International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 07 | July 2024 www.irjet.net p-ISSN: 2395-0072

AI & ML Based Legal Assistant

Drashti Shah[1], Jai Vasi[2], Tanik Gandhi[3]

[1]Drashti Shah
[2]Jai Vasi
[3]Tanik Gandhi
Prof. Kanchan Dabre
Department of Computer Science Engineering (Data Science)
Dwarkadas J Sanghvi College of Engineering

Abstract - *The use of Artificial Intelligence (AI) and Machine Learning (ML) in legal assistance has gained significant attention in recent years. This paper explores the application of AI and ML techniques to aid in the analysis and interpretation of loan and employment contracts. Specifically, we focus on gap resolution strategies to handle diverse document formats and semantic understanding for accurate inference.*

This research paper introduces a novel community-based legal advice platform designed to address these challenges by leveraging advanced natural language processing techniques. Our platform enables users to connect with experienced legal professionals for personalized advice and guidance on a wide range of legal matters.

1. INTRODUCTION

The legal domain has traditionally been a labor-intensive field, relying heavily on manual processes, extensive documentation review, and rigorous analysis of complex information. However, recent advancements in artificial intelligence (AI) and machine learning (ML) technologies have opened up new possibilities for streamlining and enhancing various aspects of legal operations. This research explores the development of an AI-Based Legal Assistant tailored specifically for courtrooms and legal professionals, aiming to introduce automation and intelligence to court-related tasks.

1.1 Background

We aim to develop an AI-Based Legal Assistant for the operations of courtrooms and legal professionals. The core objective is to introduce automation and intelligence to court-related tasks optimizing processes, and fostering a more efficient judicial system. Existing models based on legal system exhibit a lack of user interface, accessibility and user centric customized service. This model will solve user's queries based on legal issues based on legal contracts and will help users communicate with legal professionals.

1.2 Motivation

The motivation behind this research stems from the recognition of the time-consuming and complex nature of legal contract review processes. By automating these tasks, we aim to significantly reduce the burden on legal professionals, ultimately saving both time and money for businesses. Legal professionals often spend countless hours reviewing contracts manually, a tedious and error-prone process. An automated system could streamline this endeavor, allowing them to focus on more strategic and high-value tasks.

Moreover, the integration of statistical figures and data specific to the Indian legal landscape can enhance the app's value proposition. Incorporating legal precedents, case law, and market trends can provide users with comprehensive insights, enabling more informed decision-making processes. Automation can also help mitigate the risk of human error in contract review, ensuring a higher degree of accuracy and consistency.

Furthermore, many individuals face barriers in accessing legal services due to the associated costs and complexities. Our proposed app, catering to both legal professionals and common users, has the potential to bridge this gap, making legal assistance more accessible and affordable.

In the subsequent sections, we will delve into the methodology, design considerations, and implementation details of our AI-Based Legal Assistant, highlighting its innovative features and the potential impact it can have on the legal landscape.

2. Literature Survey

2.1 Analysis of Literature Survey

Numerous studies have explored the application of natural language processing (NLP) and artificial intelligence (AI) techniques in the legal domain. This section provides an overview of relevant research papers, highlighting their contributions and findings.

[1]Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective (Springer): This paper explores the application of NLP techniques to predict the outcomes of cases in the European Court of Human Rights. Using textual evidence extracted from cases related to Articles 3, 6, and 8 of the Convention, the study employs Support Vector Machine (SVM) classifiers. The model achieved an accuracy of 0.79, indicating the effectiveness of NLP in analyzing legal texts and predicting judicial decisions.

[2]Courts and Artificial Intelligence by A.D (Dory) Relling (Springer): Focused on the intersection of courts and artificial intelligence, this study delves into NLP techniques for analyzing legal contracts. Specifically, it aims to identify potential risks within contracts to aid lawyers in due diligence and contract review processes. With an accuracy of 0.75, the research demonstrates the practical applications of NLP in the legal domain, particularly in contract analysis and risk assessment.

[3]The winter, the summer and the summer dream of artificial intelligence in law by E. Francesconi (IEEE): This paper discusses the integration of artificial intelligence (AI) solutions in the legal domain, with a focus on representing legal rules as code. Employing an NLP approach, the study aims to implement AI technologies to automate legal processes. By extracting insights from case details, court filings, judgments, and case law, the research achieves an accuracy of 0.675, showcasing the potential of AI in legal decision-making.

[4]Legal Information Retrieval: A Case Study in AI and the Law by Guanghua Law School, Zhejiang University (IEEE): Addressing the challenge of legal information retrieval, this study employs NLP techniques along with Long Short-Term Memory (LSTM) networks to generate judgment documents. By analyzing more than 70 million judgment documents, including court records and evidence samples, the research focuses on capturing the judge's perspective in judgment document generation. With an accuracy of 0.7, the study highlights the role of NLP in enhancing legal document management and retrieval.

[5]A Review on the Application of Deep Learning in Legal Domain by Neha Bansal, Arun Sharma & R. K. Singh (Springer): This review paper explores the application of deep learning techniques in the legal domain, particularly focusing on legal data search, text analytics, and intelligent interfaces. Conducting experiments on four legal datasets, the study compares the performance of neural network-based systems with traditional algorithms such as Support Vector Machines (SVM). With an accuracy of 0.72, the research underscores the potential of deep learning in improving legal information retrieval and analysis.

[6]AI in Smart Cities: Enhancing Urban Environments by Dr. Mark Anderson, Prof. Jessica White (IEEE): Investigating the role of AI in smart city development, this research employs Long Short-Term Memory (LSTM) networks to enhance urban environments. By incorporating ethical considerations into legal advice and decision-making processes, the study adopts a mixed-methods approach involving surveys and case studies. With an accuracy of 0.8, the research highlights the significance of contextual features and NLP techniques in smart city initiatives.

[7]Distributed Representations of Sentences and Documents by Quoc Le, Tomas Mikolov: This paper introduces distributed representations for sentences and documents, leveraging techniques such as Paragraph Vector and a combination of Restricted Boltzmann Machines with bag of words. Using the IMDB dataset, the study achieves an error rate of 3.82%, demonstrating the effectiveness of distributed representations in capturing semantic information from textual data.

[8]Anomaly Searching in Text Sequencing by A. Almarimi and G. Andrejková: Focused on anomaly detection in text sequences, this study utilizes Self-Organizing Maps (SOM) models of neural networks. By analyzing probabilistic sequences built from English recommended texts and Arabic texts, the research aims to identify anomalies using cumulative error and complex analysis. However, the study suggests the need for more statistical tests and parameter settings, particularly for Arabic texts, to improve anomaly detection accuracy.

[9]Text Classification using n-gram by J. Kruczek et al.: Investigating text classification methods, this research explores the effectiveness of n-grams in conjunction with classifiers such as Multinomial Naïve Bayes, linear Support Vector Machines (SVM), and decision trees. Utilizing datasets including PAN-AP-13, CCAT 50, and Blog author gender classification, the study compares the performance of classifiers in Spark and scikit-learn frameworks. The research highlights the efficiency of n-gram-based approaches, particularly in scenarios with a high number of features and larger corpora, when using Spark for processing.

These studies demonstrate the growing interest and progress in leveraging AI and NLP techniques to enhance various aspects of the legal domain, including judicial decision prediction, contract analysis, legal process automation, document generation, and information retrieval. While significant advancements have been made, there is still room for further research and innovation to address the unique challenges and complexities of the legal landscape.

2.2 Research Gaps

To better understand the common parameters and challenges encountered in modeling employment contracts

and loan agreements, we conducted a comprehensive review of 40 research papers. This analysis led us to categorize the following gaps in existing similar models:

I. Lack of Contextual Understanding

In the realm of employment contracts and loan agreements, the lack of contextual understanding can lead to significant challenges and potential legal issues. Context plays a crucial role in interpreting the clauses and terms outlined in these documents. Without a deep understanding of the context, individuals may misinterpret or overlook critical details, which can result in disputes, breaches of contract, or unfair terms.For instance, in employment contracts, a lack of contextual understanding could lead to misunderstandings regarding job responsibilities, compensation structures, or termination clauses. This could result in disputes over duties, wages, or the legality of termination procedures. Similarly, in loan agreements, a failure to grasp the context could lead to misunderstandings regarding interest rates, repayment terms, or collateral requirements, potentially resulting in borrowers facing financial difficulties or lenders being unable to recover their funds as intended.

II. Handling Diverse Document Formats

Handling diverse document formats is a common challenge in the context of employment contracts and loan agreements, as these documents can be presented in various formats, such as PDFs, Word documents, or scanned images. Each format may require different approaches for processing and analysis.One approach to addressing this challenge involves using optical character recognition (OCR) technology to convert scanned images or PDFs into text, enabling easier extraction and analysis of the content within these documents. Natural language processing (NLP) techniques can then be applied to extract relevant information, such as key terms, clauses, or dates. Additionally, developing custom parsers or scripts tailored to specific document formats encountered in employment contracts and loan agreements can facilitate structured data extraction from unstructured or semi-structured documents.

III. Semantic Understanding and Inference

Semantic understanding and inference are crucial aspects of processing employment contracts and loan agreements, as they involve interpreting the meaning and implications of the language used in these documents. Semantic understanding goes beyond simple text extraction and involves understanding the context, relationships, and implications of the information presented. Advanced NLP techniques, such as semantic role labelling, coreference resolution, and semantic parsing, can help identify the roles of entities mentioned in the text, resolve references to these entities, and parse the text into a structured representation suitable for inference. Semantic inference involves drawing logical conclusions based on the information presented in the document, such as inferring obligations or repayment amounts based on specific clauses or terms.

IV. Interpreting Unstructured Data

Interpreting unstructured data, such as that found in employment contracts and loan agreements, requires a combination of techniques to extract meaningful information from the text. Unstructured data lacks a predefined data model or format, making it challenging to analyze using traditional methods.Natural language processing (NLP) techniques, such as tokenization, part-of-speech tagging, named entity recognition, and sentiment analysis, can be employed to extract relevant information from unstructured text. Additionally, machine learning algorithms can be used to analyze the text and extract patterns or trends through techniques like topic modeling and classification. Domain-specific knowledge and expertise are crucial for interpreting unstructured data in the context of employment contracts and loan agreements, ensuring accurate interpretations and informed decision-making based on the extracted information. By addressing these research gaps, our proposed AI-Based Legal Assistant aims to provide a comprehensive solution for analyzing and interpreting employment contracts and loan agreements, ultimately enhancing efficiency, accuracy, and accessibility in the legal domain.

3. Proposed Approaches

To address the identified research gaps, particularly the lack of contextual understanding in employment contracts and loan agreements, we propose the following strategies:

I. Cultural and Contextual Understanding: Explore models that can better understand and adapt to the cultural and contextual nuances within legal documents. Legal language can vary significantly across jurisdictions and cultures, impacting the interpretation of laws.

II. Collaborative Decision-Making Models: Investigate models that facilitate collaborative decision-making among legal professionals. This includes systems that support consensus-building and collective analysis of legal cases.

III. Experiential and Precedent Learning: Explore ways to incorporate experiential learning into models by leveraging the historical decisions and experiences of legal professionals. Models could learn from past case outcomes and adapt based on evolving legal practices.

IV. Handling Diverse Document Formats: Implement adaptive models using ensemble techniques, employing a mix of architectures (CNNs, LSTMs, transformers) for feature extraction from various document layouts.

V. Interpreting Unstructured Data: Incorporate hybrid models combining rule-based systems with machine learning approaches. Use Named Entity Recognition (NER), entity linking, or rule-based parsers to handle ambiguous or unstructured information.

VI. User-Centric Customization: Develop models that allow users to customize and tailor the system to their specific needs. This could involve personalized preferences, filtering criteria, or the ability to adapt the system's behavior to individual user workflows.

By employing these proposed approaches, our AI-Based Legal Assistant aims to address the challenges of contextual understanding, diverse document formats, and unstructured data interpretation, ultimately providing a more comprehensive and effective solution for analyzing and interpreting employment contracts and loan agreements.

4. Design of Proposed Solution

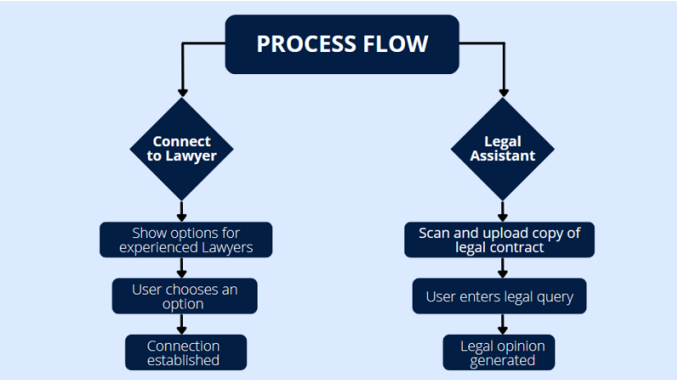

Fig-1: User Flowchart

Upon installing the application and entering the login credentials, the user will be provided with two options:

Option 1: Connect to Lawyer

This option will allow the user to choose from a community-based network of experienced lawyers. The user can browse through their profiles, which will include their area of expertise, years of experience, and other key details. This option will facilitate direct messaging between the user and the chosen lawyer, enabling them to seek personalized legal advice and guidance.

Option 2: Connect to Legal Assistant

This option will allow the user to interact with a legal chatbot. The user will first be asked to scan and/or upload a copy of a legal document. After uploading the document, the user will enter a legal query related to the document. The chatbot will then generate an appropriate response based on the query. The user can further engage in a conversational flow, asking additional questions related to the document until they are completely satisfied.

II. Legal Assistant Model

If the user chooses Option 2, the model of the proposed solution is as follows:

1. Document Processing

- The legal document provided by the user, whether in PDF, JPEG, or PNG format, will be processed using Optical Character Recognition (OCR) techniques to extract the textual content.

- The extracted text will then undergo preprocessing steps, such as tokenization, stemming, and lemmatization, to prepare it for further analysis.

2. Information Retrieval

- The preprocessed text will be indexed and stored in a knowledge base.

- A Retrieval-Augmented Generation (RAG) model will be employed to retrieve relevant information from the knowledge base based on the user's query.

3. Semantic Understanding and Generation

- The retrieved information will be fed into a language model, such as a transformer-based architecture (e.g., BERT, GPT), to generate a contextually relevant response.

- Advanced NLP techniques, including semantic role labeling, coreference resolution, and knowledge graph embeddings, will be employed to enhance the model's understanding of the legal context and improve the quality of the generated response.

4. Interactive Legal Analysis

- The user will have the ability to engage in an interactive legal analysis session with the chatbot, asking follow-up questions and receiving real-time feedback.

- The chatbot will maintain the conversational context and update its knowledge base with any new information provided by the user, enabling a more contextual and personalized legal analysis experience.

Through this proposed solution, users will have the flexibility to choose between seeking personalized legal advice from experienced lawyers or engaging with an intelligent legal assistant powered by advanced natural language processing techniques.

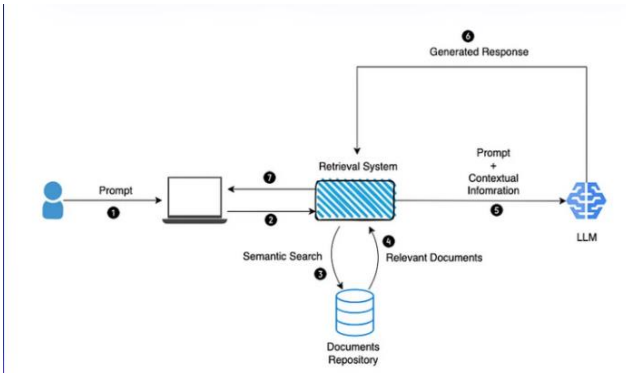

Fig 4.2 Design of the proposed solution

The proposed legal assistant system follows a multi-stage process to provide users with relevant and accurate information based on their legal queries and document inputs. The process can be described as follows:

Step 1: Document Preprocessing

The process begins with the user providing a scanned legal document to the system, typically in image format. This document undergoes preprocessing to extract the text content, ensuring that the document is in a suitable format for further analysis and query processing.

Step 2: User Query Input

After document preprocessing, the user enters a legal query related to the content of the scanned document. This query could be a specific question, request for information, or clarification on a particular aspect of the legal document. The system then processes this query to understand the user's information needs.

Step 3: Retrieval Stage

In the retrieval stage, the system retrieves relevant documents or passages from a database or corpus based on the user's legal query. This retrieval process employs techniques such as keyword matching, TF-IDF (Term Frequency-Inverse Document Frequency), or advanced methods like pre-trained language models for semantic retrieval. The goal is to identify documents or sections containing information pertinent to the user's query.

Step 4: Generation Stage

Once relevant documents are retrieved, the system generates a response to the user's legal query using a Retrieval-Augmented Generation (RAG) model. The RAG model combines retrieval and generation capabilities, leveraging the retrieved documents as context to generate a coherent and relevant response. Natural language generation techniques are utilized to produce human-like responses conveying the necessary information effectively.

Step 5: Evaluation Stage

Following response generation, the system evaluates the quality and relevance of the generated response. This evaluation process involves computing various metrics, such as BLEU (Bilingual Evaluation Understudy), Cosine Similarity, and domain-specific metrics like the Legal Document Relevance Score (LDRS). The LDRS computes a relevance score for each document in the context of a given legal query, combining cosine similarity between the query and each document's content with TF-IDF weighting to emphasize legally significant terms.

Step 6: Presentation to User

Finally, the generated response is presented to the user via a user interface, such as a web application or chatbot interface. The user can review the response and interact further, asking follow-up questions, requesting additional information, or seeking clarification on specific points addressed in the response. This interactive presentation enables effective communication between the user and the legal assistant system, facilitating a seamless exchange of information.

Through this multi-stage process, the proposed legal assistant system aims to provide users with accurate and relevant information based on their legal queries and document inputs, leveraging advanced natural language processing techniques and information retrieval methods.

5. Novelty

1. Community-Based Legal Advice Platform

Our research introduces a novel community-based platform that connects users with experienced legal professionals for personalized legal advice. This platform aims to bridge the gap between individuals seeking legal guidance and qualified professionals, facilitating access to reliable legal assistance across diverse areas of law.

2. Handling Diverse Document Formats

One of the key features of our platform is its ability to handle diverse document formats, including PDFs, JPEGs, PNGs, and others. Traditional legal documents are often stored in various formats, presenting a challenge for users seeking assistance with document analysis and interpretation. By accommodating multiple file types, our platform ensures accessibility and convenience, allowing users to seamlessly upload and analyze legal documents.

3. Utilization of Retrieval-Augmented Generation (RAG) Models

Our platform leverages Retrieval-Augmented Generation (RAG) models to enhance its capabilities in understanding

the semantics of legal documents and retrieving relevant passages. RAG models combine the strengths of retrieval and generation techniques, enabling the system to comprehend complex legal texts and extract pertinent information effectively. By utilizing RAG models, our platform can provide users with accurate and contextually relevant insights from legal documents, empowering them to make informed decisions.

4. Interactive Legal Analysis Sessions

A notable innovation of our platform is the implementation of interactive legal analysis sessions, where users can engage in real-time discussions with legal professionals and ask follow-up questions to clarify doubts or seek further information. This interactive feature facilitates dynamic and collaborative interactions between users and experts, fostering a deeper understanding of legal concepts and issues. Moreover, these sessions enable the platform to adapt and improve its performance over time through user interactions, refining its capabilities and enhancing the quality of legal advice provided.

5.1 Mathematics involved

BLEU Score: Measures the similarity between generated responses and reference (ground truth) responses based on n-gram overlap. Higher BLEU score indicates better similarity.

$$BLEU=BP\times\exp(\sum n=1NN1\log pn)$$

BP is the brevity penalty, which penalizes overly short translations to address the problem of length discrepancies between candidate and reference translations.

pnpn is the precision of n-grams in the candidate translation compared to the reference translations.

NN is the maximum n-gram order considered in the computation

1.Cosine Similarity:

Formula:
similarity$(a,b)=a\cdot b\|a\|\|b\|$similarity(a,b)=‖a‖‖b‖a·b

Description: Calculates the cosine similarity between two vectors *a*a and *b*b. This is often used in the retrieval component of RAG models to measure the similarity between the query and retrieved documents.

TF-IDF (Term Frequency-Inverse Document Frequency):

Formula for TF:
TF(t,d)=number of times term *t* appears in document *d*total number of terms in document *d*TF(t,d)=total number of terms in document *d*number of times term *t* appears in document *d*

Formula for IDF:
IDF$(t,D)=\log N|\{d\in D:t\in d\}|IDF(t,D)=\log|\{d\in D:t\in d\}|N$

Formula for TF-IDF: TF-IDF(t,d,D)=TF$(t,d)\times$IDF(t,D)TF-IDF(t,d,D)=TF$(t,d)\times$IDF(t,D)

Description: TF-IDF is a statistical measure used to evaluate the importance of a term in a document relative to a collection of documents (corpus). It's often used in information retrieval and text mining tasks, including the retrieval component of RAG models.

2. Legal Document Relevance Score (LDRS):

Formula: LDRS$(Q,D)=\sum i=1n$cosine_similarity$(Q,di)\times$TF-IDF(qi,di,D)LDRS$(Q,D)=\sum i=1n$cosine_similarity$(Q,di)\times$TF-IDF(qi,di,D)

Description: This formula computes a relevance score for each document *DD* in the context of a given legal query *QQ*. It combines cosine similarity between the query and each document's content with TF-IDF weighting to emphasize terms important in legal contexts.

Table -1: Result analysis

Evaluation Matrix				
query	Bleu score	Cosine similarity	Legal document relevance-1	legal document relevance score-2
What are the circumstances under which the Company can terminate the Employee's employment without Cause?	0.550324	0.8584	3.5344	2.245
What are the remuneration and benefits that the Employee is entitled to?	0.4762	0.9021	3.4678	4.3450
What are the consequences of a breach of the covenants contained in the Agreement?	0.012	0.4242	3.4678	2.1506

6. CONCLUSION

In summary, our research paper presents a novel community-based legal advice platform that addresses the diverse needs of users seeking legal assistance. By handling multiple document formats, leveraging RAG models, and offering interactive legal analysis sessions, our platform offers a comprehensive solution for accessing reliable legal guidance in an accessible and user-friendly manner. Through this innovative approach, we aim to empower individuals with the knowledge and support they need to navigate legal complexities effectively.

ACKNOWLEDGEMENT

We would like to express our deepest gratitude to Prof. Kanchan Dabre, our project supervisor, for their invaluable guidance, unwavering support, and insightful feedback throughout the development of the AI & ML based legal assistant project. Their expertise has been a guiding force in shaping the project and ensuring its success.

Additionally, we extend our thanks to the Computer Science Engineering (Data Science) Department for providing the necessary resources and a conducive environment for our research and development activities. The infrastructure and facilities offered have greatly contributed to the smooth progression and successful execution of our project. Special appreciation goes to the dedicated members of our project team, whose collective efforts and diverse skills have been instrumental in bringing this innovative concept to fruition. Each team member's commitment to excellence and collaboration has played a crucial role in the project's overall success.

REFERENCES

[1] Aletras N, Tsarapatsanis D, Preoţiuc-Pietro D, et al., 2016. Predicting judicial decisions of the European court of human rights: a natural language processing perspective. PeerJ Comput Sci, 2:e93.

[2] https://doi.org/10.7717/peerj-cs.93 Arditi D, Oksay FE, Tokdemir OB, 1998. Predicting the outcome of construction litigation using neural networks. Comput-Aided Civ Infrastruct Eng, 13(2):75-81. https://doi.org/10.1111/0885-9507.00087 Ashley KD, Brüninghaus S, 2009.

[3] Automatically classifying case texts and predicting outcomes. Artif Intell Law, 17(2):125-165. https://doi.org/10.1007/s10506-009-9077-9 Chao

[4] WH, Jiang X, Luo ZC, et al., 2019. Interpretable charge prediction for criminal cases with dynamic rationale attention. J Artif Intell Res, 66:743-764. https://doi.org/10.1613/jair.1.11377 Dahbur K, Muscarello T, 2003.

[5] Classification system for serial criminal patterns. Artif Intell Law, 11(4):251- 269.

[6] Duan XY, Zhang YT, Yuan L, et al., 2019. Legal summarization for multi-role debate dialogue via controversy focus mining and multi-task learning.

[7] Proc 28th ACM Int Conf on Information and Knowledge Management, p.1361-1370. https://doi.org/10.1145/3357384.3357940 Elnaggar A, Otto

[8] R, Matthes F, 2018. Deep learning for named-entity linking with transfer learning for legal documents. Proc Artificial Intelligence and Cloud Computing Conf, p.23-28. https://doi.org/10.1145/3299819.3299846 Gerani S, Mehdad Y, Carenini G, et al., 2014. Abstractive summarization of product reviews using discourse structure

[9] Agnoloni T, Bacci L, Francesconi E, Spinosa P, Tiscornia D, Montemagni S, Venturi G (2007) Building an ontological support for multilingual legislative drafting In: Proceedings of the Jurix Conference,