CLAUSE IDENTIFICATION MODEL THROUGH VECTOR SPACE MODEL

Arsh Kaul[1], Nikita Adsule[1], Rutwij Patil[1] and Vaishnavi Mane[1]

[1]*Department of Information Technology, Pune Institute Of Computer Technology, Savitribai Phule Pune University, Pune, India*

Abstract: As each and every company needs legal support and representation[4], the amount of documentation is rapidly increasing which is regularly required to be reviewed and updated so as to make sure companies are not in breach of any policies or terms . Reviewing this vast and unstructured data becomes a laborious task, hence our proposed system is designed to make this process efficient and less time-consuming. In our proposed model all documents in the form of plain text or images can be given as input which are then stored and classified through pre-processing. Some guidelines for good and bad clauses are also provided following which the system will highlight violations, breaches, terms, etc as output. This system can be expanded in different domains as well wherein evaluation of data is essential.

Keywords: Vector-Space Model, Information Retrieval, Natural Language Processing, Machine Learning, tf-idf model, Text Analytics

1. INTRODUCTION

Naol Bakala [1] had mentioned it in his paper the standard definition of Vector Space Model. He mentioned that we consider text documents which contain words as vectors. In the tf * idf weighting approach, these vectors describe the relevance of a phrase, as well as its absence or presence in the document. Every page is divided into a word frequency table, which is an inverted index data structure. Tables are termed vectors when represented in vector space and can be stored as arrays of terms. A vocabulary is created using all of the words found in all of the system's papers, as well as dictionary terms, ensuring that no phrases are repeated.

Vector Space Model has 3 main components :

i) Document Indexing

ii) Term Weighting

iii) The Cosine Similarity

Because court papers include enormous volumes of unreliable data, it is vital to extract crucial information from them in order to expedite processes. The goal of this project is to make it easier for individuals to find factual information in various legal papers in a short amount of time rather than having to read them all. People should put this into practise since it can help them boost their productivity.

Users can enter data as plain text or as documents (word, pdf, etc.) and the system will recognize good and problematic clauses. OCR will be used to translate picture documents into text that can then be utilized as a system input.

2. PRIOR WORK

- In the paper titled Vector Space Model of Information Retrieval - A Reevaluation by Wong, S. & Raghavan, Vijay[2], they point out that the methodologies utilized in today's vector-based systems are incompatible with the vector space model's assumptions. Clearly, these concerns lead to suggestions about how things may have been done differently. More significantly, it is hoped that this research will help to clarify the concerns and problems associated with employing the vector space paradigm in information retrieval.

- In the paper titled Design and Analysis Of a General Vector Space Model for Data Classification in Internet of Things by Sang, J., Pang, S., Zha, Y[3]. , We study about a novel text categorization technique based on a vector space model that has been suggested. By incorporating synonym substitution to existing text classification algorithms, this system enhances feature identification and weighting methods. The experimental findings reveal that the suggested classification method greatly improves classification precision. We can create the necessary training set by arranging and storing texts according to various categories, which is what text classification is all about. Then, as per the various

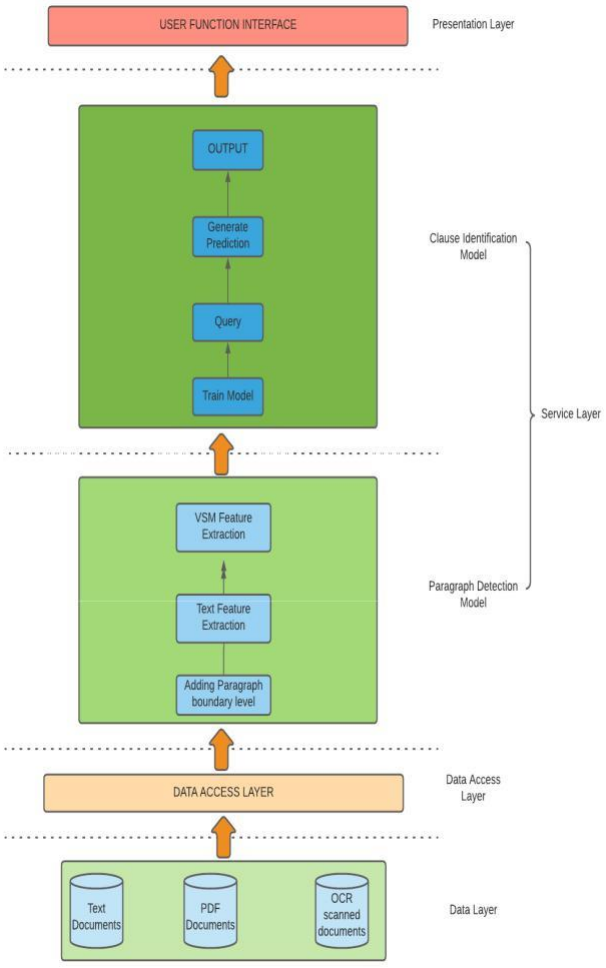
categories, the associated classification rules are specified, and a text classifier is built. We plan to include these notions into our clause identification model since they will aid us in indexing and data reduction.

3. PROPOSED SYSTEM ARCHITECTURE

3.1 Overview

Users such as legal practitioners, agents, and reviewers who need to ensure that a document is not in violation of specific terms and is suited to perform its function effectively can submit documents into the system as plain text or images. The data will then be reduced to informative words by the system, which is known as Document indexing. The users may now post their query which will lead to the building of our pre trained clause identification model. The model will use previous feedback and some external guidelines provided to figure out the good or bad clauses requested by the users and show those as the required output. Every output will send the feedback to the user and store it for the clause identification model as well. Thus we can make sure that our model is machine learned and efficient.

3.2 System Architecture & Workflow



The main workflow of the architectural diagram has been divided into four main sections, viz. Data Layer. Data Access Layer, Service Layer and Presentation Layer.

3.2.1 DATA LAYER

This layer consists of all the data which we are going to feed into our main model. As our main concern is with legal documents so these documents are mostly in text formats or pdf formats. Documents which are in form of images can also be fed into the system. They are further processed by OCR and converted to text documents. We expect the input data size to be huge considering the vastness of legal documents across various domains.

3.2.2 DATA ACCESS LAYER

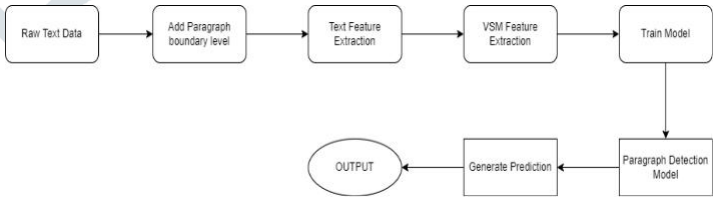
The data access layer acts as a mediator or supplier between the model and the storage device. Working data access layers are present in every software that runs on the computer and needs to access data on the hard drive. Instead of the program's main body communicating directly with the persistent storage location, it delegated the task to the data access layer, which subsequently carried out the work on the program's behalf. Its sole purpose is to shuffle data back and forth, allowing the remainder of the programme to focus on its other tasks.

3.2.3 SERVICE LAYER

Here we have bifurcated the service layer into 2 different parts based on their functionalities, viz. Paragraph Detection Model and Clause Identification Model.

3.2.3.1 Paragraph Detection Model

The most typical approach [4] for recognising a paragraph is to split the material using consecutive new lines (\n\n). Then there were certain guidelines to follow, such as what is the first word of the phrase; if the sentence begins with a word like 'AND,' 'BUT,' etc., it is not the beginning of the paragraph.



Paragraph detection using rules works for small datasets but is not practical for large datasets. Every time the dataset grows, new sophisticated rules must be added. As a result, this strategy does not work with huge datasets. For recognising sentence boundaries, OpenNLP includes a sentence detector. It may be set up to identify the end of a paragraph. However, the meaning of each sentence is not the focus of this paragraph. Because the paragraph definitions

for each domain differ, it is impossible to create our own paragraph border.

3.2.3.2 Clause Identification Model

In this part of the service layer our main model's functionality comes into action. We have defined 4 functionalities for this model, viz. Training, Query Processing, Generating Prediction and Output.

Training of this model is done by constantly feeding it numerous datasets of legal domain across different categories of agreements, letters and contracts.

Query Processing is done by the model after it receives a query which is fired by the user. After the query processing is done the model undergoes the process of generating prediction followed by giving appropriate output to the user.

3.2.4 PRESENTATION LAYER

This layer consists of User Function Interface. Here the users get their desired output from the clause identification model. In this case the users would come to know the good and the bad clauses which are there in the documents they have submitted to the model.

CONCLUSION

Thus, by utilizing a vector space model and incorporating various parts of fields such as machine learning, tf-idf model, text classification, and so on, our model intends to replace the external support that most firms use today to evaluate legal papers and the clauses included therein. As a result, our model will be able to save time and effort spent on this time-consuming process.

FUTURE SCOPE

Our system can be expanded to online websites because certain terms and conditions can be misleading or ethically inappropriate, resulting in negative consequences for the user.

When our model is implemented, it will identify these hazardous scenarios and alert the user before they agree to the terms and conditions.

ACKNOWLEDGMENTS

We would like to express our gratitude to DR Anant Bagade for his guidance.

We would also like to thank Prof Shyam Deshmukh and Prof Anuradha Yenkikar for their insightful inputs.

We thank Veritas for this golden opportunity and Mr Ishwar Patil for his immense support and ideas.

REFERENCES

[1] Naol Bakala (2019). Information Retrieval System By Using Vector Space Model. INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 8, ISSUE 10, OCTOBER 2019

[2] Wong, S. & Raghavan, Vijay. (1984). Vector Space Model of Information Retrieval - A Reevaluation.. Communications of The ACM - CACM. 167-185.

[3] Sang, J., Pang, S., Zha, Y. *et al.* Design and analysis of a general vector space model for data classification in Internet of Things. *J Wireless Com Network* **2019,** 263 (2019). https://doi.org/10.1186/s13638-019-1581-3

[4] Shah, Parth & Joshi, Sandeep & Pandey, Amaresh. (2018). Legal Clause Extraction From Contract Using Machine Learning with Heuristics Improvement. 1-3. 10.1109/CCAA.2018.8777602.