**PROJECT PROPOSAL**

**Natural Language Processing (NLP) for legal texts**.

**Chair Committee:** Dr. Sanjay Modak

**Mentor:** Dr. Yannis Karamitsos

**Student Name:** Malik Umar Hassan

**Cohort:** **7**

**UID**: 776006946

GRAMMARLY

**Table of Contents**

[1. PROBLEM STATEMENT 3](#_Toc103791333)

[2. ABSTRACT 3](#_Toc103791334)

[3. BACKGROUND INFORMATION 4](#_Toc103791335)

1. PROBLEM STATEMENT……………………………………………………………………….3

2. ABSTRACT ………………………………………………………………………………………….3

3. BACKGROUND INFORMATION ……………………………………………………….....4

4. PROJECT DEFINATION & GOALS...………………………………………………………..4

5. LITERATURE REVIEW……………………………………………………………………………5

6. TAKEAWAYS FROM LITERATURE………………………………………………………….6

7. METHODOLOGY…………………………………………………………………………………..7

8. DATA ANALYSIS ………………………………………………………………………………….8

9. PROJECT RESOURCES**….**………………………………………………………………………8

10.PROJECT DELIVERABLES **……………**…………………………………………………….8

11. TIME LINES MILESTONE…………………………………………………………………….8

12. PROJECT BUDGET……………………………………………………………………………..8

13. PROJECT GAP…….……………………………………………………………………………..8

14. BIBLIOGRAPHY ………………………………………………………………………………….9

TABLE OF FIGURES

[Figure 1:Legal research process architecture, F, M. (2010), 6](#_Toc103791401)

# PROBLEM STATEMENT

Much of the information of interest to lawyers and legal scholars is in the form of texts. Whether they are briefs, contracts, court decision, law review articles, legislative acts, treaties, newspapers or blog posts, all are either legal documents themselves or documents about the law. Finding, analyzing, commenting on, relating to, and explaining these documents has been the daily operations of legal practice and legal scholarship alike for centuries.

The process of law making in Pakistan is very consuming as technology cannot be used in lengthy processes. In this project, methods will be used to explore how natural language processing (NLP) can be used to control and structure legal content generated unstructured legal texts (such as judgments, skeleton arguments, scholarly articles, Law Commission reports).

During the research found that material for practical implementation of this study was not found especially for courts and legal firms. On the basis of our study we need to create a complete application to ease the work of legal teams.

Then will develop a system that can predict whether a feature in the legal text is of interest or not

# ABSTRACT

A legal prediction-based approach will help judges and solicitors to take judicial decisions on current cases, which are going on in courts and make predictions on new cases on the basis of existing references and judgments. This model also helps law students learn about legal references. This application was developed specifically for the “Supreme Court of Pakistan (SCP)” and the “Pakistan Bar Council (PBC)” to expedite their judgments and provide legal guidance to lawyers based on historical data and constitutions.

The project will build on the concepts of NLP, ML and Business Intelligence (BI) using Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology. Real judgments uploaded on the official website of SCP in the form of Portable Document Format (PDF) and a structured database of law firms will used in this project. ~~In addition, the project will be divided into two main phases, which are mentioned as follows.~~

1. ~~Find out the information and knowledge from PDF(s) using text mining and transform it to structured data base.~~
2. ~~Create a Graphics user interface GUI for PBC lawyers to help them in case citation~~.

# BACKGROUND INFORMATION

The Supreme Court of Pakistan (SCP) is the apex court of Pakistan, whose verdicts and orders are binding to all other courts of country. All executive and judicial authorities are bound to act in aid of the Supreme Court [1]. Furthermore, the detail verdict of the court is presented in the form of unstructured data i.e. PDF to the general public and attorneys of The Pakistan Bar Council PBC. That judgement may contain few pages or may in the form of volumes of book.

In 2017, the highest court of Pakistan announced the verdict against the disqualification of Ex. Prime minister of Pakistan “Mian Muhammad Nawaz Sharif “, in contradiction of Panama gate case [2]. Initial reference and allegations was [money laundering](https://en.wikipedia.org/wiki/Money_laundering), [corruption](https://en.wikipedia.org/wiki/Corruption) on the basis of Panama Papers issued by the [International Consortium of Investigative Journalists](https://en.wikipedia.org/wiki/International_Consortium_of_Investigative_Journalists) (ICIJ). But disqualification happened on holding the visa of other country and working as a CEO of his son company during the tenure of his prime minister ship.

For understanding such details judgement insights this study will help public prosecutors to analyse any verdict in a few minutes using NLP, furthermore, the prediction model using ML will gives lawyers to predict the newly cases result on the basis of existing knowledge base.

1. **PROJECT DEFINITION AND GOALS**

This project will use NLP platforms that can examine a case study or document and suggest other analogous cases to notaries for further consideration. These references can help lawyers understand the pattern of a case more quickly and systematically. In addition, this model will comfort legal teams to recite, recognize and examine large amounts of documents, whether that’s during a felonious inquiry or a trade matter, this system will helped lawyers to spend less time trudging through documents and labelling key clauses, and instead focus more time on understanding what the documents mean, gaining the critical insights, and providing value advice to the client in a timely manner.

1. **LITERATURE REVIEW**

Dyevre (2021) argued that gauging legal documents such as statutes, treaties, judicial decisions and law review articles is a critical and time consuming job for any legal scholar and practitioner. Many machine learning and natural language processing techniques can apply on such unstructured data for assessment. Furthermore, text mining technique like topic modelling, word embedding and transfer learning can help them to ease their work in efficient way.

Moreover, Branting (2018) explains the special issues while doing legal text analytics; IAAIL (International Association for Artificial Intelligence and Law) emphasised more on empirical and corpora-based approaches rather than argumentation and interference.

F, M. (2010), proposed the architecture which automates the legal research process through data mining. In his study; he divided the legal search process into two main components namely “The mining process” and “The research process”. The mining process responsible for analysing each document and on the basis of information creates a law report repository. Information retrieval has done on the basis of text block analysing on each law report.

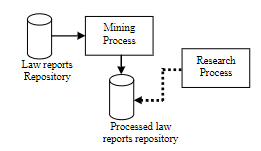


Figure 1:Legal research process architecture, F, M. (2010),

Ning (2022) suggest the study which make the judicial system artificially intelligent by using scientific research methods especially unsupervised and supervised machine learning techniques; unsupervised machine learning creates data sets without known outputs or predefined data; Word Cloud (visualization of most prominent or frequent words display in a body of text), Latent Semantic Analysis and Principal Component Analysis (a natural language processing method that analyses relationships between a set of documents and the terms contained within), Word Embedding (words that are close in meaning are grouped near to one another in vector space), Document Clustering with Word Embedding and Topic Modelling are the main concept were used in legal search analysis.

Novotna et al. (2020) describes topic modelling as process which recognizing the words from the topics present in the document or the corpus of data. This is useful because extracting the words from a document takes more time and is much more complex than extracting them from topics present in the document. Removing stop words and punctuation marks Stemming, Lemmatization, Encoding them to ML language using Count vectorizer or Tfidf vectorizer are important terms for topic modelling. Topic modelling is done using LDA (Latent Dirichlet Allocation). Topic modelling refers to the task of identifying topics that best describes a set of documents. These topics will only emerge during the topic modelling process (therefore called latent). And one popular topic modelling technique is known as Latent Dirichlet Allocation (LDA).

Lance (2020) categorized the sentiment analysis of legal documents into two parts and named that legal sentiment analysis and opinion mining (LSAOM), LSA describes the narratives and discussion, on the other hand LOM describes the identification and brightness of unambiguous or hidden opinion accompaniments immersed within legal discourse. Because of the messy nature of text, analysing, understanding, organizing, and sorting through text data is hard and time-consuming, so most companies fail to use it to its full potential.

Chhatwal (YEAR) explained text classification by predictive coding or technology assisted review (TAR) that can considerably enrich the overall quality and speed of the document review process by reducing the time it takes to review documents and help to categories the document into predefined clusters.

Zaidan et al. (2007) proposed a machine learning method to use annotated rationales in documents to boost text classification performance. Support vector machine SVM algorithm used in experiment and its results shows there is a drastic improvement in text classification.

Qader et al. (2019) explained supervised methods, which, do not require ex post validation because they seek to ‘emulate’ what humans do by discovering patterns in documents labelled by human annotators prior to training. Supervised approaches have these approaches Obtaining Labelled Documents, Bag-of-Words Methods and Transfer Learning and Transformers.

1. **TAKEAWAYS FROM LITERATURE**

This study explained the technology assisted review (TAR) of legal documents using natural language process and text mining techniques. Major part of the study discusses unsupervised machine learning techniques.

Unsupervised machine learning covers the Word Cloud, Latent Semantic Analysis & Principal Component Analysis, Topic Modelling and clustering. Furthermore, study also described the

STRONG STATEMENT :Legal sentiment analysis and opinion mining (LSAOM);

FUTURE WORK Final part of study describes the clustering of documents on the basis of legal text. Support vector machine and other clustering algorithm will used to implement this research.

**FOR PRESENTATION YOU MAKE A BULLET LIST**

1. **METHODOLOGY**

In this project we are going to use **CRISP- DM** methodology to implement for text mining and for prediction model.

CRISP-DM organizes the data mining process into six phases: business understanding, data understanding, data preparation, modelling, evaluation, and deployment. These phases help organizations understand the data mining process and provide a road map to follow while planning and carrying out a data mining project [9].

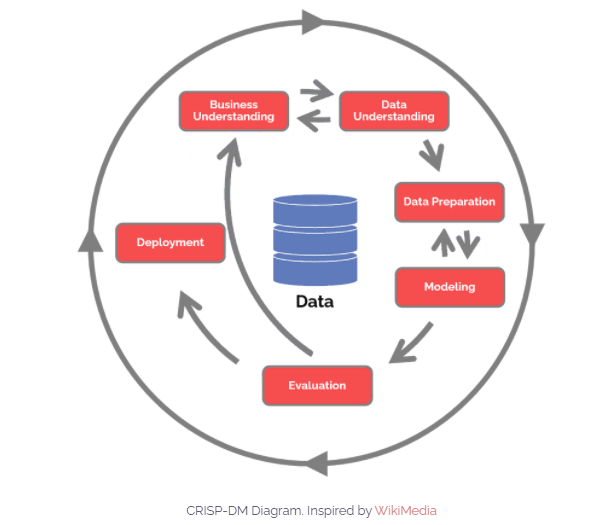


Figure 2: CRISP – DM Methodology

* **Business Understanding**

In this initial step we need to rectify what exactly issue(s) need to solve for business by gaining there business knowledge and true business understanding.

* **Data Understanding**

After gaining business understanding second step is to understand the data which customer have and rectifying the different sources of data and checking its quality.

* **Data Preparation**

This is most extensive steps and tends to make almost 80% of project time. Cleaning bad data, finding outliers, anomalies and duplicate in data are the main steps of data preparation.

* **Modelling**

Selecting machine learning models on the basis of business problem and data, building the test case. The performance of each test is then measured to see how well it performs against other potential models.

* **Evaluation**

Against the business objective result has been verified, after interpretation of results decision has be taken.

* **Deployment**

In this phase we have an option either project report will be created or software component.

1. **DATA ANALYSIS**

On following official website of The Supreme Court of Pakistan, PDF of detail judgment available

[**https://www.supremecourt.gov.pk/latest-judgements/**](https://www.supremecourt.gov.pk/latest-judgements/)

1. **PROJECT RESOURCES**

No cost required to accomplish these tasks. Open source tools like pycharm (IDE), python, NLTK, Streamlit libraries and Tableau will be used to find out the insights of these legal documents.

1. **PROJECT DELIVERABLES**

This case study supported to automate the time consuming legal and research work of SCP. The final web component consider as a framework with several text mining technique applied on it as well as speech to text recognition for judges while writing the citation.

In future enhancement this web component will capable of saving unstructured and semi structured data into relational data base for maintain the litigation history.

1. **TIME LINES MILESTONE**

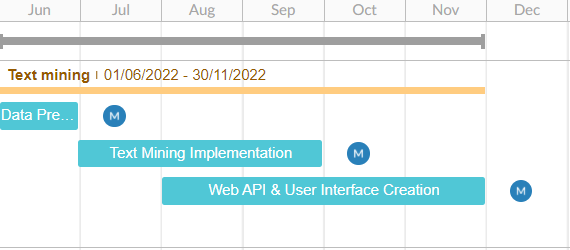


Figure 1.3 Milestones and Deliverables of Project

1. **PROJECT BUDGET**

No budget requires for this project.

**BIBLIOGRAPHY**

**Doi.org**

1. Supreme Court of Pakistan [**https://www.supremecourt.gov.pk/about/history**](https://www.supremecourt.gov.pk/about/history)**.**
2. The International Consortium of Investigative Journalists [**https://www.icij.org/**](https://www.icij.org/investigations/panama-papers/from-front-pages-to-prison-time-behind-the-scenes-of-a-panama-papers-criminal-case)**.**
3. F, M. (2010). Automating Legal Research through Data Mining. International Journal of Advanced Computer Science and Applications, 1(6). https://doi.org/10.14569/ijacsa.2010.010602
4. Conrad, J. G., & Branting, L. K. (2018). Introduction to the special issue on legal text analytics. In Artificial Intelligence and Law (Vol. 26, Issue 2, pp. 99–102). Springer Netherlands. <https://doi.org/10.1007/s10506-018-9227-z>

Conrad, J.G., Branting, L.K. Introduction to the special issue on legal text analytics. *Artif Intell Law* **26,**99–102 (2018). https://doi.org/10.1007/s10506-018-9227-z

1. Hanke, J., & Thiesse, F. (2017). LEVERAGING TEXT MINING FOR THE DESIGN OF A LEGAL KNOWLEDGE MANAGEMENT SYSTEM SERAMIS-Sensor-Enabled Real-world Awareness for Management Information Systems View project. https://www.researchgate.net/publication/320471617
2. Berosik, G., Chaudhary, M., Dozier, C., Atkinson, G., Guo, X., & Samler, S. (n.d.). Mining legal text to create a litigation history database.
3. Dyevre, A. (2021). Text-mining for Lawyers: How Machine Learning Techniques Can Advance our Understanding of Legal Discourse. Erasmus Law Review, 14(1). https://doi.org/10.5553/elr.000191
4. Ning, J. (2022). Natural Language Processing Technology Used in Artificial Intelligence Scene of Law for Human Behavior. Wireless Communications and Mobile Computing, 2022, 1–8. https://doi.org/10.1155/2022/6606588
5. Novotn´a, T., Novotn´a, N., Harašta, J., Harašta, H., & Ol, J. K. ´. (2020). Topic Modelling of the Czech Supreme Court Decisions.

<https://pypi.org/project/spacy-udpipe/>.

1. Qader, W. A., Ameen, M. M., & Ahmed, B. I. (2019). An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges. Proceedings of the 5th International Engineering Conference, IEC 2019, 200–204.

<https://doi.org/10.1109/IEC47844.2019.8950616>

1. Chhatwal, R., Gronvall, P., Huber-Fliflet, N., Keeling, R., Zhang, J., & Zhao, H. (n.d.). Explainable Text Classification in Legal Document Review A Case Study of Explainable Predictive Coding……..
2. Zaidan, O. F., Eisner, J., & Piatko, C. D. (2007). Using “Annotator Rationales” to Improve Machine Learning for Text Categorization \*. http://cs.jhu.edu/
3. Dr. Lance B. Eliot. (2020). Legal Sentiment Analysis and Opinion Mining (LSAOM). Https://Arxiv.Org/Abs/2010.02726v1.