**TECHNOLOGY ASSISTED REVIEW**

**OF LEGAL DOCUMENTS**

by

**Malik Umar Hassan**

**A Capstone Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Professional Studies: Data Analytics**

**Department of Graduate Programs & Research**

**Rochester Institute of Technology**

**09/10/2022**

**Master of Science in Professional Studies:**

**Data Analytics**

**Graduate Capstone Approval**

Student Name**: Malik Umar Hassan**

Graduate Capstone Title**: Technology Assisted Review**

**Graduate Capstone Committee:**

**Name: Dr. Sanjay Modak Date:**

**Chair of committee**

**Name: Dr. Ioannis Karamitsos**  **Date:**

**Member of committee**

# ACKNOWLEDGEMENTS

Words cannot express my gratitude to my mentor Dr. Yannis and chair of my committee for his invaluable patience and feedback. I also could not have undertaken this journey without my defence committee, who generously provided knowledge and expertise. Additionally, this endeavour would not have been possible without the generous support from the RIT Dubai, who financed me fifty percent of scholarship.

I am also grateful to my classmates and cohort members, especially my office mates, for their editing help, late-night feedback sessions, and moral support. Thanks, should also go to the librarians, research assistants, and study participants from the university, who impacted and inspired me.

Lastly, I would be remiss in not mentioning my family, especially my parents, spouse, and children. Their belief in me has kept my spirits and motivation high during this process. I would also like to thank my cat for all the entertainment and emotional support.

# 

# 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.

# TABLE OF CONTENTS

[ACKNOWLEDGEMENTS 3](#_Toc116235834)

[ABSTRACT 4](#_Toc116235835)

[TABLE OF CONTENTS 5](#_Toc116235836)

[TABLE OF FIGURES 5](#_Toc116235837)

[CHAPTER 1 6](#_Toc116235838)

[1.1 INTRODUCTION 6](#_Toc116235839)

[1.2 PROJECT GOALS 6](#_Toc116235840)

[1.3 AIMS AND OBJECTIVES 7](#_Toc116235841)

[1.4 RESEARCH METHODOLOGY 7](#_Toc116235842)

[1.5 LIMITATION OF STUDIES 9](#_Toc116235843)

[CHAPTER TWO 10](#_Toc116235844)

[2.1 LITERATURE REVIEW 10](#_Toc116235845)

[CHAPTER 3 17](#_Toc116235846)

[3.1 PROJECT DESCRIPTION 17](#_Toc116235847)

[CHAPTER 4 17](#_Toc116235848)

[4.1 DATA ANALYSIS 17](#_Toc116235849)

[CHAPTER 5 18](#_Toc116235850)

[5.1 CONCLUSION 18](#_Toc116235851)

[BIBLIOGRAPHY 19](#_Toc116235852)

# TABLE OF FIGURES

[Figure 1 :CRISP – DM Methodology 8](#_Toc116235853)

[Figure 2 : legal research process architecture, F, M. (2010), 10](#_Toc116235854)

[Figure 3 :Methods of Dimension reduction (https://towardsdatascience.com) 11](#_Toc116235855)

[Figure 4 : Text classification (Thangaraj M (2018). P.14) 12](#_Toc116235856)

[Figure 4 : Bayes Theorem (Tsangaratos & Ilia, I. (2016). P.4) 13](#_Toc116235857)

[Figure 6 : Overview of text Summarization using transformer model (Sutskever (2014), p. 2) 15](#_Toc116235858)

[Figure 7 : Textual data visualization framework (Conner et al. (2019). p. 5) 16](#_Toc116235859)

# CHAPTER 1

# INTRODUCTION

Much of the information of interest to lawyers and legal scholars is in the form of texts. Whether they are briefs, contracts, court decisions, law review articles, legislative acts, treaties, newspapers, or blog posts, all are either legal documents themselves or documents about the law. Finding, analysing, commenting on, relating to, and explaining these documents have 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 by unstructured legal texts (such as judgments, skeleton arguments, scholarly articles, and Law Commission reports).

The research found that the material for practical implementation of this study was not Found especially for courts and legal firms. Based on our study, we need to create a web application to ease the work of legal teams and give the opinion, on whether the feature in the legal text is in the interest of the appellate or not.

# PROJECT 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 modelwill 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 help 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 valuable advice to the client in a timely manner.

# AIMS AND OBJECTIVES

The Supreme Court of Pakistan ([SCP](https://www.supremecourt.gov.pk/about/history)) is the apex court of Pakistan, whose verdicts and orders are binding to all other courts of country. All executive and judicial authorities are obligated to come to the aid of the Supreme Court. 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 ([ICIJ](https://www.icij.org/investigations/panama-papers/former-pakistan-pm-sharif-sentenced-to-10-years-over-panama-papers/)) 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 published by the 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.

# RESEARCH METHODOLOGY

In this project we used both quantitative and qualitative method of research, first all the in-depth literature regarding NLP and text mining is written.

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

Schroer et al. (2021) explained CRISP-DM to organize the data mining process into six phases: business understanding, data understanding, data preparation, modelling, evaluation, and deployment. To carrying out a data mining projects, following steps will help organization to understand the mining process and provide them a road map for planning and execution.

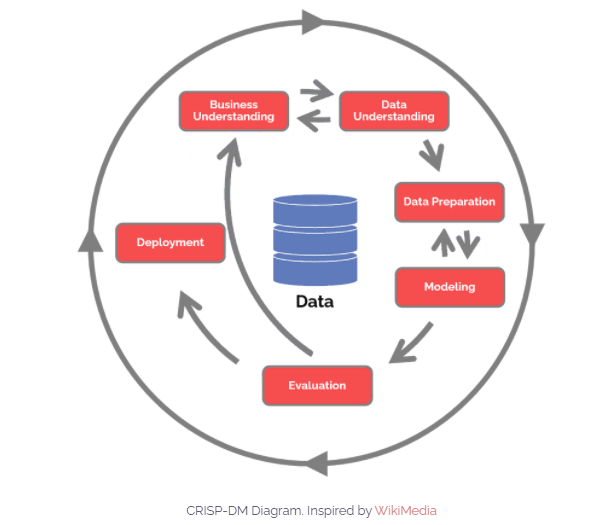


Figure 1 :CRISP – DM Methodology

* **Business Understanding**

In this initial step we need to rectify what exactly issue(s) need to solve for business by gaining their 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 been taken.

* **Deployment**

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

# LIMITATION OF STUDIES

Lack of implementation material may create issues while developing the complete model.

# CHAPTER TWO

# 2.1 LITERATURE REVIEW

According to Dyevre, A. (2021), 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. For the assessment of unstructured data many machine learning and natural language processing techniques can apply. Furthermore, text mining techniques like topic modelling, word embedding, and transfer learning can help them to ease their work efficiently.

Moreover, Branting, L. K. (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.

Firdhous. (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 is responsible for analysing each document and on the basis of information creating a law report repository. Information retrieval has been done on the basis of text block analysing on each law report.

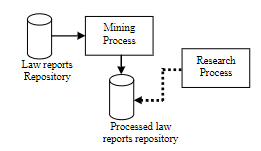


Figure 2 : legal research process architecture, F, M. (2010),

Ning, J. (2022) suggests the study which makes 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 (Process of analysing relationships amount the set of documents and token within), Word Embedding (grouping of words which are close in meaning with int the vector space), Document Clustering with Word Embedding and Topic Modelling are the main concept were used in legal search analysis.

Nanga et al. (2021), explain that linear dimensionality reduction, LSA, and PCA are the oldest and most commonly used techniques for dimension reduction. PCA aims to rectify the data points with a set of well principal components, to sum up; the idea of PCA is simple — reduces data set variables, while conserving information as possible

Hotelling, H. (1933) explains the analysis of a Complex of Statistical Variables into Principal Components according to him dimension reduction can be sub categorized into two main methods i.e., linear and non-linear methods. For linear methods, a significant low-dimensional space is proposed to be discovered in data input with space that is high-dimensional, where the embedded data in the input space has a linear structure for linear reduction methods. The linear techniques consider as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Singular Value Decomposition (SVD), Latent Semantic Analysis (LSA), Locality Preserving Projections (LPP), Independent Component Analysis (ICA) and Project Pursuit (PP).

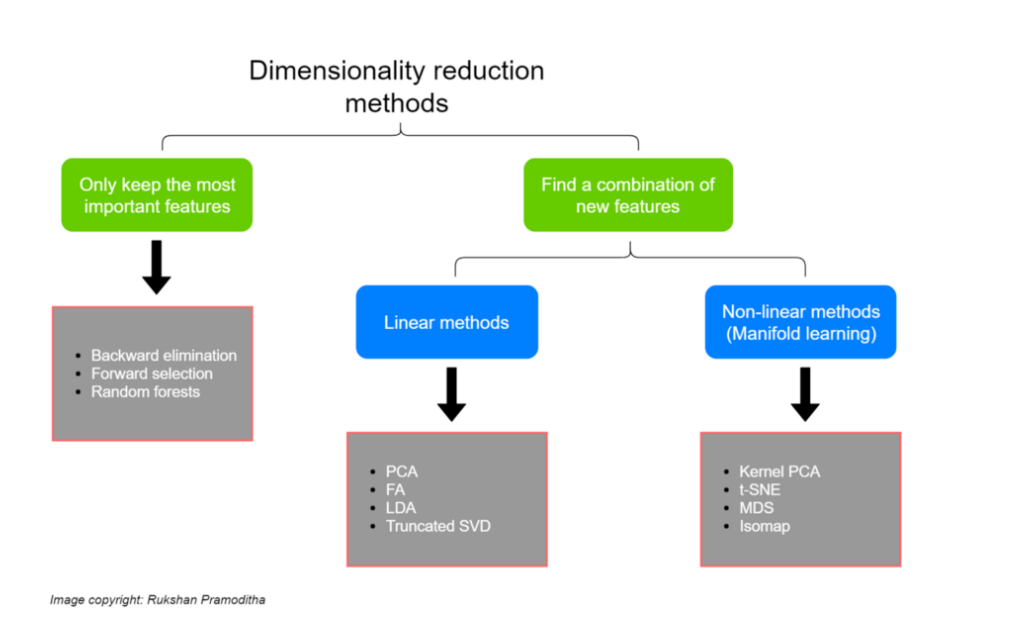
According to Pratihar, D.K. (2011) The non-linear techniques which were developed to work with applications that have complex non-linear structures considered were Kernel Principal Component Analysis (KPCA), Multi-dimensional Scaling (MDS), Isomap, Locally Linear Embedding (LLE), Self-Organizing Map (SOM), Latent Vector Quantization (LVQ), t-Stochastic neighbour embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP). 

Figure 3 :Methods of Dimension reduction (https://towardsdatascience.com)

Novotna et al. (2020) describe topic modelling as the process which recognizing the words from the topics present in the document or the corpus of data. Mining the words from is highly complex and more time consuming, and this is very helpful while extracting this information from topics that is existing in the document. Removing stop words and punctuation marks Stemming, Lemmatization, and encoding them to ML language using Count vectorizer or Tfidf vectorizer are important terms for topic modelling. The process of identifying the topics from the set of documents is known as topic modelling. This latent will appears during the process of topic modelling, and Latent Dirichlet Allocation (LDA) is one of famous modelling technique.

Chhatwal R et al. (2019) 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 helping to categorize the document into predefined clusters.

Thangaraj M (2018), further explains text classification as a process of assigning a category to any sentence or document, this mainly includes motion classification, news classification, and citation intent classification. In his literature he divides the text classification into two major groups first one is statistical and another one is machine learning.

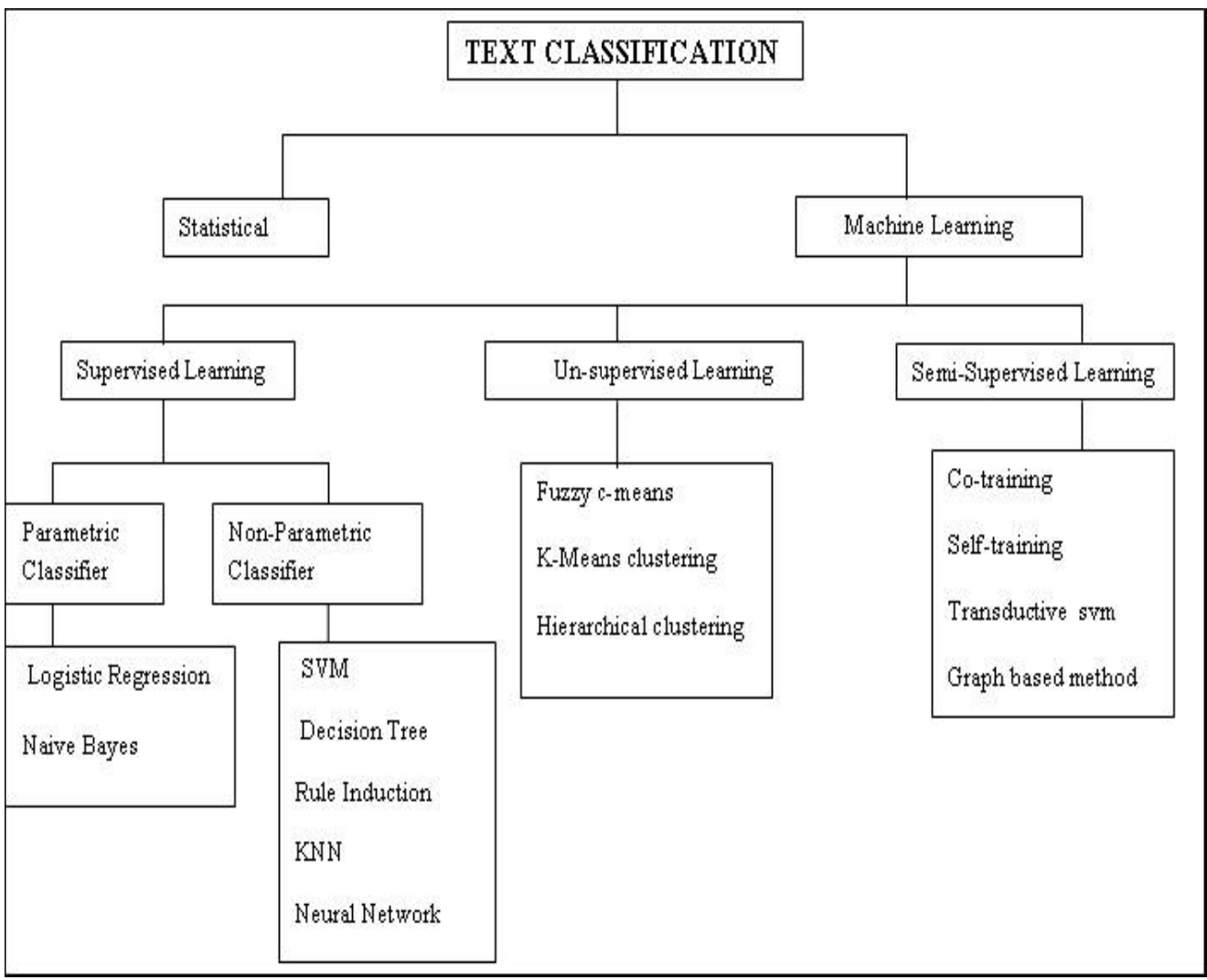


Figure 4 : Text classification (Thangaraj M (2018). P.14)

Tsangaratos & Ilia, I. (2016). Explains the text & document classification using the NaiVe Bayes Classifier when the variable is conditionally independent and needs to estimate the class conditional probability it’s based on Bayes’ theorem.

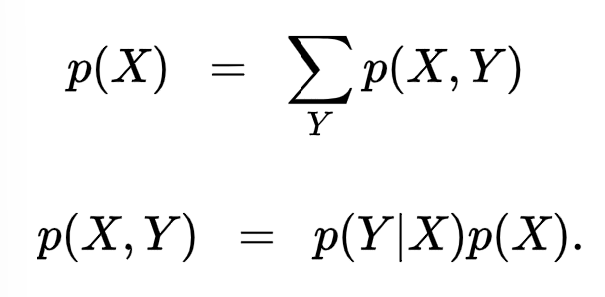
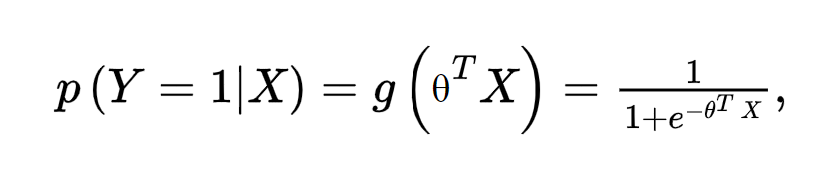


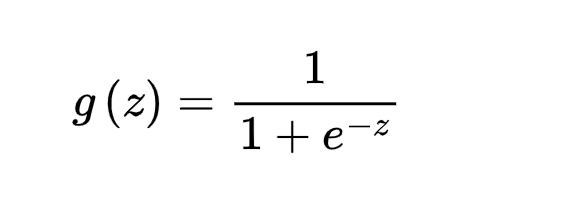
Figure 4 : Bayes Theorem (Tsangaratos & Ilia, I. (2016). P.4)

Where the first equation explains the sum of the rule and the second equation describes the product rule. p(X,Y) is considered as joint probability, p(Y|X) is conditional probability and p(X) is a marginal probability. Furthermore, one of the disadvantages of this classification is that it cannot find the interaction between the features. The result is based on relative probabilities not mathematically accurate.

Wang&Park,2017 describe a maximum entropy classifier also known as (Logistic regression), an approach to learning p(Y|X) directly in the case where Y is discrete-valued, and X = 〈X1, …, Xd〉is any vector containing discrete or continuous variables. The purpose of logistic regression is to directly calculate the distribution P(Y|X) from the training data set. The following model represents the logistic regression.



Where,



is called the logistic function or the sigmoid function.

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 was used in the experiment and its results show there is a drastic improvement in text classification.

XAI required to estimate Pr(explainable predictive coding) by following formula.

Pr(r = Rationale | x，y = Responsive)·Pr (y = Responsive | x)

Where x is a document, y is the model-labelled designation of the document (‘responsive’ or ‘not responsive,’ for instance), and r is a text snippet from x.

Dong et al. (2017), humans cannot directly understand ensemble models; it requires a more sophisticated approach to interpret the decision, decision trees, and fuzzy logic required to explain the results. Moreover, complex models, such as deep learning models like multilayer neural network models, and non-linear SVM also helps human to understand the explainable AI. Model-based explanation based on mentioned machine learning techniques. Recent research found that a prediction-based approach is often used to identify snippets of text as an explanation for the classification of a document. A text snippet that explains the classification of a document is called a ‘rationale’ for the document.

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 labeled by human annotators before training. Supervised approaches have these approaches Obtaining Labelled Documents, Bag-of-Words Methods and Transfer Learning, and Transformers.

Pang, B et al. (2022), In legal documents summarization of large judgments is the key and important success of inference from a court decision, representation of words or tokens in source documents play an important role in the understanding of any text it can be done in extractive or abstractive manner. Where extractive summarization emphasized the concatenation of important sentences or paragraphs without getting the meaning of the text, on the other hand, abstractive summarization depends on the meaning and context of the text which is semantic inference, in this study we will emphasize abstractive inference because of Seq2Seq model with encoder-decoder architecture using RNNs or transformers.

Sutskever et al. (2014) describe Sequence-to-Sequence models based on neural networks which take the sequence of words from text vocabulary as an input and output new sequence in the different summaries of vocabulary. In this model, texts are padded with same-length sequences for the formation feature matrix, and for feature learning techniques words from the vocabulary are mapped with vectors of numbers on the basis of the probability distribution. On the basis of the probability distribution matrix encoder processes, the input sequence and decoder return their own internal states that serve as the context.

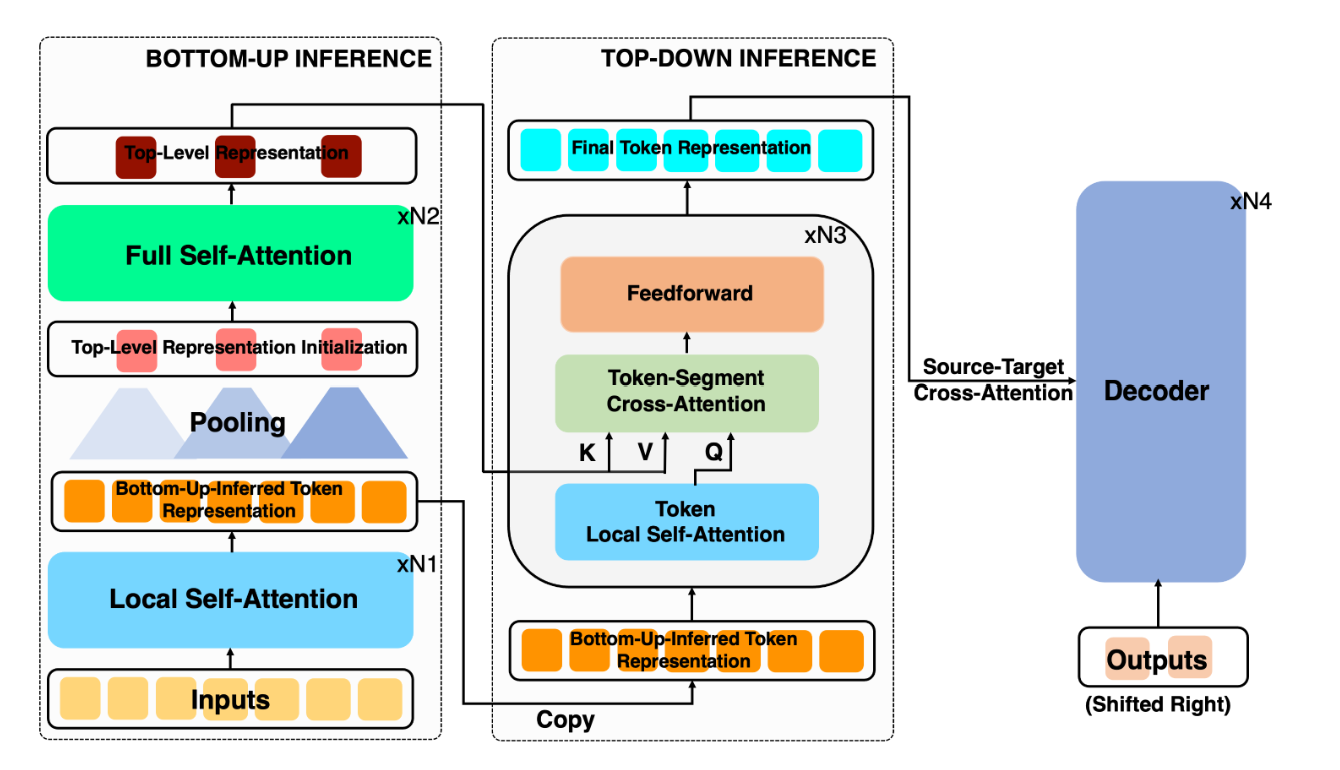


Figure 6 : Overview of text Summarization using transformer model (Sutskever (2014), p. 2)

Lance (2020) categorized the sentiment analysis of legal documents into two parts and named 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 the text, analyzing, understanding, organizing, and sorting through text data is hard and time-consuming, so most companies fail to use it to its full potential.

Gunning et al. (2019) explained the Defence Advanced Research Projects Agency (DARPA) theory that proposed a new direction for Explainable AI (XAI). In XAI systems, actions or conclusions are human explicable – “machines understand the context and environment in which they operate, and over time build underlying explanatory models that permit them to describe actual phenomena.” Similarly, in an explainable machine learning system, predictions or classifications generated by a predictive model are explainable and comprehensible to humans. Most of the explainable AI focused on two main areas of explainable machine learning models. (a) Model-based explanation. (b) Prediction-based explanation.

Conner et al. (2019), Visual inference plays an important role in finding or extracting the behavior, sentiments, and critical decision-influencing variables. while doing visualization there may face some challenges in terms of information loss, and lack of summary representation to showcase the main points.

Kabir et al. (2018), Word clouds, count frequencies, tables, and pie charts are common examples of Quantities Visualization (QViz). Sentiment analysis, Semantics, and Natural Language Processing are commonly used in Sense Visualization (SViz). When we need to explore the “Who, When, and Where” then context visualization (CViz) will work. Trends visualization (TViz) is used to evaluate the temporal data or time series data.

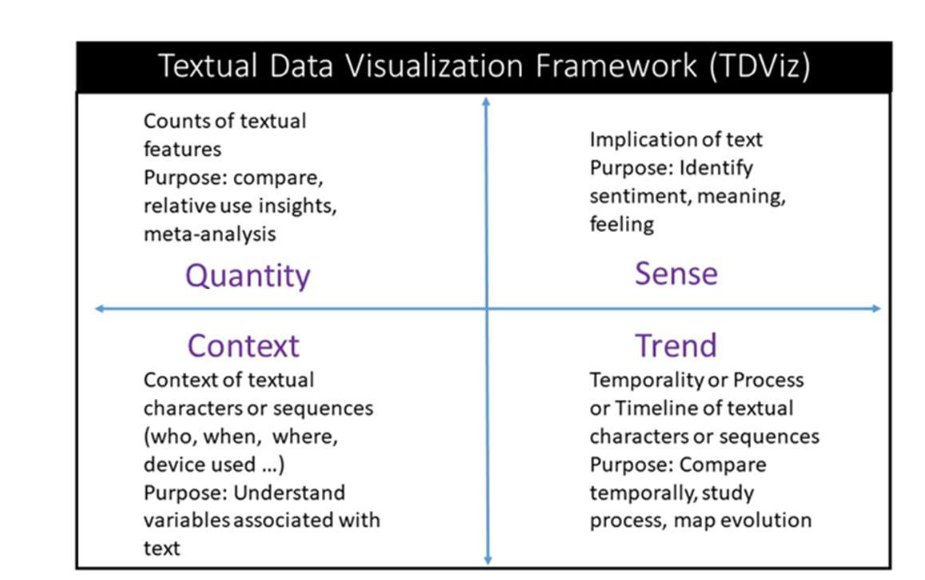


Figure 7 : Textual data visualization framework (Conner et al. (2019). p. 5)

# CHAPTER 3

# 3.1 PROJECT DESCRIPTION

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 modelwill 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.

# CHAPTER 4

# 4.1 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/)

Open-source tools like pycharm (IDE), python, NLTK, Streamlit libraries and Tableau will be used to find out the insights of these legal documents.

This case study supported to automate the time consuming legal and research work of SCP. The final web component considers 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.

# CHAPTER 5

# 5.1 CONCLUSION

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 Legal sentiment analysis and opinion mining (LSAOM).

Future work in this project is to describe the clustering of documents on the basis of legal text. Support vector machine and other clustering algorithm will used to implement this research.

# BIBLIOGRAPHY

1. “History - Supreme Court of Pakistan.” The Supreme Court of Pakistan (SCP), <https://www.supremecourt.gov.pk/about/history>.
2. “Panama Papers: Former Pakistan PM Sharif Sentenced To 10 Years - ICIJ.” The International Consortium of Investigative Journalists,

<https://www.icij.org/investigations/panama-papers/former-pakistan-pm-sharif-sentenced-to-10-years-over-panama-papers/>**.**

1. M F M Firdhous, “ Automating Legal Research through Data Mining” International Journal of Advanced Computer Science and Applications(IJACSA), 1(6), 2010.

[http://dx.doi.org/10.14569/IJACSA.2010.010602](https://dx.doi.org/10.14569/IJACSA.2010.010602)

1. 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. 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
2. 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
3. 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.). (2019).Explainable Text Classification in Legal Document Review A Case Study of Explainable Predictive Coding.
2. Dong, Y., Su, H., Zhu, J., & Bao, F. (2017). Towards Interpretable Deep Neural Networks by Leveraging Adversarial Examples.

http://arxiv.org/abs/1708.05493

1. Gunning, D., & Aha, D. W. (2019). DARPA’s Explainable Artificial Intelligence Program Deep Learning and Security.

<https://arxiv.org/abs/1904.01721>.

1. Zaidan, O. F., Eisner, J., & Piatko, C. D. (2007). Using “Annotator Rationales” to Improve Machine Learning for Text Categorization \*.

http://cs.jhu.edu/

1. Lance B. Eliot. (2020). Legal Sentiment Analysis and Opinion Mining (LSAOM). <Https://Arxiv.Org/Abs/2010.02726v1>.
2. Schroer, C., Kruse, F., & Gomez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. Procedia Computer Science, 181, 526–534. <https://doi.org/10.1016/j.procs.2021.01.199>.
3. Nanga, S., Bawah, A. T., Acquaye, B. A., Billa, M.-I., Baeta, F. D., Odai, N. A., Obeng, S. K., & Nsiah, A. D. (2021). Review of Dimension Reduction Methods. Journal of Data Analysis and Information Processing, 09(03), 189–231.

<https://doi.org/10.4236/jdaip.2021.93013>.

1. Hotelling, H. (1933) Analysis of a Complex of Statistical Variables into Principal Components. The Journal of Educational Psychology, 24, 417-441, 498-520.

<https://doi.org/10.1037/h0070888>.

1. Pratihar, D.K. (2011) Non-Linear Dimensionality Reduction Techniques. In: Wang J., Ed., Encyclopaedia of Data Warehousing and Mining, Second Edition, IGI Global, Pennsylvania,1416-1424.
2. https://towardsdatascience.com/11-dimensionality-reduction-techniques-you-should-know-in-2021-dcb9500d388b

https://doi.org/10.4018/978-1- 60566-010-3.ch219.

1. Thangaraj, M., & Sivakami, M. (2018). Text Classification Techniques: A Literature Review. Interdisciplinary Journal of Information, Knowledge, and Management, 13, 117–135.

<https://doi.org/10.28945/4066>

1. Tsangaratos, P.,& Ilia, I. (2016). Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: The influence of model’s complexity and training dataset size. Catena, 145,164–179.

https://doi.org/10.1016/j.catena.2016.06.004

1. Wang, J.,& Park, E. (2017).Active learning for penalized logistic regression via sequential experimental design. Neurocomputing, 222,183–190.

<https://doi.org/10.1016/j.neucom.2016.10.013>

1. Pang, B., Nijkamp, E., Kryściński, W., Savarese, S., Zhou, Y., & Xiong, C. (2022). Long Document Summarization with Top-down and Bottom-up Inference. <http://arxiv.org/abs/2203.07586>
2. lya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In Advances in neural information processing systems,

pp. 3104–3112, 2014.

1. Conner, C., Samuel, J., Kretinin, A., Samuel, Y., & Nadeau, L. (2019). A Picture for The Words! Textual Visualization in Big Data Analytics. APA.
2. Kabir, A. I., Karim, R., Newaz, S., & Hossain, M. I. (2018). The Power of Social Media Analytics: Text Analytics Based on Sentiment Analysis and Word Clouds in R. Informatica Economica, 22, 25-38