law. The literature review presented in the next section follows a chronological system citing works starting from 1990s to the works as recent as 2021 and picking significant researches for detailed discussion.

2. Categorization

After scrutinizing the selected works, they were broadly classified into 2 categories.

i. Models using handcrafted features
ii. Models using automatic feature selection

III. Literature Review

i. Models using handcrafted features:

Teufel and Moens [9] in 1998 proposed a partially domain independent automatic abstract generator of research publications, where the abstract was composed based on rhetoric sentences. This they did in order to help researchers decide on the worthiness of articles they would refer related to their work of interest. They split this task into two parts; first to identify abstract-worthy sentences followed by classifying those into their rhetorical units (labels considered were *background, about, related work, purpose, solution, result and conclusion*). This work was found to place more importance on automating the first step and the next step of identifying rhetorical role of each sentence was less worked upon. They employed heuristic methods which included manually crafted features like indicator phrases and headers to construct an algorithm but the classification was not close in precision compared to manual annotation carried out by human judge. Nevertheless, they did not fail at laying a strong base for future works related to rhetorics.A follow up work was done again by Tuefel and Moens in 2002 [10] to automatically summarize scientific articles concentrating on rhetorical structures of the sentences. They crafted and used more profound feature and feature value sets like para structure - initial/medial/final, Length - shorter than threshold or longer, verb syntax - active or passive or non verb, etc., and experimented with naive Bayesian models in a supervised learning environment to classify sentences across seven rhetorical labels namely *Aim, Textual, Own, Background, Contrast, Basis and Other*. They evaluated their work using 3 measures - kappa, percentage accuracy and macro-F where macro-F gave the mean of the system's performance. Results indicated that the system was able to surpass a desirous baseline with kappa coefficient, K=0.71 compared to one human annotator and K=0.59 when compared to pool of annotators but the macro-F score showed that it was still about 20% lagging compared to human performance.Leveraging the methodology of Teufel and Moens (2002), Grover et al. [11] worked on sytem that could automatically summarize jurisdictional proceedings of House of Lords (HOLJ) in 2003, a first ever attempt in legal domain to use rhetoric sentences for the purpose of summarization challenging the 'popular-at-that-time' fact extraction technique. The summary of a scientific text, as worked upon by Tuefel and Moens, should hold such sentences which can convince its audience that the author has contributed to a particular field of science while in legal domain, the goal of the summary of a case proceeding is quite different. The author (a judge here) should be summarizing a case by taking into account all applicable points of law and convincing the audience that his/her decision is unblemished. Basing there work on this context, Grover et al. focused on understanding the correlation between lingual features and rhetorical roles. They used XML based tools like Text Tokenization Toolkit (TTT) and LT XML tool sets to analyse the case texts linguistically. ltpos program was used for part of speech (POS) tagging and fsgmatch to tell apart the verb clauses in each sentence and finding its tense, aspect, voice and modality features. They then hand annotated a small part of the corpus with the help of a human annotator for a preliminary analysis who designated one rhetorical label to each sentence (primary labels considered were *background, case* and *own*). Now with lingual information and rhetorical details available, they identified connections and patterns through which they were able to identify seven finer level of labels [12] - *fact, proceedings, background, proximation, distancing, framing* and *disposal*. This work showed that tense and aspect of the verb clause can be useful features for building a rhetorical sentence classifier.In a subsequent work in 2004, Hachey and Grover [13] explored further on their work on HOLJ corpus. They hand annotated 40 case documents with the newly identified seven labels and used it to train rhetorical sentence classifiers. They tried four different classifiers: C4.5 decision trees, Naive Bayesian(NB), Winnow algorithm, a linear classifier with error-driven learning and Sequential Minimal Optimization (SMO) algorithm trained Support Vector Machine, using default parameters settings to all classifiers. Briefly, the feature set included features such as: location of a sentence within the document, sentence length, whether the sentence contains words from the title, whether it contains a citation, presence of thematic words and cue phrases (Hachey and Grover, 2004) which were similar to Teufel and Moen's. What stands out is that they have simulated identifying a few cue phrase features automatically using the linguistic information acquired from their early work. They dabbled with various combinations of feature types to test and compare their classifiers. The details can be found in the article by Hachey and Grover [14]. In terms of micro-averaged F-scores, C4.5 yielded better results (65.4%) and SVM the next best with a score of 60.6% . NB was the third best (51.8%) making Winnow as the poor performer with 41.4%.With the increasing interest in the task of legal documents summarization based on rhetoric details, Farzindar and Lapame [15],[16] in 2004 described a method for summarizing the legal proceedings of Canadian federal courts and presenting them as a tabular-style summary. They carefully studied the composition of 50 case judgments and observed that the text had a specific thematic structure to it. Exploiting this, they used hand woven features like presence of important section headings, a segment's absolute and relative positions,determining whether the style of expression is direct or narrative, and few lingual indicators like verb classes and tense to label blocks of texts as their respective semantic/rhetoric roles. The roles identified were *Decision data, Introduction, Context, Juridical analysis* and *Conclusion.* The implementation version of

this concept was called LetSum which was developed in Java and Perl. They demonstrated a promising result with 90% accuracy for correct thematic segmentation. The work by M. Saravanan et al. [17] in 2008 being one of the breakthrough works on automating the identification of rhetorical roles in legal documents explores 3 strategies for segmenting legal documents specific to the sub-domains of rent control, income tax and sales tax., across seven rhetorical structures viz., *Identifying the case*, *Establishing the facts of the case*, *Arguing the case*, *History of the case*, *Arguments*, *Ratio of decidendi* and *Final Decision* (from Bhatia's, 1993) for the purpose of summarization of cases. As a base model, the researchers implemented 'SLIPPER', back then a popular rule learning algorithm that works on divide-and-conquer strategy, proposed by Cohen (in 1999). This method checks each rule in the rule set for each instance/ sentence. Drawback was that it took longer duration for larger corpora even for a two-class problem. Considering more than two classes and also aiming to avoid over-fitting of ensemble of rules, the researchers proposed an alternative rule based method that used "chain relation", a technique that identifies co-occurrences of roles in legal judgments. Here they considered rules as conjunctions of primitive conditions. A rule set R as hypothesis that partitions the set of instance X into any one of different rhetorical roles considered. Each rule outputs 1 if its condition is met, 0 if it is not met. Further improvising, they designed the third model [18] , a CRF model-based retrieval system, a novel approach depicting it in the way a human can summarize a legal judgment by understanding the significance of roles and associated contents. "Condition Random Fields (CRFs) are undirected graphical models used to specify the conditional probabilities of possible label sequences given an observation sequence. CRFs make first-order Markov independence assumption and thus can be understood as conditionally-trained finite state machines (FSMs)"(wikipedia) which are suitable for sequence labeling. They made use of features like cue phrases (eg, "We agree with court", "Question for consideration is", etc.,) Named Entity Recognition (checking for presence or absence of named entities like 'Supreme Court', 'Lower court') and legal vocabulary features (words that appear with capitalizations, affixes, and in abbreviated texts specifically in legal jargon) to reduce the complexity of legal domain. They manually annotated a small part of the corpus and used it to train the models. The F1 scores of all three models across different domains is shown in table 1. This work shows that CRF model with special features performs much better with satisfactory results than rule based and other rule learning methods in labeling the text for legal domains.

Table 1 Micro-Average of F-measure - From M.Saravanan et al.[17]

DOMAINS	SLIPPER	CHAIN RULE BASED	CRF
Rent Control	0.53	0.752	0.849
Income Tax	0.449	0.686	0.817
Sales Tax	0.407	0.637	0.787
Accuracy	0.407	0.637	0.787

In 2012, Giulia Venturi[19] designed and built "TEMIS" a grammatically and semantically annotated corpus which comprises texts from Italian legislative. It was a heterogeneous collection of texts exemplifying documents from three different releasing agencies: European Commission, Italian State and Piedmont Region, regulating a variety of domains, ranging from environment, human rights, disability rights to freedom of expression. This work solely centered on building a well annotated legal corpus, was inspired by the promising results achieved in bio-medical field where large corpora were annotated at different levels of analysis and also the unavailability of large annotated legal text corpora up until then which could be used as domain–specific resource. For the first level of analysis, the corpus was automatically dependency-parsed by the DeSR parser that used Support Vector Machines as learning algorithm. At this level, the research was targeted on the grammatical labels of the text. Further on a semantic level of analysis, a subset of the syntactically labeled TEMIS corpus was enriched with rhetorical annotations using a tool called SALSA that worked on FrameNet framework. Based on Frame Semantics, FrameNet is a lexical database for the English language. The goal of the FrameNet project is to document the range of semantic and syntactic combinatory possibilities of each word in each of its senses (like the word "lies" can mean different in sentences like "he lies to the court" and "he lies down on the grass") . These corresponds to Frame Elements (FEs), and is evoked by Lexical Units (LUs). The tool was fine-tuned with hand crafted features that represented the linguistic profile of the considered legislative texts that ranged from basic raw text features, such as sentence length, to more complex ones like the parse tree depth and identified roles like *Obligation_scenario, Being_obligated, Being_obligatory, Imposing_obligation, Grant_Permission, Prohibit_action* and *Deny_action.* The system was evaluated in terms of standard accuracy dependency parsing measure, i.e. labeled attachment score (LAS) of the parser. It achieved highest accuracy score of 79.30 on EU corpus (European Italian legal texts). Semantically annotated TEMIS corpus overtly devoted to be used for legal text processing applications based on NLP tools found its use for several semantic processing tasks

ii. **Models using automatic feature selection**

In 2017, a work by O.Shulayeva et al. [20] showed feasibility to automatically annotate sentences into their semantic roles using supervised machine learning framework based on linguistic features into 3 broader roles namely *facts, legal principles* and *neutral* on the idea that it will further help in case outcome prediction as facts similar enough to precedent cases should receive similar decisions as the precedents. They took aid of 2 annotators to create a gold corpus of 50 reports manually annotated which was used for training their machine classifier. The annotators labeled any statement which was used to reach a conclusion as *Legal Principle* (sentences have deontic modality i.e., terms like must, may, may not etc.); statements bearing on what

uncontroversially exists, occurred, or is a piece of information as *Facts* (eg. "Miss Anna was not a party to the 1986 Transfer…"); statements that did not belong to either of the roles as *Neutral* (sentences that were noted down as mere hypothetical outcomes that were self opinionated). Now for the automation, they used Naive Bayesian Multinomial Classifier based on a set of selected features that relied on automatic feature selection to prune the feature set. The features considered were - Part of speech tags, Unigrams, Dependency pairs, length of sentences, position in text and cit. Part of speech tags were extracted using NLTK (Natural Language Tool Kit, a powerful Python package), Unigrams followed bag-of-words approach, grammatical relations and dependencies were drawn out using Standford CoreNLP (a NLP toolkit offering Java-based modules for the solution of a plethora of basic NLP tasks) and other features were derived by means of a python script. A good Precision, Recall and F scores were seen across all the 3 classes along with an overall accuracy score of 0.85.Detailed result is presented in table 2. The dataset used in this work is relatively smaller in size compared to the usually humongous corpora used in legal analytics and the paper does not discuss the impact it might pose on accuracy in case the classifier is used on a larger corpus of legal data. Still, the proposed the Naive Bayesian Multinomial classifier identified 85% of instances correctly and thus provided a suitable basis for further work.

Table 2 Per Category and Aggregated Statistics for Automatic Classifier - From O.Shulayeva et al.[20]

	Precision	Recall	F Measure
Principles	0.823	0.797	0.810
Facts	0.822	0.815	0.818
Neither	0.877	0.892	0.884
No. Of sentences	2659		
Accuracy	0.85		

Another work in 2017 by Nejadghoii et al. [21] harnessed rhetorical role identification of sentences in order to aid a system that can find Cannadian immigration cases which held factual sentences matching a query input by an user. They manually annotated 150 documents into 8 roles - *Procedure, Fact, Party position, Issue, Analysis, Conclusion, Judgement for appellant* and *judgement for respondent.* They used this annotated data to further train various binary classifiers that could classify sentences as fact-vouching or otherwise. They used word embeddings like skip-gram model to grab semantic meanings of words. Since legal terminologies can hold different meanings compared to a general language, they also built word embeddings trained on immigration corpus. For this they employed fastText, an open source NLP library for learning word embeddings and text classification created by Facebook's AI Research team which comes with an advantage of having capability to provide vectors for Out-Of-Vocabulary words with character level embedding. They worked on six variations of binary classifier - i)SVM classifier with term frequency-inverse document frequency (tfidf) features which gave 81% accuracy; ii) SVM with legal vocabulary trained word embedding that showed 83% accuracy; iii)SVM with tfidf and trained embedding with 84% accuracy; iv) fastText supervised model with random initial embedding performing at 83% accuracy; v)fastText model with pre-trained word vectors by fastText that worked with an accuracy to 86% and vi) fastText classifier trained with immigration law word embeddings that spiked the accuracy to 90%. Though this work is seen achieving results with greater accuracy, it is limited to immigration dataset which required some domain specific customization. The goal of this next work being automating argument mining from decisions on adjudicated disability claims by U.S veterans for service-related post-traumatic stress disorder (PTSD), in 2019 V.R.Walker et al. [22] developed a qualitative methodology paired with quantitative testing for developing classifiers of sentences according to their rhetorical roles. The qualitative methodology included an extension of the semantic theory of attribution analysis the team employed which in the context of argument mining, is the descriptive task of determining which actor is asserting, assuming or relying upon which propositions, in the course of presenting reasoning or argument. They took an annotated dataset of U.S. decisions as the gold standard, and used a very small sub-sample of such decisions, and built protocols for each roles identified for developing rule-based scripts, and quantitatively tested the script performance. They also compared those outcomes against the performance of standard supervised ML models trained on larger samples from the same dataset. They worked on 5 main rhetorical roles : *Finding Sentence* (an authoritative conclusion, eg."the evidence fails to link the veteran's claim of having psychiatric disorder") , *Evidence Sentence* (testimony of a witness or description of other evidences, Eg."The examiner opined that the Veteran clearly had a preexisting psychiatric disability when he entered service."), *Reasoning Sentence* (sentences explaining credibility and probative value of the evidence submitted to court. Eg."the clinician's etiological opinions are credible based on their internal consistency and her duty to provide truthful opinions"), *Legal-Rule Sentence* (sentences that hold legal implications that needs to satisfied for the claim to be valid. Eg."Establishing direct service connection generally requires medical or, in certain circumstances, lay evidence of (1) a current disability; (2) an in-service incurrence") and *Citation Sentence* (standard notations as references to legal authorities or other materials. Eg."See Dalton v. Nicholson, 21 Vet. App. 23, 38 (2007); aff'd per curiam, 78 ,F.3d 604 (Fed. Cir. 1996)."). All the remaining sentences that did not belong to any of these 5 roles were labeled Other Sentence. For building rule-based classifier, they used Finding Sentences as critical connectors and represented a legal rule as a set of propositions with AND and OR logical connectives, one of which is the conclusion and the remaining propositions being the rule conditions which in turn had nested conditions. The entire representation was called a "rule-tree". For ML implementations, they preprocessed the dataset using NLTK's packages. Using CountVectorizer class of the Scikit-learn Machine Learning library as the feature extractor, they chose individual tokens in all the sentences and the bigrams and trigrams that appear in them as the features for training the 3 ML algorithms: Naive Bayes(NB), Logistic Regression(LR) and Support Vector Machines with linear kernel(LSVM). NB classifier was implemented using GaussianNB variant of Scikit-learn Python module. LR was implemented using log-linear and one-versus-the-rest approach and SVM classifier employed one-versus-one and voting scheme. For the ML classifiers, the experiment was done in 2 sets. Firstly a multi-class experiment (all 6

classes) and secondly a two-class experiment (labeling sentences as either Finding and Non-Finding). Each ML algorithm was run 10 times, where each run used a randomly chosen training subset that contained 90% of the labeled sentences. The trained classifier was then used to predict the labels for the remaining 10% of sentences. Average accuracy scores of the 3 ML algorithms for 2 sets of experiment is shown in table 3. This paper proves that qualitative study can be adopted for achieving promising results. On one hand, the test results give a hope that some access-to-justice use cases can be addressed at much lower cost than previously believed while on the other hand, the paper shows that most of the high accuracy scores appear to come from high precision score of one particular role among all other (Non Finding Sentences in case of two-class classification and Citation Sentences in case of Multi-Class Sentences). Thus one might infer that the classifiers are likely to introduce a number of false-positives to other roles.

Table 3 Average Accuracy Statistics of ML classifiers From V. Walker et al.[22]

Algorithms/Metrics	Multi-Class Accuracy	Two-Class Accuracy
NB	81.7%	93.4%
LR	85.7%	96.3%
SVM	85.7%	96.8%

Yet another research in 2019, as a novel approach, P Bhattacharya et al. [23] used Deep Learning models for automatically identifying rhetorical roles of sentences in Indian legal documents where no handcrafted features were needed. They also performed an extensive annotation study, curated a gold standard with the help of 3 annotators and analyzed the agreement between them, as well as the agreement of the model with the annotators. They treated this problem as a 7-class sequence labeling problem, where Deep Learning models were used to predict one label (rhetorical role) for every sentence in a document. The 7 labels identified were *Facts(FAC),Ruling by Lower Court(RLC), Argument(ARG), Statute(STA), Precedent(PRE), Ratio of the decision(Ratio) and Ruling by Present Court(RPA).* As part of pre-processing, Each document was split into sentences using the SpaCy tool. Each such sentence was considered a unit for which one label (out of the seven rhetorical roles) was to be predicted. They implemented baseline approach of CRF with hand crafted features and proposed two neural models: (1)Hierarchical BiLSTM Classifier and (2)Hierarchical BiLSTM CRF Classifier. In baseline approach, they treated each document as sequence of sentences and each sentence was represented as a vector of all selected features. They observed that some dependencies existed in the corresponding sequence of labels; (e.g., RLC usually followed FAC, RPC was always the end label.). Exploiting this nature of the document structure, they used Condition Random Fields (CRF) which are best suited for sequence labeling that work on probabilistic model along with hand crafted features like parts-of-speech tags, layout features, presence of cue phrases , and named entities. In Hierarchical BiLSTM Classifier, sequence of sentences were fed into BiLSTM layer, which returned a sequence of automatically extracted feature vectors. They initialized this BiLSTM layer with sentence embeddings using another biLSTM (they tried 2 variations : one being a randomly initialized word embedding and the other being a pre-trained sentence embedding sent2vec that was trained again on 53K court cases). As for implementing the Hierarchical BiLSTM CRF Classifier, they enriched the above model by constructing a CRF layer on top. This CRF is fed with the feature vectors generated by the top-level BiLSTM using which it assigned rhetorical label to sentences. They analyzed the Inter Annotator Agreement using F-measure than using the usually preferred measures like Kappa. It has an aggregated average F score of 0.83. For DL models, they wielded 5-fold cross validation with the 50 manually annotated documents. In each fold, they have considered 40 documents for training the model, and the other 10 documents for testing the performance of the model. They have used macro-averaged Precision, Recall and F-score metrics for evaluating the performance of proposed models. Results are shown in table 4.This paper shows that deep learning models can much better identify rhetorical roles of sentences in legal documents, compared to methods using hand-crafted features. The authors also provide a detailed annotation study.

Table 4 Macro Precision, Recall and F-score - From S. Ghosh et al.[23]

Category	Method	Variations	Precision	Recall	F-score
Baselines (CRF with handcrafted features)	Feature set from Teufel & Moens and Saravanan	-	0.5070	0.4358	0.4352
Neural models	Hier-BiLSTM	Pretrained emb	0.8168	0.7852	0.7968
		Random initialization	0.5358	0.5254	0.5236
	Hier-BiLSTM-CRF	Pretrained emb	0.8396	0.8098	0.8208
		Random initialization	0.6528	0.5524	0.5784

In 2021, Roberto Aragy et al. [24] worked on rhetorical role identification for Portuguese legal documents. This work specifically dedicated to petitions filed by the civilians to the Brazilian courts, explores various machine learning and deep learning

techniques to label sentences with semantic roles. Petition is the first document submitted by plaintiffs to file cases in courts. These documents will then be ruminated meticulously to pick out various aspects like parties involved, factual information, previous case proceedings if any, etc. With an aim to automate this process so as to smarten up Brazilian legal system efficiency, the research considered assigning rhetorical labels to sentences as a fine solution. They reflected on Brazilian civil procedure which states basic requirements of a petition and identified eight roles namely *identification of parties, facts, arguments, legal basis, precedents, requests, remedy,* and *others.* With the help of a legal expert, part of the corpus was manually annotated where each sentence was assigned one of the eight roles using an annotation tool called doccano. Treating it as a sentence classification problem, they first experimented with Naive Bayes(NB) and SVM as their baseline model. They preprocessed the data with NLTK library. They tried two variations of text representation on these model - Bag of Words(BoW) and term frequency-inverse document frequency(TFIDF) using skikit-learn library. Among the four baseline models, SVM with TFIDF was found to be best performing with an F-score of 60.66 which was only a tad bit better than NB with TFIDF whose F-score was 60.47. Next they experimented with BERT-based models. BERT is a recent revolutionary tool in the field of NLP. It is an advanced language model trained on massive amounts of data which uses transformer architecture. Using pretrained BERT model from Hugging Face library which were trained on Portuguese texts as a base layer , they trialed with two spinoffs. One coupled with linear layer and softmax, another with Multilayer Perceptron(MLP) and softmax. They used Keras library to train these models with default parameters. The top performing model was BERT with linear layer which yielded an F-score of 80.50 nearly 20 points higher than the best baseline model. The other model with MLP produced an F-score of 75.50 which as well proved better than baseline models. Some of the other noteworthy works that reaped the benefit of rhetoric roles in legal field are : work by P Bhattacharya et al.[25], an extension of their previous work, where they additionally experimented with United Kingdom supreme court case dataset using variations of transformer based models. Work by H Yamada et al. [26] for summarizing Japanese judgement documents using argumentation structure. Work by K Jasim et al. [27] on Arabic legal documents for detecting claims as part of argument mining using supervised learning with various binary classifiers and Work by Kavila et al. [28] where they propose a hybrid system for text summarization in legal domain which uses key phrase matching technique, to name a few.Next, a table is presented to give a bird's-eye view of all the elaborately discussed works which lists Authors, Conference and the place it was held at, the domains worked upon, techniques used, rhetorical roles identified and the results of the best performing models.

Table 5 A Bird's-eye View of All Works Discussed

Method	**Domains**	**Basic Technique**	**Rhetorical Roles identified**	**Evaluation Metric and Results of the best performing model**
Teufel and Moens, 1998 Madrid	General research publications	Hand crafted features : indicator phrases and headers; heuristic method	*background, about, related work, purpose, solution, result* and *conclusion*	-
Teufel and Moens, 2002, Cambridge	Scientific articles	Advanced Hand crafted features : indicator phrases, header, para structure - initial/medial/final, Length - shorter than threshold or longer, verb syntax - active or passive or non verb; naive Bayesian model.	*Aim, Textual, Own, Background, Contrast, Basis* and *Other*	Kappa co-eff, K=0.71
Grover et et.al., ACL 2003, Canada	House of Lords Judgements of Civil law cases of UK	Feature set from Teufel and Moens; XML based tools	*backgound, case* and *own*	-
Hachey et.al., ACL 2004, Barcelona Spain	House of Lords Judgements of Civil law cases of UK	Hand crafted features : location, sentence length, cue phrases, thematic words, citations; C4.5, NB, Winnow and SVM.	*fact, proceedings, background, proximation, distancing, framing* and *disposal*	Micro avg F score, C4.5-65.4
Farzindar and Lapame, ACL 2004, Barcelona Spain	Canadian federal courts	hand woven features : section headings, a segment's positions, style of expression ,verb classes and tense ; Thematic structure of document.	*Decision data, Introduction, Context, Juridical analysis* and *Conclusion.*	Accuracy, 0.90

Method	Domains	Basic Technique	Rhetorical Roles identified	Evaluation Metric and Results of the best performing model
Saravanan et.al., JURIX 2008, Florence, Italy	Rent Control, Income Tax, Sales Tax	Features: Indicator / cue phrases, Label transition featuresRule learning algorithms and Conditional Random Fields (CRF) -, avg F measure - 0.78	*Identifying the case, Establishing the facts of the case, Arguing the case, History of the case, Arguments, Ratio of decidendi* and *Final Decision*	Accuracy, 0.78
Giulia Venturi, SPLeT 2012, Istanbul, Turkey	Environment, Human rights, Disability rights to freedom of expression	Features : basic raw text features such as sentence length, to more complex ones like the parse tree depth ;FrameNet framework, Lexical Units (LUs)	*Obligation_scenario, Being_obligated, Being_obligatory, Imposing_obligation, Grant_Permission, Prohibit_action* and *Deny_action.*	LAS, EU Corpus - 79.3
O. Shulayeva et.at., ICAIL 2017, London, UK.	British and Irish Legal cases	Features (automatic): Part of speech tags, Unigrams, Dependency pairs, length of sentences, position in text and cit ; supervised machine learning: Naive Bayesian Multinomial Classifier.	*Legal Principle, Facts* and *Neutral*	Accuracy, 0.85
Nejadghoii et.al., JURIX 2017,Luxembourg	Immigration	Word embeddings, Skip gram ; SVM and fastText classifiers	*Procedure, Fact, Party position, Issue, Analysis, Conclusion, Judgement for appellant* and *judgement for respondent*	Accuracy, fastText with pre-trained embedding - 0.90
Walker et.al., ICAIL 2019, Montrael, Canada	U S vetrans legal claims for post-traumatic stress disorder	Scikit-learn library for automatic feaure extration : bigrams and trigrams ; Comparing rule-based & ML methods- NB, LR, LSVM.	*Finding Sentence, Evidence Sentence, Reasoning Sentence, Legal-Rule Sentence* and *Citation Sentence*	Accuracy, SVM and LR for multi-class - 85.7% SVM for Two-class - 96.8%
P Bhattacharya et.al., JURIX 2019, Madrid, Spain	Legal judgments from the Supreme Court of India	Word and Sentence embeddings, Automatic feature extraction by BiLSTM; Base Model - CRF with hand crafted features. Neural net models - Hier-BiLSTM and Hier-BiLSTM-CRF;	*Facts(FAC),Ruling by Lower Court(RLC), Argument(ARG), Statute(STA), Precedent(PRE), Ratio of the decision(Ratio)* and *Ruling by Present Court(RPA).*	F-score, Hier-BiLSTM with pre-trained embedding - 0.82
Roberto Aragy et al., BRACIS 2021, Sao Paulo, Brazil	Portuguese Petition Documents	NLTK and Keras libraries for preprocessing and training; Base Model - NB and SVM with Bow and TFIDF Proposed Models - BERT-MLP and BERT-Linear	*identification of parties, facts, arguments, legal basis, precedents, requests, remedy*, and *others*	F-score, BERT with linear layer - 80.50

IV. Conclusion

The advantage of Rhetorical roles in Legal domain has been leveraged since two decades and is still being actively worked upon by experimenting with new advents in NLP. The two broad categories of models discussed in this paper have their own advantages and drawbacks. The models with hand crafted features are coherent as the scientists who build the features have a complete sense and control on the features that are influencing their models. On the downside, hand-crafted features are largely dependent on legal-expert knowledge which is expensive to obtain. Contrarily the models implementing automatic feature selection, uses machine learned features that have no real world interpretation and eliminates the need of expensive legal expertise but, they become opaque in terms of explaining as to why a particular sentence is classified as that specific role and not other. Thus, transparency is compromised in exchange for eliminating the trouble of manually engineering features.

The concept of Rhetorical roles ought not be restricted to legal domain but can find its use in various other fields like education to evaluate answer scripts, sports to categorize commentaries on sportsperson, medical to annotate doctor's notes on patients etc., It can also be taken a notch up and used in speech recognition researches.

REFERENCES

[1] L. Loevinger, "Jurimetrics--The Next Step Forward", in *Minnesota Law Review*, 1948, pp.3-41

[2] A. Layman, "Symbolic logic: A razor-edged tool for drafting and interpreting legal documents," Yale LJ 66, 1956, pp.833-879.

[3] L. Mehl, "Automation in the legal world: from the machine processing of legal information to the" Law Machine," Mechanisation of Thought Processes (2 vols), London: HMSO,1958.

[4] G. Buchanan Bruce and E Headrick Thomas, "Some speculation about artificial intelligence and legal reasoning," Stanford Law Review, 1970, pp.40-62.

[5] D. Hafner Carole, "Representing knowledge in an information retrieval system," Information Retrieval Research, London, 1981.

[6] A. Gardner, "The design of a legal analysis program," AAAI-83, 1983.

[7] L. Rissland Edwina, "Examples in Legal Reasoning: Legal Hypotheticals", IJCAI, 1983.

[8] List Archived December 17, 2014, at the Wayback Machine of past ICAIL conferences up until 2014. http://www.iaail.org/page/past-icails

[9] S Teufel and M Moens, "Sentence extraction and rhetorical classification for flexible abstracts," In AAAI Spring Symposium on Intelligent Text summarization, 1998, pp. 16-25.

[10] S. Teufel and M. Moens, "Summarising scientific articles-experiments with relevance and rhetorical status," Computational Linguistics, 28(4), 2002, pp.409--446.

[11] C. Grover, B. Hachey, and C. Korycinski, "Summarising legal texts: Sentential tense and argumentative roles," in Proceedings of the HLT-NAACL 03 Text Summarization Workshop, 2003, pp. 33-40.

[12] C. Grover, B. Hachey,, and I. Hughson. "The HOLJ Corpus. Supporting summarisation of legal texts," in Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora,Geneva, Switzerland, 2004, pp. 47-54.

[13] B. Hachey, and C. Grover, "A rhetorical status classifier for legal text summarisation," in Text Summarization Branches Out, 2004, pp. 35-42.

[14] B. Hachey, and C. Grover, "Sentence classification experiments for legal text summarisation," in Proceedings of the 17th Annual Conference on Legal Knowledge and Information Systems (Jurix), 2004.

[15] A. Farzindar and G Lapalme, "Legal text summarization by exploration of the thematic structure and argumentative roles," in Text Summarization Branches Out, 2004, pp. 27-34.

[16] A. Farzindar and G Lapalme, "LetSum, an automatic Legal Text Summarizing," in Legal Knowledge and Information Systems: JURIX 2004, the Seventeenth Annual Conference, vol. 120, IOS Press, 2004, pp.11.

[17] 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.

[18] M. Saravanan and B. Ravindran. "Identification of rhetorical roles for segmentation and summarization of a legal judgment," Artificial Intelligence and Law, 2010, pp. 45-76.

[19] G. Venturi, "Design and Development of TEMIS: a Syntactically and Semantically Annotated Corpus of Italian Legislative Texts," in Proceedings of the LREC 2012 4th Workshop on Semantic Processing of Legal Texts, 2012, pp. 1-12.

[20] O. Shulayeva, A. Siddharthan and A. Wyner, "Recognizing cited facts and principles in legal judgements," in Artificial Intelligence and Law 25(1), 2017, pp.107-126.

[21] I. Nejadgholi, R. Bougueng, and S. Witherspoon. "A Semi-Supervised Training Method for Semantic Search of Legal Facts in Canadian Immigration Cases," In JURIX, 2017, pp. 125-134.

[22] V. R. Walker, K. Pillaipakkamnatt, A. M. Davidson, M. Linares, and D. J. Pesce, "Automatic Classification of Rhetorical Roles for Sentences: Comparing Rule-Based Scripts with Machine Learning," in Proc. Workshop on Automated Semantic Analysis of Information in Legal Texts (with ICAIL), 2019.

[23] P. Bhattacharya, Paheli, S. Paul, K. Ghosh, S. Ghosh, and A. Wyner, "Identification of Rhetorical Roles of Sentences in Indian Legal Judgment," in Legal Knowledge and Information Systems: JURIX 2019: The Thirty-second Annual Conference, IOS Press, 2019, Vol. 322. pp. 3-12

[24] R. Aragy, E. R. Fernandes, and E. Norberto Caceres, "Rhetorical Role Identification for Portuguese Legal Documents," in Brazilian Conference on Intelligent Systems, 2021, pp. 557-571.

[25] P. Bhattacharya, Paheli, S. Paul, K. Ghosh, S. Ghosh, and A. Wyner, "DeepRhole: deep learning for rhetorical role labeling of sentences in legal case documents," Artificial Intelligence and Law, 2021 pp.1-38.

[26] Yamada, Hiroaki, S. Teufel, and T. Tokunaga, "Building a corpus of legal argumentation in Japanese judgement documents: towards structure-based summarisation," Artificial Intelligence and Law, 2019, pp.141-170.

[27] Jasim, Khudhair, A.T. Sadiq, and H. S. Abdullah. "A Framework for Detection and Identification the Components of Arguments in Arabic Legal Texts," in 2019 First International Conference of Computer and Applied Sciences (CAS), 2019, IEEE, pp. 67-72.

[28] S.D. Kavila,P. Vijayasanthi, G. S. V. Prasada Raju, and R. Bandaru. "An automatic legal document summarization and search using hybrid system," in Proceedings of the international conference on frontiers of intelligent computing: Theory and applications (FICTA), 2013, pp. 229-236.