**NLP-Powered Legal Document Generation**

Ridit Jain1, Lavanya Anand1, Ashi Jain1, Akshat Manohar1

1  BML Munjal University, Computer Science Department, Gurgaon, Haryana, India

Contributing authors: [ridit.jain.21cse@bmu.edu.in](mailto:ridit.jain.21cse@bmu.edu.in); [lavanya.anand.21cse@bmu.edu.in](mailto:lavanya.anand.21cse@bmu.edu.in); [ashi.jain.21cse@bmu.edu.in](mailto:ashi.jain.21cse@bmu.edu.in); [akshat.manohar.21cse@bmu.edu.in](mailto:akshat.manohar.21cse@bmu.edu.in)

**Abstract**  
This paper covers the conceptualization and production of a NLP- impelled Document creating System having legal issues in import. The system is based upon the high NLP technologies that fully streamlines the process of writing and editing legal documents as well as their analysis and maintenance. The system under consideration merges these innovations into a user-friendly, solidimented platform, which is its objective is to revolutionize the style of legal documentation. The methodology outlined involves several key phases: User inputs registration, document template maintenance, data validating, reporting documents generation, and working loop feedback. By designing the three phases to cater for the unique challenges in legal documentation across distinct legal jargon, possible compliance risks and bureaucratic nature of traditional document drafting, they provide a simplified, efficient way to handle all legal documentation. The system not only improves the efficacy and accessibility in the process of legal document processing but also boosts the notion of educated and helpful legal documents, as a result the administration system seems clearer and easier to deal with for all users.

**Keywords** Natural Language Processing, Legal Document Automation, Document Generation, Dynamic Templates

**1 Introduction**

This projects puts an AI Document Generation System into play working in the legal domain. The system can take the profit of the latest NLP advancements and in this way NLP can solve the many issues related to legal documents. While integrating the system of technically advanced features into an online user-centric platform is the central goal of the initiative, this initiative targets the main problem in the conventional approach to legal documentation by creating a dramatically improved and superior version.

The purpose of the project will be to provide to the business community, legal community, and individuals with tools so that the legal documents can be generated, analyzed, and properly managed. We achieve this by offering customizable templates, which can be tailored to specific situations, and dynamic document generation based on inputting natural language inputs into the system. This simplifies complexity during legal documentation. The purpose is to simplify the process, allowing individuals to easily access the resources that are needed so that they would go with the legal processes.

This project addresses key challenges in the legal field: the tedious job of papers preparation, legal gobbledygook (which is something too complicated), and risk the compliance with the law. Through automation os these processes, it saves times and resources while contributing to enhanced document accessibility and clarity for both specialists and laymen.

The motivation stems from so many opportunities where NLP can become an invaluable tool for ensuring the aforementioned factors for legal document processing become more efficient, accurate, and accessible. Inspiration takes place from the experience of being part of the digital transfiguration of the legal industry and of possibility to help reduce the cognitive and time burden on law professionals and to make the legal services democratically accessible. It is my ambition to create an interactive, simplified document review system that makes legal understanding easy, brings legal matters close to everyone's doorstep, and provides a better understanding of the legal documents.

The main thing which makes project novel is the fact that it is holistic in that it is involves the booking of the entire challenge of legal document management. Unlike the current existing solutions which may be targeted to a specific aspect, our system is different in that it integrates into one integrated platform, a number of advanced natural language processing techniques.

The paper is organized such that various legal document processing operations will be examined in detail with an application of NLP technology. The first part begins with a familiarization of this project, highlighting the enabling participatory politics which we aim to achieve by using AI through the system of generation of documents. The second part, the literature review, focuses on applying available methodologies and highlighting spaces our system intends to address. Part 3 is about our methodology. After describing each stage within the system, from data acquisition to dynamic document creation and feedback mechanisms. The section 4 covers the system's performance and performance metrics, like computational as well as human make an analysis of the effectiveness and areas for improvement. In general, the fifth title reviews outcome of our finding, in which it provides the orientation of the future course and gives the recommendations for improvements on the systems that process legal document. To conclude in Section 6, all the arguments are summarized, and by highlighting the core contribution, a recommendation for further investigations and interpretations on this emerging topic is offered.

**2 Literature review**

The process of documents assembly is one of the most important steps in judicial and administrative procedures, thus one needs to commission experts to do the task or one needs to know the law themselves. This job can be quite frustrating for non-lawyers as well as professionals including law students and practitioners who have to deal with the complexities and specific aspects of legal writing. Generally, the process involves consulting precedent documents, or rather either using them in entirety or great help from more experienced colleagues so as to arrive at a final product. The essence of this process lies in its reliance on dual forms of knowledge: formal, resulting from legal education, which is structured to master, and practical, when a person gains it by his own experience. Nevertheless, the existing systems lack the capacity to articulate deterministic knowledge and do not support the deductive reasoning inherent to intricate legal issues.

Additionally, the exactitude required by the law in the preparation and use of legal documents is another significant element of the field that determines the successful of the whole system. These texts are influenced by a style of writing featuring specialized vocabulary and expressions known as legalese—which is unintelligible to the layperson—and, at the same time, requires advanced training to conceive and understand them. The macromolecular structure and the specific language of these bodied documents are utilized to provide transparency to the law by ensuring that it is impenetrable to any misunderstanding that may lead to exploitable loopholes.

Daniel Martin Katz and colleagues wrote a study that involved reading more than 600 legal texts to investigate the impact of NLP on the law. Their research, through the timeline of Legal NLP over the past decade, clearly expresses the development of Legal NLP over the past decade, namely, growth in the number of tasks performed, languages supported, and the data added to the fields knowledge [1]. Towards the future, the future paper "NLP-Machines Powered Automated Machines (LAMs) in India" shows the transformative influence of NLP in automation of India’s judiciary. It explains how NLP technologies can change the conventional processes of legal proceedings, taking into account the technological state of readiness of Indian judiciary, main impacts of this change, and the public acceptance of it [2].

Nay’s John’s research smartly uses both supervised and unsupervised learning models to tackle the tasks such as predicting the result of any given legal process and comprehension of sentiments, and the results of his research was very accurate with the overall accuracy rate above 80% [3]. Likewise, the study Julain Nyarko and Jens Frankenreiter undertake inserts these learning models as well as NLP techniques such as tokenisation, parsing and semantic analysis in relations to legal texts, is another example. The research team's work captured a wide range of techniques to facilitate better analysis and processing of legal texts using machine learning and NLP. Their approaches enable precision tasks, which is particularly helpful for document classification and named entity recognition and also helps to minimize the problem of legal term ambiguity and making efforts towards partly achieving comprehension and context over 90% and 85%, respectively [4].

The crucial need for Natural Language Processing (NLP) in the legal domain is better understood when looking at the usual approaches to document creation. Normally we make the dynamic templates that contain run-past points, which can be populated with data and extended to additional clauses and sentences as needed. This approach has been the subject of various studies, such as "Reusable Templates and Guides For Documenting Datasets and Models for Natural Language Processing and Generation: A Case Study of the HuggingFace and GEM Data and Model Cards" and "Learning Neural Templates for Text Generation." These papers criticise the current weaknesses in documentation of NLP datasets and models, thereby highlighting the urgent necessity of apostrophized documentation that could satisfy different data owners and users in both NLP and Natural Language Generation (NLG). They provide new methodologies which are important in overcoming these gaps [5][6].

Although the studies mainly pertain to the general document generation purpose in NLP, the work by Sam Wiseman, Stuart M. Shieber, and Alexander M. Rush in "Challenges in Neural Data-to-Document Generation" pinpoint the specific challenges of transforming data into comprehensible documents [7]. "Template Filling with Generative Transformers" by Xinya Du, Alexander M. Rush, and Claire Cardie delves deeper into the mechanics of document assembly. It highlights the need for not only filling out templates but also transforming user-provided information to meet standards and enhance comprehensibility, thus advancing automated document generation [8].

Uniquely, the paper "Knowledge-Based Legal Document Assembly" by Marko Marković and Stevan Gostojić closely examines only the legal sector. The paper suggests an approach that integrates law professionals’ expertise to a machine-readable format for more practical legal document creation. The legal application system administers complex statutory provisions during multiple stages. Firstly, it digitalized the traditionally legal norms and the professionals' expert knowledge into structured document. The rules beliefbase is developed next and the formal rule base which supports defeasible reasoning is developed and adopted that allows exceptions in legal judgments. In the synthesis phase, through the help of the interactive interface, the information is actually being applied to the documents, which are going to have an interactive component that guides the user through the process of data input. In the End this system is summed up in the web-based app performing all mentioned components and acting seamlessly for the user [9].

While the templates are filled by users with answers, those answers are not usually written in distinctive legal vocabulary and therefore should be turned into official legal language. There are many papers that do research on translating legalese into plain English. One of these papers is from Laura Manor and Junyi Jessy Li that uses unsupervised extractive summarization models to do their summarization This work has disclosed the blurred boundaries involved in the multi-faceted nature of legal language demonstrated by no ordinary standardized data set incorporating 446 contract section summary pairs, thus calling for a dramatic changes in the approach to continued legal texts clarifications [10]. On the contrary, there has been less written about adaptation of the regular English into legalese. "Lawyer LLaMA: Enhancing LLMs with Legal Knowledge" by Quzhe Huang and colleagues develops a novel framework namely "Lawyer LLaMA" which is intended for use in the legal domain. It describes a three-step process for domain-specific adaptation of large language models (LLMs): injecting domain knowledge, learning from expert data, and integrating relevant external information while excluding the irrelevant [11].

In addition to that, the method of proofreading should be dedicated to checking additional document items like grammar, writing style, or sentence structure, also to the standards running. To deal with this problems, Orlando Amaral Cejas, Muhammad Ilyas Azeem, Sallam Abualhaija, and Lionel C. Briand have come up with NLP-based algorithmic solution for easily Data Processing Agreements (DPAs) compliance checks under General Data Protection Regulation (GDPR). The methodology outlined in their study involves the creation of two primary artifacts: a set of "shall" essential requirements extracted from the provisions of GDPR that are genuinely critical for DPA requirements and the legal terms in these requirements explained in a glossary. The authors thus elaborate on the implementation for automated compliance checking that involves NLP technologies delivering phrasal level representations of DPA content and comparing these with predefined GDPR requirements [12].

Overcoming the hurdles that agencies are facing in deploying such systems, when supported by knowledgeable and experienced individuals, will enable the transition of complex legal processes to automation and facilitate legal work while maintaining an acceptable level of quality. These inventions, not for just improving the quality and accuracy of legal documents, but they are also bringing the extreme burdens of legal services to the area of accessibility, capability that will make every living people in the planet who do not have any experiences in this arena capable and conversant. To help with the complexities of legal language, the development of NLP and machine learning techniques is hugely important, which is key to providing legal documents with the required precision and clarity – a vital aspect in law.  The sophistication of these innovations becomes a transforming factor in the field of legal practices and accessibility, which makes it more significant to conduct a continuous research and development in this field.

**3 Methodology**

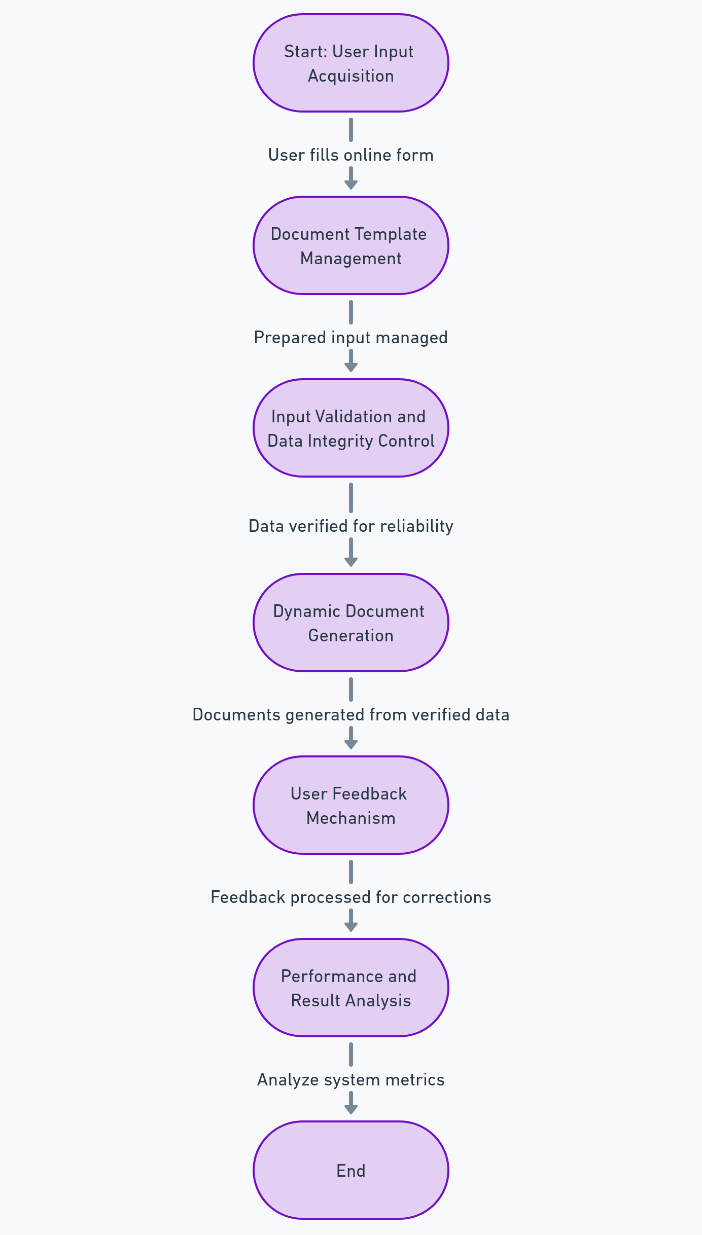


Figure 1. Flow diagram of methodology

**3.1 User Input Acquisition**

The initial stage is defined by capturing of user information via an online form accessible through a web interface, designed to be friendly and make ample utilization of interactive elements to enhance the user experience. This can be done by typing data, using downloaded reverse engineering and/or command line options. The system is subsequently be fed with this input and then it goes ahead to select and retrieve the right documents which will be filled. This step is indeed the most crucial activity in the formation of the logical structure for the documents to be generated

**3.2 Document Template Management**

Then, the system consumes the prepared input which is delivered to archive library stacked by modular LaTeX templates. Each template provides fields with variables and optional sentences integrated into the program code thus generating dynamic documents based on the user's input. Management of these templates involves identifying the user’s requirements and feeding it into the structure of the template. This procedure sits on the fundamental approaches of NLP relevant to the parser of the placeholders into the templates so as to link the produced reports with the thoughts of the users.

**3.3 Input validation and data integrity control.**

Protocols that focus on the reliability of the data entered by users for the system implement the verified mechanisms. This part also involves identification and authentication of named entities that encompass person names, dates, and places, all of which are prerequisites for the precise justification of documents. The identification part is associated with this entity, and then, it is checked against the predefined format and value to secure data integrity. Moreover, regular expressions and locations may be made use of to enforce adherence to legal and logical forms (for instance, validation of number inputs and date formats).

**3.4 Dynamic Document Generation**

Our approach is based on this technique which is known as dynamic document generation. Relying on the verified data, the system is able to generate the forms with the appropriate content. Additionally, to allow the users to add other clauses that need to be included, the system offers an interface where the users input the clauses in English. These inputs are then passed into a neural network-based system that carries out abstractive summarization by employing a custom-built legal dictionary that maps common terms to their legal equivalents. It is a necessity to carry out this transformation so the documents will be legally valid but also understandable and enforceable.

**3.5 User Feedback Mechanism**

The following step is to give a piece of electronic draft for the user’s review, through an online platform one can access the website. Users can make corrections which were missed or include the rest of the modifications. At the end, the document is subjected to one last spell-checking and grammar correction using NLP algorithms, in order to ensure that the document keeps a professional and a readable format. The program gives the suggestion of corrections in the interface which users have the right to either to reply in the favor of corrections or provide additional inputs to the correction given. Such loop is the key to reach the goal of 100% customer satisfaction and the highest data authenticity.

**4 Performance and Result**

The analysis of different data process systems like date recognition, name validation, monetary amount formatting, day validation, and email format validations have brought about the most important findings on their operation and some areas that need to improve. The date recognition system, while effective with common date formats, encounters difficulties with ambiguous formats, resulting in an accuracy of 0.7, precision of 0.714, recall of 0.679, and an F1 score of 0.694. These results indicate moderate effectiveness, compounded by an average edit distance of 3.9, which points to frequent misalignments in handling less standard date expressions. Name validation presents further challenges, particularly with unusual or complex names embedded in sentences, yielding an accuracy of 0.481, precision of 0.405, recall of 0.421, and an F1 score of 0.406. The mentioned number illustrates that we need to improve the entity recognition to get the name detection and formatting accuracy better. Conversely, the monetary amount formatting system shows robust performance with standard inputs, achieving an accuracy of 0.804, precision of 0.737, recall of 0.797, and an F1 score of 0.755. Despite the accurate handling it shows with the standard formats of numbers, it still fails to handle the sentences that contain commas placed at mistaken spots

or non-standard numerical inputs; thus the need for improvement to be able to handle a variety of formats is obvious. The date validation system demonstrates high reliability, effectively filtering out non-numeric inputs and values outside the 1-31 range, as evidenced by its high metrics: with the accuracy of 0. 979, precision of 0. 96, recall of 0. 958, and a F1 score of 0. 959. This provides it with the unique advantage of the applications whose day input needs to be authenticated in a strict manner. The email format validation system stands out with its high recall of 1.0 and an overall accuracy of 0.852, precision of 0.765, and an F1 score of 0.867. It successfully detects valid email structures however it may sometimes have inaccuracy because of misclassification of other non-standard but potentially valid formats which emphasize on a requirement of being more precise in differentiating between valid email format and invalid email formats.

Human validation, a multivariate metric that encompasses a series of tests, including quality assessment, coherency, and responsiveness, mirrors the human readable standards. The criterion involves

Table 1. Detailed Assessment of Text Quality Based on Linguistic and Translation Metrics

|  |  |  |
| --- | --- | --- |
| **Criteria** | **Value** | **Rationale** |
| BERT Similarity | 0.9398 | This clearly implies that the words used are almost identical and that the content touched on the respective contextual background while capturing the targeted audience. |
| Readability | Flesch: 46.51, FK: 12.9 | Readability level according to this is on the college level that is rather difficult for less educated people but good enough for originally educated people. |
| Translation Edit Rate (TER) | 0.895 | The greater TER reflect the fact that the system should process the sentence many times in order to be similar to the reference which shows lower translation quality or accuracy. |
| BLEU Score | 0.112 | The low BLEU score suggests some problems with the alignment to the reference translation, and thus implies the lexical accuracy or the vocabulary choice behind the translation is not that good. |

utilizing computational means that determine the degree of textual similarity, reading levels, compliance with

predefined rules, and presence of proper grammar. Each measurement poses as a critical indicator to making the communication strategy of the text more effective and in good agreement with the desired standards.

Evaluation of these systems provided indication of their benefits and drawback, giving lessons about the right design or how to technological enhancement can be done to ensure accuracy and function of the systems in different situations.

**5 Discussion**

The assessment of data processing systems such as date recognition, name validation, formatting of monetary amount, day validation, email format validation gives a detailed insight about them with the help metrics like it accuracy, precision, recall and F1 score. That makes it possible to evaluate the performance of systems to operate them in a manner that delivers data of different types correctly and quickly.

Date validation displays exceptional performance across all metrics, with an accuracy of 0.979 and precision of 0.96. These figures show very high accuracy and reliability, therefore it is an effective solution for systems that keep the standards high for input verification. In contrast, the name validation system struggles significantly, with poor accuracy (0.481) and precision (0.405). This error shows that the system is struggling to correct associating and formatting appellations of character properly, there is nearly half of the instances which are not processed well, which means that a higher advanced recognition must be implemented. The monetary amount formatting system shows good performance with an accuracy of 0.804 and a recall of 0.797, showing it can reliably identify correct monetary amounts. However, its precision of 0.737 marks the fact that we might deal with the problem of false positives, especially in case we are dealing with non-standard comma placement. The email format validation system being free of false misses gives 100% recall (no valid emails are missed) but the specificity of 0. 765 describes how it could misidentify correct formats sometimes and classify them as invalid. The fact of such a gap requires many-folded clarification and differentiation on real and fake email addresses.

The human-oriented validation metrics are the way relevant to this because the systems' effectiveness evaluation is supplemented with them. To ensure that the outputs of the systems correspond to natural language with a high accuracy and are relevant to the humans’ life by keeping it simple, quality assessment, coherency, and responsiveness are promoted. These evaluations determine whether or not the technical aspect, such as data integrity, coherence, as well as timeliness, is correct. They also help create high user satisfaction and increase the efficiency of the system. In addition to that, modern systems evaluate how likely are textual similarity, grammatical correctness, or rule compliance, quantifying the overall performance of the systems in human-oriented terms.

Through the process of this appraisal, all deficiencies for the betterment of the system highlight and then indicate what further optimizations are needed to make these systems more accurate and consistent in various applications. Besides, it is significant to compere all the metrics with standardised values so as to see the positives and minuses of the different systems to finally point at the right direction for the future improvements.

**6 Conclusion**

Legal documents automation and management will become one the important and significant tasks in the scope of legal documents generation system using the NLP. The application of the proper NLP techniques in the production of a comprehensive platform can be followed by the solution of many complex problems with legal documentation. This UI has simplified the interaction with the user who chose pre-built templates, and inputs the patient's medical data in natural language. It streamlines the process and makes the process seemingly accident free along with being highly acceptable to everyone in the ecosystem. It not only simplifies the process of drafting documents but also ensures that it is developed according to current laws and supports multiple languages. It in turn reduces the process for lots of other languages and hence it has a wider impact on the legal domain.

The successful implementation of this system provides an emerging picture of NLP technology and its capability to change legal practices, therefore, the future research and development in this field should be considered a promising direction. These systems are constantly being perfected and improved from one version to another and they eventually make justice more accessible and support greater digitalization in the legal sector.

**7 Acknowledgments**

The authors wish to express their deep appreciation to Dr. Atul Mishra for the wonderful consultancy and mentoring he provided throughout this research work. Knowledge of his and the contribution to the work was therefore extremely vital. The authors also gratify BML Munjal University for providing the academic facilities that ultimately aid in the success of this research study.

**8 References**

1. Katz, D. M., Hartung, D., Gerlach, L., Jana, A., & Bommarito II, M. J. (2023). Natural language processing in the legal domain. *arXiv preprint arXiv:2302.12039*.
2. Alexopoulos, C., Saxena, S., & Saxena, S. (2023). Natural Language Processing (NLP)-Powered Legal A (t) Ms (LAMs) in India: Possibilities and Challenges. *Journal of the Knowledge Economy*, 1-21.
3. Nay, J. (2021). Natural Language Processing and Machine Learning for Law and Policy Texts. In D. M. Katz, R. Dolin, & M. Bommarito, Legal Informatics. Cambridge University Press
4. Frankenreiter, J., & Nyarko, J. (2022). Natural Language Processing in Legal Tech. In D. Engstrom, Legal Tech and the Future of Civil Justice
5. McMillan-Major, A. a. (2021). Reusable Templates and Guides For Documenting Datasets and Models for Natural Language Processing and Generation: A Case Study of the HuggingFace and GEM Data and Model Cards. In Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics. Association for Computational Linguistics.
6. Wiseman, S., Shieber, S. M., & Rush, A. M. (2019). Learning Neural Templates for Text Generation.
7. Wiseman, S., Shieber, S. M., & Rush, A. M. (2018). Challenges in Data-to-Document Generation.
8. Du, X., Rush, A., & Cardie, C. (2021). In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, Y. Zhou, Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 909-914). Association for Computational Linguistics.
9. Gostojić, M. M. (2020). Knowledge-Based Legal Document Assembly.
10. Li, L. M. (2019). Plain English Summarization of Contracts.
11. Feng, Q. H. (2023). Lawyer LLaMA Technical Report.
12. Cejas, O. A. (2023). NLP-Based Automated Compliance Checking of Data Processing Agreements Against GDPR. IEEE Transactions on Software Engineering, 4282-4303.