iJRASET
International Journal For Research in Applied Science and Engineering Technology



# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 10 **Issue:** VIII **Month of publication:** August 2022

**DOI:** https://doi.org/10.22214/ijraset.2022.46368

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 10 Issue VIII Aug 2022- Available at www.ijraset.com*

# Legal Text Mining Using Legal Judgements

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:*** ***Rapid advancements in digital data acquisition techniques have led to massive amounts of data. Above 80 percent of today's data is composed of unstructured or semi-structured data. The discovery of appropriate patterns and trends to analyse the text documents from massive volume of data is a big issue. Thus, text mining is believed to have high commercial value. Text mining is the process of extracting useful patterns from text data which plays an effective role in decision making task. Here, we will use the legal judgements as the input document then we will apply cosine similarity method to see which judgements are similar and finally we will develop a web application in which the user can search for a particular judgement.***
***Keywords: Text mining, natural language processing, legal system, legal data, unstructured data***

## I. INTRODUCTION

Text mining may be a process of extracting interesting and nontrivial patterns from huge amount of text documents. There exist different techniques and tools to mine the text and find out valuable information for future prediction and decision-making process. The choice of right and appropriate text mining technique helps to enhance the speed and decreases the time and effort required to extract valuable information. "Text mining" covers a variety of techniques that allow software to extract information from text documents. It's not a new technology, but it's recently received spotlight attention due to the emergence of Big Data. Data is usually susceptible to missing values, noisy data, incomplete data, inconsistent data and outlier data. So, it's important for these data to be processed before being mined. Pre-processing data is an important step to enhance data efficiency. Data pre-processing is one among the most data mining steps which deals with data preparation and transformation of the dataset and seeks at the same time to make knowledge discovery more efficient. Pre-processing includes several techniques like cleaning, integration, transformation and reduction. The applications of text mining are very diverse and span multiple disciplines, starting from biomedicine to legal, business intelligence and security. From a legal perspective, text mining touches upon several areas of law, including jurisprudence, copyright law and database law. This study shows an in-depth description of data pre-processing techniques which are used for data mining.

## II. RELATED WORK

Arafat Hossain et al. [1] aimed to examine the pattern of words that appeared on the front page of a well-known daily English newspaper in Bangladesh, The Daily Star, in 2018 and 2019. The elucidation of that era's possible social and political context was also attempted using word patterns. The study employs three widely used and contemporary text mining techniques: word clouds, sentiment analysis, and cluster analysis. The word cloud reveals that election, kill, cricket, and Rohingya-related terms appeared quite 60 times in 2018, whereas BNP, poll, kill, AL, and Khaleda appeared quite 80 times in 2019. These indicated the country's passion for cricket, political turmoil, and Rohingya-related issues. Furthermore, sentiment analysis reveals that words of fear and negative emotions appeared quite 600 times, whereas anger, anticipation, sadness, trust, and positive-type emotions came up quite 400 times in both years. Finally, the clustering method demonstrates that election, politics, deaths, digital security act, Rohingya, and cricket-related words exhibit similarity and belong to an identical group in 2019, whereas rape, deaths, road, and fire-related words clustered in 2018 alongside a similar-appearing group. Generally, this analysis demonstrates how vividly the text mining approach depicts Bangladesh's social, political, and law-and-order situation, particularly during election season and therefore the country's cricket craze, and also validates the importance of the text mining approach to understanding the overall view of a country during a particular time in an efficient manner. Shahmin Sharafat et al. [2] in their paper developed smart legal systems which carry immense potential to supply legal community and public with valuable insights using legal data. These systems can consequently help in analyzing and mitigating various social issues. In Pakistan, since last few years, courts are reporting judgments online for public consumption. This public data, once processed, are often utilized for betterment of society and policy making in Pakistan. This study takes the primary step to realize smart legal system by extracting various entities such as dates, case numbers, reference cases, person names, etc. from legal judgments. To automatically extract these entities, the first requirement is to construct dataset using legal judgments.

©IJRASET: All Rights are Reserved 

Hence, firstly annotation guidelines are prepared followed by preparation of annotated dataset for extraction of varied legal entities. Experiments conducted using sort of datasets, multiple algorithms and annotation schemes, resulted into maximum F1-score of 91.51% using Conditional Random Fields. Said A. Salloum et al. [3] in their study, collected and textually analyzed various text mining techniques, 300 refereed journal articles in the field of mobile learning from six scientific databases, namely: Springer, Wiley, Science Direct, SAGE, IEEE, and Cambridge. The selection of the collected articles was based on the criteria that all these articles should incorporate mobile learning as the main component in the higher educational context. Experimental results indicated that Springer database represents the most source for research articles in the field of mobile education for the medical domain. Moreover, results where the similarity among topics couldn't be detected were due to either their interrelations or ambiguity in their meaning. Furthermore, findings showed that there was a booming increase within the number of published articles during the years 2015 through 2016. Additionally, other implications and future perspectives are presented within the study.

## III.PROPOSED SYSTEM

The following is the proposed system architecture:

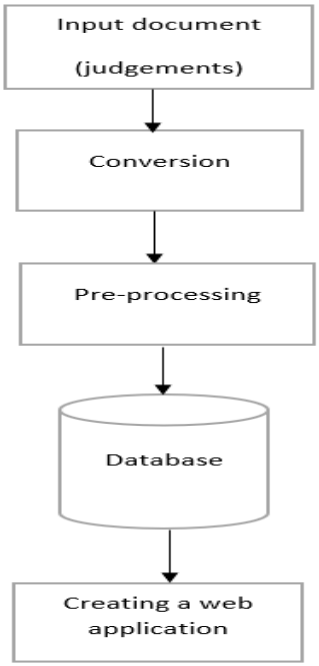



Fig. 1 System architecture

### *A. Input Document*

The input document consists of legal judgements taken from the indiankanoon.org. The judgements are in pdf format.

### *B. Conversion*

The downloaded judgements in the pdf form are converted into text (txt) file.

### *C. Pre-processing*

This is the data mining technique which is used to transform the raw data in useful and efficient format. The following two 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) *Removing stopwords:* The stopwords are removed from the raw data.

### *D. Data storing*

In this step, the cleaned data which is obtained after pre-processing is imported into the database and stored.

### *E. Creating a Web Application*

Here, we have developed a web application by using the python flask web framework. This application is designed to provide a search engine to the user. The user has to provide the input into the search bar and this entry will be passed on to the database and the desired result or the judgements will be retrieved.

## IV. RESULTS

The developed web application looks as in the following figure. Here the user has to enter the input for which a particular judgement will be provided.

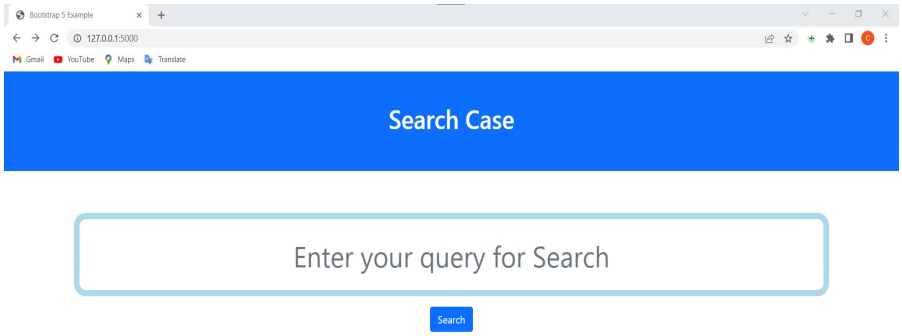



Fig. 2. Developed web-based application

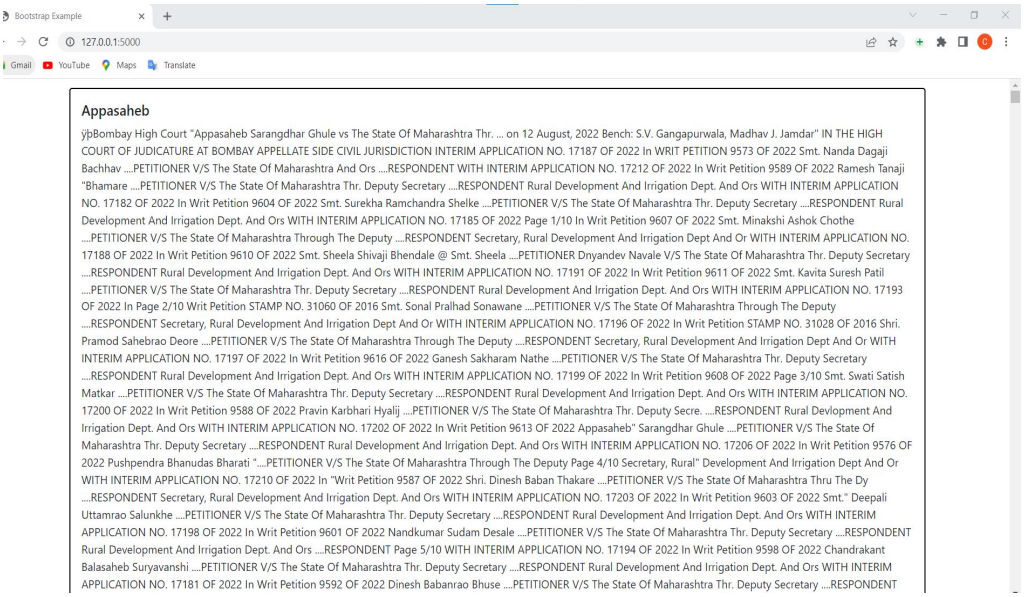


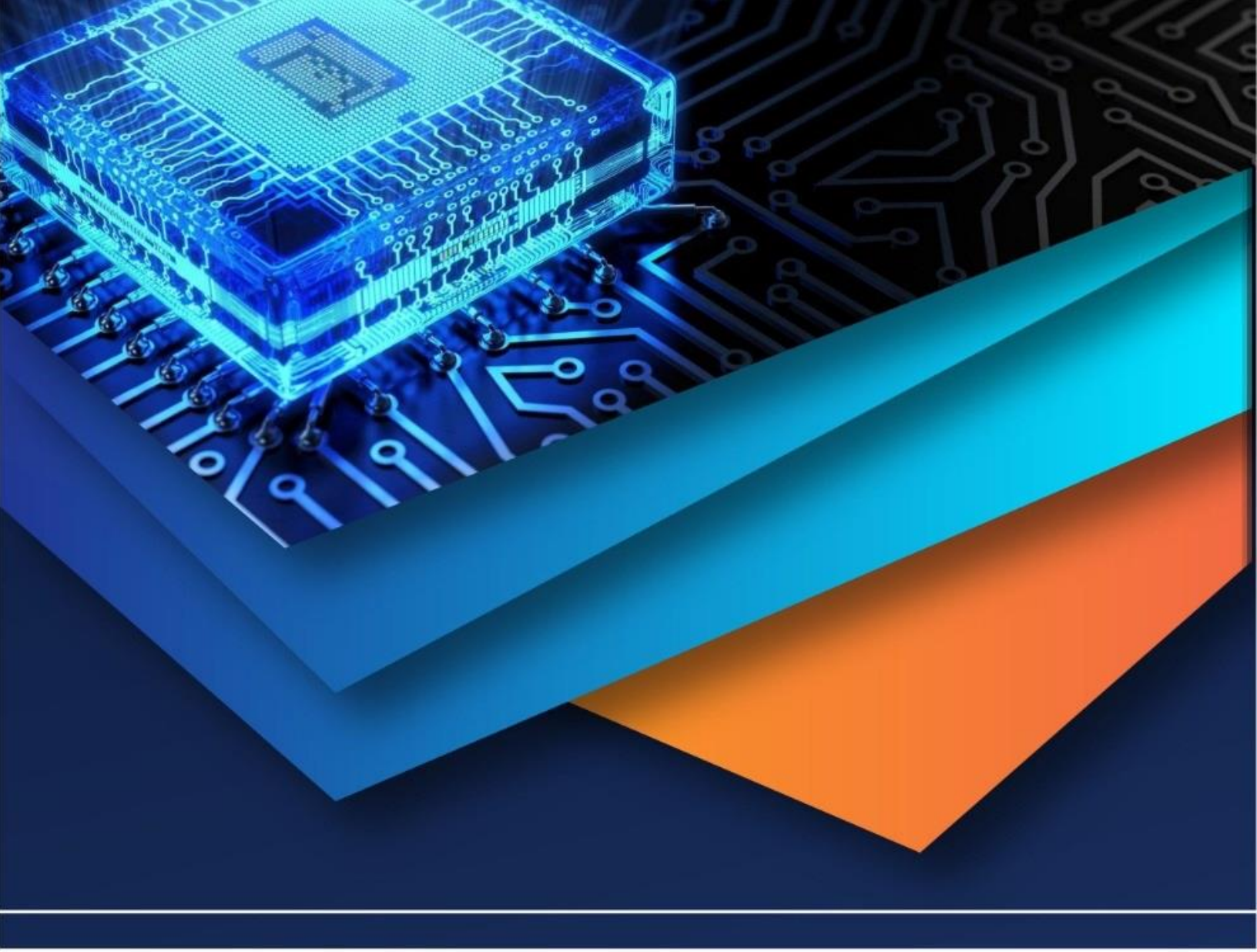

Fig. 3 Retrieving judgement

## V. CONCLUSIONS

Here, a web-based application is proposed and implemented. This application allows the user to search for particular judgements. The user will be able to find similar judgements based on his/her input. This is made possible by using the countvectorizer and cosine similarity method.

## REFERENCES

[1] Text Mining and Sentiment Analysis of Newspaper Headlines by Arafat Hossain, Md. Karimuzzaman, Md. Moyazzem Hossain and Azizur Rahman October 2021 DOI:10.3390/info12100414
[2] Data mining for smart legal systems by Shahmin Sharafat, ZaraNasar, Syed WaqarJaffry https://doi.org/10.1016/j.compeleceng.2019.07.017
[3] Using Text Mining Techniques for Extracting Information from Research Articles by Said A. Salloum, Mostafa Al-Emran, Azza Abdel Monemand Khaled Shaalan doi.org/10.1007/978-3-319-67056-0_18 (2017)
[4] Text Mining: Techniques, Applications and Issues by Ramzan Talib, Muhammad Kashif Hanif, Shaeela Ayesha, and Fakeeha Fatima November International Journal of Advanced Computer Science and Applications 7(11) DOI:10.14569/IJACSA.2016.071153(2016)
[5] Text Mining Scientific Papers: A Survey on FCA-Based Information Retrieval Research by Jonas Poelmans, Dmitry I. Ignatov, Stijn Viaene, Guido Dedene & Sergei O. Kuznetsov January 2011 DOI:10.1007/978-3-642-31488-9_22
[6] 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)
[7] Automating Legal Research through Data Mining by Mohammed Firdaus (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 1, No. 6, December 2010
[8] Ms. Anjali Ganesh Jivani, "A Comparative Study of Stemming Algorithms," ISSN:2229-6093. IJCTA NOV-DEC 2011.
[9] Kaiz Merchant and Yash Pande, "NLP Based Latent Semantic Analysis for Legal Text Summarization" International Conference on Advances in Computing, Communications and Informatics (ICACCI)-Bangalore, India (2018.9.19-2018.9.22) IEEE 2018
[10] Suad A. Alsadi and Wesam S. Bhaya, "Review of Data Preprocessing Techniques in Data Mining", Journal of Engineering and applied Sciences, 12:4102-4107(2017) doi:10.36478/jeasci.2017.4102.4107



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)