International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 06 | June - 2021 ISSN: 2582-3930

Legal Text Summrization Using Machine Learning

Harshada Laxman Devadhe [1], Bhakti Uday Dunakhe [2], Simran Uddhav Sarkale [3]

Harshada Laxman Devadhe, Department of Information Technology, VPKBIET, Baramati
Bhakti Uday Dunakhe, Department of Information Technology, VPKBIET, Baramati
Simran Uddhav Sarkale, Department of Information Technology, VPKBIET, Baramati

---

Abstract - It is difficult task for lawyers and Judges to read and anaylyze the judgment within shorter time span. Judgement is a huge text document. We are designing an automated text summarizer to focus on key part of brief text. It not only reduces the size of brief documents but also highlights the effective and essential sentences required for judgement analysis.. Index Terms—component, formatting, style, styling, insert

1.INTRODUCTION

In present days, where time is very important it is difficult task for lawyers to read all judgment documents and make conclusions on it. So the need of document summarization is increasing day by day due to digitalization. It is easy to retrieve a document that is related to the case. Due to the Document summarization it becomes easy to fetch relevant information, reducing the complexity and unnecessary lenghty parts of the documents so, text summrization is imporatnt. Text summarization is one of the method of identifying the important meaningful information from a document or set related document and compressing them into a shorter version preserving its overall meaning. To extract important information from the long judgement, it takes single document as input and accordingly output is generated in the form of summary. Text Summarization: 1)Abstractive text summarization 2) Extractive text summarization text summrization technique understand the input documents and make its output as few undestandable wordsor lines.

2. Related Work

This section tells about the methodologies which have been used for the summarization purpose. We are going to focus on extractive text summarization which deals with choosing necessary sentences from the text.Where,abstractive summarization deals with understanding the main concept and meaning of the document or text given. It finds the new concept from the document by using various methods.

1] To explore different types of summaries,substantial amount of research has been directed .Mostprominent methods of summarization provided by Nenkova and McKeown (2012).

2] An algorithm based on LSA (Latent Semantic Analysis) was proposed by Wang and Ma (2013).

3] Various graph based approaches presented by Hirao et al. and Hamidand Tarau.

4] Research at IBM on Extractive summarization has been done by Baxendale in 1958. By using the position of text he extracted important sentences.The author has tested 200 paragraphs towards his goal to find that in 85 percent of the sentences which author has taken first topic which is main topic sentence and the last sentence came 7 percent.From these two sentences the most accurate sentence would be selected.

5] J.N.Madhuri in 2018,has done her work on Automatic Text Summarization Technique.In which she is generating a summary based on frequency of Keywords.Highest Frequently occurring words are used to generate summary.

6] Christian M. Meyer,Benjamin H attasch,Carsten Binnig worked on Sherlock system where interactive summary is generated by taking user's feedback during multiple iterations.

7]Research on Multi-document summarization Based on LDA topic model. The author Jinqiang Bian; Zengru Jiang Qian Chen;2014 Sixth International Conference on Intelligent Human-Machine Systems and Cybernetics. Compared to the VSM(Vector Space Model) and graph stand models, the LDA (Latent Dirichlet Allocation)model can find hidden topics in the corpus and hidden topics have the advantage of using sentence structure to build a good summary. On paper, based on the LDA Model, a new way of putting a sentence is suggested. How combines the topic distribution of each sentence and the topic-corporate significance together for the next calculation sentence opportunities, and then, depending on the background, select sentences to form a summary.

8]LDA Analyzer: A Tool for Exploring Topic Models by Chunyao Zou; Daqing Hou 2014 IEEE International Conference on Software Maintenance and EvolutionOnline technology forums are an important source of mining useful information for engineering software. LDA (Latent Dirichlet Allocation) is an unregulated machine learning method that can be used for basic extraction articles from these great forums.

However, the main effect of reading the LDA forum is usually a large metric contains millions of numbers, which it is impossible for investigators to directly monitorthe distribution of numbers and psychologically examine the relationship between published articles and the larger collection of informations texts. In this paper, we introduce the LDAAnalyzer, an LDA diagnostic tool that makes the capabilities of hidden title texts higher.LDA Analyzer contains (1) LDA modeling (2) LDA release analysis and (3) new corpus training. With the help of LDA Analyzer, our semantic subject modeling testing based on major technical forums is possible.

9]A Survey of the State-of-the-Art Models in Neural Abstractive Text Summarization Ayesha Ayub Syed; Ford Lumban Gaol; Tokuro Matsuo IEEE Access 19 January 2021 Dealing with vast amounts of textual data. Automatic summarization systems are capable of addressing this issue. A review of various neural networks based abstractive summarization models have been presented. The proposed conceptual framework includes five key elements identified as encoder decoder architecture, mechanisms, training strategies and optimization algorithms, dataset, and evaluation metric. A description of these elements is also included in this article. The purpose of this research is to provide an overall understanding and familiarity with the elements of recent neural networks based abstractive text summarization models. Analysis has been performed qualitatively with the help of a concept matrix indicating common trends in the design of recent neural abstractive summarization systems. Models employing a transformer-based encoder decoder architecture are found to be the new state-of-the-art. Based on the knowledge acquired from the survey, this article suggests the use of pre-trained language models in complement with neural network architecture for abstractive summarization task.

3.Approarch

In this proposed approach, we are using extractive method to get summary of given input.

1) At first,test which is our input is tokenized so that tokens are generated.

2) After tokenization stop words are removed.The words which are remained are considered as a key word.

3) Part of tag is attached to each key word,as key words are taken as input.

4)Apply Topic Modelling Algorithm to detect the topics occur in document.

5) output generated by topic modelling algoritham is given input to summary generation

6) Now as per our wish how many topics we want to generate give no as input

7) by using cosine similarity matrix sentence ranking is done

8)Then summary get generated.

4. ALGORITHM

• Text Document is taken as input to the system.
• The input file is tokenized and tokens are generated.
Fig. 1. Architecture
• Remove stop words from tokens and assume remaining words are keywords.
• Give tag to the keywords.
• Above process is involved in pre-processing.
• Latent Dirichlet Allocation (LDA) used for Topic Modelling.
• Topics are generated as output by LDA.
• Calculate the sentence scoring by cosine distance of that keywords.
• Extract sentences which containing a high scoring keywords.
• Summary

5. SYSTEM IMPLEMENTATION

• Important Functions
• Document Uploading This function reads the documents uploaded by client and save it on server disk.
• PDF2Text Converting If the document is in PDF format we need to convert it in text format.
• Cleaning Convert a document into a list of tokens. This lowercases, tokenizes. The output are final tokens = Unicode strings, that won't be processed any further.
• Lemmatization Lemmatisation in linguistics is the process of grouping together the inflected forms of a word so they can be analysed as a single item, identified by the word's lemma, or dictionary form.
• BagofWords Model Training Bag of Words model is used to pre-process the text by converting it into a bag of words, which keeps a count of the total occurrences of most frequently used words.
• LDA Model Training LDA model estimation from a training corpus and inference of topic distribution on new, unseen documents.
• Sentence Extraction This function is used to calculate score of topic v/s tokens
• Document Summarization All extracted sentence are ordered using cosine distance.

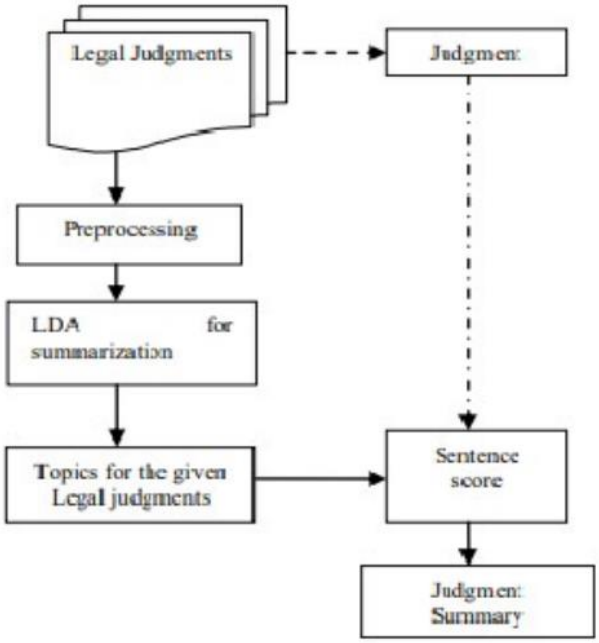

Fig -1: Figure

ACKNOWLEDGEMENT

This is to acknowledge and thank all the individuals for their contribution specially our respected and learned guide Dr. S. A. Takale for her help and guidance to this paper.Without their coordination and help this task could not be completed alone.We avail this opportunity to express our gratitude.

REFERENCES

[1] Nenkova, A.(2011).“Automatic summarization, Foundations and Trends in Information Retrieval”,5(2),103- 233 Gupta,V and Lehal,G.s (2010). “A survey of text summarization extractive techniques.” Journal of Emerging Technologies in Web Intelligence,2(3),258-268 [3] Goldstein.J,carbonell.J,Kantrowitzt.M(1998).“Multiple document summarization by sentence Extraction”40-48 [4] Weigo Fan, Linda Wallace, Stephanie Rich and Zhongju Zang,: “Tapping the power of text mining”, Journal of ACM,Blacksburg 2005. Baxendale,P.(1958). “Machine-made index for technical literature” –an experiment. IBM Journal of Research developement354-361 Vishal Gupta,G.s. Lehal. “A survey of text mining techniques and applications”, Journal of Emerging Technologies in Web intelligence,VOL 1,NO 1,6076,August 2009 G.Erkan, Dragomir R.Radev. “LexRank:graph based centrality as salience in Text summarization”,Journal of Artificial intelligence Research,Re-search,vol.22,pp.457-479 2004 Luhn, H (1958). “The automatic creation of literature abstracts”. IBM Journal of Research Development, 2(2):159-1