2-12-2022

Technology Assisted Review of Legal Documents

Malik Umar Hassan
mh3619@rit.edu

RIT

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

RIT Dubai

02/12/2022

RIT

Master of Science in Professional Studies:

Data Analytics

Graduate Capstone Approval

Student Name**: Malik Umar Hassan**

Graduate Capstone Title**: Technology Assisted Review of Legal Documents**

Graduate Capstone Committee:

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

Chair of committee

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

Member of committee

ACKNOWLEDGEMENTS

I am incredibly grateful to Dr. Yannis, my mentor and the chair of my committee Dr. Sanjay Modak, for his input and patience. This trip would not have been possible without the extensive knowledge and experience of the Défense Committee. Additionally, this project would not have been feasible without the RIT Dubai's generous funding of my scholarship, which covered fifty percent of my costs.

I am also grateful of the excision help, late-night response meetings, and emotional care I received from my undergraduates and cohort associates, particularly my workplace companions. Thanks, should also go to the curators, research subordinates, and study members from the institute, who affected and inspired me.

I should not disremember to mention my family, mainly my parentages. Their sureness in me has continued my eagerness and optimistic boldness during this progression. I also want to express my thankfulness to my friends for providing me with so much advise and guidelines.

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.

Keywords: Technology Assisted Review (TAR), NLP, Legal Documents, NER, Text Classification, Sentiment Analysis

TABLE OF CONTENTS

TABLE OF FIGURES

CHAPTER 1

1.1 INTRODUCTION

Texts make up a large portion of the information that interests attorneys and legal academics. All legal documents, whether they are documents about the law or documents themselves, include briefs, contracts, court decisions, law review articles, legislative acts, treaties, newspapers, and blog entries. Legal practice and legal academia have both spent centuries finding, analyzing, commenting on, reacting to, and explaining these documents.

Pakistan's legislative process is particularly time-consuming because technology cannot be utilised during protracted procedures. This project will investigate several techniques for controlling and structuring legal material produced by unstructured legal documents using natural language processing (NLP) (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.

1.2 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 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 technique will enable attorneys to focus more on comprehending the meaning of the papers, acquiring the necessary insights, and promptly offering the client relevant counsel rather than wasting time slogging through documents and labelling significant clauses.

1.3 AIMS AND OBJECTIVES

The Supreme Court of Pakistan (SCP) is the highest court in Pakistan, and all other courts must follow its judgments and directives. All executive and judicial branches have a duty to support the Supreme Court. The Pakistan Bar Council PBC attorneys and the general public are also given access to the court's detailed ruling in the form of unstructured data, or PDF, which may be a few pages or a number of volumes long.

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) initial reference and allegations was money

laundering, corruption on the basis of Panama The International Consortium of Investigative Journalists' papers (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.

Public prosecutors will benefit from this study's ability to quickly analyze any verdict using NLP for understanding such judgmental insights. Additionally, the prediction model created using machine learning will allow attorneys to forecast the outcome of upcoming cases using the body of existing knowledge.

1.4 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, the data mining process will be divided into six steps using CRISP-DM: business understanding, data understanding, 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.

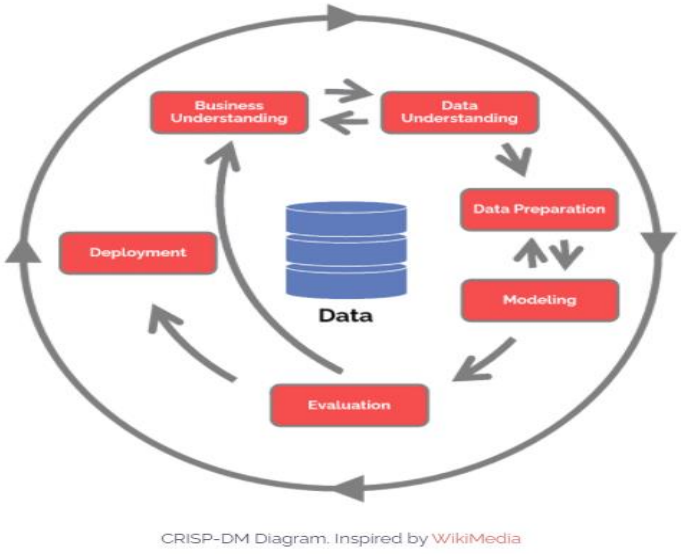

Figure 1 :CRISP – DM Methodology

- **<u>Business Understanding</u>**
  We need to identify the precise problem(s) that need to be resolved for business in this first step by learning about and truly comprehending their business.
  Set objectives – This entails outlining your main goal from a commercial standpoint. You might also want to address any further connected questions. For instance, your main

objective might be to keep present customers by predicting at what time they are probable to shift to a rival. Other pertinent business inquiries include "Does the channel used affect whether customers stay or go?" and " Minimum ATM charges will reduce the number of customers who did not do business with company?"

Produce project plan –In this step you need to highlight the business strategy and data mining objective. including the initial tool and technique selection, should be laid out in the plan.

Business success criteria –In this step organization needs to set the key constraints whether that project will successful and mentioned the outlined. These should ideally be precise and quantifiable, such as a set degree of customer churn reduction. However, there may be occasions when more arbitrary measures, such as "deliver valuable understandings into the associations," are required. Check whether that is the case, it must remain made vibrant who is making the subjective valuation.

Assess the current situation - This entails gathering more information about all of the existing assets, limitations, presumptions, and other elements that you have to look into matter when choosing your information examination goal and development strategy.

1. **Inventory of resources** – mentioned the main content at your disposal for the project, such as:
    - Workers (professional specialists, statistics specialists, technician, data mining specialists)
    - Data (Attributes, velocity, variety, veracity, Value)
    - Computation resources (Virtual machine, servers)
    - Software (Software for data mining, other related software program)
2. R**equirements, assumptions and constraints** – Include entirely project necessities, such as the deadline, the level of comprehension and value that must be achieved, and if any data leakage or security hazard, and any legal considerations. Verify that you authorize to use that data. List the project's underlying presumptions. These may contain non-verifiable expectations about the business as well as verifiable expectations about the data that may be tested during data mining. If the validity of the results would be impacted, it is especially crucial to list the latter. List the project's limitations. There may be limitations on the resources that are available, but there may also be technical boundaries, size of the practical data-set for machine learning model.
3. **Risks and contingencies** – List the potential hazards or occurrences that could harm the project or create delays. List the related backup plans, including the steps you'll take if the risks or occurrences materialize.
4. **Terminology** – Accumulate a appendix of vocabulary applicable to the plan. This will generally have two components:
    1. lexicon of pertinent business terms, which is a component of the project's business knowledge. A helpful "knowledge elicitation" and educational exercise was the creation of this lexicon.
    2. A dictionary of data mining terms that includes instances related to the business issue at hand.
5. **Costs and benefits** – Create a cost-benefit analysis for the project that weighs its expenses against the possible rewards for the company if it succeeds. This

comparison needs to be as precise as is practical. For instance, in a business setting, you ought to employ financial measurements.

6. In business jargon, objectives are stated in a goal. Technical project objectives are stated in a data mining aim. The company's objective can be to "increase catalogue sales to established customers," for instance. Predicting how many widgets a client will purchase based on their past three years of purchases, demographic data (stage, income, city, etc.), and the item's pricing is one example of a data mining goal.
7. **Business success criteria –** Describe the project's intended outputs that will allow the business goals to be met.
8. **Data mining success criteria** – formally state the necessities for a positive project consequence, such as a exact stage of forecast exactness or a propensity-to-buy contour with a specific quantity of "boost." Similar to corporate success criteria, it could be required to express these in individual relations; in such case, it is important to identify the individual or people who are making the subjective evaluation.
9. Yield development strategy
10. Define the proposed strategy for attaining the data mining objectives and, consequently, the occupational objectives. Your strategy must outline the procedures to be followed for the next phases of the project, including the preliminary tool and method selection.
11. **Project plan** – The project's stages should be listed along with their length, resources needed, inputs, outputs, and dependencies. In the data mining process, large-scale iterations, such as repeated modelling and evaluation steps, should be made as explicit as feasible. Analysing the connections between the timetable and the hazards is crucial for the project plan. Indicate the findings of these studies clearly in the development plan, preferably with activities and proposals in case the hazards materialize. Choose the evaluation strategy that will be applied at this point in the assessment stage. Your development strategy will change as the project progresses. You will evaluate results at the conclusion of each phase and make the necessary updates to the project plan. The development plot should include exact evaluation facts for these changes.
12. **Initial assessment of tools and techniques** – You should do an preliminary instrument and technique assessment after the first phase. At instance, in this case, you choose a data mining device that supports a variability of practises for distinct stages of the procedure. Since the choice of tools and procedures may have an impact on the entire project, it is crucial to evaluate them early in the process.

- **<u>Data Understanding</u>**

To proceed to the subsequent stage, you must collect the information definite in the development properties. This initial collection contains it if data loading is needed for data thoughtful. It styles seamless wisdom to weight your data into a program you use specifically for comprehending data, for instance. You must think about how and when you will combine multiple data sources if you get them.

- **Initial data collection report** – List the data bases used, their places, the methods used to obtain them, and any problems that were encountered. Keep track of the issues you ran into and any fixes you made. This will facilitate both the repetition of this project in the future and the implementation of related projects in the forthcoming.

Describe data

- Analyze the "surface" or "gross" characteristics of the attained information and present the findings.
- **Data description report** − Define the composed data, including its volume, structure, names of the fields, names of the records in each table, and any additional external structures that have been exposed. Verify that the information you have composed pleases your requirements.

Explore data

- In this phase, you'll use strategies for querying, data visualization, and reporting to respond to data mining queries. These may include:
- essential characteristics are distributed (for example, the mark characteristic of a forecast job).
- Associations among a couple of characters or a few tiny characteristics.
- Outcomes of modest accumulations.
- Belongings of important sub-populations.
- Meek algebraic studies.
- Your data mining objectives may be directly addressed by these analyses. Additionally, they might advance the information excellence and description reports, as well as the alteration and other facts preparation actions required for extra investigate.
- **Data exploration report** − Label the results of your data study, including the initial conclusions or hypotheses and how they affected the rest of the project. If applicable, you might add graphs and plots to this section to highlight data variables that call for extra examination of exciting data subsections.

Verify data quality

- Check the data's accuracy by asking questions like:
- Is the material comprehensive (does it contain all the essential issues)?
- Is it correct or is it cover mistakes, and if so, how frequent are they?
- Is data containing any lost value? If so, in what way are they represented, wherever do they occur, and how often do they happen?

Data quality report

- List the outcomes of the data quality assurance. Provide potential remedies if there are issues with the quality. Problems with data quality can typically only be solved with a combination of data and business expertise.

- **<u>Data Preparation</u>**

  Select your data

  - In this project stage, you choose the information that will be used for analysis. When making this choice, you may take into consideration a variety of criteria, including the data's quality, relevance to your information mining objectives, and any technological limitations on the amount or types of data that are applicable. Be careful that choosing data also means choosing a table's rows for records and attributes for characteristics.
  - **Rationale for inclusion/exclusion** − Motioned the information to be involved or omitted, together with the justifications for those choices.

  Clean your data

- The task at hand is to improve the information excellence to the level vital by the analytical methods you've nominated. In order to do this, one may select clean subclasses of the information, add appropriate defaults, or use more striving approaches like modelling to approximate misplaced information.
- **Data cleaning report** – Define the options and steps you decide to deal with data quality issues. Think about any data transformations used for cleaning and how they might affect the analysis's findings.

Construct required data

- This job involves useful data selection procedures like creation of derived variables, completely fresh data rows.
- **Derived attributes** – These are latest variables that are built from single or more than one prior variable in the same record. For instance, you could generate a brand-new attributes called shape using the variables dimensions.
- **Generated records** – You've just described the production of any wholly original records. For consumers who haven't made a purchase in the past year, for instance, you might need to generate records. Such entries in the raw data were unnecessary, but it would make sense to explicitly indicate a specific customer's lack of purchases when using the information in models.

Integrate data

- These are methods for compounding information from various databases, data base tables, or records to produce fresh attribute.
- **Merged data** – composed two or extra tables that cover various pieces of information about the same objects is referred to as merging tables. For instance, a chain of retail stores might include three tables: one with summary sales data (such as revenue, percent variation in sales from the previous tenure), one with evidence about the geography of the neighbourhood, and one with generic store information (such as floor space, kind of mall). For each store, there is a single record in each of these tables. These tables can be combined, integrating fields from the source tables, creating a single table with one entry for each shop.
- **Aggregations** – Aggregations are operations that compute new values by combining data from various data sets and/or set of data tables. For instance, creating a new table from a table of customer purchases with one row for each purchase and fields like the amount of acquisitions, average buying worth, percentage of orders cost to credit cards, percentage of items on sale, etc.

- **<u>Modelling</u>**

Select modelling method

You'll select the real modelling method you'll employ as the original modelling phase. In spite of the possibility that you already choose a tool during the business understanding stage, you will now select the specific modelling method, such as “decision-tree” construction using C5.0 or neural network formation with back propagation. If several procedures are used, carry out this job for each practice independently.

- **Modelling technique** – Record the real modelling approach that will be utilized.

- **Modelling expectations** – Numerous modelling strategies rely on certain data presumptions, such as the requirement that class attributes be symbolic, that all variables have similar disseminations, and that no values be missing. Note whatever inferences you make.

Generate test design

Create a method or technique to evaluate the model's accuracy and validity before you start structure it. For instance, using error frequency as a gauge of the quality of data mining models is popular in supervised data mining models like classification. In order to create the model on the training dataset and determine its value on the independent test set, the dataset is often divided into train and test sets.

Test design – Describe the anticipated strategy for developing, putting to use, and assessing the models. Selecting in what way to divide the given dataset into training, test, and testing datasets is a key aspect of the plan.

Build model

- Utilizing the prepared dataset, construct one or more models using the modelling tool.
- **Parameter settings** – Every modelling device typically has a huge amount of variables that can be altered. Note the variables, the values selected for them, and the reasoning behind those choices.
- **Models** – These models, not a report on the models, are what the modelling tool really produced.
- **Model descriptions** – Label the resultant models, discuss how they were interpreted, and note any issues you had understanding what they meant.

Assess model

Adapt the models to your preferred test design, data mining accomplishment principles, and domain knowledge. Analyse the technical success of the modelling and discovery methodologies, and then get in touch with business analysts and domain specialists to talk about the data mining findings in the context of the business. This assignment just takes into account models; however, the evaluation step additionally evaluates all other project-produced deliverables.

The models should now be ranked and evaluated using the evaluation criteria. As far as possible, you should consider the business aims and achievement standards in this situation. In the majority of data mining developments, a single approach is used multiple times, and numerous separate techniques are used to get the data mining outcomes.

- **Model assessment** – List the features of the models you created (such as accuracy) and rank them in terms of quality to summarize the outcomes of this job.
- **Revised parameter settings** – Update parameter values and fine-tune them for the upcoming modelling run in accordance with the model evaluation. Build models iteratively and evaluate them until you are confident that you have create the finest model(s). Keep track of all such reviews and evaluations.

- **Evaluation**

Evaluate your results

- In past review rounds, the model's generalizability and accuracy were two shortcomings that were resolved. You will evaluate the model's potential to assist you in achieving your business objectives and look for any applicable arguments against its validity in this step. If time and monetary limits allow, another choice is to test the model(s) on demo applications before implementing them in the live application. As part of the review process, you must also assess any extra data mining outcomes you may have shaped. Models that need to be connected to the initial business objectives as well as all other findings are the results of data mining.
- **Assessment of data mining results** – Include a final comment indicating whether the development previously pleases the project's initial business objectives when summarizing assessment results in terms of business achievement standards.
- **Approved models** – After being assessed in relation to business success elements, the created models that encounter the specified requirements are the appropriate models.

Review process

- The models created thus far seem to be appropriate and satisfy commercial needs. You should now conduct a more thorough analysis of the data mining engagement to determine if anything significant was overlooked. This assessment also addresses issues with quality control, such as whether the model was constructed properly. Did we only employ the traits that were permissible and available for future analysis?
- **Review of process** – Recapitulate the procedure evaluation and list the tasks that were neglected or that needed to be completed again.

Determine next steps

- Based on the outcomes of the valuation and procedure evaluation, you now agree how to continue. Do you complete this development and go on to deployment, launch fresh data mining projects, or begin new iterations? Additionally, you should evaluate the money and resources you have at your disposal because this could influence the decisions you make.
- **List of possible actions** – Write down any prospective next actions, along with the advantages and disadvantages of each choice.
- **Decision** – Give a brief explanation of your decision.

- **Deployment**

Plan deployment

- Using the findings of your assessment, you will create a plan for their deployment during the deployment step. If an recognised method for developing the pertinent model or models has been found, it is given here for future use. It seems logical to take deployment into account throughout the business understanding phase as well, given how important deployment is to the project's success.
- Extrapolative analytics can greatly enhance your organization's operational side in this situation.

- **Deployment plan** – Label your deployment plan in detail, outlining the stages that must be taken and in what way to take them.

Plan monitoring and maintenance

- If the results of the data mining are incorporated into the routine operations and environment of the business, monitoring and maintenance become substantial challenges. To minimize overly extended periods of inappropriate use of data mining results, a maintenance strategy should be carefully devised. A thorough monitoring process plan is necessary for the project to keep track of how the data mining results are being applied (s). This strategy accounts for the particular deployment strategy.
- **Monitoring and maintenance plan** – Outline the tasks that must be completed and how to do them in the monitoring and care plan.

Produce final report

- A concluding statement will be written by you at the development's conclusion. This statement may merely be a summary of the development and its involvements (if they haven't previously been documented as an ongoing activity), depending on the deployment plan, or it may be a thorough presentation of the data mining outcome (s).
- **Final report** – The most recent written report on the data mining assignation is available here. It contains all of the earlier deliverables, reviews them, and organises the results.
- **Final presentation** – A meeting where the results are presented to the client will frequently take place at the project's conclusion.

Review project

- Analyse what went well and what didn't, what was complete well and what may have been ended well.
  Experience documentation – List the most significant learnings from the project. For instance, any issues you ran into, deceptive strategies, or advice for choosing the top data-mining practises in comparable circumstances could be included in this material. In perfect developments, involvement citations also include any insights that particular project participants may have authored during earlier project phases.

1.5 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 decrees, agreements, jurisdictional 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 ML & NLP processing techniques can apply. Furthermore, text mining techniques like subject modelling, expression embedding, and transmission of education can help them to ease their work efficiently.

Moreover, Branting, L. K. (2018) explains the special issues while doing legal text analytics; IAAIL emphasised added 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 key mechanisms namely text analytics and diagnostic research. The analytics procedure 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.

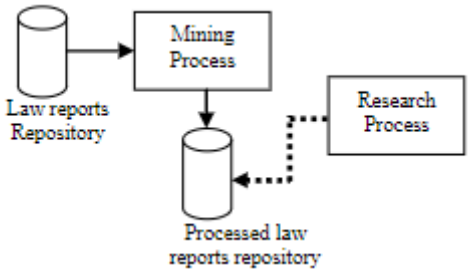

Figure 2 : legal research process architecture, F, M. (2010),

Ning, J. (2022) suggests the study which makes the judicial system artificially intelligent by using scientific research methods especially USML and SML techniques; data set is not label in unsupervised machine learning in other word we can say output is not defined in this case; 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) discusses how to analyze a set of statistical variables into principal components. According to him, there are two basic ways for dimension reduction: linear and non-linear methods. When using linear methods, it is suggested that a substantial low-dimensional space be found in high-dimensional data input when the implanted information in the input planetary has a linear construction. “PCA”, “LDA”, “SVD”, “LSA”, “LPP”, “ICA”, and Development Hunt are among the linear techniques that are taken into consideration (PP).

According to Pratihar, D.K. (2011) Kernel Principal Component Analysis (KPCA), Multi-dimensional Scaling (MDS), Isomap, Locally Linear Embedding (LLE), Self-Organizing Map (SOM), Latent Vector Quantization (LVQ), t-Stochastic Neighbor Embedding (t-SNE), and Unvarying Diverse Estimate and Forecast were among the non-linear methods developed to work with solutions that have multipart non-linear assemblies taken into thought(UMAP).

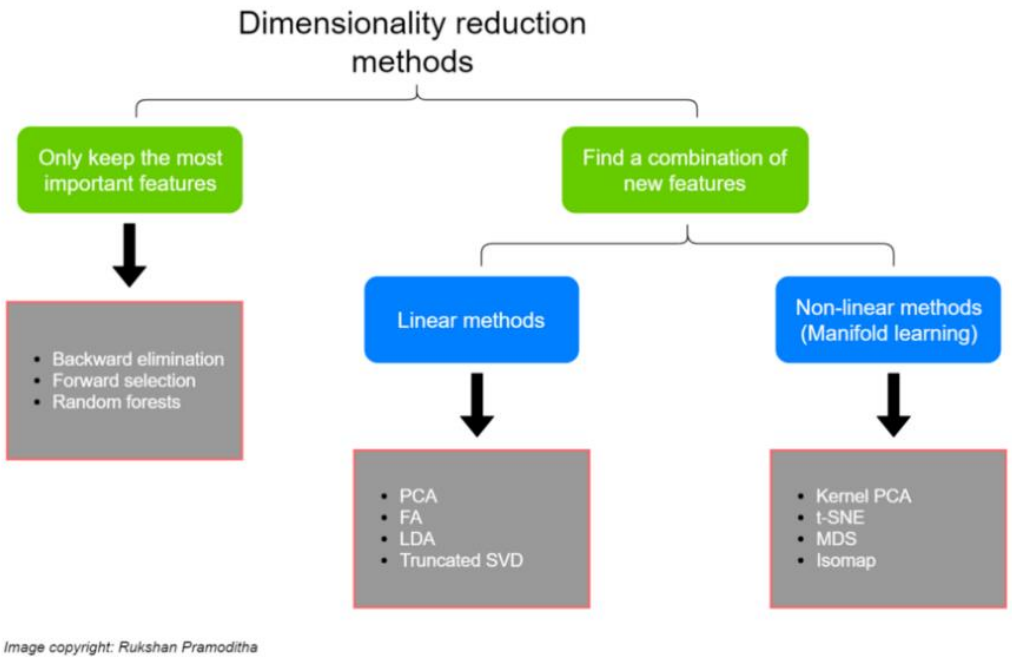

Figure 3 :Methods of Dimension reduction (https://towardsdatascience.com)

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 total excellence and rapidity of the text review procedure by plummeting the period it receipts to evaluation papers serving to classify 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.

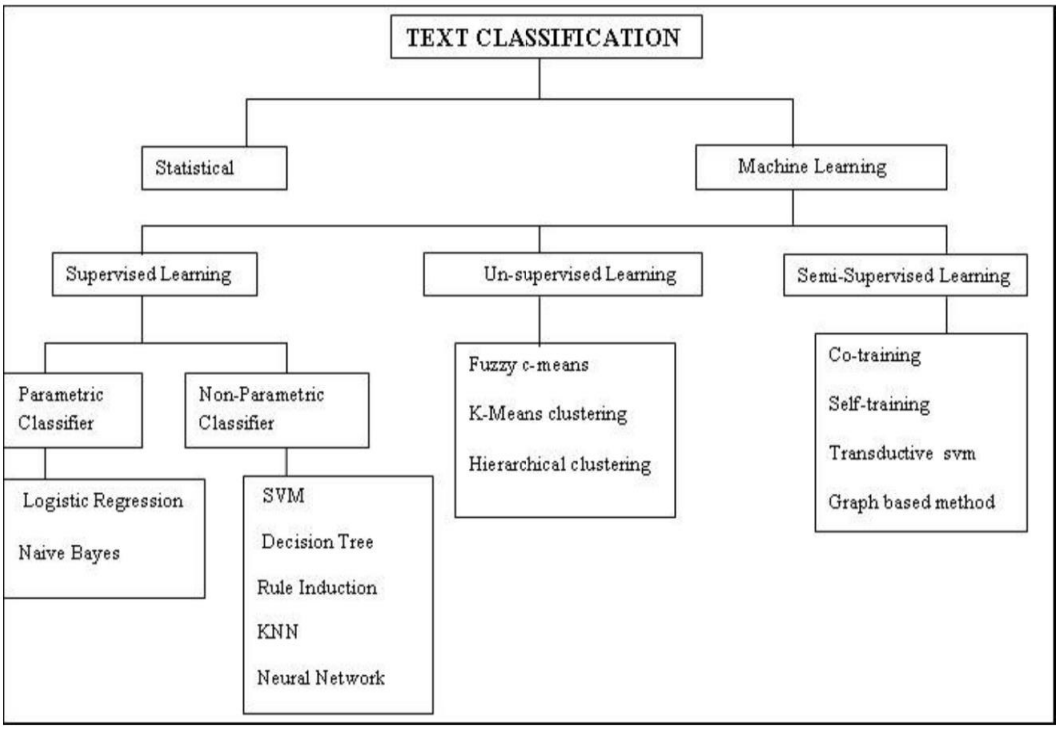

Figure 4 : Text classification (Thangaraj M (2018). P.14)

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.

$$p(X) = \sum_{Y} p(X, Y)$$

$$p(X, Y) = p(Y|X)p(X).$$

Figure 5 : Bayes Theorem (Tsangaratos & Ilia, I. (2016). P.4)

Where the initial calculation explains the amount of the rule and the second equation describes the product rule. p(X,Y) is considered as combined likelihood, p(Y|X) is restricted likelihood and p(X) is a marginal likelihood. 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 method to knowledge p(Y|X) directly in the circumstance wherever Y is discrete-valued, and X = 〈X1, …, Xd〉 is any trajectory covering separate or non-stop attributes. The purpose of logistic regression is to unswervingly calculate the distribution P(Y|X) from the training data set. The following machine learning application represents the logistic regression.

$$p\left(Y=1|X\right)=g\left(\theta^{T}X\right)=\frac{1}{1+e^{-\theta^{T}X}},$$

Where,

$$g\left(z\right)=\frac{1}{1+e^{-z}}$$

is called the logistic job or the sigmoidal function.

Zaidan et al. (2007) proposed a machine learning technique that improves text categorization performance by using the annotations in documents. The experiment used the Support Vector Machine SVM technique, and the outcomes demonstrate a considerable improvement in text classification.

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

$$\Pr(r = \text{Rationale} \mid x,\ y = \text{Responsive}) \cdot \Pr(y = \text{Responsive} \mid x)$$

Here x consider as text or doument, y is the model-labelled title of the file (‘responsive’ or ‘not responsive,’ for example), and r is a typescript extract from x.

Dong et al. (2017), humans cannot directly understand ensemble models; it needs a additional urbane method to understand the result, decision trees, and fuzzy logic required to clarify the outcomes. Moreover, multipart mock-ups, such as deep learning representations 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. According to recent studies, it is frequently used to detect text snippets as an elucidation for the sorting of a file. The term "rationale" for a document refers to a passage of text that enlightens the organisation of the text.

Qader et al. (2019) explained supervised methods, which, don't need ex-post authentication because they pursue to 'emulate' what persons do by determining outlines in papers labelled 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.

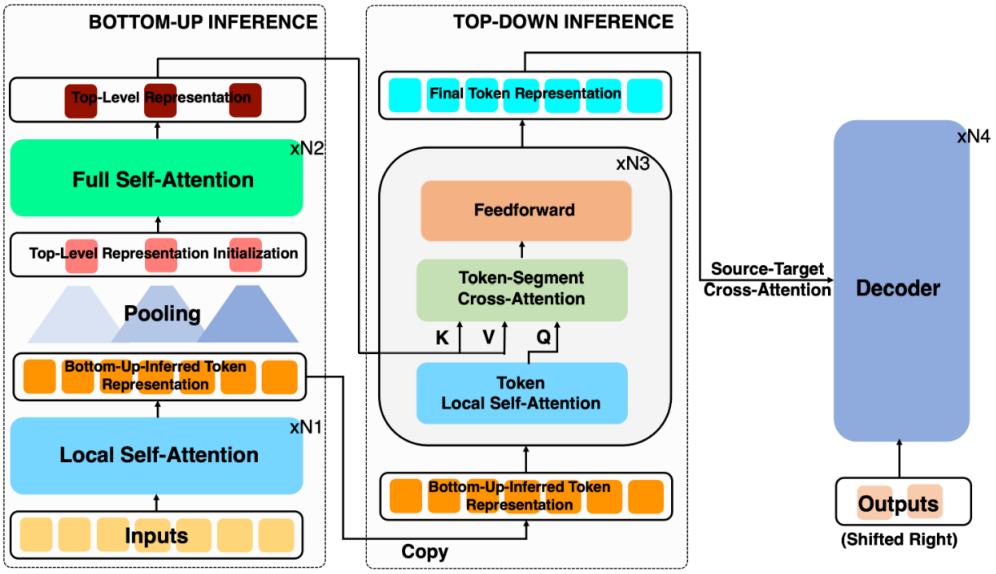

Figure 6 : Overview of text Summarization using transformer model (Sutskever (2014), p. 2)

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. The jumbled nature of text makes it difficult and time-consuming to analyse, analyse, organize, and filter through text data, which prevents most businesses from making the most of it.

Gunning et al. (2019) explained the Defence Advanced Research Projects Agency (DARPA) theory that wished-for a new way for Explicable AI (XAI). A similar concept applies to explicable machine learning application, where forecasts or organisations produced by a analytical machine learning application or model are explicable and understandable to humans. In XAI systems, actions or conclusions are humanoid explicable - "machines comprehend the situation and atmosphere in which they function, and over time shape fundamental descriptive models that licence them to define real marvels." Explainable machine learning models were the main focus of the majority of explainable AI. (A) An explanation using a model. (b) An explanation based on a prediction.

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

Kabir et al. (2018), Word clouds, count frequencies, tables, and pie charts are common examples of Quantities Visualization (QViz). Sentimentality examination, Semantics, and NLP 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.

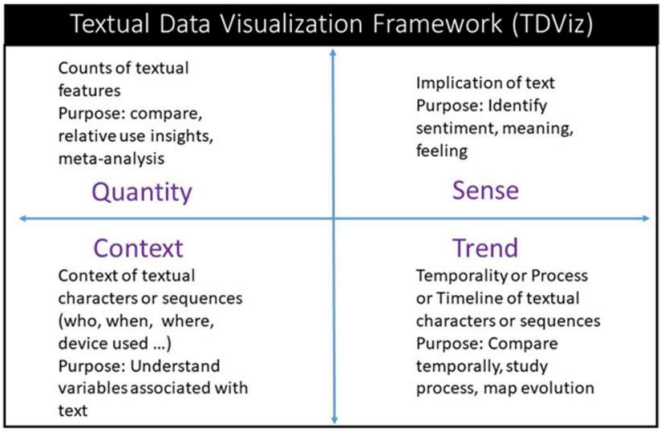

Figure 7 : Textual data visualization framework (Conner et al. (2019). p. 5)

CHAPTER 3

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 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 allow lawyers to focus more on comprehending what the documents imply, acquiring the necessary insights, and giving the client valuable advice in a timely manner rather than wasting time slogging through documents and labelling significant clauses.

3.1 Information Extraction (Pre-Processing Text)

S.Cerietal. et al. (2013), The amount of unstructured data has increased in recent years. The tremendous amount of data comprises a massive amount of information, which can be overwhelming at times and cause us to miss out on important information. However, if this data is used correctly, it can lead to a wide variety of beneficial outcomes.

Given the rapid growth of unstructured information owed to the massive capacity of data, information extraction techniques face new difficulties. Traditional information extraction algorithms could not handle this amount of unstructured huge data. These Information Extraction methods required to be updated due to the large volume and diversity of data. Recent advancements in technology have made it possible to identify and summarize Extraction problems using natural language processing methods.

NLP is an area of artificial intelligence (AI) that allows robots to read, comprehend, and extrapolate sense from social tongues. It is a crucial element of a vast array of software programs that we utilize on a regular basis.

Data abstraction is the procedure of automatically extracting organised data from unstructured and/or semi-structured (XML, Json) machine-readable papers and other automatically represented

sources(IE).

```
 THE SUPREME COURT OF PAKISTAN
(Appellate Jurisdiction)


PRESENT:
MR. JUSTICE UMAR ATA BANDIAL, HACJ
MR. JUSTICE SYED MANSOOR ALI SHAH
MR. JUSTICE MUHAMMAD ALI MAZHAR




CIVIL PETITION NO.285 5 OF 2017.
(Against the Judgment of Lahore High
Court, Bahawalpur Bench, Bahawalpur
dated 11.04.20107 passed in Civil Revision
No. 383 /2010 )

Muhammad Sarwar and others
```

Figure 8 : Text Extraction Using NLTK

The utmost common podium for fashioning Python programming language that use human language data is called NLTK. Along with wrappers for powerful NLP libraries, a lively discussion board, and a variety of text processing libraries for different labelling or classification procedures, tokenization, stemming, tagging, parsing, and semantic perceptive, it offers simple interfaces to more than fifty corpora and lexical data, including WordNet.

Appreciations to a hands-on guide covering software design rudiments together with themes in computational linguistics and thorough (Application programable Interface) API documentation, NLTK is acceptable for linguists, technologists, scholars, teachers, researchers, and industry workers equally. For Windows, Mac OS X, and Linux, NLTK is accessible. The fact that NLTK is a community-driven, open-source, free development is its sturdiest feature.

NLTK has been acclaimed as "an outstanding collection to show with natural language" and "a wonderful tool for teaching, working in computational linguistics using Python."

3.1.2 Tokenization

Siddalingappa et al. (2015), The technique of breaking up a big sample of text into words is known as tokenization. This is necessary for jobs involving natural language processing, where each word must be recorded and submitted to additional analysis, such as classification and counting for a specific sentiment, etc. To do this, a library called the Natural Language Tool kit (NLTK) is used. Before starting the Python program for word tokenization, install NLTK.

Tokenization is a easiest way that alters unprocessed input into a meaningful information string. Even though it is most recognised for its uses in cybersecurity and the formation of NFTs, tokenization is an essential stage in the natural language processing procedure. Tokenization is a technique used in natural language dealing out to break down phrases and paragraphs hooked on simpler language-assignable elements.

The initial stage of the NLP procedure is collecting the information (a sentence) and break it into logical parts (words).

This may seem clear, but by break a sentence down into its essential pieces, a machine is able to realise all the parts and the whole. This will style it easy for the program to understand both the individual words and how they suitable into the complete transcript. This is vital as it allows the computer to count the occurrence of exact arguments and the places in the text where they typically seem. Future phases of natural language processing will hinge on on this.

3.1.3 Kinds of Tokenization

Bird, et al. (2015), There are numerous diverse approaches that are used to discrete arguments to tokenize them, and these approaches will basically change future stages of the NLP process.

a. Word Tokenization

The most used type of tokenization is word tokenization. It employs delimiters (characters like ", " or ";" or "","") to divide the data into its corresponding words when there are natural disruptions, like as gaps in voice or spaces in text. Although this is the simplest method for breaking up voice or text into its component pieces, it has several disadvantages.

Word tokenization has a hard time distinguishing unidentified or out of vocabulary (OOV) words. This is frequently resolved by substituting mysterious words with a straightforward nominal that indicates the absence of a word. This is a haphazard solution, especially as the five "unknown" word tokens could represent five entirely distinct unknown words or just one word.

The lexis that arguments tokenization is trained with governs how accurate it is. The best loading of confrontations for these representations must attack a cooperation among correctness and efficiency. An NLP based application would be more correct if it had the lexis of a full vocabulary, but that's not continuously the finest method. This is mainly legal for mock-ups who are being trained for particular usages.

```
['C.P.No.2855', 'of', '201', '7', '-:4', ':', '-', 'perverse', 'or', 'contrary', 'to', 'the', 'law', 'but', 'the', 'interferenc
e', 'for', 'the', 'mere', 'fact', 'that', 'the', 'apprais', 'al', 'of', 'evidence', 'may', 'suggest', 'another', 'view', 'of',
'the', 'matter', 'is', 'not', 'possible', 'in', 'revisional', 'jurisdiction', '.', 'So', 'far', 'as', 'challenge', 'to', 'the',
'concurrent', 'find', 'ings', 'of', 'the', 'courts', 'below', 'in', 'the', 'revisional', 'jurisdiction', 'of', 'the', 'High',
'Court', ',', 'this', 'Court', 'has', 'held', 'in', 'the', 'case', 'of', 'Ahmad', 'Nawaz', 'Khan', 'Vs.', 'Muhammad', 'Jaffar',
'Khan', 'and', 'others', '(', '2010', 'SCMR', '984', ')', ',', 'that', 'High', 'Court', 'has', 'very', 'limited', 'jurisdictio
n', 'to', 'interfere', 'in', 'the', 'concurrent', 'conclusions', 'arrived', 'at', 'by', 'the', 'courts', 'below', 'while', 'exe
rcising', 'power', 'under', 'section', '115', ',', 'C.P.C', '.', 'Similar', 'view', 'was', 'taken', 'in', 'the', 'case', 'of',
'Sultan', 'Muhammad', 'and', 'another', '.', 'Vs.', 'Muhammad', 'Qasim', 'and', 'others', '.', '(', '2010', 'SCMR', '1630',
')', 'that', 'the', 'concurrent', 'findings', 'of', 'three', 'courts', 'below', 'are', 'not', 'opened', 'to', 'question', 'at',
'the', 'revisional', 'stage', '.', '7', '.', 'In', 'our', 'considerate', 'view', ',', 'the', 'order', 'passed', 'by', 'the', 'H
igh', 'Court', 'does', 'not', 'suffer', 'from', 'any', 'misreading', 'or', 'non-reading', 'of', 'evidence', 'nor', 'any', 'othe
r', 'illegality', 'and', 'or', 'irregularity', 'was', 'called', 'our', 'attention', 'for', 'justifying', 'any', 'interference',
'.', 'This', 'Civil', 'Petition', 'for', 'Leave', 'to', 'Appeal', 'was', 'dismissed', 'a', 'nd', 'leave', 'was', 'refused', 'b
y', 'our', 'short', 'order', '.', 'Above', 'are', 'the', 'reason', 's', 'in', 'the', 'aid', 'of', 'short', 'order', '.', 'Actin
g', 'Chief', 'Justice', 'Judge', 'Judge', 'Islamabad', ',', '17.09.2021', 'Khalid', 'Approved', 'for', 'reporting']
```

Figure 9 : Word Tokenization

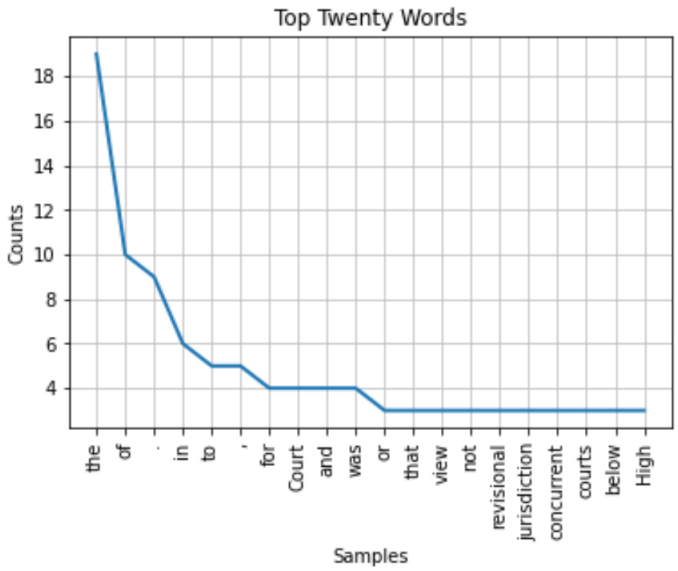

Figure 10 : Top Twenty Words

b. Character Tokenization

Tokenization of characters was developed to resolve some of the glitches related with word tokenization. It entirely splits text into characters rather than dividing it up into words. In contrast to word tokenization, this enables the tokenization process to maintain information about OOV terms.

Because the amount of the "vocabulary" is limited to the number of letters required by the language, character tokenization doesn't ache the same language glitches as word tokenization. A character tokenization vocabulary for English, for instance, would contain roughly 26 characters.

Although character tokenization addresses OOV problems, it is not deprived of its own drawbacks. Even short statements can be broken down into characters rather than words, greatly lengthening the output. Our earlier case, "what restaurants are close," is separated into four tokens using word tokenization. Contrarily, character tokenization reduces this to 24 tokens, giving the user a 6X boost in available tokens.

Understanding how the characters relate to the words' meanings is another step added by character tokenization. Sure, character tokenization can draw further conclusions, such the fact that the phrase above has five "a" token. The goal of NLP, meaning interpretation,

is further eschewed by this tokenization technique.

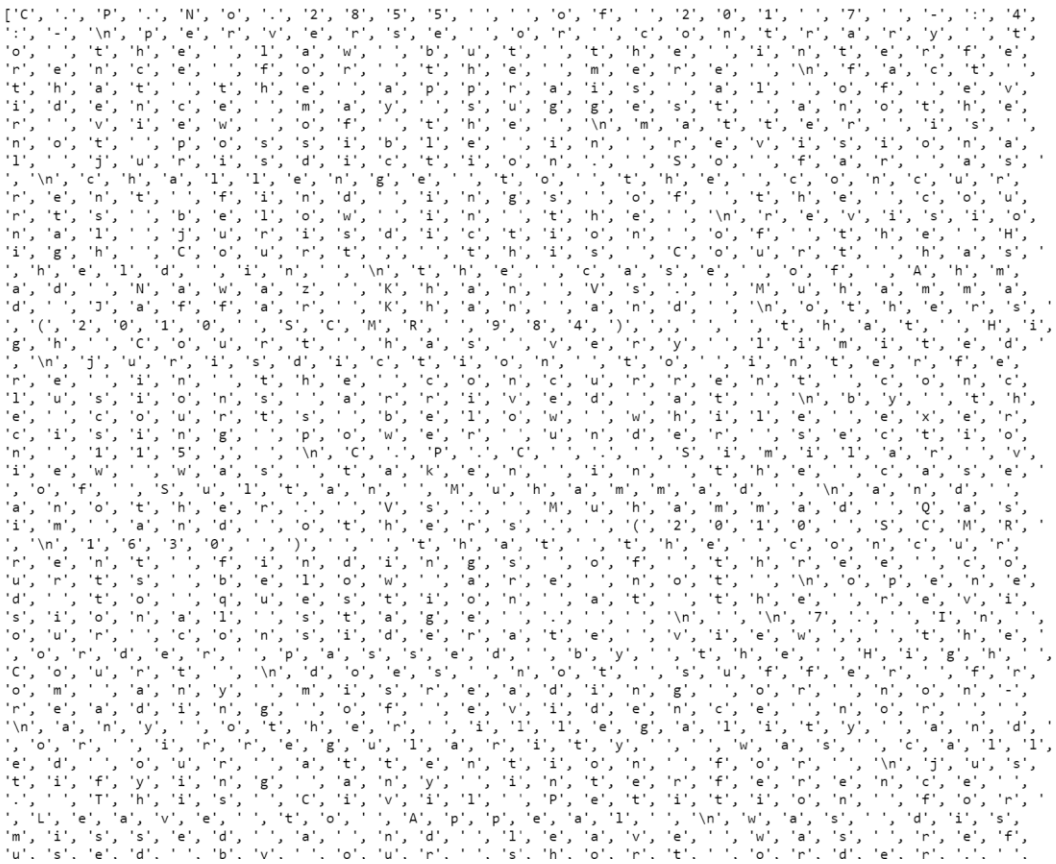

Figure 11: Character Tokenization

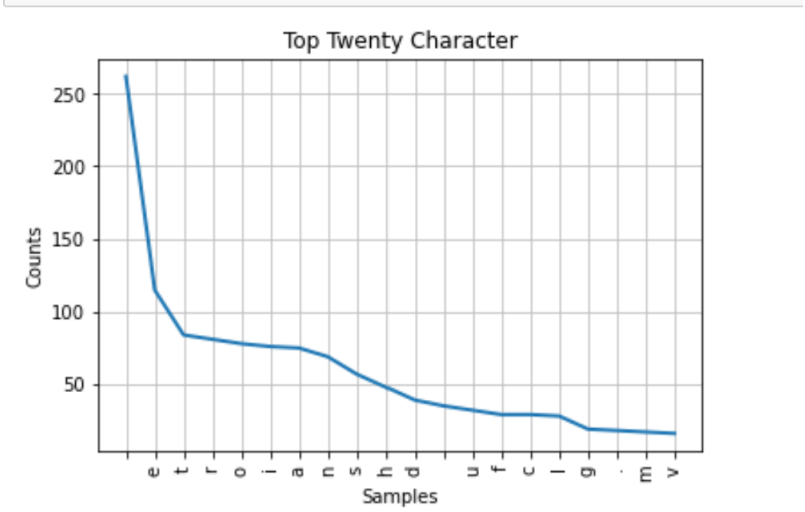

Figure 12 : Top Twenty Characters

c. Sentence Tokenization

Similar to word tokenization, sub word tokenization uses certain linguistic rules to further dissect individual words. They often use cutting off affixes as one of their key tools. Prefixes, suffixes, and infixes can aid programs in comprehending the function of a word because they alter the basic meaning of words. This can be particularly helpful for terms that are not in your lexicon because figuring out an attach can deliver a package more understanding of how unfamiliar words work.

The sub word machine learning model will look for these sub arguments and distinct the words that include them into their individual components. For instance, the question "What is the uppermost structure?" would be divided into "what is the tallest building" and "ing."

How does this technique assistance the problem of OOV words? Let's look at an example:

Maybe a extra complex word for a machine would be "plotting," the present tense of the verb "plot," which means to plot or scheme. It's improbable that many basic vocabularies include the word "machinating".

```
['C.P.No.2855  of 201 7 -:4:-\nperverse or contrary to the law but the interference for the mere \nfact that the apprais al of
evidence may suggest another view of the \nmatter is not possible in revisional jurisdiction.',
 'So far as \nchallenge to the concurrent find ings of the courts below in the \nrevisional jurisdiction of the High Court, thi
s Court has held in \nthe case of Ahmad Nawaz Khan Vs. Muhammad Jaffar Khan and \nothers (2010 SCMR 984),  that High Court has
very limited \njurisdiction to interfere in the concurrent conclusions arrived at \nby the courts below while exercising power
under section 115, \nC.P.C .',
 'Similar view was taken in the case of Sultan Muhammad \nand another.',
 'Vs. Muhammad Qasim and others.',
 '(2010 SCMR \n1630 )  that the concurrent findings of three courts below are not \nopened to question at the revisional stage
.',
 '7.',
 'In our considerate view, the order passed by the High Court \ndoes not suffer from any misreading or non-reading of evidence
nor  \nany other illegality and or irregularity  was called our attention for \njustifying any interference .',
 'This Civil Petition for Leave to Appeal \nwas dismissed a nd leave was refused by our short order.',
 'Above are \nthe reason s in the aid of short order.',
 'Acting Chief Justice  \n \n \n \n         Judge  \n \n \n \nJudge  \n \nIslamabad ,  \n17.09.2021  \nKhalid    \nApproved for
reporting']
```

Figure 13 : Sentence Tokenization

This term would be changed into an unidentified token if the NLP application used word tokenization. The word might be divided into a "unknown" token and a "ing" token if the NLP model had used sub word tokenization. From there, it can draw insightful conclusions about the word's role in the phrase.

	Phrases	tokenized
0	C.P.No.2855 of 201 7 -:4:-\nperverse or contr...	[C.P.No.2855, of, 201, 7, -:4, :, -, perverse,...
1	So far as \nchallenge to the concurrent find i...	[So, far, as, challenge, to, the, concurrent, ...
2	Similar view was taken in the case of Sultan M...	[Similar, view, was, taken, in, the, case, of,...
3	Vs. Muhammad Qasim and others.	[Vs., Muhammad, Qasim, and, others, .]
4	(2010 SCMR \n1630) that the concurrent findi...	[(, 2010, SCMR, 1630,), that, the, concurrent...

Figure 14 : Phrases Tokenization

However, what data can a machine derive from a single suffix? For instance, the common “ing” suffix serves a few obviously well-defined purposes. It could change a verb into a noun, as in the case of the verb "shape" which became the noun "structure." Additionally, it can change a verb into its present participle, such as turning the verb "run" into "running."

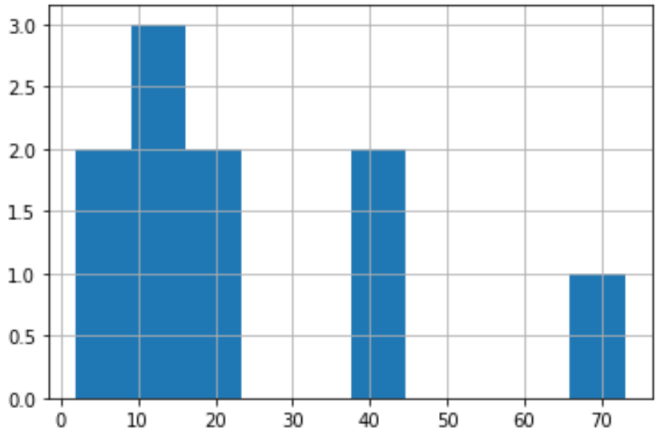

Figure 15 : Character frequency in sentence

Given this knowledge about the 'ing' suffix, an NLP model can infer a number of useful things about any word that has the subword 'ing. If a word contains the prefix "ing," it is aware that it is either a current verb or a verb that has been transformed into a noun. This significantly reduces the possible uses of the ambiguous word "machinating" in a statement.

There are numerous techniques to tokenize text or speech, but the success of each one depends greatly on the value of the original NLP code. The first step is tokenization, which breaks down a complex data input into workable building pieces for the natural language processing algorithm.

Individuals and machineries are able to interconnect more effectively as deep learning models for natural language processing continue to advance. This is just one example of the various ways tokenization is laying the groundwork for ground-breaking technical advances.

3.1.4 Stop Words

Sarica, et al. (2021), A stop word is a usually select phrase that a search engine has been set up to ignore, both while indexing entries for searching and when retrieving them as the result of a search query. "The," "a," "an," and "in," for instance, are stop words.

We don't want these expressions to use any additional database processing time or stock planetary. We can rapidly eliminate the arguments you reflect to be stop words by possession track of them. 16 different languages' worth of stop words are stored in the NLTK (Natural Language Toolkit) database in Python. The NLTK data directory is where you may find them.

Types of Stop Words

Blanchard, et al. (2007), Stop words are essentially a "single group of words" that, depending on the context, can mean a number of different things. For instance, in some circumstances, removing all stop words—including determiners like the, a, and an, prepositions like above, across, and before, and some adjectives like good and nice—could be a suitable stop word list. However, this might not be a desirable thing for other applications. For instance, eliminating adjectives like "good" and "lovely" as well as negations like "not" can cause sentiment analysis algorithms to malfunction. Depending on the requirements of the application, one may opt to employ a basic stop list that only includes determiners, determiners with prepositions, or only coordinating conjunctions. Examples of minimal stop word lists that you can use:

- **Determiners** – The, a, an, and another are examples of determiners that frequently follow nouns in sentences.
- **Coordinating conjunctions** – Examples of coordinating conjunctions that join clauses, phrases, and words include: for, an, nor, but, or, yet, and so.
- **Prepositions** – Examples of prepositions that communicate spatial or temporal links are in, under, toward, and before. We may desire a completely distinct collection of stop words in other contexts, such as clinical literature, which are more specialized. In contrast to phrases like "heart failure" and "diabetes," terms like "mcg," "doctor," and "patient" may have less discriminating power when used to create intelligent applications. Instead of using a published stop word list in these circumstances, we can alternatively create domain-specific stop words.

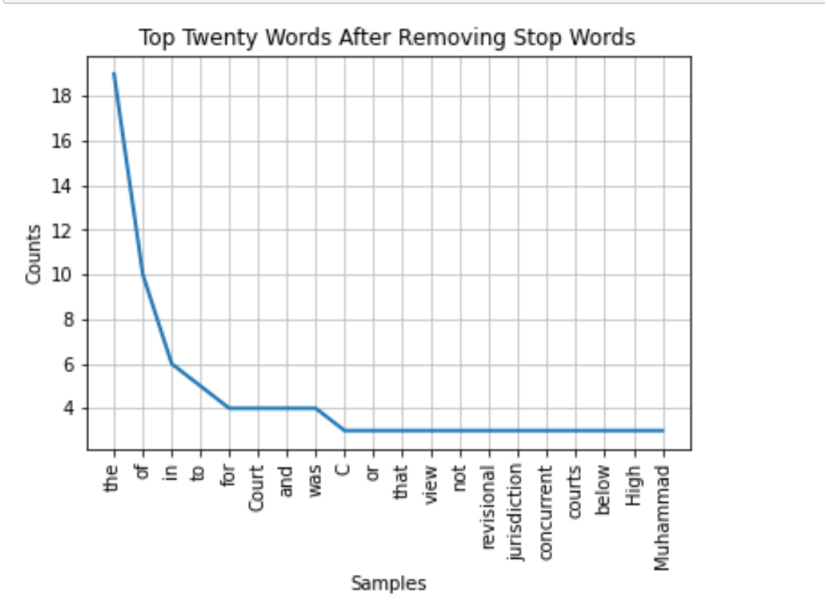

Figure 16 : Stop Words Removal

3.1.5 Part of Speech Tagging

Chiche, et al. (2022). The procedure of categorising a word in a text (corpus) as belonging to a specific chunk of speech built on its definition and context—that is, how it is used in relation to other words in a phrase, sentence, or paragraph—is known in mass linguistics as part-of-speech tagging call “POS”. This process is also known as grammatical tagging or word-category disambiguation. An abbreviated form of this is typically taught to school-age children when classifying words as nouns, verbs, adjectives, adverbs, etc.

Toward simply map words to their “POS” tags is much easier than to determine part of speech tags. This is because POS labelling is not a universal idea. It is absolutely possible for the same word to be given a different “POS” in successive phrases, depending on the context. There can therefore be no standard mapping for POS tags.

“POS” tagging might not be the answer to a specific NLP problem on its own. It is a step that is performed to simplify a number of different concerns, though. Think about a few applications for POS tagging in numerous NLP jobs.

- ## Rule-based POS Tagging

Pham B (2022), Rule-based POS tagging is one of the earliest methods of classification. Rule-based tagging program search a lexicon or lexicon for likely labels for respective word. When a word has more than one suitable tag, rule-based taggers use hand-written rubrics to select the appropriate label. Rule-based tagging also permits disambiguation by looking at a word's linguistic traits and those of its related terms. Consider the example of assuming that a word must be a noun if it comes after an article.

As the name implies, all of this material is oblique in the method of rules in rule-based POS tagging. These rules may be either −

- Rules if context pattern
- Or else, as the representation of lexically ambiguous sentences intersected with Regular expressions built into finite-state automata.

By virtue of its two-stage architecture, Rule-based POS tagging is also understandable –

The primary phase involves assigning each word a list of potential parts of speech using a dictionary.

Second step: In the second stage, it sorts the list down to a single part-of-speech for each term using extensive lists of hand-written disambiguation criteria.

- ## Properties of Rule-Based POS Tagging

The following characteristics are shared by rule-based POS taggers.

These tagging programs are knowledge driven taggers. Rule base POS tagging's rules are created by hand.

- Rules are used to code the information.
- There are just a small number of rules—roughly 1,000.
- In rule-based taggers, smoothing and language modelling.

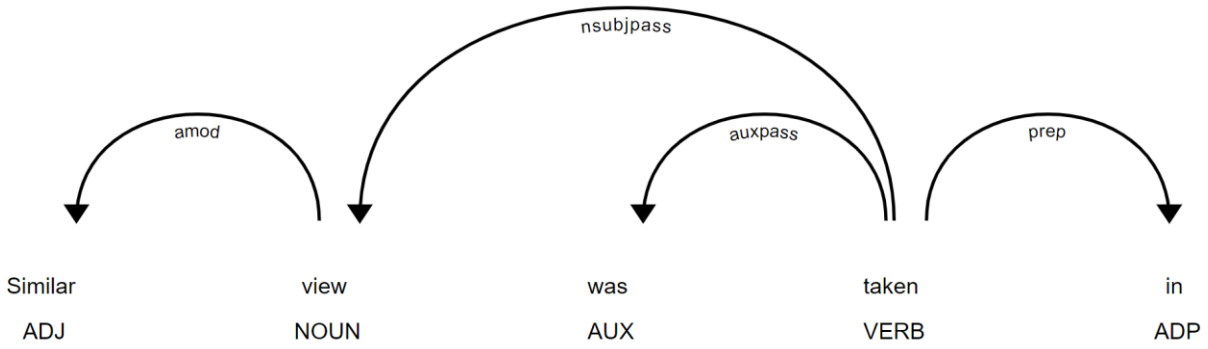

Figure 17 : Part Of Speech Tagging (POS)

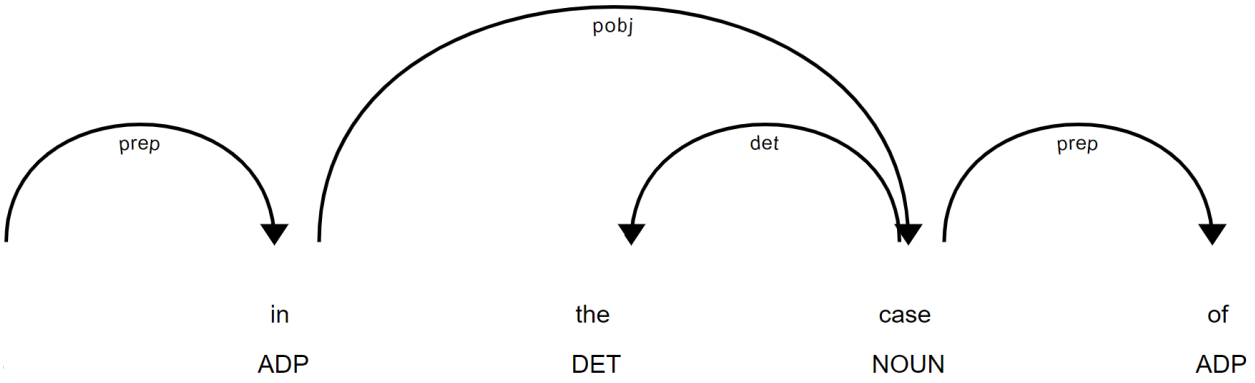

Figure 18 : Part Of Speech Tagging (POS)

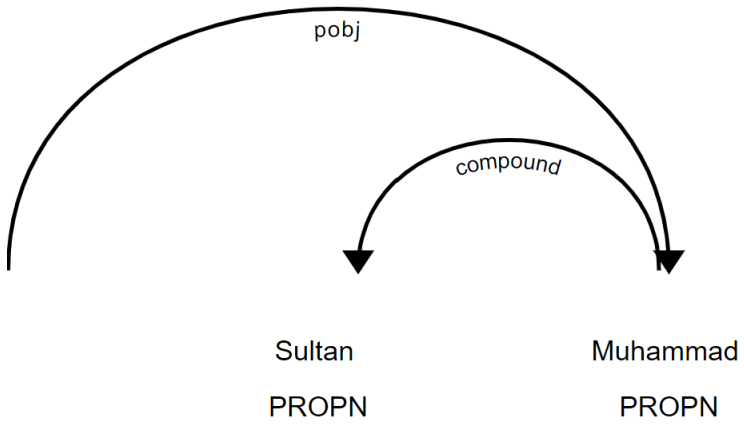

Figure 19 : Part Of Speech Tagging (POS)

3.1.6 Stemming

Adam, G et al. (2010). The NLP method recognised as stemming assists in preparing text, words, and documents for text normalization by reducing word inflection to its root forms.

According to Wikipedia, modulation is the procedure done which a word is changed to convey a range of linguistic groups, together with tenses, situation, speech, feature, individual, quantity, gender, and attitude. So, if a term may have numerous inflected forms, when numerous varied forms appear in the similar transcript, the NLP process becomes more difficult. Thus, we use stemming to break down words into their most elementary method, or stem, which might or might not be a valid term in the linguistic.

For example, "connect" is the stem of the terms "connections," "connected," and "connects." On contrary, the unknown word "troubl" is the foundation of trouble, troubled, and problems.

The initial paper of stemmer was published by Julie Beth Lovin in 1968. This article was revolutionary at the time and had a big impact on advanced effort in this part. She cites three major stemming algorithm attempts that were made previously: one by Princeton

University's John W. Tukey, another by Harvard University's Michael Lesk working under Professor Gerard Salton, and a third procedure created by Los Altos, California's R and D Consultants under the leadership of James L. Dolby.

Another stemmer by Martin Porter was included in the matter of the paper Package for July 1980. Because of its widespread use, this stemmer eventually came to be accepted as the standard for English stemming. Dr. Porter received the Tony Kent Strix prize in 2000 for his contributions to information retrieval and stemming.

- ## Stemming Importance

Bouras et al. (2008), As was already mentioned, a single term in the English language contains many different variations. These differences in a transcript corpus produce redundant data for building NLP or ML models. This application may not perform as expected. To figure a robust application, it is vital to regularise transcript by eliminating recurrence and stemming arguments to their basic forms.

- ## Types of Stemmers

Giorgos et al. (2010), There are numerous types of stemming procedures, and all of them are encompassed in Python NLTK. Let us have a look at them below.

1. Porter Stemmer

In 980 Martin porter wrote the stemmer algorithm for porter stemmer. The method employs reduction phases of five words, respectively through its actual set of mapping norms. The Porter Stemmer, the unique stemmer, is famous for its comfort of use and swiftness. The resultant stem is frequently a shorter phrase with the same root meaning.
Porter Stemmer is a unit in NLTK that gears the Porter Stemming method. Let us inspect this with the help of a sample.

```
High  :  high
Court  :  court
has  :  ha
very  :  veri
limited  :  limit
jurisdiction  :  jurisdict
to  :  to
interfere  :  interfer
in  :  in
the  :  the
concurrent  :  concurr
conclusions  :  conclus
arrived  :  arriv
at  :  at
by  :  by
the  :  the
courts  :  court
below  :  below
while  :  while
```

Figure 20 : Porter Stemmer

2. Snowball Stemmer

Another achievement of Martin Porter is Snowball Stemmer. In comparison to the first Porter Stemmer, the technique utilized here is further precise and is stated to as "English Stemmer" or "Porter2 Stemmer" Giorgos et al. (2010). The Snowball stemming method is implemented by the NLTK module “SnowballStemmer” Let's study this stemming technique with an illustration.

```
High ---> high
Court ---> court
has ---> has
very ---> veri
limited ---> limit
jurisdiction ---> jurisdict
to ---> to
interfere ---> interfer
in ---> in
the ---> the
concurrent ---> concurr
conclusions ---> conclus
arrived ---> arriv
at ---> at
by ---> by
the ---> the
courts ---> court
below ---> below
while ---> while
exercising ---> exercis
```

Figure 21 : Snowball Stemmer

3. Lancaster Stemmer

Lancaster Stemmer is a upfront tool, even though it frequently crops results with extreme stemming technique, makes stems unintelligible or non-linguistic. The Lancaster stemming method is implemented by the NLTK.

```
High ---> high
Court ---> court
has ---> has
very ---> very
limited ---> limit
jurisdiction ---> jurisdict
to ---> to
interfere ---> interf
in ---> in
the ---> the
concurrent ---> concur
conclusions ---> conclud
arrived ---> ar
at ---> at
by ---> by
the ---> the
courts ---> court
below ---> below
while ---> whil
exercising ---> exerc
```

Figure 22 : Lancaster Stemmer

4. Regexp Stemmer

Regular expressions are used by the regex stemmer to recognize structural fixes. Sub part of string that contest a regular expression will be eliminated. The NLTK module reg exp stemmer implements the Regex stemming method.

```
High ---> High
Court ---> Court
has ---> has
very ---> very
limited ---> limited
jurisdiction ---> jurisdiction
to ---> to
interfere ---> interfer
in ---> in
the ---> the
concurrent ---> concurrent
conclusions ---> conclusion
arrived ---> arrived
at ---> at
by ---> by
the ---> the
courts ---> court
below ---> below
while ---> whil
exercising ---> exercis
```

Figure 23 : Regexp Stemmer

3.1.7 Lemmatization

Khyani, et al. (2021), The process of integration of a word's numerous modulated procedures into a sole component for examination is recognised as lemmatization. Lemmatization gives the arguments additional context, just like stemming does. As a result, it creates a single term from terms having related meanings.

Text pre-processing includes lemmatization and stemming. Users frequently get these two terms mixed up. Some individuals conflate these two. Since lemmatization does morphological analysis on the words, it performs better than stemming.

Various Approaches to Lemmatization:

Sun et al. (2016), Lemmatization will be covered using nine different methods, several examples, and code implementations. Word net, word net (having POS tags) , Blob of text , Blob of text with POS tags, SPACY, Tree Tagger, Pattern, Gensim and at last Stanford cored NLP.

1. Wordnet Lemmatize

Wordnet is a verbal knowledge base of more than two hundred dialects that is available to the over-all community and offers semantic correlations between its terms. One of the first and most popular lemmatize techniques is this one.

- It can be found in Python's nltk library
- Wordnet creates semantic relationships between words. such as synonyms
- Synsets are used to organize synonyms.
- synsets : a group of data elements that are semantically equivalent.

```
High ---> High
Court ---> Court
has ---> ha
very ---> very
limited ---> limited
jurisdiction ---> jurisdiction
to ---> to
interfere ---> interfere
in ---> in
the ---> the
concurrent ---> concurrent
conclusions ---> conclusion
arrived ---> arrived
at ---> at
by ---> by
the ---> the
courts ---> court
below ---> below
while ---> while
```

Figure 24 : Wordnet Lemmatizer

2. Wordnet Lemmatizer (having POS tag)

We found that the Wordnet results from the aforementioned method fell short of expectations. After lemmatization, words like "sitting," "flying," etc. remained unchanged. This is due to the fact that these words are not treated as verbs but rather as nouns in the provided sentence. POS (Part of Speech) tags help us get around this.

We include a tag describing its type with a specific word (verb, noun, adjective etc). As An Example,

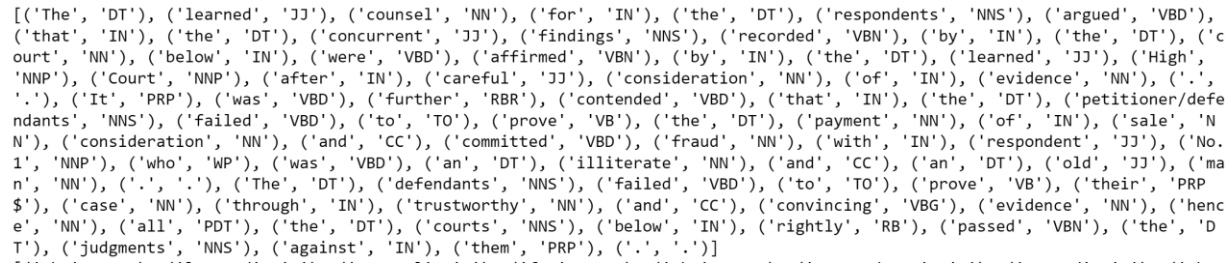

Figure 25 : Wordnet Lemmatizer (with POS tag)

3. Text Blob

Python's TextBlob public library is castoff to procedure word-based information. It offers a straightforward API to use its methods and carry out fundamental NLP operations.

4. TextBlob (with POS tag)

The constraints of this strategy are the same as those of the Wordnet approach when the proper POS tags are not used. In order to solve this issue, we make advantage of one of the TextBlob module's more potent features, "Part of Speech" tagging.

5. SPACY

Large amounts of text can be parsed and "understood" with the free and open-source Python library known as spaCy. There are different models that support various languages available (English, French, German, etc.).

6. Tree Tagger

The Tree Tagger is a tool for adding part-of-speech and lemma information to text annotations. The Tree Tagger can be modified to tag more languages in the presence of a training corpus that has been manually tagged. Over 25 languages have been successfully tagged using it.

7. Pattern

A popular Python module for web mining, NLP, machine learning, and network analysis is called Pattern. It has several practical NLP features. Additionally, it has a unique feature that we'll cover shortly.

8. Gensim

Large text collections can be handled by Gensim via data streaming. Based on the pattern package we loaded previously, it has lemmatization capabilities.
Lemmatization can be done using the function gensim.utils.lemmatize(). In Python, this function is part of the utils module.

We may extract UTF8-encoded tokens in their fundamental form=lemma using this lemmatizer from the pattern. defaults to only taking into account nouns, verbs, adjectives, and adverbs (all other lemmas are discarded).

9. Stanford CoreNLP

- Some of the examples of stand ford core NLP are sentence borders, POS, NER, number and temporal data, reliance and electorate analyses, sentimentality, citation credits, and relations means Using Core NLP, linguistic annotations for text can be extracted for token and sentence.
- Your single platform for natural language processing is Core NLP.
- Multiple language support such Arabic, Chinese, English, French, German, and Spanish are amongst the six dialects that Core NLP now provisions.

3.1.8 Word Cloud

Heimerl, et al. (2014), Individually word's magnitude in a word cloud, a text data visualization tool, specifies its occurrence or relevance. To highlight high text data in a graphical way then word cloud may use for it. Word clouds are a popular tool for analysing information from social networking sites.

Figure 26 : Word Cloud

3.1.9 Name Entity Recognition (NER)

Yang. et al. (2022), One of the most frequent tasks in data preprocessing is named entity recognition (NER). It entails locating important data in the manuscript and categorizing it into a number of predetermined groups. A constant subject of discussion or reference in a book is referred to as an entity. NER is an example of NLP.

The two steps involved in NLP's fundamental two-step procedure are listed below:

- Identifying items in the text
- Sorting them into several groups

Person, Organization and place / location are some of the categories that make up NER more important.

This may include few important tasks such as time and date, expression of text, email address and numeric like percentage, money etc.

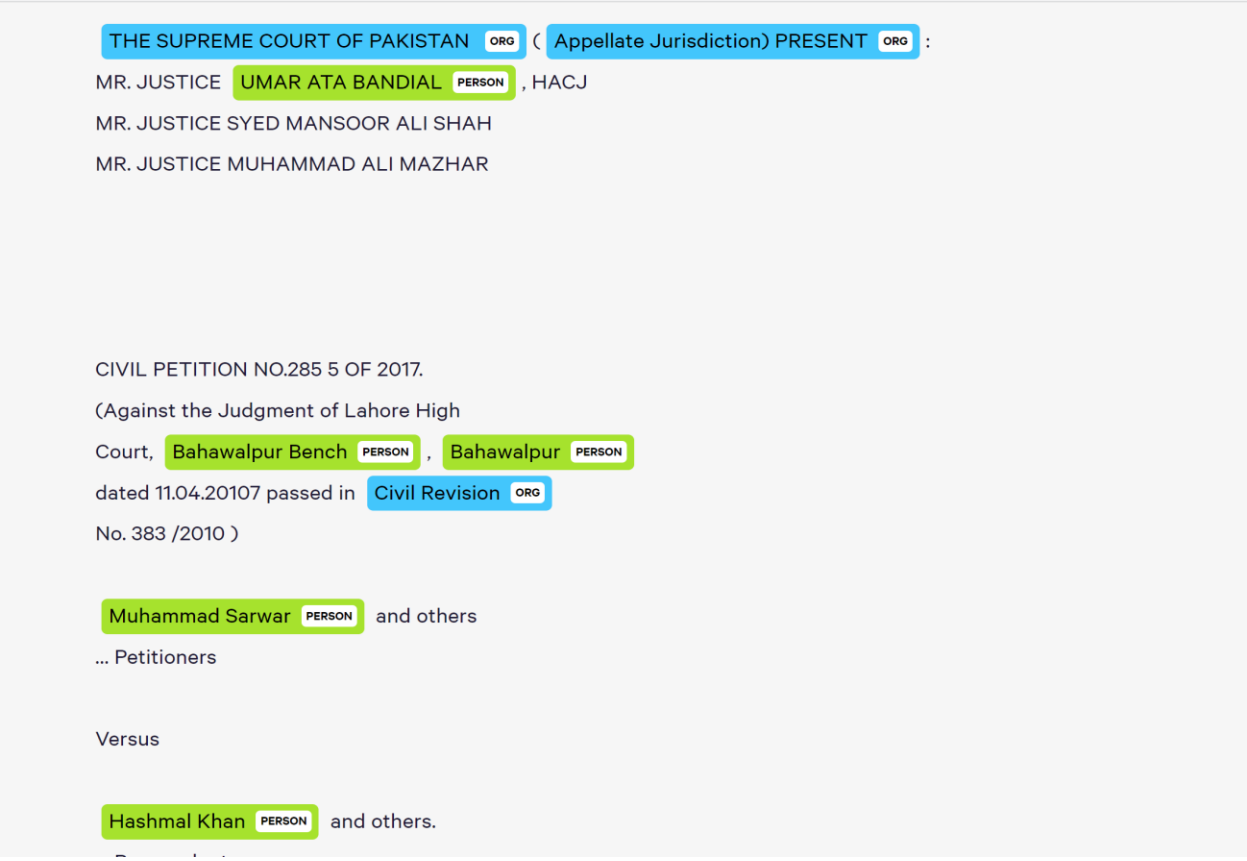

Figure 27 : Name Entity Recognition (NER)

3.1.10 Sentiment Analysis

Karamitsos. et al. (2022), Natural Language Processing, often known as NLP, is the method used by sentiment analysis systems to evaluate whether a text is good, negative, or neutral. Some methods for emotion analysis can even categorize a piece of text with more specific sentiment markers like disappointment, enthusiasm, or contempt. On manuscript information derived from social media, such as web analyses, emails, customer service chats bots, review or surveys replies, web blogs, and newscast articles, sentiment analytics are frequently conducted. Additionally, some sentiment analysis technologies use Natural Language Processing to analyze the speech tone of voice recordings or even actual phone calls.

The analytical algorithm that the sentiment analytics software employs determines its accuracy. It's vital to remember that no sentiment analysis technology, whether it's free or so pricey you can hardly explain it, is 100% error-proof. Even the most sophisticated sentiment analysis technologies have trouble picking up on people's usage of sarcasm and slang in conversation. However, automating sentiment analysis is usually lot more effective than manually going through them when you're dealing with hundreds or thousands of messages from different sources every day. Literally, sentiment analysis is the process of using sentiment analysis software to determine the emotions, feelings, or sentiment.

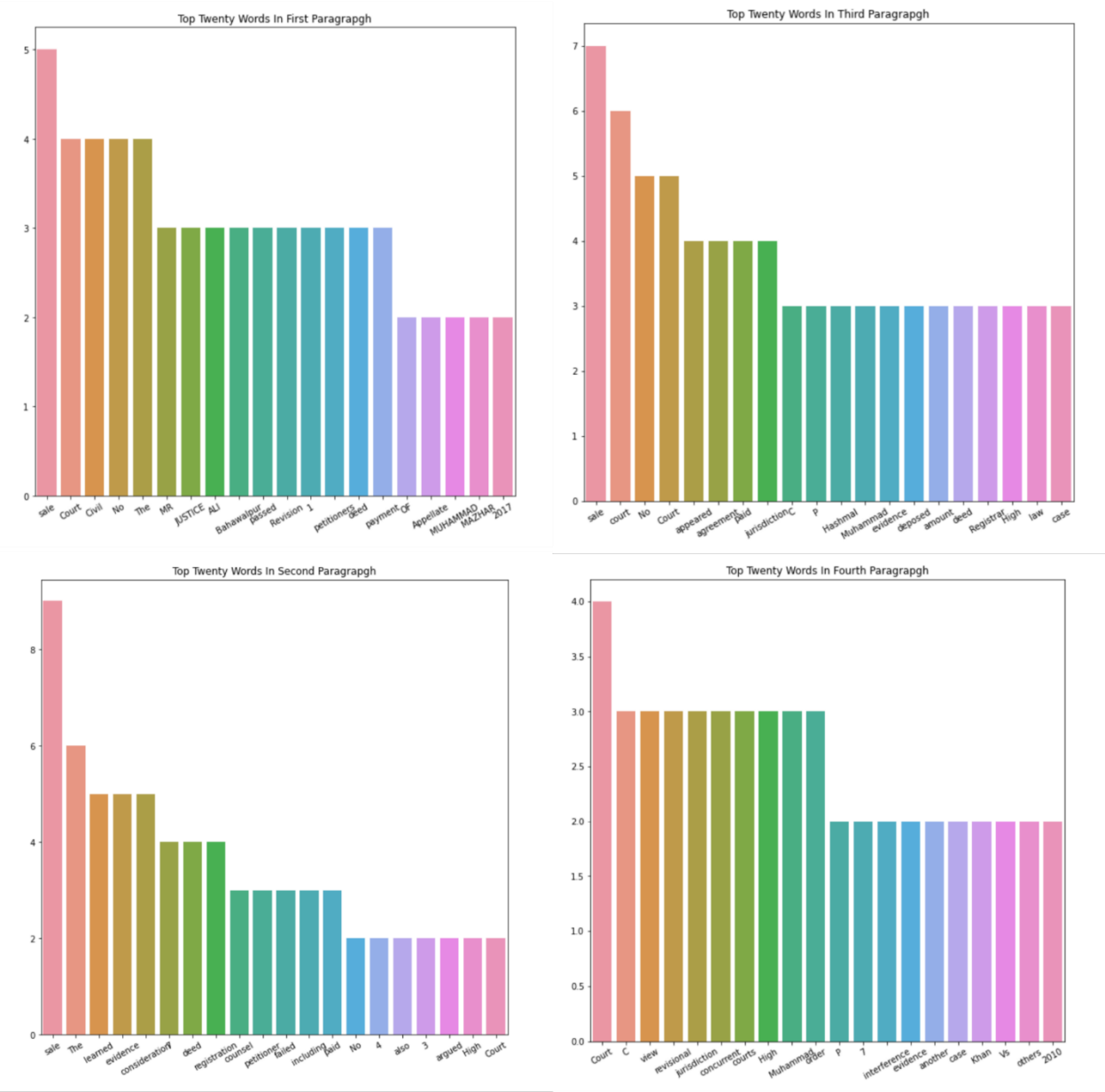

Figure 28 : Word Frequency Distribution

3.1.11 Polarity & Emotion Detection

The technique of identifying whether a word, phrase, or document is positive, negative, or neutral is known as sentiment analysis, and it is one of the most significant and often used functions in text analytics and NLP. Understanding the meaning of specific words and phrases is the first step in teaching computers to comprehend the sentiment of text. At the most fundamental level, a sentiment analysis engine must recognize that adjectives like "appalling," "morally bankrupt," and "awful" are negative, while adjectives like "brilliant," "wonderful," and "perfect" are positive. Unfortunately, sentiment and emotion are rarely expressed with such clarity.

Figure 29 : Polarity & Emotion Detection

CHAPTER 4

4.1 DATA ANALYSIS

A PDF of the full judgment is accessible on the Supreme Court of Pakistan's website.
"https://www.supremecourt.gov.pk/latest-judgements/"

Following data needs to be extract from the document

Data	Description
Court name	The court where hearing was held
Judges Name	Judges who are involved in the case
Petition No	Unique id of case
Case Type	Case type represent either is civil case or criminal
Petitioners	Person who put the case
Respondents	Case against the person
Date Of Petition	Date on which case filed
Lawyers Name	Name of lawyers of both parties
Date Of Hearing	Date when case listen by judge
Legal references	Laws reference against judgement taken

Table 1 : Data Extraction and Description

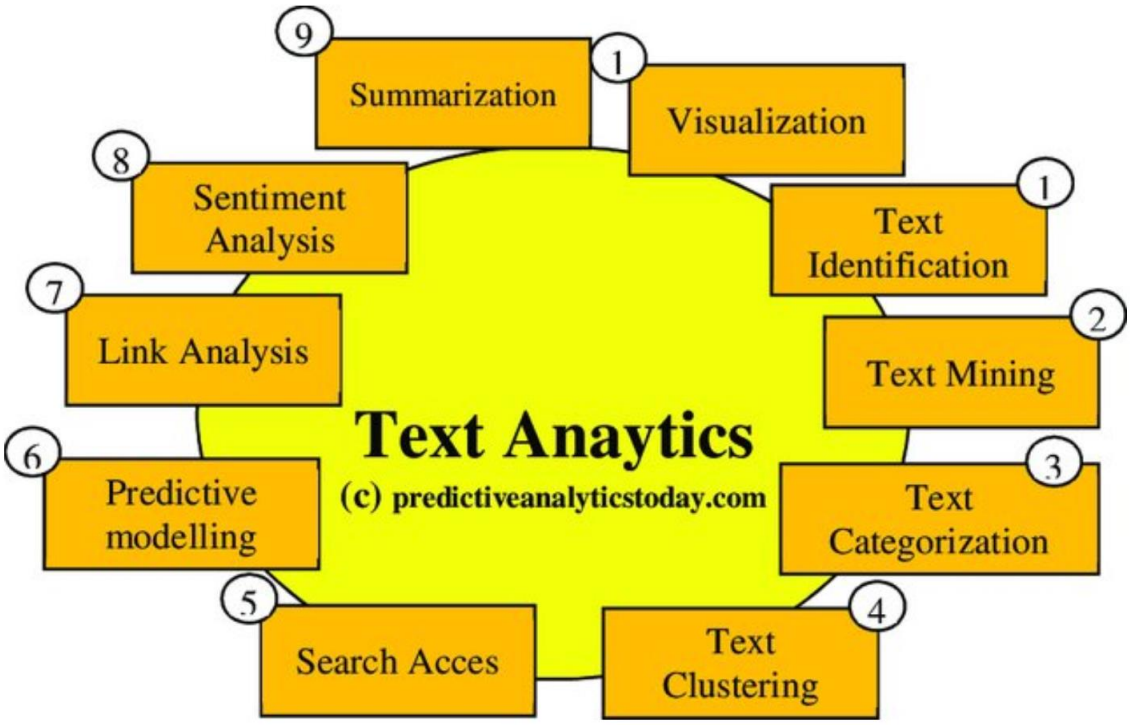

Figure 30 : Text analytics workflow (Dyevre, A. (2021) p.5).

pen-source tools like pycharm (IDE), Jupiter notebook, 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.

5.2 Future Studies & Recommendations

This NLP based model will consume as web API as well as public web portal where lawyers and judiciary can analyse each verdict which produce by SCP. Flask python framework will used to implement the web API. And Django another python framework will be used to implement the portal.

I will recommend the judicial system of Pakistan to use this model to ease their work without any flows and delay.

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/.
3. 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
4. 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
5. 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
6. 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
7. 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/.
8. 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
9. 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.
10. Dong, Y., Su, H., Zhu, J., & Bao, F. (2017). Towards Interpretable Deep Neural Networks by Leveraging Adversarial Examples. http://arxiv.org/abs/1708.05493
11. Gunning, D., & Aha, D. W. (2019). DARPA's Explainable Artificial Intelligence Program Deep Learning and Security. https://arxiv.org/abs/1904.01721.
12. Zaidan, O. F., Eisner, J., & Piatko, C. D. (2007). Using "Annotator Rationales" to Improve Machine Learning for Text Categorization *. http://cs.jhu.edu/
13. Lance B. Eliot. (2020). Legal Sentiment Analysis and Opinion Mining (LSAOM). Https://Arxiv.Org/Abs/2010.02726v1.
14. 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.
15. 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.
16. 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.
17. 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.

18. https://towardsdatascience.com/11-dimensionality-reduction-techniques-you-should-know-in-2021-dcb9500d388b
https://doi.org/10.4018/978-1- 60566-010-3.ch219.
19. 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

20. 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

21. 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

22. 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

23. 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.

24. Conner, C., Samuel, J., Kretinin, A., Samuel, Y., & Nadeau, L. (2019). A Picture for The Words! Textual Visualization in Big Data Analytics. APA.

25. 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

26. Siddalingappa, R., & Mohan, J. N. (2015). Dimensionality reduction for text preprocessing in text mining using NLTK Precision Medicine for cancer from human genome using artificial intelligence and machine learning techniques View project. https://www.researchgate.net/publication/309263381

27. Bird,Steven,EdwardLoperandEwanKlein(2009),NaturalLanguageProcessingwithPython.O"ReillyMediaInc.

28. S.Cerietal.,WebInformationRetrieval (2013),"InformationalRetrievalProcess"Data-CentricSystemsandApplications, DOI10.1007/978-3-642-39314-3_2

29. Sarica, S., & Luo, J. (2021). Stopwords in technical language processing. PLoS ONE, 16(8 August). https://doi.org/10.1371/journal.pone.0254937

30. BlanchardA.Understandingandcustomizingstopword listsforenhanced patentmapping. WorldPatInf.2007;29:308–316. https://doi.org/10.1016/j.wpi.2007.02.002

31. Chiche, A., & Yitagesu, B. (2022). Part of speech tagging: a systematic review of deep learning and machine learning approaches. In Journal of Big Data (Vol. 9, Issue 1). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1186/s40537-022-00561-y
32. Pham B. Parts of Speech Tagging : Rule-Based (2022). https:// digitalcom mons. harrisburgu.edu/cisc_student-coursework/.
33. Adam, G., Asimakis, K., Bouras, C., & Poulopoulos, V. (2010). An efficient mechanism for stemming and tagging: The case of Greek language. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6278 LNAI(PART 3), 389–397. https://doi.org/10.1007/978-3-642-15393-8_44

34. Bouras, C., Poulopoulos, V., Tsogkas, V.: PeRSSonal's core functionality evaluation: En-hancing text labeling through personalized summaries. Data and Knowledge Engineering Journal 64(1), 330–345 (2008).
35. Giorgos Adam, Konstantinos Asimakis, Christos Bouras, and Vassilis Poulopoulos. 2010. An e•cient mechanism for stemming and tagging: the case of Greek language. In International Conference on Knowledge-Based and Intelligent Information and Engineering Systems. Springer, 389–397.
36. Khyani, D., S, S. B., M, N. N., & M, D. B. (n.d.). An Interpretation of Lemmatization and Stemming in Natural Language Processing. https://www.researchgate.net/publication/348306833
37. Sun, Shiliang & Luo, Chen & Chen, Junyu. (2016). A Review of NaturalLanguage Processing Techniques for Opinion Mining Systems. InformationFusion. 36. 10.1016/j.inffus.2016.10.004
38. Heimerl, F., Lohmann, S., Lange, S., & Ertl, T. (2014). Word cloud explorer: Text analytics based on word clouds. Proceedings of the Annual Hawaii International Conference on System Sciences, 1833–1842. https://doi.org/10.1109/HICSS.2014.231
39. Yang, T., He, Y., & Yang, N. (2022). Named Entity Recognition of Medical Text Based on the Deep Neural Network. Journal of Healthcare Engineering, 2022. https://doi.org/10.1155/2022/3990563
40. Karamitsos, I., Albarhami, S., & Apostolopoulos, C. (2019). Tweet Sentiment Analysis (TSA) for Cloud Providers Using Classification Algorithms and Latent Semantic Analysis. Journal of Data Analysis and Information Processing, 07(04), 276–294. https://doi.org/10.4236/jdaip.2019.74016

41. 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