Artificial Intelligence and Law (2024) 32:165–200
https://doi.org/10.1007/s10506-023-09345-y

ORIGINAL RESEARCH

# A sentence is known by the company it keeps: Improving Legal Document Summarization Using Deep Clustering

**Deepali Jain**[1] · **Malaya Dutta Borah**[1] · **Anupam Biswas**[1]

Accepted: 3 January 2023 / Published online: 1 February 2023


## Abstract

The appropriate understanding and fast processing of lengthy legal documents are computationally challenging problems. Designing efficient automatic summarization techniques can potentially be the key to deal with such issues. Extractive summarization is one of the most popular approaches for forming summaries out of such lengthy documents, via the process of summary-relevant sentence selection. An efficient application of this approach involves appropriate scoring of sentences, which helps in the identification of more informative and essential sentences from the document. In this work, a novel sentence scoring approach DCESumm is proposed which consists of supervised sentence-level summary relevance prediction, as well as unsupervised clustering-based document-level score enhancement. Experimental results on two legal document summarization datasets, BillSum and Forum of Information Retrieval Evaluation (FIRE), reveal that the proposed approach can achieve significant improvements over the current state-of-the-art approaches. More specifically it achieves ROUGE metric F1-score improvements of (1−6)% and (6−12)% for the BillSum and FIRE test sets respectively. Such impressive summarization results suggest the usefulness of the proposed approach in finding the gist of a lengthy legal document, thereby providing crucial assistance to legal practitioners.

**Keywords** Legal document summarization · Extractive summarization · Legal BERT · Deep clustering

✉ Deepali Jain
deepali_rs@cse.nits.ac.in

Malaya Dutta Borah
malayaduttaborah@cse.nits.ac.in

Anupam Biswas
anupam@cse.nits.ac.in

[1] Department of Computer Science and Engineering, National Institute of Technology Silchar, Silchar, Assam, India

## 1 Introduction

Nowadays legal documents such as legal bills, case judgments, legal reports, etc., are easily available due to the recent growth of the world wide web. But such documents are quite lengthy and in most of the cases, these documents are unstructured also (Bhattacharya et al. 2019a; Kanapala et al. 2019). In such scenarios, it becomes very difficult even for legal practitioners to understand them completely and focus on the important aspect of such documents. For common citizens, legal documents are far from comprehensible and they are not able to draw any conclusions which can benefit them. Considering these situations, there should be some automatic techniques which summarize these lengthy documents and these summaries must contain all the important information. There have been several summarization approaches proposed in the literature for various kinds of legal documents (Jain et al. 2021d). In this work, an effective sentence scoring approach is proposed via the application of Legal BERT and deep clustering techniques, which significantly improves the scores of summary relevant sentences. This clustering enhanced sentence scoring approach enables high quality extractive summarization of legal documents.

An automatic summary of any document can be generated in two ways: (1) Extractive (2) Abstractive. In case of the former, summary formation is done via the selection of salient sentences from the document itself. Whereas in case of the later, natural language generation techniques are employed in order to produce novel sentences. The key idea involved in the process of extractive summarization, which is the primary focus of this work, is to perform appropriate sentence scoring and identification of summary relevant sentences by considering high scoring ones. Several sentence scoring approaches have been proposed in the literature (Edmundson 1969; Luhn 1958; Mihalcea and Tarau 2004; Nenkova and Vanderwende 2005), among which the study of sentence-level information content is one of the most common approaches. Before the recent developments of Deep Learning (DL) based techniques, traditional Natural Language Processing (NLP) based handcrafted sentence features like named entities, noun-adjective phrase counts, etc. were considered for the purpose of vectorized representation of sentences. Such representations were then used as part of a summary relevance learning problem. However, with the advent of effective DL techniques in the recent years, more appropriate abstract sentence representations with the help of pretrained embedding approaches, are becoming more common. Recently, BERT based sentence embedding techniques have been utilized by several researchers for various NLP tasks which has been able to provide considerable improvements (Eidelman 2019; Cohan et al. 2019). It is important to note here however, that for a long document summarization task, such sentence level representation alone can't capture the global informativeness of the individual sentences (Xiao and Carenini 2019). This is especially crucial in case of lengthy documents like those in the legal domain. Therefore, a novel summarization approach is proposed in this work to address this problem, that enhances domain-specific Legal BERT based sentence scores via deep embedded sentence clustering.

Clustering based improvement of extractive summarization has been explored widely in the literature (Akter et al. 2017; Alguliyev et al. 2019; Mallick et al.

2021; Wang et al. 2011), for the task of general text summarization. But most of these approaches utilize clustering techniques for grouping sentences and identifying informative cluster representative ones. However, such kind of approaches do not contain any supervised learning component and ignores the possible availability of labeled training data. We deal with this problem by considering sentence clustering techniques for improving supervised sentence-level scores, thereby realizing an effective combination of supervised as well as unsupervised learning. Moreover, since traditional clustering techniques ignore the task of representation learning, we consider the idea of deep embedding based clustering (Xie et al. 2016) which simultaneously improves the cluster assignments and the representation of sentences.

A Deep Clustering Enhanced Summarization approach (DCESumm) is proposed in this work, where firstly we find the summary relevance scores of each sentence in a document via Legal BERT based Multi-Layer Perceptron (MLP) model. Then, these scores are improved via deep clustering approach (Xie et al. 2016) to get the modified scores of the sentences. Finally, using the modified scores, we pick the sentences as per the desired summary length and sort them according to their original order in the document to get a summary. The main idea with respect to the clustering based score enhancement is that if a sentence is having high summary relevance and also belongs to a cluster that contains other summary relevant sentences, then the particular sentence should have a higher probability of inclusion in the predicted summary. Whereas, if a sentence individually scores high, but belongs to a cluster containing low scoring sentences, then the particular sentence should also be scored lower.

The key contributions of this work are:

- Summarization of legal documents is formulated as a sentence classification task in which we find the summary relevance scores of each sentence individually, via a domain-specific Legal BERT based MLP model.
- A novel sentence scoring approach is proposed using deep clustering to further enhance sentence-level summary relevance scores, thereby enabling the incorporation of global document-level information.

The organization of the rest of the paper is as follows: Sect. 2 presents a detailed literature review of previous related works. The methodology of this work is described in Sect. 3. Section 4 presents a detailed description of the dataset along with evaluation metrics, comparison methods and experimental setup. The experimental results are presented in Sect. 5, while Sect. 6 gives a detailed discussion on the findings from the experimental analysis. Finally we conclude the paper in Sect. 7 along with potential future research directions.

## 2 Related work

The problem of text summarization is well explored in the field of Natural Language Processing. To address this problem, several approaches have been proposed in the literature such as frequency based approaches (Luhn 1958; Edmundson 1969;

Nenkova and Vanderwende 2005), meta heuristic approaches (Clarke and Lapata 2008; Saini et al. 2019), Bayesian approaches (Vanderwende et al. 2007; Ma and Nakagawa 2013), etc. There are several supervised approaches also for solving text summarization problem, but these approaches require a large amount of labeled data (Louis et al. 2010). Graph-based approaches have recently been applied to solve the text summarization problem, where sentences and their pairwise similarity scores are considered as the nodes and edge weights respectively. Some of the graph based approaches are Textrank (Mihalcea and Tarau 2004) and Lexrank (Erkan and Radev 2004).

When it comes to summarizing legal documents, most of the work revolves around extractive summarization techniques, in which researchers make use of rhetorical labels for the formation of summaries (Hachey and Grover 2004; Saravanan et al. 2006; Bhattacharya et al. 2019b). There have been research where optimization based approaches are proposed for legal document summarization (Kanapala et al. 2019; Bhattacharya et al. 2021). Several comparative analysis works on legal judgment documents, considering extractive summarization techniques, have been carried out in the literature (Bhattacharya et al. 2019a; Jain et al. 2021d; Gupta et al. 2022). Most of the work have been carried out on case judgment documents in the past, but recently researchers have been exploring other legal documents as well, which includes legal bills, legal reports, patents, etc., (Eidelman 2019; Huang et al. 2021; Jain et al. 2020, 2021a, b). With the development of deep learning approaches, researchers nowadays are proposing solutions which involves deep learning components for solving the legal document summarization problem (Duan et al. 2019; Jain et al. 2021c).

Solving the summarization problem with the help of clustering technique is not new. There have been several works done which has used clustering based approaches ( Mallick et al. 2021; Wang et al. 2011) for text summarization in general. There have been works which proposes only the clustering based approach such as K-means approach for finding the salient sentences and hence for performing text summarization (Akter et al. 2017; Shetty and Kallimani 2017). Whereas there are other lines of research where researchers have used clustering based approach along with optimization techniques for the summarization task (Mishra et al. 2022). Recently researchers have utilized the contextualized embeddings for the formation of the clusters, following which the important sentences are picked from each cluster to form a summary (Moradi and Samwald 2019). Alguliyev et al. (2019) firstly discovered all the topics using the K-means clustering approach. In order to create optimal summary, a modified differential evolution (DE) technique is developed which allows to cover all the important topics in a summary and removes redundancy. Akter et al. (2017) calculated word score using the tf-idf approach and sentence score is the summation of its individual words' scores with its position along with the help of cue words and skeleton words. Finally the summary is created by using the K-means algorithm. Srikanth et al. (2022) proposed a dynamic way for determining the number of sentences from each cluster for summary generation. For the formation of clusters, K-means algorithm is used which clusters the BERT-based embeddings. Shetty and Kallimani (2017) used K-means algorithm to form clusters, which depend upon the cosine similarity measure. Finally important sentences are

chosen from each cluster selected to build a summary. Alqaisi et al. (2020) proposes a clustering-based approach along with multi-objective optimization techniques inspired by evolutionary algorithms. This optimization based approach addresses the key aspects of relevancy, diversity/redundancy and coverage for building the final summary. Recently, a novel sentence clustering approach is proposed by Mishra et al. (2022) which is based on multi-objective differential evolution technique.

Although clustering based summarization has been explored widely in the literature, very little or no work is done on feature learning. Clustering algorithms depend upon the notion of distance or dissimilarity among the data points in the feature space. For example, in case of the K-means clustering algorithm, we consider the $L_2$ norm between data points represented in a particular feature space. But how this feature space is chosen, is left as a research problem for the specific domain under consideration. Moreover, in case of lengthy documents, considering only sentence level information is not sufficient. We should also consider how does a sentence contribute towards the context of the entire document. In order to deal with these research gaps and also considering our domain specific dataset, we propose a Deep Clustering Enhanced Summarization approach (DCESumm) in this work. The main idea is to combine sentence-level summary relevance classification with deep sentence clustering based score enhancement. This kind of combination improves sentence scoring, thereby capturing both the global as well as the local context of the sentences. Moreover, due to the application of Deep Embedded Clustering approach, we overcome the limitations of traditional clustering techniques, thereby achieving better score enhancements.

## 3 Methodology

In this work, an improved sentence scoring mechanism DCESumm is proposed for performing extractive summarization of legal documents. The individual sentence scores in each document are firstly calculated by utilizing a trained MLP model based on Legal BERT embeddings. This probability score gives the summary relevance of each of the sentence of a document, based on the contents of individual sentences. Since these scores do not capture the global information of a document, therefore we propose to enhance these scores further with the help of deep embedded clustering based approach (DEC) which tries to capture the global information as well. The proposed approach is pictorially depicted in Fig. 1.

The DCESumm approach is firstly focused on the task of summary relevance prediction for individual sentences, as shown in Fig. 1. In this step we first find the contextual representation of the sentences using a pretrained language model and then create an extractive summarization dataset. This dataset can be utilized to train a sentence-level summary relevance prediction model. Once we have the summary relevance predictions for individual sentences, then we move on to the next step of clustering. In the second step, we perform Deep Embedded Sentence Clustering to form sentence clusters. Once the sentence clustering is done, in the third step we enhance the individual summary relevance scores of the sentences by using the cluster scores. Finally, using the enhanced scores we rank the document sentences for

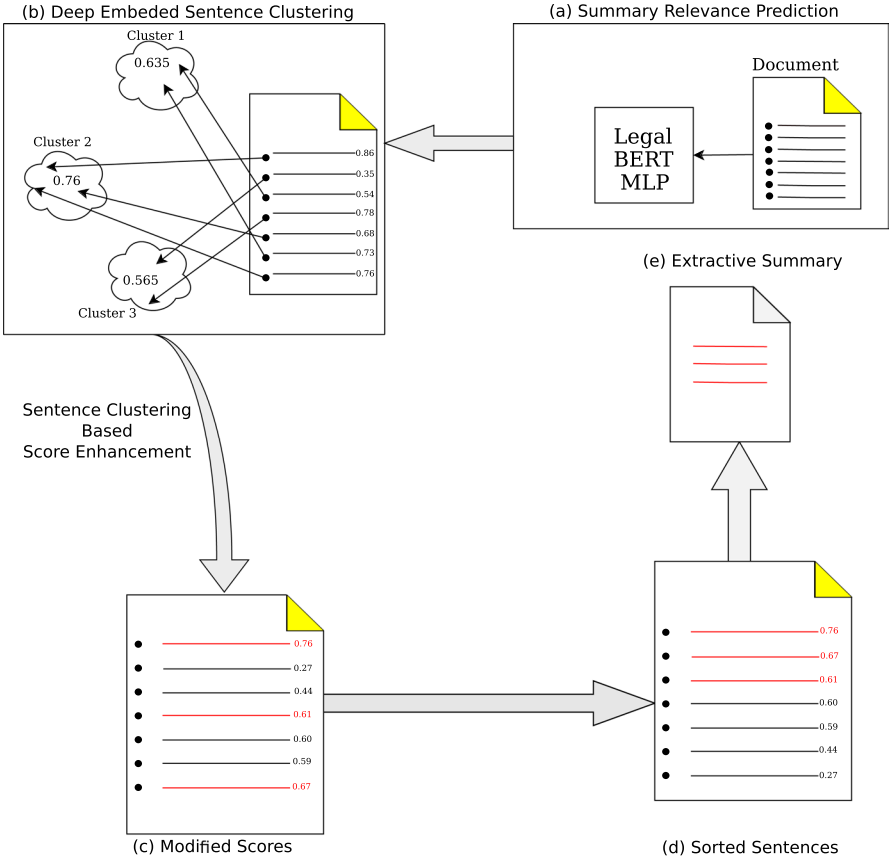



**Fig. 1** Illustration of getting predicted summaries using DCESumm

generating the final predicted summary. In this way, with the help of the proposed approach, we are able to achieve improved legal document summarization by combining the power of supervised sentence level predictions and unsupervised sentence clustering. The detailed description of each of these steps is presented in the subsections below.

### 3.1 Contextual representation

Given a document, the first task is to figure out which of the sentences in the document are individually informative, without considering the global context. In order to achieve this, firstly sentence embeddings are created for the individual sentences

present in the document under consideration. Using a labeled training dataset we can learn a binary classification model that can give us sentence level predictions for summary relevance. A BERT based pre-trained model is utilized to obtain the sentence representations. The domain specific application of BERT based pre-trained models are being explored by several researchers since these approaches are able to achieve state-of-the-art results on many different tasks (Devlin et al. 2018). Recently, a legal domain specific BERT model has been developed by Chalkidis et al. (2020), where the authors have shown significant performance improvements on several Legal NLP tasks. The same Legal BERT model consisting of 12-hidden layers with 768 units has been utilized in this work for finding sentence representations. The average of all the tokens in a sentence is considered to finally obtain a sentence contextual representation/features. In this way, our input sentence is represented in a $n$-dimensional vector consisting of 768 features. If the $i^{th}$ document consists of $k$ sentences, then the general representation can be written as:

$$\begin{aligned} d_i &= \{LB(s_1), LB(s_2), ....LB(s_k)\} \\ &= \{[s_1^1, s_1^2, ..., s_1^n], [s_2^1, s_2^2, ..., s_2^n]\}, \{[s_k^1, s_k^2, ..., s_k^n]\} \end{aligned} \tag{1}$$

where, $LB(s_k)$ is a pretrained Legal BERT model which is utilized to obtain the individual sentence representation of a document under consideration. We consider $n = 768$, which represents the dimension of each sentence embedding as resultant from Legal BERT pretrained model.

### 3.2 Extractive dataset building

In order to build an extractive summarization dataset we need to process the initial training samples appropriately. Given a set of documents $d_i$'s ($i = 1, 2...., m$) and their corresponding ground truth summaries $s_i$'s ($i = 1, 2...., m$), a new training dataset can be built with the help of sentence scores obtained by Algorithm 1. The main idea is that once we obtain the sentence scores (*d_scores*) for each sentence inside each document, we then pick the top 20% of the high scoring sentences to be considered as being relevant of inclusion in summary. Apart from this, we also include all the sentences from the reference summaries as part of the extractive summary for the particular documents. This kind of extractive dataset building approach has been utilized by other researchers as well (Anand and Wagh 2019). In this way, we get the extractive training dataset as shown below:

$$D^{TrExt} = \{[d_1, y_1], [d_2, y_2], ....., [d_m, y_m]\} \tag{2}$$

where, each individual $d_i$ consists of sentence embedding vectors of $n$-dimensions, and each $y_i$ consists of 0's and 1's representing sentence relevance for summary generation.

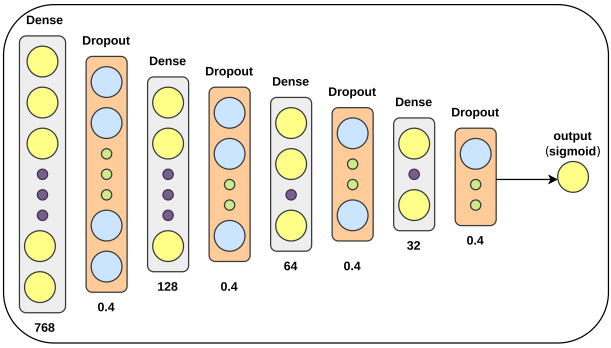



**Fig. 2** MLP model for sentence-level summary relevance prediction

**Algorithm 1** Sentence scoring for Extractive Dataset Building

**Input:** D=$\{d_1, d_2, d_3, ....d_m\}$, S=$\{s_1, s_2, s_3, ....s_m\}$, where D is a set of training document and S is a set of summary

**Output:** Sentence scores for each document $d\_scores$

$l\_list \leftarrow [\,][\,]$
$se\_list \leftarrow [\,][\,]$
$ds\_list \leftarrow [\,]$
$d\_scores \leftarrow [\,]$
**for** $i = 1\ to\ \mid D \mid$ **do**
    tokenize $D[i]$ to create a sentence list s_list
    $scores \leftarrow []$
    $ds\_list[i] \leftarrow s\_list$
    **for** $j = 1\ to\ \mid s\_list \mid$ **do**
        $score \leftarrow \frac{(r_1 F1(s\_list[j],S[i])+r_2 F1(s\_list[j],S[i])+r_l F1(s\_list[j],S[i]))}{3}$ ▷ Compute ROUGE scores
        $scores[j] \leftarrow score$
    **end for**
$d\_scores[i] \leftarrow scores$
**end for**
**return** $d\_scores$

### 3.3 Summary relevance classification task

For the purpose of learning the binary classification task of sentence summary relevance prediction, we use an MLP model ($M_1$) as shown in Fig. 2. The MLP model takes sentence embedding as an input and outputs the probability of a sentence to be involved in summary. For training $M_1$, $D^{TrExt}$ is flattened such that $D^{TrExtMLP}$ is obtained as shown below:

$$
\begin{aligned}
D^{TrExtMLP} = \{&([x_1, x_2, ...., x_n]^{(1)}, y^{(1)}),\\
&([x_1, x_2, ...., x_n]^{(2)}, y^{(2)}),\\
&\vdots\\
&\vdots\\
&([x_1, x_2, ...., x_n]^{(q)}, y^{(q)})\}
\end{aligned} \tag{3}
$$

where, $n$ is the sentence embedding dimension, and the total no. of sentences is represented by $q$ for all the documents.

We then train the MLP model on $D^{TrExtMLP}$, which finally generates a summary relevance probability score with the help of the sigmoid function as shown in Eq. 4. This probability score lies in the range of [0, 1].

$$
Summary\ Relevance\ Score(s') = \frac{1}{1 + e^{-M_1(LB(s'))}} \tag{4}
$$

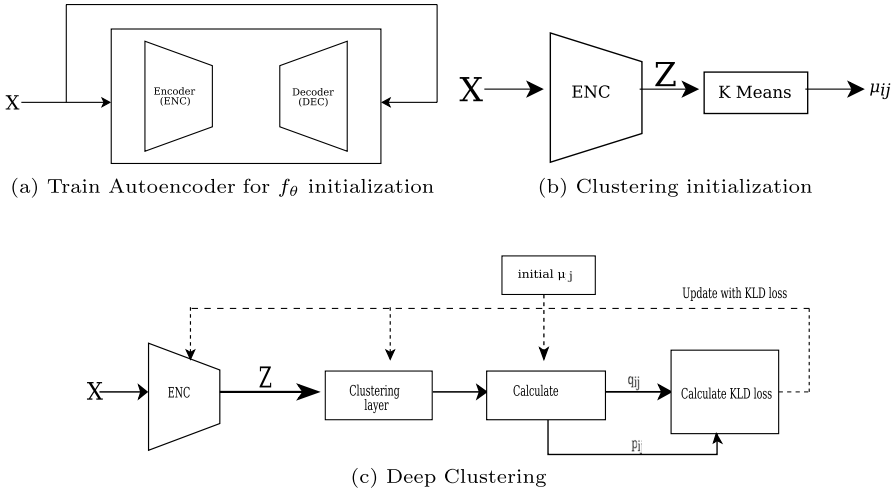
where, $s'$ is a sentence for which its summary relevance is to be determined, $LB(s')$ is the Legal BERT based $n$-dimensional vector representation of $s'$, and $M_1(LB(s'))$ gives the summary relevance score for $s'$ which is scaled to the range of [0, 1] using the sigmoid function.

This kind of scoring is used by previous researchers for text summarization in general but they are limited by the use of summary relevance at sentence level locally. In order to deal with this limitation, we propose a clustering based sentence score enhancement, which can give better indication of the summary relevance of sentences, considering the global document-level context.

### 3.4 Deep clustering approach

The overall deep clustering approach is shown in Fig. 3, which is used for building sentence clusters for the enhancement of individual scores. In this approach, firstly we find the feature representation using deep autoencoder as shown in Fig. 3(a), and then we initialize the cluster centroids via K-means algorithm as depicted in Fig. 3(b). Once the initialization is done, cluster centroids are iteratively refined by minimizing the Kullback-Lieber (KL) divergence loss between the soft cluster assignments and the auxiliary target distribution as shown in Fig. 3(c). This minimization is performed with the help of Stochastic Gradient Descent (SGD) optimizer with 0.01 learning rate. Each of these steps is explained in a detailed manner below:

- In order to map between the data space and the feature space, we initialize the DEC with stacked autoencoder (SAE). SAE consists of several layers of denoising autoencoders, which further contain encoder-decoder pairs. Training is performed for the reconstruction of previous layer's output after random corruption as done by Xie et al. (2016). Once the training is done, all the layers of denoising encoder are concatenated in such a way, so that a deep autoencoder is formed. From this deep autoencoder, decoder layer is discared and only the encoder layer is utilized for obtaining the low dimensional features, as shown in Fig. 3(a).




(a) Train Autoencoder for $f_\theta$ initialization (b) Clustering initialization

(c) Deep Clustering

**Fig. 3** Overall model for finding sentence clusters using DEC (Xie et al. 2016) approach

- Initialization of cluster centroids is done using the obtained feature embeddings via K-means clustering, as shown in Fig. 3(b). The basic idea of K-means approach is to group similar data points by means of some similarity measure between them. In this work, we consider Euclidean distance for finding the distance between two data points.
- Once the initialization is done, we iteratively refine the cluster centers by minimization of the KL divergence loss between the true target distribution and the soft assignments of clusters, using SGD optimizer. The reason for choosing SGD optimizer is that, it is one of the most widely used optimizers in the deep learning literature (Acharya et al. 2019; Tajaddodianfar et al. 2020; Umer et al. 2021) that is capable of optimizing neural network weight parameters effectively. This process is repeated until convergence or the tolerance value is reached, as shown in Fig. 3(c).

The main idea of deep clustering approach is that, feature representation and clustering assignments are improved simultaneously with the help of KL divergence loss function. More specifically after initialization of cluster centers, soft assignments are computed between embedded points and cluster centroids. After this, embedded points are updated and using an auxiliary target distribution along with the current high confidence assignments, cluster centroids are also refined. The updation process is done through KL divergence loss between the auxiliary distribution $p_i$ which represents the high confidence predictions and the soft assignments $q_i$. This loss formulation is performed in the same way as Xie et al. (2016). The soft assignment for the $j^{th}$ cluster ($q_{ij}$) is done using the student's $t$ distribution which represents the soft probability of assigning $i^{th}$ data point to the $j^{th}$ cluster. Whereas, the individual components for the true/auxiliary target distribution ($p_i$) are defined in Eq. 5, corresponding to the $j^{th}$ cluster assignment.

$$p_{ij} = \frac{q_{ij}^2 \,/\, f_j}{\sum_{j'} q_{ij'}^2 \,/\, f_{j'}} \tag{5}$$

where, $p_{ij}$ is the individual component of the true/auxiliary target distribution ($p_i$) for the $j^{th}$ cluster, with soft frequencies of clusters $f_j = \sum_i q_{ij}$ and $q_{ij}$ is the soft probability of assigning the $i^{th}$ data point to the $j^{th}$ cluster.

The idea of true target distribution or the auxiliary target distribution is that, these are the high confidence assignments based on which we iteratively refine the soft assignments. Therefore the choice of target distribution is very crucial for DEC approach to work. This distribution should have some properties such as it should strengthen predictions, put more emphasis on high confidence points and normalize loss contribution of each centroid to prevent large clusters from distorting the hidden feature space. The main intuition behind the $p_{ij}$ formulation mentioned in Eq. 5 is to have larger probability values for highly confident cluster assignments as governed by the $q_{ij}$ value. For example, if a particular point has a $q_{ij}$ value of 0.9, then according to Eq. 5 the numerator will still have a larger value (0.81 before normalization by $f_j$). This will ensure that such an assignment is having high probability in the true/auxiliary distribution. Whereas, for a less confident cluster assignment with smaller $q_{ij}$ (for example $q_{ij} = 0.5$), the numerator will get reduced significantly. Thereby ensuring that such an assignment has a lower probability in the true distribution. With such an appropriate true/auxiliary distribution, the training process can be pushed towards better cluster assignments by reducing the KL-divergence between the soft assignments and the current high confidence assignments as shown below:

$$KL\ Loss = \sum_i \sum_j p_{ij}\, log \frac{p_{ij}}{q_{ij}} \tag{6}$$

The whole process of assigning each sentence of a document to a cluster is done through deep embedding based clustering which we refer to as $M_2$ model as shown below:

$$M_2([s^1, s^2, ..., s^n]) = c_j \tag{7}$$

where, $M_2$ is the deep clustering model which is applied on a sentence embedding vector, $n$ is the sentence embedding dimension and $c_j$ represents the cluster to which the sentence is assigned (with $j = 1, 2, ..., k$). The value of $k$ is chosen based upon the $\hat{k}$, which is the cluster ratio percentage determined on the basis of number of sentences in a document. The value of $\hat{k}$ is calculated as 35% of the document length in terms of number of sentences. This value is chosen via empirical evaluation of a set of six possible choices- {20%, 25%, 30%, 35%, 40%, 45%}.

### 3.5 Cluster based score

Using the deep clustering model ($M_2$) that has been explained in Subsection 3.4, we find the cluster score as follows:

For each sentence $s'_{ij}$ in cluster $c_j$, we find the cluster score ($Score(c_j)$) as the average of all the individual summary relevance scores of the sentences in that cluster, as depicted in the equation below:

$$Score(c_j) = \frac{1}{\mid c_j \mid} \sum_{j=1}^{m} M_1(s'_{ij}) \tag{8}$$

Now using these scores, we modify the individual sentence $s'_{ij}$ scores as shown below:

$$Final\ Score = \frac{(1\ +\ Score(c_j))}{2} \times M_1(s'_{ij}) \tag{9}$$

This cluster score based enhancement of summary relevance scores can be thought of as a community based weighting scheme for sentences. The intuition is that if a sentence belongs to a cluster that mostly consists of high scoring sentences, then the sentence should get a higher *Final Score*, whereas if a sentence belongs to a cluster that is mostly consisting of low scoring sentences, then the sentence should also get a lower score. This essentially conveys the fact that individually a sentence might seem very informative and relevant of inclusion in the predicted summary, but if it is similar to many low scoring sentences in the document, then probably our initial assessment of the sentence was an overestimation and we need to update our knowledge. This kind of community based influence is very commonly found in the social network analysis literature, where a social entity (node) can get influenced significantly by its surroundings (Bonhard and Sasse 2006; Carmel et al. 2009). In the best case scenario for a sentence $s_{ij}$, the cluster score based weight component $\frac{(1 + Score(c_j))}{2}$ can have little to no impact on the sentence score $M_1(s'_{ij})$, if the sentence belongs to a very high scoring cluster. This is because, the expression $\frac{(1 + Score(c_j))}{2}$ goes towards the value 1 as the cluster scores increases. On the other hand, in the worst case scenario, a sentence's summary relevance score can be scaled down to almost half of its value when it belongs to a very low scoring cluster. This kind of weighting ensures that we get local sentence-level importance as well as the quality of the sentence in a more global manner at the document level.

### 3.6 Summary formation

With these final scores, we sort all the sentences of a document in decreasing order. We then pick top sentences from this sorted list to get the sentences to finally form a summary. More specifically, we pick top $l$% of the sentences from each document, where $l$ represents the average ratio of number of words in the gold summaries to the number of words in the documents in the training dataset. Considering the experimental evaluation of the proposed approach on the BillSum (Eidelman 2019) and FIRE (Parikh et al. 2021) datasets, these average ratios are 15% and 40% respectively.

At the time of inference, we can generate the summary of a particular test document by following the steps as shown in Algorithm 2. The trained models $M_1$

and $M_2$ can be utilized to generate the predicted summary of the test document as depicted in the algorithm below:

**Algorithm 2** Summary Prediction Algorithm

**Input:** $M_1$, $M_2$ where $M_1$ is a trained Legal BERT MLP model and $M_2$ is a trained DEC clustering model, Testing document D

**Output:** Predicted Summary for each document in D

$pred\_summ \leftarrow [\,]$
$f\_score \leftarrow [\,]$
**for** $i = 1\ to\ \mid D \mid$ **do**
    $sents \leftarrow$ sent_tokenize($D[i]$)
    $m\_score \leftarrow [\,]$
    **for** $j = 1\ to\ \mid sents \mid$ **do**
        $score \leftarrow M_1$.predict($sents[j]$) ▷ $score$: Summary relevance score
        $m\_score[j] \leftarrow getClusterScore(M_2, sents[j], D[i], score)$ ▷ Using Cluster model as in Eq. 8
    **end for**
    $f\_score[i] \leftarrow m\_score$
**end for**
$s\_sents \leftarrow\ getSortedDocSents(f\_score, D)$ ▷ Sort sentences based on their final score
$l\_sents \leftarrow\ pickTopL(s\_sents, D)$ ▷ Pick top $l\%$ sentences for summaries.
$pred\_summ \leftarrow\ joinTopSents(l\_sents, D)$ ▷ Join sentences in original documents' order.
**return** pred_summ

## 4 Experimental setup and evaluation strategy

### 4.1 Evaluation metrics

The automatic assessment of the proposed approach is carried out using the ROUGE metric, which is a standard metric for evaluating the system generated summaries in the text summarization literature. ROUGE stands for Recall Oriented Understudy for Gisting Evaluation which counts the number of word sequences, word counts, n-gram overlaps between the reference summaries (ideally created by humans) and system generated (predicted) summaries. Several variants of ROUGE are there such as ROUGE-S, ROUGE-W, ROUGE-L and ROUGE-N. Each of the variants of ROUGE evaluation measure generates three scores (Precision, Recall and F1-measure). In this work, ROUGE-N (where N= 1, 2) and ROUGE-L are considered for evaluating the predicted summaries which are discussed below:

- **ROUGE-N:** ROUGE-N measures the n-gram overlaps between the reference summaries and the predicted summaries as shown in the equation below:

$$ROUGE - N = \frac{\sum_{S \in summ_{ref}} \sum_{N-gram \in S} Count_{match}(N - gram)}{\sum_{S \in summ_{ref}} \sum_{N-gram \in S} Count(N - gram)} \tag{10}$$

  where N is the N-gram length, $Count_{match}(N - gram)$ is the overlapping of N-grams between predicted summaries and reference summaries, $Count(N - gram)$ is the total possibe N-grams in reference summaries. ROUGE-1 counts the unigrams whereas ROUGE-2 counts the bigrams between predicted summaries and reference summaries.
- **ROUGE-L:** It is the longest common subsequence (LCS) based metric which measures the longest sequence of words that are common between predicted summary and reference summary. In LCS, there are no consecutive matches, but the matches are in-sequence, and therefore sentence-level word order is reflected as n-grams. ROUGE-L between reference summary X of length m and predicted summary Y of length n is calculated as shown below:

$$ROUGE - L = \frac{LCS(X, Y)}{m} \tag{11}$$

  where LCS(X,Y) is the longest common subsequence length between X and Y.

### 4.2 Data

In this work, two legal benchmark datasets have been used for performing the experiments whose details are given below:

- **BillSum:** BillSum dataset is firstly introduced by Eidelman (2019), which has 22,218 United States (US) Congressional bills and 1,237 California (CA) state bills. The US Congressional bills are further split into 18,949 training and 3,269 testing samples. In the case of US Congressional bills, both the training and testing documents contain 62 average number of sentences. Whereas the US training and testing summaries contain 6 average number of sentences. Whereas the CA testing documents and summaries contain 46 and 9 average number of sentences respectively.
- **FIRE dataset:** This dataset is part of a legal judgement summarization competition organized by the event- Artificial Intelligence for Legal Assistance (AILA), which is collocated with the FIRE-2021 (Parikh et al. 2021b) conference. We have participated in this competition as the “nits_legal" team (Jain et al. 2021c), and thus have gained access to this dataset (Parikh et al. 2021a). The training data consists of 500 document-summary pairs (Parikh et al. 2021). These are the judgement documents delivered by the Supreme Court of India. Moreover, the pre-processed and sentence tokenized versions of the documents as well as the summaries are provided by the organizers. In addition to the summaries, each document is accompanied by the sentence relevance labels (binary labels) either 0 or 1 and rhetorical labels (multiclass

labels). They also provided 50 documents during the testing time but since due to the non-availability of ground truth labels, we made use of training dataset only for performing our experiments. We randomly split our training dataset into training, validation and testing dataset (400/50/50) five times. In this way, we get five different train-val-test splits and we report the average of those five different random splits in our paper.

### 4.3 Baseline/SOTA approaches for comparison

A comparison is performed with seven widely used baseline methods on BillSum and FIRE datasets as shown in Tables 1, 2, 3 and 4. Apart from baseline comparison, we also perform state-of-the-art (SOTA) comparisons which are shown in Tables 5, 6, 7 and 8.

#### 4.3.1 Baseline extractive summarization methods

- **Textrank** (Mihalcea and Tarau 2004): It is a graph-based approach in which the document is converted into a graph. This graph has the sentences as nodes and edges are built on the basis of similarity between two sentences.
- **Sumbasic** (Nenkova and Vanderwende 2005): It is a greedy search approximation approach in which the scoring of the sentence is done based upon average probability of words in that sentence with a re-weight component to minimize redundancy.
- **Latent Semantic Analysis (LSA)** (Steinberger and Jezek 2004): The idea of this approach is to apply Singular Value Decomposition (SVD) for obtaining the salient sentences present in a document.
- **KLSum** (Haghighi and Vanderwende 2009): The main idea in this approach is to keep on adding sentences to the summary in a greedy manner as long as a decrease can be observed in the KL divergence between the summary and document sets.
- **Reduction** (Jing 2000): The idea of this approach is to remove the extraneous phrases from the extracted sentences from the document. The output of the reduction approach is the reduced sentence which is used to produce the summary.
- **Restricted Boltzmann Machines (RBM)** (Verma and Nidhi 2017): It is an unsupervised deep learning approach in which summary is created by firstly extracting the features followed by feature enhancement.
- **Casesummarizer** (Polsley et al. 2016): It is a legal specific baseline approach which produces extractive summary based on the frequency of words along with some extra domain specific knowledege.

#### 4.3.2 State-of-the art approaches for comparison

- **Legal-Pegasus**[1]: The Pegasus (Zhang et al. 2020) model is fine-tuned on the US securities litigation dataset to obtain the Legal Pegasus model.

[1] https://huggingface.co/nsi319/legal-pegasus.

**Table 1** ROUGE metric based comparison against abstractive groundtruth summaries on BillSum test dataset

| | Approach | US testing dataset | | | CA testing dataset | | |
|---|---|---|---|---|---|---|---|
| | | ROUGE-1 | ROUGE-2 | ROUGE-L | ROUGE-1 | ROUGE-2 | ROUGE-L |
| Baseline | ContextEmbed | 0.3804 | 0.2234 | 0.3886 | 0.4210 | 0.2083 | 0.3498 |
| | Reduction (Jing 2000) | 0.3473 | 0.1757 | 0.3305 | 0.3996 | 0.1844 | 0.3256 |
| | KLSum (Haghighi and Vanderwende 2009) | 0.2638 | 0.0927 | 0.2139 | 0.2800 | 0.1035 | 0.2265 |
| | LSA (Steinberger and Jezek 2004) | 0.3277 | 0.1289 | 0.2891 | 0.3364 | 0.1314 | 0.2971 |
| | Sumbasic (Nenkova and Vanderwende 2005) | 0.2398 | 0.0811 | 0.2275 | 0.3280 | 0.1277 | 0.2977 |
| | Textrank (Mihalcea and Tarau 2004) | 0.3270 | 0.1794 | 0.3384 | 0.4069 | 0.2016 | 0.3457 |
| | RBM (Verma and Nidhi 2017) | 0.2971 | 0.1080 | 0.2397 | 0.3166 | 0.1007 | 0.2470 |
| | CaseSummarizer (Polsley et al. 2016) | 0.3402 | 0.1449 | 0.2851 | 0.3632 | 0.1552 | 0.2948 |
| Proposed | DCESumm | 0.4200 | 0.2428 | 0.3887 | 0.4366 | 0.2389 | 0.3915 |

**Table 2** ROUGE metric based comparison against abstractive groundtruth summaries on FIRE testing dataset

| | Approach | FIRE test dataset | | |
|---|---|---|---|---|
| | | ROUGE-1 | ROUGE-2 | ROUGE-L |
| Baseline | ContextEmbed | 0.5039 | 0.3350 | 0.4488 |
| | Reduction (Jing 2000) | 0.4691 | 0.2816 | 0.3815 |
| | KLSum (Haghighi and Vanderwende 2009) | 0.4319 | 0.2589 | 0.3771 |
| | LSA (Steinberger and Jezek 2004) | 0.3930 | 0.2330 | 0.3714 |
| | Sumbasic (Nenkova and Vanderwende 2005) | 0.4093 | 0.2550 | 0.3913 |
| | Textrank (Mihalcea and Tarau 2004) | 0.4974 | 0.2805 | 0.3913 |
| | RBM (Verma and Nidhi 2017) | 0.4387 | 0.1956 | 0.3521 |
| | CaseSummarizer (Polsley et al. 2016) | 0.4909 | 0.2662 | 0.3944 |
| Proposed | DCESumm | 0.5092 | 0.3599 | 0.4878 |

- **BO-Textrank** (Jain et al. 2020): A Bayesian Optimization (BO) based approach is proposed by the authors for improving the Textrank algorithm for extractive summarization. The fine-tuned Textrank is then used for summarization task.
- **SummaRuNNer** (Nallapati et al. 2017): The SummaRuNNer model is based upon Recurrent Neural Networks (RNNs), which formulates the problem of extractive summarization as a binary sequence labeling task.

### 4.4 Experimental environment

The DCESumm approach is compared against baseline and state of the art approaches which has been shown in Sect. 5. A ROUGE based evaluation is done where the F1 scores corresponding to the ROUGE-L, ROUGE-2, ROUGE-1 metrics are reported in Sect. 5. The experimental analysis has been carried out on publicly available BillSum dataset and competition-based FIRE dataset which are described in Sect. 4.2. In case of BillSum dataset, summary length is set to 15%. This is the ratio of reference summary word counts to the document word counts in the training dataset. In case of the FIRE dataset, the same ratio is found to be 40% which is used as a summary length. For training the MLP model presented in Fig. 2, the BillSum training dataset has been split into an 80:20 ratio, where 80% is used for training and 20% is used for validation. In case of the FIRE dataset, we randomly split the training dataset five times into 400/50/50 training, validation and testing documents respectively. Our MLP model consists of four fully connected layers with 768, 128, 64 and 32 units followed by a sigmoid activation layer. We consider a batch size of 32, an Adam optimizer (Kingma and Ba (2014)) with 0.001 learning rate. Whereas, for the training of the deep clustering model, firstly autoencoder is pretrained for 500 epochs with a batch size of 64, a mean-squared error loss function is used. The training is performed using an SGD optimizer with a learning rate of 0.01. K-means clustering is performed with 20 restarts and the best initialization of centroids is chosen for the subsequent deep clustering process.

**Table 3** ROUGE metric based comparison against extractive groundtruth summaries on BillSum test dataset

| | Approach | US testing dataset | | | CA testing dataset | | |
|---|---|---|---|---|---|---|---|
| | | ROUGE-1 | ROUGE-2 | ROUGE-L | ROUGE-1 | ROUGE-2 | ROUGE-L |
| Baseline | ContextEmbed | 0.7789 | 0.6952 | 0.7622 | **0.7068** | 0.5859 | 0.6708 |
| | Reduction (Jing 2000) | 0.6810 | 0.5952 | 0.6835 | 0.6601 | 0.5738 | 0.6655 |
| | KLSum (Haghighi and Vanderwende 2009) | 0.2212 | 0.1306 | 0.2481 | 0.3365 | 0.1719 | 0.2931 |
| | LSA (Steinberger and Jezek 2004) | 0.4311 | 0.2869 | 0.4453 | 0.4327 | 0.2615 | 0.4094 |
| | Sumbasic (Nenkova and Vanderwende 2005) | 0.2564 | 0.1431 | 0.3235 | 0.4219 | 0.2428 | 0.4127 |
| | Textrank (Mihalcea and Tarau 2004) | 0.7008 | 0.5820 | 0.6723 | 0.6467 | 0.5039 | 0.6029 |
| | RBM (Verma and Nidhi 2017) | 0.3022 | 0.1455 | 0.2881 | 0.3790 | 0.1797 | 0.3285 |
| | CaseSummarizer (Polsley et al. 2016) | 0.3711 | 0.2367 | 0.3616 | 0.4398 | 0.2561 | 0.3796 |
| Proposed | DCESumm | **0.7808** | **0.6974** | **0.7663** | 0.7056 | **0.5893** | **0.6752** |

**Table 4** ROUGE metric based comparison against extractive groundtruth summaries on FIRE testing dataset

| | Approach | FIRE test dataset | | |
|---|---|---|---|---|
| | | ROUGE-1 | ROUGE-2 | ROUGE-L |
| Baseline | ContextEmbed | 0.5716 | 0.4892 | 0.5882 |
| | Reduction (Jing 2000) | 0.4374 | 0.4157 | 0.5381 |
| | KLSum (Haghighi and Vanderwende 2009) | 0.4411 | 0.3996 | 0.5192 |
| | LSA (Steinberger and Jezek 2004) | 0.4467 | 0.4095 | 0.5312 |
| | Sumbasic (Nenkova and Vanderwende 2005) | 0.5514 | 0.3989 | 0.5272 |
| | Textrank (Mihalcea and Tarau 2004) | 0.5158 | 0.4381 | 0.5527 |
| | RBM (Verma and Nidhi 2017) | 0.5206 | 0.3351 | 0.4611 |
| | CaseSummarizer (Polsley et al. 2016) | 0.5660 | 0.4328 | 0.5381 |
| Proposed | DCESumm | 0.6741 | 0.6437 | 0.7227 |

All the experiments are performed using a Linux based 64-bit machine equipped with 16 GB RAM, i7 processor and RTX-2070 GPU (8 GB RAM). All the implementations are done with the help of the Python programming language using packages like Tensorflow (Abadi et al. (2016)), Scikit-learn (Pedregosa et al. 2011), Numpy (Harris et al. 2020), Gensim (Rehurek and Sojka 2010) and spaCy (Honnibal et al. 2020). The experimental results are evaluated against reference summaries using the ROUGE score Python package (Lin 2004).

## 5 Results and analysis

In this section, the experimental evaluation of the proposed approach has been presented in a detailed manner. More specifically, we present comparisons against baseline and SOTA approaches in Sects. 5.1 and 5.2 respectively, along with the N-gram overlap analysis presented in Sect. 5.3. We also present an inference time analysis to understand the average per sample prediction time, during testing in Sect. 5.4. Moreover, we also present a comparison of the proposed approach with a K-means based clustering approach for different cluster ratios in Sect. 5.5.

### 5.1 Baseline comparison

Table 1 shows the ROUGE-F1 scores on US Testing and CA Testing datasets when compared against the abstractive groundtruth summaries. From this table, we can see that our proposed approach DCESumm outperforms all the baseline approaches. More specifically, our proposed approach significantly improves ROUGE-1 and ROUGE-2 F1 scores by 4% and 2% respectively, on

**Table 5** Comparison of DCESumm approach against state-of-the-art approach on BillSum test dataset with respect to abstractive groundtruth summaries

| | Approach | US Testing dataset | | | CA Testing dataset | | |
|---|---|---|---|---|---|---|---|
| | | ROUGE-1 | ROUGE-2 | ROUGE-L | ROUGE-1 | ROUGE-2 | ROUGE-L |
| State-of-the-art | BO-Textrank (50) (Jain et al. 2020) | 0.3560 | 0.1720 | 0.3080 | 0.4010 | 0.1850 | 0.3170 |
| | BO-Textrank (2000) (Jain et al. 2020) | 0.3480 | 0.1740 | 0.3140 | 0.4040 | 0.1910 | 0.3240 |
| | BO-Textrank (5000) (Jain et al. 2020) | 0.3410 | 0.1750 | 0.3160 | 0.4030 | 0.1940 | 0.3270 |
| | SummaRUnner (Nallapati et al. 2017) | 0.4160 | 0.2245 | **0.3915** | 0.3862 | 0.1747 | 0.3281 |
| | Legal Pegasus | 0.3419 | 0.1625 | 0.3016 | 0.3412 | 0.1536 | 0.2981 |
| Proposed | DCESumm | **0.4200** | **0.2428** | 0.3887 | **0.4366** | **0.2389** | **0.3915** |

**Table 6** Comparison of DCESumm approach against state-of-the-art approach on FIRE test dataset with respect to abstractive groundtruth summaries

| | Approach | FIRE test dataset | | |
|---|---|---|---|---|
| | | ROUGE-1 | ROUGE-2 | ROUGE-L |
| State-of-the-art | BO-Textrank (50) (Jain et al. 2020) | 0.3991 | 0.2381 | 0.3726 |
| | BO-Textrank (2000) (Jain et al. 2020) | 0.4014 | 0.2408 | 0.3752 |
| | BO-Textrank (5000) (Jain et al. 2020) | 0.4011 | 0.2406 | 0.3758 |
| | SummaRUnner (Nallapati et al. 2017) | 0.4457 | 0.227 | 0.3263 |
| | Legal Pegasus | 0.3391 | 0.1654 | 0.297 |
| Proposed | DCESumm | 0.5092 | 0.3599 | 0.4878 |

the US testing dataset. Whereas our proposed approach improves the ROUGE-1, ROUGE-2, ROUGE-L F1 scores by nearly 2%, 3% and 4% respectively on the CA testing dataset. This proves that incorporating deep clustering approach along with the ContextEmbed approach helps in achieving significant improvements in terms of the extractive summarization quality. One of the noticeable points from these results is that, our approach has been able to beat ContextEmbed approach which means that adding the deep cluster scores actually enhanced the scores of the sentences and hence the summarization results.

Table 2 shows the average ROUGE-F1 scores of five random splits on the FIRE test dataset when compared against the abstractive groundtruth summaries. From this table, we can see that our proposed approach DCESumm outperforms all the baseline approaches. Specifically, our approach improves ROUGE-1, ROUGE-2 and ROUGE-L F1 scores by 1%, 2.5%, and 4% respectively. The importance of deep clustering based enhancement of sentence scores is once again quite evident from these results.

Tables 3 and 4 show the ROUGE-F1 scores for the case when the predicted summaries are compared against the noisy extractive ground truth labels for the BillSum and FIRE test datasets. From these results, we see that our proposed approach is performing extremely well across all the test sets, achieving ROUGE-1, ROUGE-2 and ROUGE-L F1 scores that are greater than or equal to 0.65, 0.55 and 0.65 respectively. Another important observation from this analysis is that, the ROUGE-F1 scores against extractive summaries is much better than the scores against abstractive summaries. This is due to the fact that, we are performing extractive summarization of the document and forming our predicted summaries by directly using the sentences that are present in the document. However, it is important to note here that comparing our predicted summaries against extractive ground truth summaries gives a better indication of the quality of the predicted summaries and helps with the understanding that the proposed approach is actually able to select important summary relevant sentences.

**Table 7** Comparison of DCESumm against state-of-the-art on BillSum test dataset with respect to extractive groundtruth summaries

| | Approach | US Testing dataset | | | CA Testing dataset | | |
|---|---|---|---|---|---|---|---|
| | | ROUGE-1 | ROUGE-2 | ROUGE-L | ROUGE-1 | ROUGE-2 | ROUGE-L |
| State-of-the-art | BO-Textrank (50) (Jain et al. 2020) | 0.3818 | 0.2421 | 0.3234 | 0.3912 | 0.2420 | 0.3385 |
| | BO-Textrank (2000) (Jain et al. 2020) | 0.3835 | 0.2436 | 0.3388 | 0.3763 | 0.2420 | 0.3239 |
| | BO-Textrank (5000) (Jain et al. 2020) | 0.3844 | 0.2441 | 0.3446 | 0.3787 | 0.2427 | 0.3310 |
| | SummaRUnner (Nallapati et al. 2017) | 0.4186 | 0.2937 | 0.4345 | 0.4339 | 0.2802 | 0.4113 |
| | Legal Pegasus | 0.3553 | 0.2338 | 0.3906 | 0.4016 | 0.2504 | 0.4022 |
| Proposed | DCESumm | 0.7808 | 0.6974 | 0.7663 | 0.7056 | 0.5893 | 0.6752 |

**Table 8** Comparison of DCESumm approach against state-of-the-art on FIRE test dataset with respect to extractive groundtruth summaries

| | Approach | FIRE test dataset | | |
|---|---|---|---|---|
| | | ROUGE-1 | ROUGE-2 | ROUGE-L |
| State-of-the-art | BO-Textrank (50) (Jain et al. 2020) | 0.4441 | 0.3751 | 0.3841 |
| | BO-Textrank (2000) (Jain et al. 2020) | 0.4743 | 0.3663 | 0.4505 |
| | BO-Textrank (5000) (Jain et al. 2020) | 0.4682 | 0.3621 | 0.4542 |
| | SummaRUnner (Nallapati et al. 2017) | 0.4646 | 0.3988 | 0.5153 |
| | Legal Pegasus | 0.3769 | 0.1874 | 0.3234 |
| Proposed | DCESumm | 0.6741 | 0.6437 | 0.7227 |

### 5.2 State-of-the-art comparison

A comparison of the proposed approach with the current state-of-the art approaches for the BillSum and FIRE datasets are presented in Tables 5, 6, 7 and 8. Here, Tables 5 and 6 present the comparison results with respect to the abstractive groundtruth summaries. Whereas, Tables 7 and 8 present the comparison results with respect to the extractive groundtruth summaries (generated with the help of Algorithm 1). From these comparisons, it is quite evident that the proposed DCESumm approach outperforms all the state-of-the-art methods considered, by a great margin with respect to both the abstractive as well as the extractive groundtruth summaries. Moreover, with respect to almost all the variants of ROUGE metric, the proposed approach is able to achieve better summarization results. This establishes the fact that the proposed DCESumm approach can be utilized for achieving improved summarization of legal documents.

### 5.3 N-gram overlap analysis

N-gram overlap based analysis is another important dimension along which the proposed DCESumm approach can be evaluated with respect to the summarization task. Such an evaluation will shed light on the amount of overlap between the predicted and the abstractive gold summaries, thereby enabling further assessment of the proposed approach. Fig. 4 shows the percentage of n-gram overlap of the DCESumm predictions with respect to the abstractive ground truth summaries and also the extractive ground truth summaries which has been created by us. The x-axis of the bar-charts present the n-grams (with n = 1, 2, 3 and 4), whereas the y-axis gives the amount of overlap in terms of percentage. From these graphs, we see that, the overlapping of predicted summaries and noisy ground truth is larger in comparison with the overlapping between predicted summaries and abstractive ground truth summaries, across all the different test sets. The reason being that, predicted summaries are of extractive nature in which summary relevant sentences are directly picked up from the document to form the summary. Therefore, there is more overlap

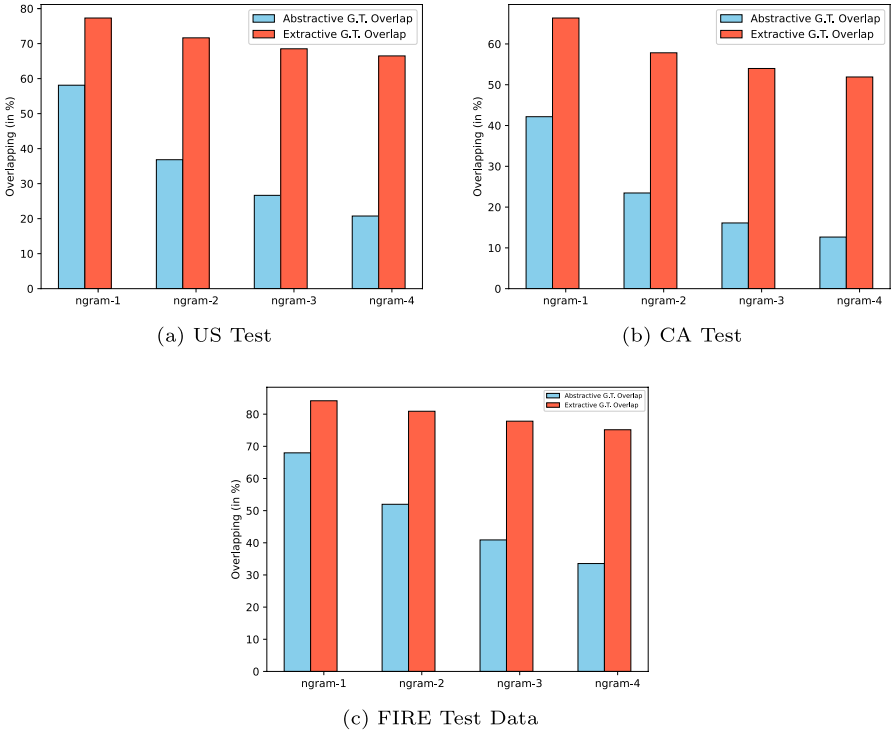



**Fig. 4** Percentage (%) of n-gram overlapping with respect to abstractive and noisy extractive ground truth summaries

between predicted summaries and extractive ground truth summaries as compared to the overlap with respect to the abstractive ground truth summaries. However, it is important to note here that the proposed approach is still able to achieve decent overlap with respect to the abstractive gold summaries, for all the different n-grams under consideration.

### 5.4 Inference time analysis

The average inference time per document for each component of the DCESumm approach with respect to all three of the testing datasets is shown in Table 9. This table shows that apart from the deep clustering score, every component takes a fraction of a second to complete the computation. Even the deep clustering score for huge documents can still be calculated in a matter of a few seconds. Therefore, the total time taken by one test sample, on an average, is significantly less. The inference time doesn't consider the training time required for the Legal-BERT related steps, since at the time of prediction on a new sample these steps are not needed to be carried out. The inference times are calculated only considering the test time

**Table 9** Average Inference time for the proposed approach

| Score fusion components | Inference time (in sec) | | |
|---|---|---|---|
| | US test | CA test | FIRE test |
| Legal BERT embeddings | 0.700 | 1.192 | 10.52 |
| MLP score | 2.670 | 1.731 | 0.110 |
| Deep clustering | 10.55 | 9.143 | 87.53 |
| Summary generation | 0.0005 | 0.0005 | 0.004 |
| Total time | 13.920 | 12.067 | 98.164 |

**Table 10** Comparison of DCESumm with K-means approach for different cluster ratios in terms of ROUGE metrics

| Cluster ratios | Approach | US testing dataset | | | CA testing dataset | | |
|---|---|---|---|---|---|---|---|
| | | ROUGE-1 | ROUGE-2 | ROUGE-L | ROUGE-1 | ROUGE-2 | ROUGE-L |
| 20% | DCESumm | 0.4192 | 0.2417 | 0.3879 | 0.4059 | 0.2224 | 0.3654 |
| 25% | | 0.4194 | 0.2425 | 0.3882 | 0.4058 | 0.2221 | 0.3651 |
| 30% | | 0.4196 | 0.2422 | 0.3883 | 0.4048 | 0.2221 | 0.3645 |
| 35% | | 0.4200 | 0.2428 | 0.3887 | 0.4366 | 0.2389 | 0.3915 |
| 40% | | 0.4198 | 0.2423 | 0.3885 | 0.4059 | 0.2225 | 0.3654 |
| 45% | | 0.4196 | 0.2422 | 0.3883 | 0.4045 | 0.221 | 0.3639 |
| 20% | K-Means | 0.3818 | 0.2009 | 0.3592 | 0.3737 | 0.1895 | 0.3312 |
| 25% | | 0.3901 | 0.2024 | 0.3487 | 0.3707 | 0.2031 | 0.3294 |
| 30% | | 0.3892 | 0.2125 | 0.3588 | 0.3721 | 0.1988 | 0.3315 |
| 35% | | 0.3801 | 0.2025 | 0.3458 | 0.3728 | 0.2098 | 0.332 |
| 40% | | 0.3807 | 0.2130 | 0.3594 | 0.3724 | 0.1794 | 0.3267 |
| 45% | | 0.3798 | 0.2031 | 0.3258 | 0.3627 | 0.1805 | 0.3220 |

steps to show the effectiveness of the proposed approach when a new document is to be summarized. Given extremely lengthy legal documents, if an automatic system provides a summary in a matter of a few seconds, it will be very beneficial for the end-users. Among all the datasets, the average inference time on the FIRE dataset is larger since it contains extremely lengthy documents.

### 5.5 DCESumm vs. K-means approach (with different cluster ratios)

A comparison of DCESumm approach with K-means based approach is shown in Tables 10 and 11, with respect to different cluster ratio percentages. The main goal of such a comparison is to understand that if we replace just the clustering algorithm from the entire summarization process, then what kind of impact on the summarization performance can be expected. From Table 10, we see that the proposed approach outperforms the K-means approach for different cluster ratio percentages in case of the BillSum dataset. This behavior is expected as the representation of samples do not iteratively improve during the working of K-means algorithm.

**Table 11** Comparison of DCESumm approach at different cluster ratios on FIRE testing dataset

| Cluster ratios | Approach | FIRE test dataset | | |
|---|---|---|---|---|
| | | ROUGE-1 | ROUGE-2 | ROUGE-L |
| 20% | DCESumm | 0.4952 | 0.3427 | 0.4724 |
| 25% | | 0.4971 | 0.3447 | 0.4745 |
| 30% | | 0.5054 | 0.3495 | 0.4813 |
| 35% | | 0.5092 | 0.3599 | 0.4878 |
| 40% | | 0.5 | 0.3484 | 0.4778 |
| 45% | | 0.4975 | 0.3446 | 0.475 |
| 20% | K-Means | 0.4042 | 0.2378 | 0.3734 |
| 25% | | 0.4033 | 0.2373 | 0.3728 |
| 30% | | 0.4038 | 0.2387 | 0.3731 |
| 35% | | 0.4064 | 0.2397 | 0.3804 |
| 40% | | 0.4036 | 0.2369 | 0.3727 |
| 45% | | 0.4056 | 0.239 | 0.3751 |

Table 11 shows a similar comparison for the FIRE dataset, where the superiority of the proposed approach can again be observed. From both of these tables we see that, the proposed DCESumm approach achieves the best performance of summarization when we consider the cluster ratio percentage as 35% of the total number of sentences in the document.

## 6 Discussion

An improved sentence scoring technique DCESumm is proposed in this work which can effectively enhance the extractive summarization of legal documents. From experimental results, it is evident that the proposed approach outperforms the SOTA as well as all the baseline approaches with respect to the ROUGE metrics under consideration.

The experimental results suggest that when scores are modified using the deep clustering based approach, it helps in improving the summarization of legal documents as shown in Tables 1-8. We consider ContextEmbed approach as the baseline in which MLP based scores are predicted for each sentence of the document. When we modify these scores with the help of cluster based scores, results are significantly improved. The primary reason for this phenomenon is that, MLP-based scores might not be able to capture the global context of the sentences and clustering based weights applied on these scores can help with that. Sentences that might not be summary relevant can get high scores via MLP based scoring, so there is a need for re-scaling the MLP scores such that the quality of these sentences with respect to the entire document is properly reflected. This is handled by the deep clustering technique where grouping of similar types of sentences is performed to assess the true quality of a sentence. Based on the sentence cluster scores the MLP scores are

modified. In this way, we get the modified scores for each of the sentences inside a document and hence better summarization performance can be achieved.

One of the important aspects to consider while evaluating the proposed approach is to compare the predicted summaries with the extractive ground truth summaries, since the generated summaries are of extractive nature. It is important to know how the predicted summaries fair, when compared against the extractive ground truth summaries as shown in Tables 3, 4, 7 and 8.

We also evaluate the performance of the proposed approach by replacing the deep clustering component with the simple K-means based clustering. From the results presented in Tables 10 and 11, we observe that the deep clustering component is quite important for the entire summarization process as it significantly outperforms the K-means based approach. Moreover, we also find out experimentally that the choice of cluster ratios also plays a key role in fine-tuning the performance of the proposed approach. Considering a cluster ratio of 35% gives the best summarization performance, irrespective of the clustering algorithm applied.

For the purpose of qualitative analysis of the summary predictions, consider the samples present in Tables 12 and 13 in Appendix A. In these tables, we present the predicted and reference summaries for the best and the worst ROUGE score samples for BillSum test dataset. By analyzing these samples, the predictive capabilities of the proposed DCESumm approach can be assessed in a qualitative manner. There are several key observations that can be made considering the high scoring predicted summaries of both the datasets.

- Summaries which have the maximum ROUGE scores show the presence of similar sentences between the reference and the predicted summaries.
- The predicted summaries are quite fluent as well, which shows that the proposed approach effectively improves the summarization of legal documents.
- There is not much difference between the lengths of the high scoring predicted summaries and the respective reference summaries, which result in balanced precision and recall values for the various ROUGE metrics.

Considering the worst predictions, we observe that the reference summaries in such situations are quite small in size. However, due to the fixed length approach of summary generation, the predicted summaries are quite large in case of both the datasets. Due to this, the ROUGE scores obtained for these samples have very high recall with very low precision values. Since the ROUGE metric matches n-grams between reference and predicted summaries, more matching could not occur due to shorter reference summaries in case of the worst predictions. More specifically, we can see that for CA test dataset, the worst prediction is characterized by a reference summary which has a summary-to-document length ratio of 0.006. This ratio is more than 3 standard deviations away from the mean summary-to-document length ratio of 0.15 in the training set. Such kind of discrepancy inevitably causes the model prediction to be poor while using a fixed size summary generation approach. Similar observations can be made for both the US test and FIRE test sets in case of the worst predictions. However, it is important to note here that if we consider the

DCESumm based sentence level scores, we can see from the examples given in Tables 14, 15 and 16 that many of the top scoring sentences have <0.2 scores assigned to them. Even with such low scores, these sentences get included in the predicted summaries when we fix the predicted summary-to-document ratio. This suggests that a proper score-thresholding based post-processing of predicted summaries could potentially result in much higher overall ROUGE scores. For example, if we consider only the top scoring sentences (one or two top scoring sentences) for summary formation as shown in Table 17, the ROUGE scores of the predicted summary improves significantly, as shown in Table 18. This analysis helps us understand that the proposed approach is able to score important sentences effectively. However due to the fixed length summary generation step, some low scoring sentences are also getting included in the predictions, resulting in lower ROUGE scores. As part of the future work, document-specific prediction of ideal summary length can be taken into consideration to overcome this limitation of the proposed approach.

Although decent summarization results could be obtained using the proposed approach, there are still certain areas where more extensive experimental studies can be performed to potentially achieve further improvements. Firstly, the supervised sentence relevance prediction step can be explored in more depth by considering improved BERT variant based sentence representations. In the NLP literature, there have been several research works that have tried to improve the vanilla BERT model, like Big Bird (Zaheer et al. 2020), Longformers (Beltagy et al. 2020), etc. Such approaches can be considered for pretraining in the legal domain, so that even better sentence representations can be obtained. Secondly, the deep clustering technique utilized in this work only considers deep autoencoders for the purpose of abstract representation learning of higher-dimensional sentence embeddings. However, several improved variants of deep autoencoders have been proposed in the literature, like DCEC (Guo et al. 2017), SDEC (Ren et al. 2019), etc., which can potentially improve the sentence clustering quality, thereby resulting in even better summarization performances. Such kind of techniques need to be explored in future research works. Finally, the proposed approach has only been evaluated on US and Indian legal documents, that correspond to legal bills and case judgements. However, in order to accurately assess the predictive capabilities of the proposed approach, it can be evaluated on other country specific datasets as well. This kind of an analysis will give a better understanding of the quality of predictions produced by the proposed DCESumm approach.

## 7 Conclusion

A novel sentence scoring based summarization approach DCESumm is proposed in this work, that combines supervised sentence-level summary relevance prediction with sentence clustering based document-level score enhancement, for the purpose of legal document summarization. Extensive experimental analysis on the publicly available legal bill dataset BillSum and the privately

accessed FIRE dataset reveals that the proposed approach is able to achieve impressive summarization performances. The proposed approach could outperform the current SOTA works by (1-6)% and (6-12)% on the BillSum and FIRE test datasets respectively. A qualitative analysis of the predicted summaries suggest that the proposed sentence scoring approach can identify summary-relevant sentences very effectively, even in case of the samples with the worst ROUGE scores. Such kind of results suggest that DCESumm approach can be very helpful in providing crucial assistance to legal practitioners, thereby enabling fast processing of lengthy legal documents.

As part of the future work, further studies can be performed by considering more advanced sentence embedding approaches, along with improved deep clustering techniques, in order to potentially increase the robustness of the proposed approach.

## Appendix A. Qualitative analysis

### A.1 Best and worst sample predictions

**Table 12** Sample Predicted summary of maximum ROUGE score from CA test

*Actual Summary*: Existing law provides that a pupil is deemed to have complied with the residency requirements for school attendance in a school district if the pupil satisfies one of the specified requirements. Until July 1, 2017, existing law authorizes a school district within the boundaries of which at least one parent or the legal guardian of a pupil is physically employed for a minimum of 10 hours during the school week to allow that pupil to attend a school in that school district through grade 12 if the parent or legal guardian of the pupil so chooses and if the parent or legal guardian of the pupil continues to be physically employed by an employer situated within the attendance boundaries of the school district. This bill would indefinitely extend the operation of the provision authorizing the school district within the boundaries of which a parent or legal guardian of a pupil is physically employed for a minimum of 10 hours during the school week to allow that pupil to attend a school in that school district.

*Predicted Summaries:* (1) A school district may deem a pupil to have complied with the residency requirements for school attendance in the school district if at least one parent or the legal guardian of the pupil is physically employed within the boundaries of that school district for a minimum of 10 hours during the school week. Once a pupil is deemed to have complied with the residency requirements for school attendance pursuant to this subdivision and is enrolled in a school in a school district the boundaries of which include the location where at least one parent or the legal guardian of a pupil is physically employed, the pupil does not have to reapply in the next school year to attend a school within that school district and the governing board of the school district shall allow the pupil to attend school through grade 12 in that school district if the parent or legal guardian so chooses and if at least one parent or the legal guardian of the pupil continues to be physically employed by an employer situated within the attendance boundaries of the school district, subject to paragraphs (2) to (7), inclusive

*ROUGE Score:* ROUGE-1: F1: 0.76294, P: 0.80000, R: 0.72916, ROUGE-2: F1: 0.58082, P: 0.60919, R: 0.55497, ROUGE-L: F1: 0.66666, P: 0.70588, R: 0.631578947368421

**Table 13** Sample Predicted summary of minimum ROUGE score from CA test

*Actual Summary:* This bill would enact the third validating act of 2016, which would validate the organization, boundaries, acts, proceedings, and bonds of the state and counties, cities, and specified districts, agencies, and entities.

*Predicted Summary:*The people of the state of California do enact as follows: This act shall be known, and may be cited, as the third validating act of 2016. As used in this act: “Public Body" means all of the following: the state and all departments, agencies, boards, commissions, and authorities of the state. except as provided in paragraph (2), “Public Body" also means all cities, counties, cities and counties, districts, authorities, agencies, boards, commissions, and other entities, whether created by a general statute or a special act, including, but not limited to, the following: agencies, boards, commissions, or entities constituted or provided for under or pursuant to the joint exercise of powers act (Chapter 5. community development commissions in their capacity to act as a housing authority for other community development purposes of the jurisdiction in which the commission operates, except for any action taken with respect to the commission’s authority to act as a community redevelopment agency. notwithstanding paragraph (1), a “Public Body” does not include any of the following: a community redevelopment agency formed pursuant to the community redevelopment law (part 1. a joint powers authority that includes a community redevelopment agency or a community development commission as a member, with respect to its exercise of the powers of a community redevelopment agency. “Public Body" includes both of the following: the successor agency to the redevelopment agency of the city and county of San Francisco, solely for the purpose of issuing bonds or incurring other indebtedness pursuant to the provisions of section 34177.7 of the health and safety code. A successor agency, as defined in subdivision (j) of Section 34171 of the health and safety code, solely for the purpose of issuing bonds or incurring other indebtedness pursuant to the provisions of section 34177. 5 of the health and safety code. “bonds" means all instruments evidencing an indebtedness of a public body incurred or to be incurred for any public purpose, all leases, installment purchase agreements, or similar agreements wherein the obligor is one or more public bodies, all instruments evidencing the borrowing of money in anticipation of taxes, revenues, or other income of that body, all instruments payable from revenues or special funds of those public bodies, all certificates of participation evidencing interests in the leases, installment purchase agreements, or similar agreements, and all instruments funding, refunding, replacing, or amending any thereof or any indebtedness. All public bodies heretofore organized or existing under any law, or under color of any law, are hereby declared to have been legally organized and to be legally functioning as those public bodies. All acts and proceedings heretofore taken by any public body or bodies under any law, or under color of any law, for the annexation or inclusion of territory into those public bodies or for the annexation of those public bodies to any other public body or for the detachment, withdrawal, or exclusion of territory from any public body or for the consolidation, merger, or dissolution of any public bodies are hereby confirmed, validated, and declared legally effective. This shall include all acts and proceedings of the governing board of any public body and of any person, public officer, board, or agency heretofore done or taken upon the question of the annexation or inclusion or of the withdrawal or exclusion of territory or the consolidation, merger, or dissolution of those public bodies. All acts and proceedings heretofore taken by or on behalf of any public body under any law, or under color of any law, for, or in connection with, the authorization, issuance, sale, execution, delivery, or exchange of bonds of any public body for any public purpose are hereby authorized, confirmed, validated, and declared legally effective. This shall include all acts and proceedings of the governing board of public bodies and of any person, public officer, board, or agency heretofore done or taken upon the question of the authorization, issuance, sale, execution, delivery, or exchange of bonds. All bonds of, or relating to, any public body heretofore awarded and sold to a purchaser and hereafter issued and delivered in accordance with the contract of sale and other proceedings for the award and sale shall be the legal, valid, and binding obligations of the public body. all bonds of, or relating to, any public body heretofore authorized to be issued by ordinance, resolution, order, or other action adopted or taken by or on behalf of the public body and hereafter issued and delivered in accordance with that authorization shall be the legal, valid, and binding obligations of the public body. All bonds of, or relating to, any public body heretofore authorized to be issued at an election and hereafter issued and delivered in accordance with that authorization shall be the legal, valid, and binding obligations of the public body. Whenever an election has heretofore been called for the purpose of submitting to the voters of any public body the question of issuing bonds for any public purpose, those bonds, if hereafter authorized by the required vote and in accordance with the proceedings heretofore taken, and issued and delivered in accordance with that authorization, shall be the legal, valid, and binding obligations of the public body. This act shall operate to supply legislative authorization as may be necessary to authorize, confirm, and validate any acts and proceedings heretofore taken pursuant to authority the legislature could have supplied or provided for in the law under which those acts or proceedings were taken. This act shall be limited to the validation of acts and proceedings to the extent that the same can be effectuated under the California constitution and the united states constitution. This act shall not operate to authorize, confirm, validate, or legalize any act, proceeding, or other matter being legally contested or inquired into in any legal proceeding now pending and undetermined or that is pending and undetermined during the period of 30 days from and after the effective date of this act. Any action or proceeding contesting the validity of any action or proceeding heretofore taken under any law, or under color of any law, for the formation, organization, or incorporation of any public body, or for any annexation thereto, detachment or exclusion therefrom, or other change of boundaries thereof, or for the consolidation, merger, or dissolution of any public bodies, or for, or in connection with, the authorization, issuance, sale, execution, delivery, or exchange of bonds thereof upon any ground involving any alleged defect or illegality not effectively validated by the prior provisions of this act and not otherwise barred by any statute of limitations or by laches shall be commenced within six months of the effective date of this act, otherwise each and all of those matters shall be held to be valid and in every respect legal and incontestable. This act shall not extend the period allowed for legal action beyond the period that it would be barred by any presently existing valid statute of limitations. nothing contained in this act shall be construed to render the creation of any public body, or any change in the boundaries of any public body, effective for purposes of assessment or taxation unless the statement, together with the map or plat, required to be filed pursuant to Chapter 8 of part 1 of division 2 of title 5 of the government code, is filed within the time and substantially in the manner required by those sections.

*ROUGE Score:* Rouge-1: F: 0.03943, ’p’: 0.78125, R: 0.02022, Rouge-2: F: 0.01422, P: 0.29032, R: 0.00728, Rouge-L: F: 0.09066, P: 0.68000, R: 0.04857

### A.2 Sample predicted summaries after postprocessing

Table 14 shows the scores obtained by the sample with the lowest ROUGE score among all the samples in the case of US Test data. More specifically, it shows the scores of top 15% sentences. These sentences are then sorted as they appear in the

**Table 14** Top 15% sentences from US Test data which has obtained the lowest ROUGE scores

| Sentence index | Sentence scores |
|---|---|
| 30 | 0.5788204073905945 |
| 6 | 0.5248245000839233 |
| 28 | 0.5113955736160278 |
| 9 | 0.4997960329055786 |
| 14 | 0.4372066259384155 |
| 2 | 0.4351777136325836 |
| 18 | 0.2944013774394989 |
| 24 | 0.24545374512672424 |
| 64 | 0.21612945199012756 |
| 34 | 0.20679223537445068 |

**Table 15** Top 15% sentences from CA Test data which has obtained the lowest ROUGE scores

| Sentence index | Sentence scores |
|---|---|
| 130 | 0.41866530118305256 |
| 145 | 0.41664032951816665 |
| 11 | 0.3466029355127631) |
| 127 | 0.32600701946529) |
| 137 | 0.3163495391441158) |
| 128 | 0.3000419411115314) |
| 142 | 0.29110082306498963) |
| 153 | 0.26672881766895845) |
| 148 | 0.2630357916268786) |
| 144 | 0.2467627939909227) |
| 129 | 0.24653769484027066) |
| 139 | 0.2448077948407117) |
| 0 | 0.2315547028059064) |
| 138 | 0.2300514103230773) |
| 1 | 0.20679222910059325 |
| 134 | 0.20679222910059325 |
| 140 | 0.20679222910059325 |
| 143 | 0.20679222910059325 |
| 146 | 0.20679222910059325 |
| 151 | 0.20679222910059325 |
| 150 | 0.17937822413300442 |
| 149 | 0.1734691510502273 |
| 147 | 0.1668768800452423 |

**Table 16** Top 40% sentences from FIRE Test data which has obtained the lowest ROUGE scores

| Sentence index | Sentence scores |
|---|---|
| 60 | 0.4948425590991974 |
| 64 | 0.48566123843193054 |
| 12 | 0.1370852142572403 |
| 59 | 0.04879920184612274 |
| 61 | 0.010069362819194794 |
| 74 | 0.004273678176105022 |
| 66 | 0.002872639801353216 |
| 0 | 0.0028726381715387106 |
| 1 | 0.0028726381715387106 |
| 2 | 0.0028726381715387106 |
| 3 | 0.0028726381715387106 |
| 4 | 0.0028726381715387106 |
| 5 | 0.0028726381715387106 |
| 6 | 0.0028726381715387106 |
| 7 | 0.0028726381715387106 |
| 8 | 0.0028726381715387106 |
| 9 | 0.0028726381715387106 |
| 10 | 0.0028726381715387106 |
| 11 | 0.0028726381715387106 |
| 13 | 0.0028726381715387106 |
| 14 | 0.0028726381715387106 |
| 15 | 0.0028726381715387106 |
| 16 | 0.0028726381715387106 |
| 17 | 0.0028726381715387106 |
| 18 | 0.0028726381715387106 |
| 19 | 0.0028726381715387106 |
| 20 | 0.0028726381715387106 |
| 21 | 0.0028726381715387106 |
| 22 | 0.0028726381715387106 |
| 23 | 0.0028726381715387106 |

original document to form a summary. From these scores, we see that not all the scores are good to be included into the summary. Similar trend has been shown in the case of US Test and CA Test dataset as shown in Tables 15 and 16. Table 17 shows the predicted summaries of worst sample after postprocessing step. This shows that, postprocessing step can actually be very helpful to improve the quality of the predicted summaries further. Table 18 shows the ROUGE scores on those samples which has obtained the minimum ROUGE scores. From this table, we see that after postprocessing which includes picking the one ore two top scoring sentences helps in improving the quality of summary and hence ROUGE scores.

**Table 17** Postprocessing Summary sample for the worst Summary

*Postprocessing CA Predicted Summary:*"Bonds" means all instruments evidencing an indebtedness of a public body incurred or to be incurred for any public purpose, all leases, installment purchase agreements, or similar agreements wherein the obligor is one or more public bodies, all instruments evidencing the borrowing of money in anticipation of taxes, revenues, or other income of that body, all instruments payable from revenues or special funds of those public bodies, all certificates of participation evidencing interests in the leases, installment purchase agreements, or similar agreements, and all instruments funding, refunding, replacing, or amending any thereof or any indebtedness. whenever an election has heretofore been called for the purpose of submitting to the voters of any public body the question of issuing bonds for any public purpose, those bonds, if hereafter authorized by the required vote and in accordance with the proceedings heretofore taken, and issued and delivered in accordance with that authorization, shall be the legal, valid, and binding obligations of the public body.

*Postprocessing US Predicted Summary:* Section 8440b of title 5, United States code, is amended in subsection (b)(4) by amending subparagraph (b) to read as follows: Section 8433(b) of this title applies to any bankruptcy judge or magistrate who elects to make contributions to the thrift savings fund under subsection (a) of this section and who retires before attaining age 65 but is entitled, upon attaining age 65, to an annuity under section 377 of title 28 or Section 2(c) of the retirement and survivors annuities for bankruptcy judges and magistrates act of 1988. By amending paragraph (6) (as redesignated under paragraph to read as follows: notwithstanding paragraph (4), if an employee or member separates from government employment and such employee's or member's nonforfeitable account balance is $3,500 or less, the executive director shall pay the nonforfeitable account balance to the participant in a single payment unless the employee or member elects, at such time and otherwise in such manner as the executive director prescribes, one of the options available under Section 8433(b) of this title.

*Postprocessing FIRE Predicted Summary:* This statement of the law, with which we agree, may be supplemented by three other well-settled principles, these being firstly, that the illegitimate son does not acquire by birth any interest in his father's estate and he cannot therefore demand partition against his father during the latter's lifetime; secondly, that on his father's death, the illegitimate son succeeds as a coparcener to the separate estate of the father along with the legitimate son(s) with a right of survivorship and is entitled to enforce partition against the legitimate son(s); and thirdly, that on a partition between a legitimate and an illegitimate son, the illegitimate son takes only one-half of what he would have taken if he was a legitimate son. there can be no doubt that though the illegitimate son cannot enforce partition during the father's lifetime and though he is not entitled to demand partition where the father has left no separate property and no legitimate son but was joint with his collaterals, he can enforce partition in a case like the present, where the father was separate from his collaterals and has left separate property and legitimate sons.

**Table 18** ROUGE scores with the proposed approach on the worst sample after postprocessing

| Dataset | Proposed approach | | | After postprocessing | | |
|---|---|---|---|---|---|---|
| | ROUGE-1 | ROUGE-2 | ROUGE-L | ROUGE-1 | ROUGE-2 | ROUGE-L |
| US Test | 0.01669 | 0.0056 | 0.0452 | 0.0847 | 0.0236 | 0.0847 |
| CA Test | 0.0394 | 0.0142 | 0.0907 | 0.0714 | 0.0116 | 0.0535 |
| FIRE Test | 0.0954 | 0.0532 | 0.1119 | 0.4307 | 0.2156 | 0.3692 |

## References


Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, . . . others (2016). Tensorflow: A system for large-scale machine learning. 12th USENIX symposium on operating systems design and implementation (OSDI 16) (pp. 265-283)

Acharya A, Goel R, Metallinou A, Dhillon I (2019). Online embedding compression for text classification using low rank matrix factorization. Proceedings of the aaai conference on artificial intelligence (Vol. 33, pp. 6196-6203)

Akter S, Asa AS, Uddin MP, Hossain MD, Roy SK, Afjal MI (2017). An extractive text summarization technique for bengali document (s) using k-means clustering algorithm. 2017 ieee international conference on imaging, vision & pattern recognition (icivpr) (pp. 1-6)

Alguliyev RM, Aliguliyev RM, Isazade NR, Abdi A, Idris N (2019) Cosum: text summarization based on clustering and optimization. Expert Syst 36(1):e12340

Alqaisi R, Ghanem W, Qaroush A (2020) Extractive multi-document arabic text summarization using evolutionary multi-objective optimization with k-medoid clustering. IEEE Access 8:228206–228224

Anand D, Wagh R (2019) Effective deep learning approaches for summarization of legal texts. J King Saud University-Computer Inf Sci 2:51

Beltagy, I., Peters, M.E., Cohan, A. (2020). Longformer: the long-document transformer. http://arxiv.org/abs/2004.05150

Bhattacharya, P., Hiware, K., Rajgaria, S., Pochhi, N., Ghosh, K., Ghosh, S. (2019). A comparative study of summarization algorithms applied to legal case judgments. European conference on information retrieval (pp. 413-428)

Bhattacharya, P., Paul, S., Ghosh, K., Ghosh, S., Wyner, A. (2019). Identification of rhetorical roles of sentences in indian legal judgments. http://arxiv.org/abs/1911.05405

Bhattacharya, P., Poddar, S., Rudra, K., Ghosh, K., Ghosh, S. (2021). Incorporating domain knowledge for extractive summarization of legal case documents. Proceedings of the eighteenth international conference on artificial intelligence and law (pp. 22-31)

Bonhard P, Sasse MA (2006) knowing me, knowing you-using profiles and social networking to improve recommender systems. BT Technol J 24(3):84–98

Carmel, D., Zwerdling, N., Guy, I., Ofek-Koifman, S., Har'El, N., Ronen, I., . . . Chernov, S. (2009). Personalized social search based on the user's social network. Proceedings of the 18th acm conference on information and knowledge management (pp. 1227-1236)

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., Androutsopoulos, I. (2020). Legal-bert: The muppets straight out of law school. http://arxiv.org/abs/2010.02559

Clarke J, Lapata M (2008) Global inference for sentence compression: an integer linear programming approach. J Artif Intell Res 31:399–429

Cohan, A., Beltagy, I., King, D., Dalvi, B., Weld, D.S. (2019). Pretrained language models for sequential sentence classification. http://arxiv.org/abs/1909.04054

Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. http://arxiv.org/abs/1810.04805

Duan X, Zhang Y, Yuan L, Zhou X, Liu X, Wang T, Wu F (2019) Legal summarization for multi-role debate dialogue via controversy focus mining and multi-task learning. Proceedings of the 28th acm international conference on information and knowledge management (pp. 1361-1370)

Edmundson HP (1969) New methods in automatic extracting. J ACM 16(2):264–285

Eidelman V (2019) Billsum: a corpus for automatic summarization of us legislation. Proceedings of the 2nd workshop on new frontiers in summarization (pp. 48-56)

Erkan G, Radev DR (2004) Lexrank: graph-based lexical centrality as salience in text summarization. J Artif Intell Res 22:457–479

Guo X, Liu X, Zhu E, Yin J (2017) Deep clustering with convolutional autoencoders. International conference on neural information processing (pp. 373-382)

Gupta S, Narayana N, Charan VS, Reddy KB, Borah MD, Jain D (2022) Extractive summarization of indian legal documents. Edge analytics (pp. 629-638). Springer

Hachey B & Grover C (2004) A rhetorical status classifier for legal text summarisation. Text summarization branches out (pp. 35-42)

Haghighi A, & Vanderwende L (2009) Exploring content models for multi-document summarization. Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics (pp. 362-370)

Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Oliphant TE (2020) Array Programming with NumPy. Nature 585(7825):357–362
Honnibal M, Montani I, Van Landeghem S, Boyd A (2020) spaCy: Industrial-strength Natural Language Processing in Python. Zenodo
Huang L, Cao S, Parulian N, Ji H, Wang L (2021) Efficient attentions for long document summarization. Proceedings of the 2021 conference of the north American chapter of the association for computational linguistics: Human language technologies (pp. 1419-1436)
Jain D, Borah MD, Biswas A (2020) Fine-tuning textrank for legal document summarization: A bayesian optimization based approach. In: Forum for information retrieval evaluation (pp. 41–48)
Jain D, Borah MD, Biswas A (2021a) Automatic summarization of legal bills: A comparative analysis of classical extractive approaches. In: 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) (pp. 394–400)
Jain D, Borah MD, Biswas A (2021b) Cawesumm: A contextual and anonymous walk embedding based extractive summarization of legal bills. In: Proceedings of the 18th International Conference on Natural Language Processing (ICON) (pp. 414–422)
Jain D, Borah MD, Biswas A (2021c) Summarization of indian legal judgement documents via ensembling of contextual embedding based mlp models. FIRE
Jain D, Borah MD, Biswas A (2021d) Summarization of legal documents: Where are we now and the way forward. Computer Sci Rev 40:100388
Jing H (2000) Sentence reduction for automatic text summarization. Sixth applied natural language processing conference (pp. 310-315)
Kanapala A, Jannu S, Pamula R (2019) Summarization of legal judgments using gravitational search algorithm. Neural Comput Appl 31(12):8631–8639
Kanapala A, Pal S, Pamula R (2019) Text summarization from legal documents: a survey. Artif Intell Rev 51(3):371–402
Kingma DP, & Ba J (2014) Adam: A method for stochastic optimization. http://arxiv.org/abs/1412.6980
Lin C-Y (2004) Rouge: A package for automatic evaluation of summaries acl. Proceedings of workshop on text summarization branches out post conference workshop of acl (pp. 2017-05)
Louis A, Joshi AK, Nenkova A (2010) Discourse indicators for content selection in summaization
Luhn HP (1958) The automatic creation of literature abstracts. IBM J Res Develop 2(2):159–165
Ma T, & Nakagawa H (2013) Automatically determining a proper length for multi-document summarization: A bayesian nonparametric approach. Proceedings of the 2013 conference on empirical methods in natural language processing (pp. 736-746)
Mallick C, Das AK, Ding W, Nayak J (2021) Ensemble summarization of bio-medical articles integrating clustering and multi-objective evolutionary algorithms. Appl Soft Comput 106:107347
Mihalcea R, Tarau P (2004) Textrank: Bringing order into text. Proceedings of the 2004 conference on empirical methods in natural language processing (pp. 404-411)
Mishra SK, Saini N, Saha S, Bhattacharyya P (2022) Scientific document summarization in multi-objective clustering framework. Appl Intell 52(2):1520–1543
Moradi M, & Samwald M (2019) Clustering of deep contextualized representations for summarization of biomedical texts. http://arxiv.org/abs/1908.02286
Nallapati R, Zhai F, Zhou B (2017) Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. Thirty-first aaai conference on artificial intelligence
Nenkova A, & Vanderwende L (2005) The impact of frequency on summarization. Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005 , 101
Parikh V, Bhattacharya U, Mehta P, Bandyopadhyay A, Bhattacharya P, Ghosh K, Majumder P (2021a) Fire 2021 aila track: Artificial intelligence for legal assistance. Proceedings of the 13th forum for information retrieval evaluation
Parikh V, Bhattacharya U, Mehta P, Bandyopadhyay A, Bhattacharya P, Ghosh K, Majumder P (2021b, December) Overview of the third shared task on artificial intelligence for legal assistance at fire 2021. Fire (working notes)
Parikh V, Mathur V, Mehta P, Mittal N, Majumder P (2021) Lawsum: A weakly supervised approach for indian legal document summarization. http://arxiv.org/abs/2110.01188v3
Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Duchesnay E (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
Polsley S, Jhunjhunwala P, Huang R (2016) Casesummarizer: a system for automated summarization of legal texts. Proceedings of coling 2016, the 26th international conference on computational linguistics: System demonstrations (pp. 258-262)

Rehurek R, & Sojka P (2010) Software framework for topic modelling with large corpora. In proceedings of the lrec 2010 workshop on new challenges for nlp frameworks
Ren Y, Hu K, Dai X, Pan L, Hoi SC, Xu Z (2019) Semi-supervised deep embedded clustering. Neurocomputing 325:121–130
Saini N, Saha S, Chakraborty D, Bhattacharyya P (2019) Extractive single document summarization using binary differential evolution: Optimization of different sentence quality measures. PloS One 14(11):e0223477
Saravanan M, Ravindran B, Raman S (2006) Improving legal document summarization using graphical models. Front Artif Intell Appl 152:51
Shetty K, & Kallimani JS (2017) Automatic extractive text summarization using k-means clustering. 2017 international conference on electrical, electronics, communication, computer, and optimization techniques (iceeccot) (pp. 1-9)
Srikanth A, Umasankar AS, Thanu S, Nirmala SJ (2020) Extractive text summarization using dynamic clustering and co-reference on bert. 2020 5th international conference on computing, communication and security (icccs) (pp. 1-5)
Steinberger J, Jezek K et al (2004) Using latent semantic analysis in text summarization and summary evaluation. Proc ISIM 4:93–100
Tajaddodianfar F, Stokes JW, Gururajan A (2020) Texception: a character/word-level deep learning model for phishing url detection. Icassp 2020-2020 ieee international conference on acoustics, speech and signal processing (icassp) (pp. 2857-2861)
Umer M, Ashraf I, Mehmood A, Kumari S, Ullah S, Sang Choi G (2021) Sentiment analysis of tweets using a unified convolutional neural network-long short-term memory network model. Comput Intell 37(1):409–434
Vanderwende L, Suzuki H, Brockett C, Nenkova A (2007) Beyond sumbasic: task-focused summarization with sentence simplification and lexical expansion. Inf Process Manage 43(6):1606–1618
Verma S, & Nidhi V (2017) Extractive summarization using deep learning. http://arxiv.org/abs/1708.04439
Wang D, Zhu S, Li T, Chi Y, Gong Y (2011) Integrating document clustering and multidocument summarization. ACM Trans Knowl Discov Data (TKDD) 5(3):1–26
Xiao W, & Carenini G (2019) Extractive summarization of long documents by combining global and local context. http://arxiv.org/abs/1909.08089
Xie J, Girshick R, Farhadi A (2016) Unsupervised deep embedding for clustering analysis. International conference on machine learning (pp. 478-487)
Zaheer M, Guruganesh G, Dubey KA, Ainslie J, Alberti C, Ontanon S et al (2020) Big bird: transformers for longer sequences. Adv Neural Inf Process Syst 33:17283–17297
Zhang J, Zhao Y, Saleh M, Liu P (2020) Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. International conference on machine learning (pp. 11328-11339)