iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** II **Month of publication:** February 2023

DOI: https://doi.org/10.22214/ijraset.2023.49027

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue II Feb 2023- Available at www.ijraset.com

Legal Text Mining

Crystal Coral Martins[1], Dr. Gajanan Gawde[2]
[1]*M.E in Computer Science & Engineering, Goa College of Engineering*
[2]*Associate Professor in Computer Science & Engineering, Goa College of Engineering*

Abstract: ***The law is a vast and complicated body of knowledge, and being able to access the right information quickly and accurately can make the difference. Having access to information is essential to providing the best possible legal advice and representation to clients. It is very essential for legal practitioners and ordinary citizens to do exhaustive research related to their case. For this, they have to read extremely long judgements and try to pick out the useful information from them. To do this, the search engine presently available provides judgements but to find out a particular judgement from this list of judgements is very difficult. So, here we have proposed and developed a search engine that will make it easier to find a particular judgement.***
Keywords: Text mining, natural language processing, legal information system, unstructured data

I. INTRODUCTION

The law is an ever-changing entity that is constantly adapting to the needs of society. As a result, access to up-to-date information is essential for legal professionals in order to stay ahead of the curve. For example, when a new law is passed, it is important for lawyers to know the details of the law in order to provide their clients with the best possible legal advice. Without access to the most up-to-date information, lawyers could be providing their clients with outdated advice that could negatively affect their cases. Furthermore, access to information is important for legal professionals to be able to investigate past cases and precedent in order to decide the best way to approach their client's cases. Access to information is also important for legal professionals to be able to identify and address legal issues that their clients may not be aware of. Finally, access to information is important to legal professionals because court cases and legal proceedings can be very complex, and without the right access to information, the legal process could be unnecessarily long and difficult to manage. By having the right access to information, legal professionals can obtain the right information quickly and efficiently, saving time and money in the long run.

II. LITERATURE SURVEY

Alfirna Rizqi Lahitani et al.[1] in their paper titled "Cosine Similarity to Determine Similarity Measure: Study Case in Online Essay Assessment" talk about Automated Essay Scoring (AES) system for determining a score automatically from text document source to facilitate the correction and scoring by utilizing applications that run on the computer. AES process is used to help the lecturers to score efficiently and effectively.

They implemented the weighting of Term Frequency – Inverse Document Frequency (TF-IDF) method and Cosine Similarity with the measuring degree concept of similarity terms in a document. The tests were carried out on a number of Indonesian text-based documents that have gone through the stage of pre-processing for data extraction purposes.

Then ranking of the document weight must be done so as to check its closeness level with the expert's document. Hence, the similarity measures were applied which showed that the document with the first rank have low degree value than others but have high cosine similarity value than others. Based on the cosine similarity principle, when the document vector has similarity closeness to 1 it is said that both documents are similar.

Maake Benard Magara et.al [2] in their paper titled "A Comparative Analysis of Text Similarity Measures and Algorithms in Research Paper Recommender Systems" aimed to establish algorithms and similarity metric combinations which can be used to optimise the search and recommendation of articles in a research paper recommender systems. The purpose of their paper was to test the performance of data mining algorithms that are going to be used to develop their research paper recommender system. They tested 3 algorithms (Random Forest, Recursive Partitioning and Boosted tree) for their accuracy and efficiency. And when they evaluated the performance of all the three algorithms, the rpart algorithm proved to be more efficient and accurate when compared to the other two counterparts by getting an average accuracy and time efficiency of 80.73 and 2.354628 seconds respectively. Farther accuracy of the prediction was conducted, and the rpart machine learning algorithm was selected along with the cosine similarity which performed best when compared with the other similarity metrics.

Veena G et.al [3] in their paper speaks about Information retrieval (IR) which is a method of obtaining necessary information from a collection of large data set. User usually searches by providing the keyword or query. Queries can be a single or multiple keywords. In information retrieval, search for a query will not show single result instead many results which match the query will be shown. The authors discussed about how fuzzy search and Levenshtein distance can be made use of in their proposed architecture. In their proposed paper initially preprocessing of query keywords is done, this involves removal of stop words from the keyword list. For each query keyword stemming is performed. To find list of similar words to each query keyword threshold distance is calculated based on length of query keyword. Similar words are found using Levenshtein distance. Inverted lists are intersected to determine documents which contain all query keywords. Proximity ranking is applied to find documents with phrases. Documents without phrases that is documents containing few query keywords but they do not form a phrase will be identified. The result will be displayed as union of documents with phrases and documents without phrases.

Vasile Rus et.al [4] in their paper "Similarity Measures based on Latent Dirichlet Allocation", discussed about semantic similarity measures at word and sentence-level based on two fully-automated approaches to deriving meaning from large corpora. The two approaches are Latent Dirichlet Allocation, a probabilistic approach, and Latent Semantic Analysis, an algebraic approach. They explore two types of measures based on Latent Dirichlet Allocation: measures based on distances between probability distribution that can be applied directly to larger texts such as sentences and a word-to-word similarity measure that is then expanded to work at sentence-level. The experiments conducted on MSRP test data obtained using the threshold for similarity that corresponds to the threshold learned from training data led to best accuracy. The threshold varied from method to method. The results obtained using the word-to-word similarity measures are labeled as Greedy and Optimal. The LSA shows results obtained when text-level LSA vectors were used, The Baseline method indicates performance when labeling all instances with the dominant label of a true paraphrase. The rest of the results are obtained when the text-to-text similarity measures based on various distribution distances are used: IR (Information Radius), Hellinger, and Manhattan. Finally, the results conclude that the LDA-offers competitive results and provides best precision and kappa score when compared with LSA.

III.METHODOLOGY

A legal search engine is a tool that allows users to search for legal documents and cases by keyword, phrase, or topic. The goal of a legal search engine is to make legal information more accessible and searchable for legal professionals, researchers, and the general public. The steps followed in developing a search engine are as follows:

1) Data Collection: Collect a large dataset of legal texts, such as court judgements, legislation, and legal briefs.
2) Pre-processing: The next step is to pre-process the text data. This includes tokenizing the text, removing stop words and punctuation, and performing stemming or lemmatization. These steps are necessary to ensure that the text is in a format that can be used for analysis.
3) Text Vectorization: Convert the text into numerical vectors, which can be used to represent the text in a numerical form.
4) Similarity calculation: Use four algorithms (Cosine Similarity, Levenshtein Distance, Latent Dirichlet Allocation (LDA), and Vector Space Model (VSM)) to calculate the similarity between the legal documents. Compare the results of each algorithm, and select the one that performs the best based on the characteristics of the problem and the nature of the data.
5) Retrieval: Use the index to retrieve the legal documents that match a user's query.
6) Search result: Return the top-ranking judgements as the search results to the user.

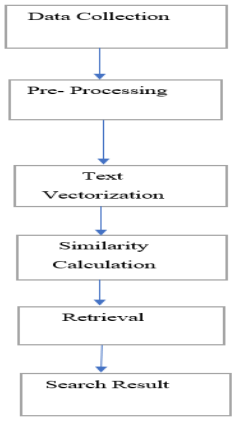


Fig.1 Methodological Steps

IV. PROPOSED ARCHITECTURE

The architecture follows the following steps:

A. *Document Extraction*

Extraction of legal documents is done from the indiankanoon.org website. The extracted judgements are in pdf format.

B. *Conversion*

The extracted judgement data is then converted into text format.

C. *Pre-processing*

This is the data mining technique which is used to transform the raw data in useful and efficient format. The following steps were followed to clean the data:

1) Tokenization: This is the process by which a large quantity of text is divided into smaller parts called tokens. Here, the sentences are split into words.
2) Stop word removal: The words like “the”, “in”, “an”, etc are removed from the raw data.
3) Lemmatization: The process of reducing the different forms of a word to one single form.

D. *Text Vectorization*

Convert the text into numerical vectors, which can be used to represent the text in a numerical form.

E. *Similarity calculation*

Using the four algorithms (Cosine Similarity, Levenshtein Distance, Latent Dirichlet Allocation (LDA), and Vector Space Model (VSM)) to calculate the similarity between the legal documents.

F. *Retrieval*

This step uses the index created to retrieve the legal documents that match the user’s query.

G. *Search result*

The flask framework is used to create a web application that allows users to search for legal documents.

The proposed architecture is as follows:

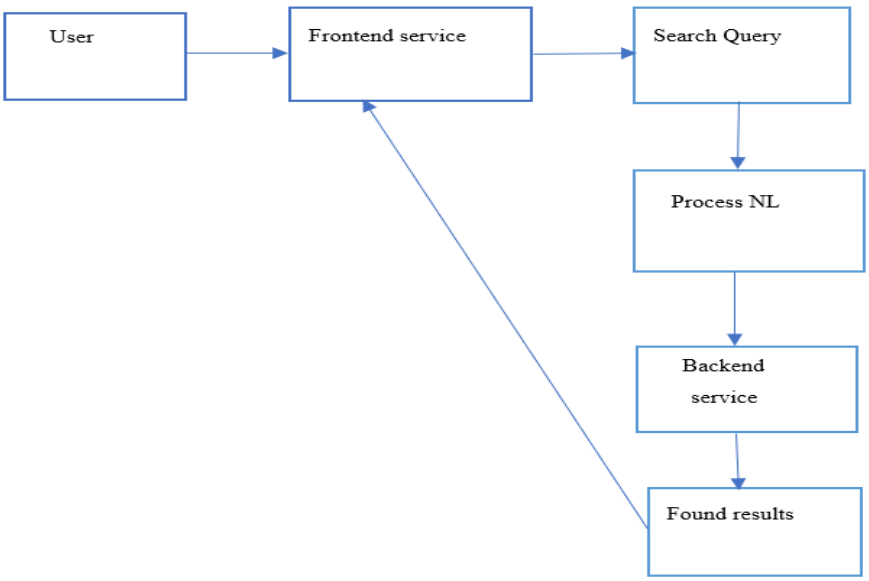


Fig. 2 System Architecture

V. RESULT

The following search engine was developed wherein the user can obtain the information by using the search bar. The search bar directs to result page which contain documents related to the entered search query as shown in below figures.

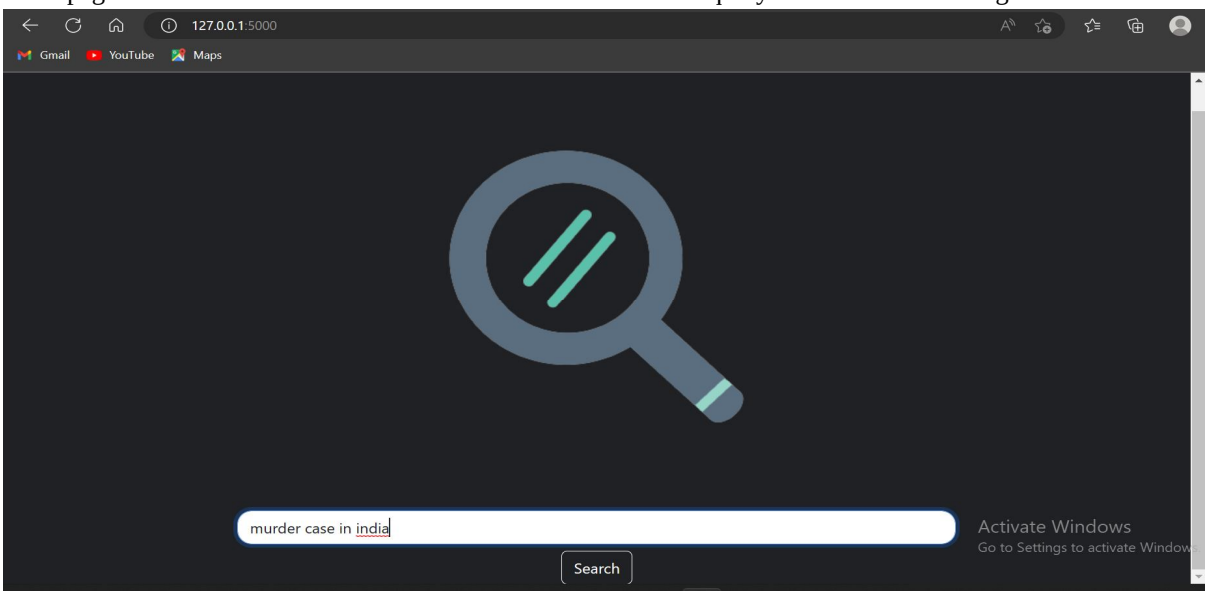


Fig. 3 Searching query using developed application

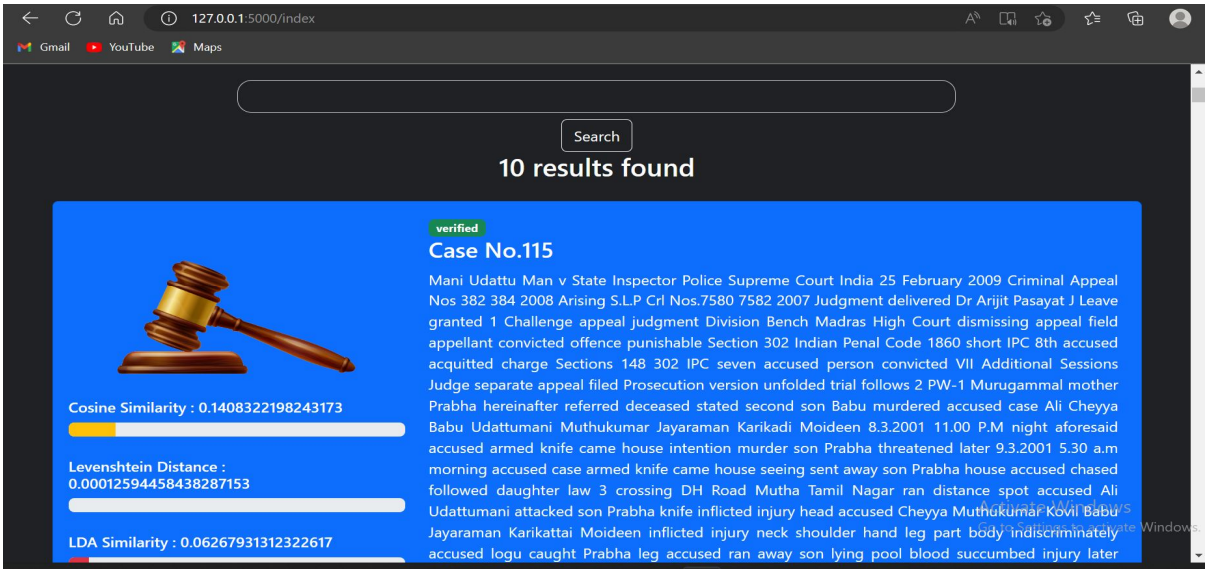


Fig. 4 Showing result for entered query

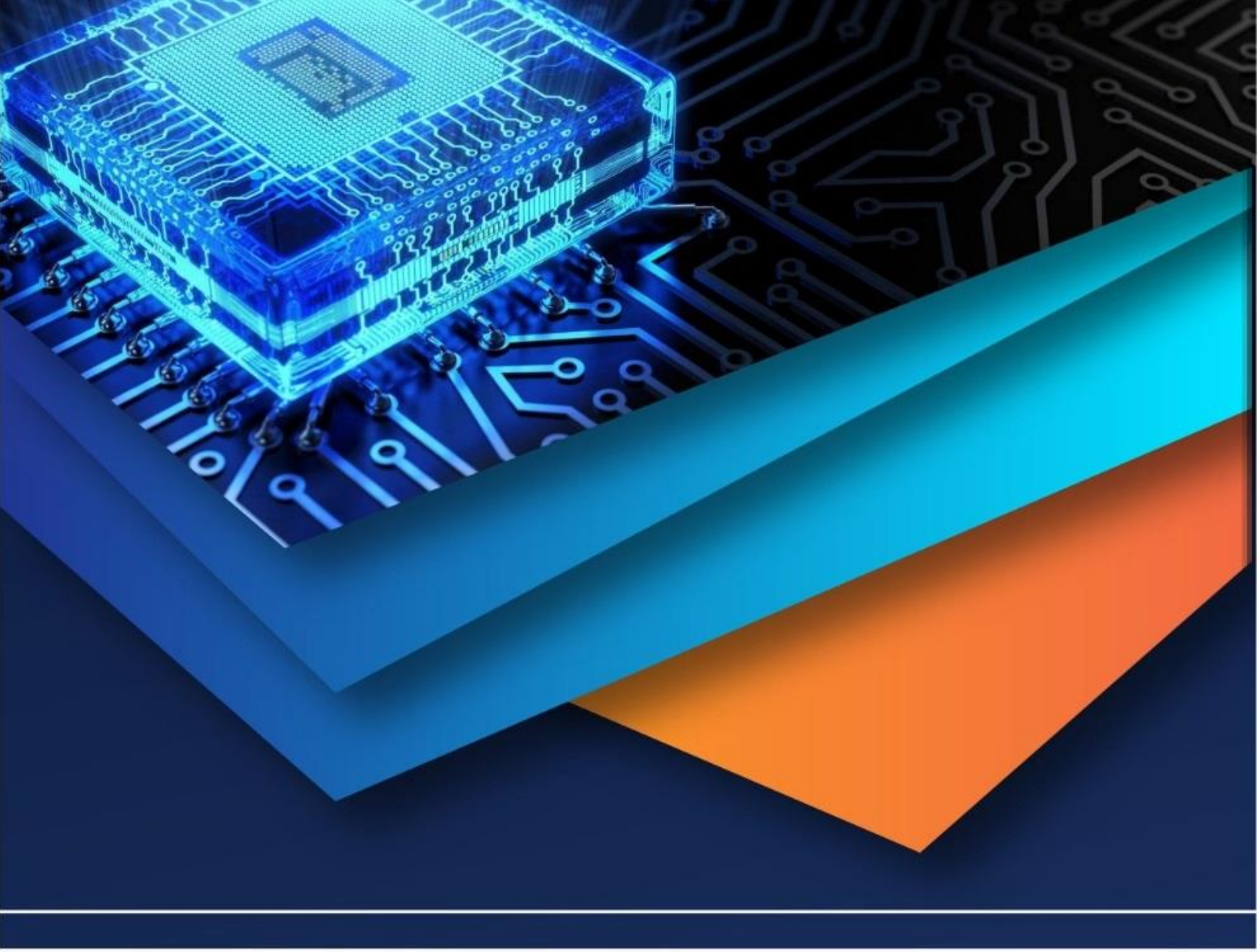
As we can see in the above figures, when the user enters a search query the resulting search offers a list of documents that are similar to that query. The results of the search include the legal judgements which contains the legal information pertaining to a particular case. These judgements can be used for thoroughly studying a particular case and for acquiring more legal knowledge.

VI. CONCLUSION

Here, a search engine is being developed. The developed search engine allows its users to access legal judgements which are specific to the legal framework of India. It gives access to a wide variety of legal judgements and is perfectly suited for usage by legal professionals, scholars, students and general citizens who are in need of legal information. This will help the legal professionals and users to keep themselves up-to-date with the legal information. Also, helping the legal professionals to provide a quick response to their client's queries.

REFERENCES

[1] Alfirna Rizqi Lahitani; Adhistya Erna Permanasari; Noor Akhmad Setiawan "Cosine similarity to determine similarity measure: Study case in online essay assessment" DOI: 10.1109/CITSM.20
[2] Maake Benard Magara; Sunday O. Ojo; Tranos Zuva "A comparative analysis of text similarity measures and algorithms in research paper recommender system" DOI: 10.1109/ICTAS.2018.8368766
[3] Veena G, Jalaja G " Levenshtein Distance based Information Retrieval" International Journal of Scientific & Engineering Research, Volume 6, Issue 5, May-2015 113 ISSN 2229-5518
[4] Vasile Rus, Nobal Niraula, Rajendra Banjade "Similarity Measures based on Latent Dirichlet Allocation", The University of Memphis USA 2013
[5] Levenshtein Distance, Sequence Comparison and Biological Database Search, Bonnie Berger; Michael S. Waterman; Yun William Yu DOI: 10.1109/TIT.2020.2996543
[6] Automating Legal Research through Data Mining by Mohammed Firdhous (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 1, No. 6, December 201
[7] Shahmin Sharafat, Zara Nasar and Syed Waqar Jaffry, " Data mining for smart legal systems" Computers & Electrical Engineering Volume 78, September 2019, Pages 328-342 https://doi.org/10.1016/j.compeleceng.2019.07.017
[8] V. Vaissnave and P. Deepalakshmi, "An Artificial Intelligence based Analysis in Legal domain" International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-2S2, December 2019 Retrieval Number: B11131292S219/2019 doi: 10.35940/ijitee.B1113.1292S219
[9] Tiedan Zhu "The Similarity Measure Based on LDA for Automatic Summarization", doi.org/10.1016/j.proeng.2012.01.419
[10] A review on text mining by Yu Zhang; Mengdong Chen; Lianzhong Liu September 2015 DOI:10.1109/ICSESS.2015.7339149 Conference: 2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS)



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY