Legal Holding Extraction from Italian Case Documents using Italian-LEGAL-BERT Text Summarization

Daniele Licari*
Praveen Bushipaka*
daniele.licari@santannapisa.it
praveen.bushipaka@santannapisa.it
Scuola Superiore Sant'Anna
Pisa, Italy

Gabriele Marino
gabriele.marino@santannapisa.it
Scuola Superiore Sant'Anna
Pisa, Italy

Giovanni Comandé
giovanni.comande@santannapisa.it
Scuola Superiore Sant'Anna
Pisa, Italy

Tommaso Cucinotta
tommaso.cucinotta@santannapisa.it
Scuola Superiore Sant'Anna
Pisa, Italy

ABSTRACT

Legal holdings are used in Italy as a critical component of the legal system, serving to establish legal precedents, provide guidance for future legal decisions, and ensure consistency and predictability in the interpretation and application of the law. They are written by domain experts who describe in a clear and concise manner the principle of law applied in the judgments.

We introduce a legal holding extraction method based on Italian-LEGAL-BERT to automatically extract legal holdings from Italian cases. In addition, we present ITA-CaseHold, a benchmark dataset for Italian legal summarization. We conducted several experiments using this dataset, as a valuable baseline for future research on this topic.

CCS CONCEPTS

• **Applied computing** → **Law**; • **Information systems** → **Summarization**; • **Computing methodologies** → **Information extraction**.

KEYWORDS

Holding Extraction, Italian-LEGAL-BERT, Extractive Text Summarization, Benchmark Dataset

ACM Reference Format:
Daniele Licari, Praveen Bushipaka, Gabriele Marino, Giovanni Comandé, and Tommaso Cucinotta. 2023. Legal Holding Extraction from Italian Case Documents using Italian-LEGAL-BERT Text Summarization. In *Nineteenth International Conference on Artificial Intelligence and Law (ICAIL 2023), June 19–23, 2023, Braga, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3594536.3595177

*Both authors contributed equally to this research.

ICAIL 2023, June 19–23, 2023, Braga, Portugal

ACM ISBN 979-8-4007-0197-9/23/06…$15.00
https://doi.org/10.1145/3594536.3595177

1 INTRODUCTION

Legal holdings are considered the most essential part of a legal decision because they summarize it without going into the merits of the specific case, establish a legal principle and set a legal precedent (which is binding in some legal systems). Judgments are often cited in future cases as authoritative evidence of the law and serve as the basis for future legal arguments. Furthermore, lawyers and judges themselves rely on the holdings to select the relevant documents for the case at stake when exploring datasets of previous cases.

In Italy, as in many other legal systems, they play a crucial role in the development and interpretation of the law by providing a clear and authoritative statement of the court's decision on a particular issue, which helps to ensure consistency and predictability in the application of the law. It goes without saying that they help serving the rule of law ensuring an equal treatment to similar cases and a different treatment to different ones. The holdings writing is carried out by legal experts who, starting from a judgment, set out the applied principle of law in a clear, precise, and concise manner. Extracting holdings is a lengthy and expensive process that requires a high degree of specific expertise and extreme care, as any error, emphasis, inaccuracy, underestimation or omission can affect subsequent cases.

We approached the problem of extracting legal holdings as a text summarization task. There are different classes of text summarization algorithms. One of them is extractive summarization, which involves identifying the most important sentences or passages from the original text and combining them to create the summary [14]. Another one is abstractive summarization, which instead involves generating a summary from scratch using generative language models. While the latter method may be more flexible, it is not a perfect fit for our use case, as it may generate text that is not present in the original document [14, 32]. We then focused on the extractive method (i) as highlighting the most relevant sentences in a given document is a more effective way to support and speed up the judges' work and (ii) because summarizing long documents is often an extractive job in its very nature [23]. Extractive methods can indeed take advantage of the discourse structure [17] to generate factually consistent summaries, thus preserving the meaning of the original document [15], and allowing to identify and analyze the

most important sentences in their specific context. Highlighted information can then be reorganized by the judge for the final writing of the legal holding.

In this work, we propose an automatic holding extraction model, based on the use of Italian-LEGAL-BERT [26], to support judges in this delicate activity, trying to save their time and reduce errors. Our approach provides a concise and accurate summary of the key points of a legal decision, making it easier for legal practitioners to quickly understand and analyze the case. This can be particularly useful in a large corpus of legal documents, where reading and summarizing each case would be really time-consuming and error-prone.

Our model was trained on the new ITA-CaseHold dataset, which we built and released to foster the development and improvement of legal NLP applications for the Italian language. It includes more than 1100 case-holdings pairs from publicly available Italian administrative cases. The legal cases concern disputes between citizens and the Italian government. The holdings have been extracted by members of the State Council, which is the highest administrative court in Italy.

To summarize, the main contributions of this paper are: (i) the introduction of the first LEGAL-BERT-based system for legal holdings extraction; (ii) the release of ITA-CaseHold, the first benchmark dataset for text summarization in the Italian legal context; (iii) the proposal of a simple but effective methodology to identify the most relevant sentences of a document, based on the computation of the harmonic mean of ROUGE R-1 and R-2 scores between the document sentences and their holding. Our code and dataset are available in github.[1]

2 RELATED WORKS

In this section we overview the most relevant prior research works in legal summarization, focusing on a) the use of extractive summarization techniques in law, and b) the already existing legal summarization datasets.

2.1 Legal Extractive Summarization

Extractive summarization involves selecting the most relevant sentences or sections of a document as a summary. The most popular methods are often graph-based and tend to be domain and language-agnostic, which means that no training is required. LexRank [18], and TextRank [33] are some of the most common graph-based extractive methods. These algorithms generate summaries that are factually consistent with the source document [15] but usually contain redundant information [20].

Difficulties in dealing with legal documents are due to their length and multiple documents: current benchmarks of long documents count up to 3.000 tokens [23, 31, 43] while an average legal document consists of more than 4.500 tokens [6, 35, 36]. ITA-CaseHold documents consist on average of 4.700 tokens, and their holdings of 800 tokens. Legal research aims at deploying tools that can quickly summarize such long texts by extracting their key points, saving time and effort. For this purpose, several domain-specific legal summarization methods have been proposed.

In 2014, Farzindar proposed LetSum [19], which assigns rhetorical roles to sentences and uses TF-IDF to rank them. A fixed percentage of sentences for each rhetorical role is then selected to be part of the summary, according to the given ranking. In 2016, CaseSummarizer [37] system was published to incorporate TF-IDF with legal-specific features like the number of legal entities present in each sentence. Liu et al. [28] used machine-learning methods (MLPs, GBD trees, and LSTMs) to rank sentences according to their likelihood to be included in the summary. Zhong et al. [46] used an iterative selection of predictive sentences incorporating a CNN and a RF to distinguish reasoning and evidential support sentences from others. Finally, DELSumm [7] selects the summary sentences by ingesting the informativeness of sentences and context words from the legal domain knowledge into an objective function to be maximized with ILP, suggesting the importance of effective domain knowledge incorporation.

Since the advent of the era of self-attention-based transformers [41], approaches based on models like BERT [16] and RoBERTa [47] have outperformed previous models in most scenarios, deploying pre-trained large models (PLM) on domain-specific corpora (BioBERT [25], SciBERT [4], FinBERT [2]). In the legal context, Chalkidis et al. [12] proposed LegalBERT, a BERT-based model trained on UK and EU legislation and US and EU court documents. Zheng et al. developed CaseLaw BERT [45], another BERT-based PLM, trained on US legal documents. Hendersen et al. introduced Pile of Law [22], a BERT large model trained on a huge dataset of EU and US legal documents. With regard to the Italian legal domain, Licari et al. proposed Italian-LEGAL-BERT [26], training the XXL Italian BERT-base model[2] on large Italian civil law corpora and its pre-trained variant from scratch on Italian legal documents[3] based on the CamemBERT architecture. Similarly, Tagarelli et al. [40] fine-tuned a BERT model using the Italian civil code.

Miller [34] summarized class lectures by selecting the BERT-embedded sentences closest to the centroids in a K-means clustering. Similarly, PacSum [44] revises popular graph-based algorithms using BERT to compute the similarity between sentences and incorporates the relative position of sentences in the weights of the graph. Liu et al.'s BERTSum [29] effectively employs BERT for abstractive and extractive summarization and is improved by Yuan et al. [42], by the addition of a hierarchical graph mask to incorporate structure constraints. Agarwal et al. [1] proposed a hierarchical multi-task learning approach leveraging rhetorical role labeling to improve the summarizer, using SBERT, BiGRU, and Maximal Marginal Relevance (MMR) [8] to embed, rank and select sentences.

2.2 Legal Datasets for Text Summarization

Albeit many freely available datasets exist for classification tasks [10, 11, 13, 36], the only datasets addressing the legal text summarization problem, to the best of our knowledge, are EUR-Lex-Sum [3], BillSum [24] and those from Shukla et al. [39]. EUR-Lex-Sum by Aumiller et al. [3], is a multi- and cross-lingual dataset containing 1.500 EU legislation documents in 24 different European languages, including Italian. They also experimented with cross-lingual summarization in Legal domain. BillSum [24] consists of 22.000 pairs of

[1] Github: https://github.com/dlicari/ITA-CASEHOLD

[2] publicly available on huggingface.co/dbmdz/bert-base-italian-xxl-cased

[3] publicly available on huggingface.co/dlicari/Italian-Legal-BERT

US congressional bills and summaries. Shukla et al. [39] provided three legal text summarization datasets addressing both extractive and abstractive summarization for UK and Indian court cases. Still, when it comes to holdings, the only relevant dataset to our knowledge is CaseHOLD [45], which consists of 53.000 multiple-choice questions about holdings of US court cases from the Harvard Law Library case law corpus, but it is available only in English. Being unable to find any dataset or experiments done on Italian legal holdings, we conclude that ITA-CaseHold, presented below, is the first dataset for this use case.

3 THE ITA-CASEHOLD DATASET

We present the ITA-CaseHold dataset as the first benchmark Italian legal holdings dataset. It consists of 1101 pairs of judgments and holdings between the years 2019 and 2022 from the archives of Italian Administrative Justice[4]. The Administrative Justice system in Italy covers a wide range of issues, including public contracts, environmental protection, public services, immigration, taxes, and compensation for damages caused by the State. It also provides citizens with the opportunity to challenge administrative decisions in an independent and impartial trial. The most relevant judgments are analyzed by the State Council, which extracts their legal holdings to create a legal precedent that can be easily searched for and cited. Figure 1 shows the distribution of legal entities on the ITA-CaseHold dataset.

3.1 Data Filtering

Our initial scraping yielded 1326 different judgment-holding pairs. To filter out the documents with incomplete and unrepresentative holdings, we derived the compression ratio for each pair of the dataset. This is the word count ratio between documents and holdings [21]. A higher compression ratio means documents are longer and their respective holdings are shorter. A manual check of documents and holdings was done from the 75th percentile to the 100th percentile. Documents above the 90th percentile do not have complete holdings, so we chose to remove them completely. Between the 75th and 90th, we picked the documents which have complete holdings Any redundant pair was also removed, leaving a final dataset of 1101 samples. The dataset consists of URL (link to the document and holdings), documents, holdings, and their legal subject. These were finally split into 80% training (10% for validation) and 20% test sets. The split was stratified according to the legal subjects of the documents to have a uniform distribution between train, test, and validation sets.

3.2 Descriptive Statistics

ITA-CaseHold consists of 1101 documents, including 792 in the training set (73631 non-blank sentences), 88 in the validation set (7716 non-blank sentences), and 221 in the test set (19865 non-blank sentences). The documents and holdings were tokenized using NLTK Italian tokenizer [30] to derive the statistics shown in Table 1. The compression ratio is the token count ratio between a document and its holding. We observe a high standard deviation across all datasets both with respect to documents and holdings length. Each dataset is quite skewed to the left, but the mean compression ratio and those corresponding to the highlighted percentiles are fairly useful. From 1, we observed that the Administration process and COVID-19 legal subjects contribute the highest, the reason for it is because the dataset was between the years 2019-2022.

[4]Data-source: https://www.giustizia-amministrativa.it/web/guest/focus-giurisprudenza-e-pareri

4 METHODOLOGY

Our approach is based on fine-tuning the Italian-BERT, Italian-LEGAL-BERT, and Italian-LEGAL-BERT-SC models to predict the most relevant sentences in a document. The scores were generated by Rouge R-1 and R-2. ROUGE[27] scores are commonly used metrics for evaluating the quality of automatic summaries. The Rouge scores measure the degree of overlap between two summaries in terms of shared n-grams, which provides a measure of how similar they are. It considers both Precision and Recall and the F-1 score are calculated based on this. R-1 Precision and Recall compare the similarity of uni-grams between reference and candidate summaries and R-2 Precision and Recall compare the similarity of bi-grams (2 consecutive words) between reference and candidate summaries. The formula of Rouge N, where N is the number of grams and S is a set of references is:

$$\text{ROUGE-N} = \frac{\sum_{S \in ref} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in ref} \sum_{gram_n \in S} Count(gram_n)}$$

In this experiment, we calculated Rouge R-1 score and Rouge R-2 score. After calculating Rouge R-1 and R-2, the score for each sentence is given by the harmonic mean of ROUGE R-1 and ROUGE R-2 scores between the sentence and the corresponding document holding. We called this model Extractive Harmonic Mean-BERT (HM-BERT). The idea behind using the harmonic mean is to reduce the number of false positives in comparison with the arithmetic mean. The formula of the harmonic mean is

$$HM = 2(R1 * R2)/(R1 + R2) \quad (1)$$

The higher the score of a sentence, the higher the similarity between the sentence and the holding, and so its relevance. These scores were used to fine-tune BERT models as a regression task. The methodology can be summarized as follows (Figure 2):

(1) The documents were split into sentences.
(2) R-1 and R-2 scores between each sentence and its respective document holding were computed.
(3) To retrieve a single score out of the R-1 and R-2 ones, we computed their harmonic, for each sentence.
(4) Italian-LEGAL-BERT was fine-tuned in the regression task of predicting the score of a given sentence.
(5) The validation dataset was used to determine the optimal number of top k sentences to compose the final holding. We tried $k = 3, 5, 7$ and found that $k = 5$ yielded the best results.

Before fine-tuning Italian-LEGAL-BERT, the input text sequence is tokenized into subword units using WordPiece tokenization of Italian-Legal-BERT. Each token is then converted into a vector representation using an embedding matrix and fed into a multi-layer bidirectional Transformer encoder. The output of the last layer of the Transformer encoder is pooled into a special classification token [CLS] that is the representation of the whole sequence. Finally, the [CLS] vector representation is then fed into a regression head,

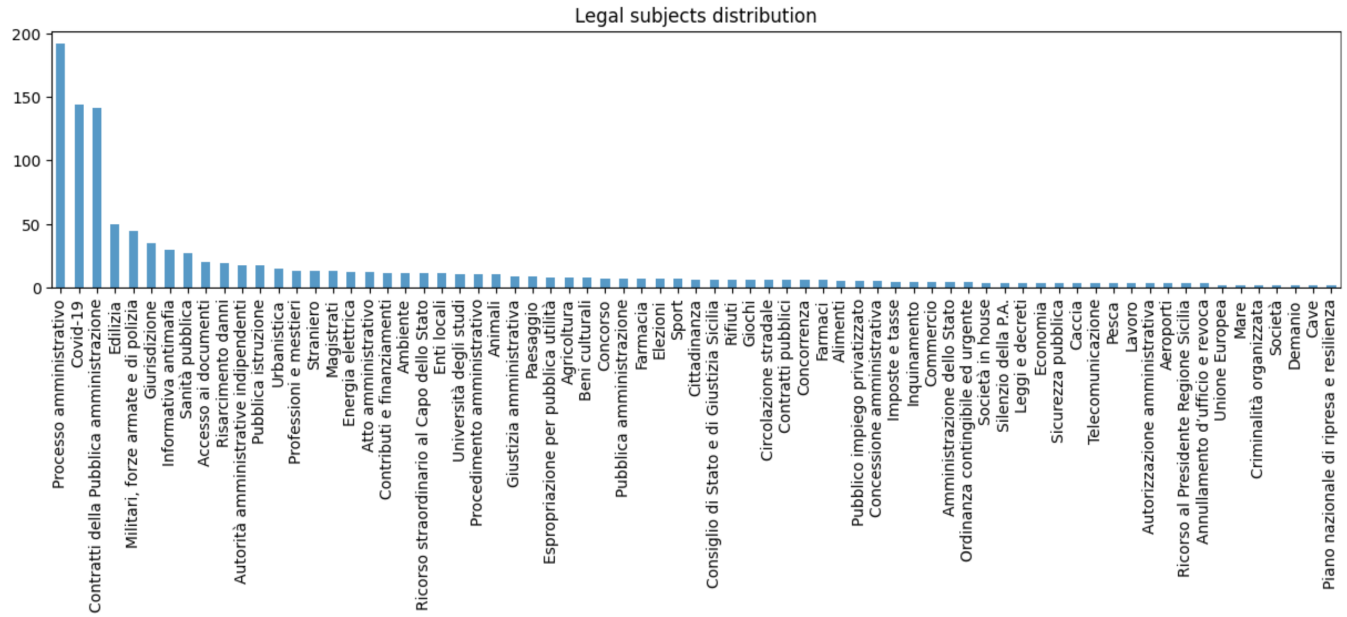


Figure 1: Legal subject distribution on ITA-CaseHold dataset. The most common legal subjects are Administrative Processes, Covid-19, Public Administration Contracts, Construction, Military, armed forces and police, Jurisdiction, Anti-Mafia Reporting, Public Health, Access to Documents, Compensation for damages, Independent administrative authorities, Public education, Urban planning, Professions and trades, and Foreigner.

Table 1: Token count distribution of the dataset

Statistics	Documents			Holdings			Compression ratio		
	Train	Val	Test	Train	Val	Test	Train	Val	Test
Min	268	293	309	54	104	78	1.35	1.47	1.6
Max	19549	18189	16389	3372	2709	2906	24.02	23.81	28.2
Std	3313.67	3434.10	3226.69	613.04	670.65	580	4.84	4.63	5.13
Mean	4715.32	4819.10	4588.37	812.17	857	734	7.12	6.95	7.53
25%	2245.25	2279.25	1857	357.5	335.5	310	3.56	3.47	3.9
50%	4133	4180	4219	655	649.5	582	5.46	5.58	5.86
75%	6423	6613.25	6111	892.25	1184.25	1014	9.6	8.96	9.45

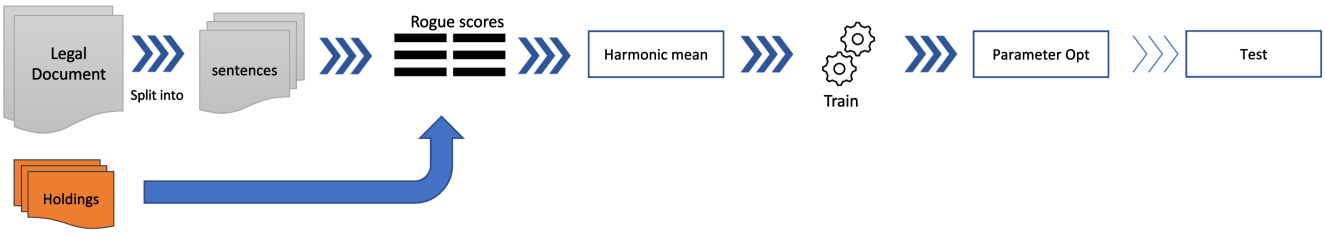


Figure 2: The steps of our training pipeline. First, the documents were split into sentences, and R-1 and R-2 scores are computed for each of them with respect to their document holding. Then, the harmonic means are computed and used to train Italian-LEGAL-BERT. Finally, the parameter describing the number of summarizing sentences to select is optimized.

which is a fully connected layer that maps the pooled vector to a scalar value. The output of the regression head is the predicted value for the regression task.

In order to determine the optimal number of sentences (parameter k) to select for our summaries, we conducted a thorough validation process. We used the validation data set and both the HM-BERT and AM-BERT models for this process. The steps to find the optimal k value follows:

(1) Validation documents were split into sentences.
(2) Trained model was used to compute the score of each sentence.
(3) The sentences were sorted by predicted scores.
(4) With k = 3,5,7, we extracted three different summaries.
(5) ROUGE scores (R1, R2, Rl, RW) were evaluated between the extracted summaries and original summaries.
(6) Based on these scores we chose the optimal k value, and in our case, it's 5.

For testing, we followed similar steps as validation which starts with splitting a document into sentences. The trained model was then used to predict scores for these sentences. The sentences were then sorted in descending order. The top 5 sentences were chosen, and these top 5 sentences were rearranged based on sentence index id and then extracted.

For each testing document, we followed these steps for each document (Figure 3)

(1) The document was split into sentences.
(2) The trained model was used to compute the score of each sentence.
(3) The sentences were sorted by predicted scores.
(4) Top 5 sentences were selected and sorted according to their position in the original document to compose the final holding.
(5) ROUGE scores were evaluated between the extracted and original holding.

5 EXPERIMENTS

We deployed the ITA-CaseHold dataset in nine different experiments, comparing the performances of a baseline zero-shot extractive method and six BERT-based models. The experiments were conducted on an NVIDIA-DGX system equipped with a 32GB TeslaV100 GPU and were evaluated with ITA-ROUGE, a modified version of the ROUGE [27] to compute the Rouge metrics for the Italian language which uses the Italian NLTK models for tokenization and stemming. Specifically, we stuck to R-1, R-2, R-L, and R-W scores.

5.1 Zero-Shot Baseline

LexRank [18] is a popular zero-shot model, which uses a graph-based approach for automatic text summarization and does not require any training. It is an unsupervised and language-agnostic algorithm, meaning it does not require domain fine-tuning, which makes it a convenient choice for a baseline model.

We replicated a similar version of EUR-Lex-Sum [3, 38] LexRank baseline model. To compute the centrality scores, we used Italian-LEGAL-BERT [26] embeddings. Even though Italian-LEGAL-BERT is not intended to be a sentence transformer, it outperformed other competitive models in this use case. LexRank ranks the sentences of a document based on similarity scores between them. To choose the number of top-scoring sentences to be used in the final summary of a document, we computed the median compression ratio on the training dataset. This is the paragraph-level average ratio between the length of a document paragraph and the length of its summarizing section in the document holding. By multiplying the median compression ratio for the number of paragraphs in a document we get the number of sentences to be used in the holding of that given document.

Table 2: Models comparison on ROUGE scores

Model	Encoder	R-1	R-2	R-L	R-W
LexRank	ITA BERT	47.92	24.79	27.3	9.49
	ITA LEGAL BERT	46.79	23.49	27.25	9.42
	ITA LEGAL BERT SC	46.25	23.03	27.01	9.28
AM-BERT	ITA BERT	51.48	30.77	28.61	10.27
	ITA LEGAL BERT	52.34	32.60	28.89	10.53
	ITA LEGAL BERT SC	51.74	31.04	29.56	10.78
HM-BERT	ITA BERT	53.12	33.71	29.72	10.88
	ITA LEGAL BERT	**54.41**	**36.28**	30.33	**11.24**
	ITA LEGAL BERT SC	53.66	34.34	**30.64**	11.21

5.2 BERT-Extractive

This section provides the details of BERT Extractive models presented in detail in Section 4. For ease of discussion, we will refer to the model trained using the harmonic mean of the R-1 and R-2 scores of the document sentences as HM-BERT, and to the model trained with the arithmetic one as AM-BERT. The models' architecture is the same, only the objective function differs.

For each sentence, we computed the harmonic and arithmetic mean of R-1 and R-2 scores given the correspondent document holding, and use these scores as the target variables respectively for HM-BERT and AM-BERT.

Our software stack included PyTorch, Hugging Face transformers, Simple transformers, and Py-Rouge. We used Italian-LEGAL-BERT, Italian-BERT, and Italian-LEGAL-BERT-SC as the encoder. The Italian-LEGAL-BERT model has an embedding dimension of 768, an input token size of 512, 12 hidden layers with 12 attention heads, and an attention dropout of 0.1. A sequence regression head (i.e. a linear layer) was added to the pooled output. The training was carried out with an AdamW optimizer and a linear scheduler. We trained HM-BERT and AM-BERT for 4 epochs, using a batch size of 16 and setting 256 as the maximum sequence length.

5.3 Results

The results in Table 2 show that HM-BERT with Italian-LEGAL-BERT outperforms the other two encoders. LexRank's unsupervised approach is the least effective and only serves as a baseline. BERT-based models outperform it by at least 1.6 on all ROUGE scores. The better performances of HM-BERT show the importance of using the harmonic mean between R-1 and R-2 scores as a target for fine-tuning the regressor. The core idea of Rouge is lexical overlap between extracted summaries and original summaries, recent findings have shown that ROUGE score does not correlate well with

Figure 3: The steps of our test pipeline for a single test document. First, the document is split into sentences, and the trained model is used to compute the score for each of them. The sentences are then sorted by their scores and the top 5 are selected and sorted according to their position in the original document to compose the final holding. ROUGE scores between the extracted and the original holding are finally evaluated.

how humans assess the quality of a candidate summary [5, 9]. To overcome this, we have also approached human evaluation.

We evaluated our model's performance by having a legal expert, including a Law professor, qualitatively assess it. We randomly selected ten legal judgments and their corresponding extracted holdings, resulting in a total of fifty extracted sentences (i.e., 5 sentences x 10 documents). The expert annotated these extracted sentences with a label indicating their relevance: either 'High' or 'Low'. The majority of the extracted sentences were labeled as 'High' relevance, accounting for 66% of the total. In figure 4, we provide a snippet of the expert's validation.

6 LIMITATIONS

The score that our model assigns to a sentence is representative of the syntactic overlap between that sentence and the corresponding document holding. This could lead the model to select redundant information. In the context of administrative justice and for our case study, experts analyzing our system have not pointed out this kind of problem, but this could be different in other legal contexts (e.g., civil law). For this reason, we intend to improve and evaluate our model by integrating sentence clustering, centrality graph-based, or trigram blocking methods [29] to avert redundancy in sentence selection, and possibly investigate the performances of our approach to other legal contexts, eventually extending the training data. Other contexts may also require fine-tuning again the parameter describing the number of sentences to compose the final summary. In providing the most relevant sentences of the document, our model extrapolates them from their context, which may result in a hardly interpretable holding. For this reason, we are currently developing an application to highlight the selected sentences in the original judgment, for them to be easily contextualized. Another limitation comes from the difficulty of an extractive model in providing effective summaries of complex and tangled documents. To this aim, abstractive methods may become handy, due to their greater flexibility. Finally, as with any other black-box deep learning model, HM-BERT decisions have difficulty being explained.

7 CONCLUSIONS AND FUTURE WORK

7.1 Conclusions

In this paper, we presented ITA-CaseHold, a new benchmark dataset for legal text summarization and holdings extraction in the Italian domain. We set new baselines for this dataset by introducing HM-BERT, a new extractive summarization tool based on Italian-LEGAL-BERT, a pre-trained BERT model for the Italian legal domain. HM-BERT is trained to select the most relevant sentences according to their similarity with the training holdings. The similarity function was defined as the harmonic mean of the overlap of unigrams (ROUGE R-1) and bigrams (ROUGE R-2) between the sentence and the corresponding document holding. Experiments shows that the harmonic mean improves the effectiveness of the model with respect to the arithmetic mean. In our case study, the optimal number of sentences to be extracted turned out to be 5. Our model achieved excellent results in terms of ROUGE scores. The holdings extracted were further validated by experts who qualitatively confirmed the usefulness of our tool.

7.2 Future work

Search Engine We are currently working on the deployment of our tool as a web application to help jurists and practitioners in selecting the most relevant sentences in legal judgments, saving their time and effort. Such an application can be easily coupled with a search engine to identify groups of judgments with similar holdings. At the same time, we plan to increase the size and quality of the released dataset to encourage and improve future research in this field.

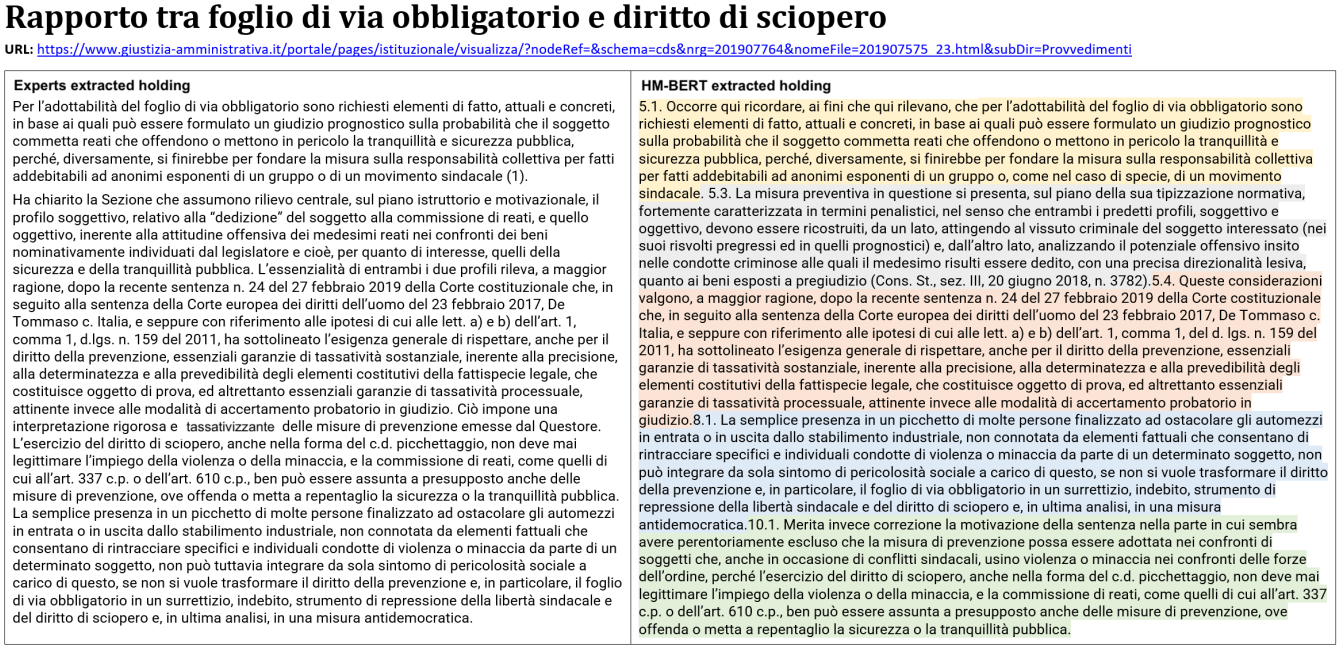
Neural QA system Next, we want to leverage the potential of pre trained large language generative models to develop a Question-Answering (QA) neural system for administrative justice, which could prove to be a valuable tool for people trying to better understand complex Italian rules and regulations.

Text Summarization Furthermore, the released dataset contains both the complete holding and its title. This information could be used to develop models of 'extreme' text summarization capable of synthesizing a text into a single sentence.

Classification Task Finally, the presented dataset could also be explored as a Multi-class classification task with Legal subjects serving as the classes. As the first Italian legal dataset, any model developed would serve as a baseline model for this task within the legal domain. This could aid the development of new legal tools and technologies in the future.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their constructive comments. They express our gratitude to domain expert Vanessa Battiato for her excellent help in the qualitative analysis of the results. This research is part of an Italian Nationwide



Rapporto tra foglio di via obbligatorio e diritto di sciopero

URL: https://www.giustizia-amministrativa.it/portale/pages/istituzionale/visualizza/?nodeRef=&schema=cds&nrg=201907764&nomeFile=201907575_23.html&subDir=Provvedimenti

Experts extracted holding

Per l'adottabilità del foglio di via obbligatorio sono richiesti elementi di fatto, attuali e concreti, in base ai quali può essere formulato un giudizio prognostico sulla probabilità che il soggetto commetta reati che offendono o mettono in pericolo la tranquillità e sicurezza pubblica, perché, diversamente, si finirebbe per fondare la misura sulla responsabilità collettiva per fatti addebitabili ad anonimi esponenti di un gruppo o di un movimento sindacale (1).

Ha chiarito la Sezione che assumono rilievo centrale, sul piano istruttorio e motivazionale, il profilo soggettivo, relativo alla "dedizione" del soggetto alla commissione di reati, e quello oggettivo, inerente alla attitudine offensiva dei medesimi reati nei confronti dei beni nominativamente individuati dal legislatore e cioè, per quanto di interesse, quelli della sicurezza e della tranquillità pubblica. L'essenzialità di entrambi i due profili rileva, a maggior ragione, dopo la recente sentenza n. 24 del 27 febbraio 2019 della Corte costituzionale che, in seguito alla sentenza della Corte europea dei diritti dell'uomo del 23 febbraio 2017, De Tommaso c. Italia, e seppure con riferimento alle ipotesi di cui alle lett. a) e b) dell'art. 1, comma 1, d.lgs. n. 159 del 2011, ha sottolineato l'esigenza generale di rispettare, anche per il diritto della prevenzione, essenziali garanzie di tassatività sostanziale, inerente alla precisione, alla determinatezza e alla prevedibilità degli elementi costitutivi della fattispecie legale, che costituisce oggetto di prova, ed altrettanto essenziali garanzie di tassatività processuale, attinente invece alle modalità di accertamento probatorio in giudizio. Ciò impone una interpretazione rigorosa e tassativizzante delle misure di prevenzione emesse dal Questore. L'esercizio del diritto di sciopero, anche nella forma del c.d. picchettaggio, non deve mai legittimare l'impiego della violenza o della minaccia, e la commissione di reati, come quelli di cui all'art. 337 c.p. o dell'art. 610 c.p., ben può essere assunta a presupposto anche delle misure di prevenzione, ove offenda o metta a repentaglio la sicurezza o la tranquillità pubblica. La semplice presenza in un picchetto di molte persone finalizzato ad ostacolare gli automezzi in entrata o in uscita dallo stabilimento industriale, non connotata da elementi fattuali che consentano di rintracciare specifici e individuali condotte di violenza o minaccia da parte di un determinato soggetto, non può tuttavia integrare da sola sintomo di pericolosità sociale a carico di questo, se non si vuole trasformare il diritto della prevenzione e, in particolare, il foglio di via obbligatorio in un surrettizio, indebito, strumento di repressione della libertà sindacale e del diritto di sciopero e, in ultima analisi, in una misura antidemocratica.

HM-BERT extracted holding

5.1. Occorre qui ricordare, ai fini che qui rilevano, che per l'adottabilità del foglio di via obbligatorio sono richiesti elementi di fatto, attuali e concreti, in base ai quali può essere formulato un giudizio prognostico sulla probabilità che il soggetto commetta reati che offendono o mettono in pericolo la tranquillità e sicurezza pubblica, perché, diversamente, si finirebbe per fondare la misura sulla responsabilità collettiva per fatti addebitabili ad anonimi esponenti di un gruppo o, come nel caso di specie, di un movimento sindacale. 5.3. La misura preventiva in questione si presenta, sul piano della sua tipizzazione normativa, fortemente caratterizzata in termini penalistici, nel senso che entrambi i predetti profili, soggettivo e oggettivo, devono essere ricostruiti, da un lato, attingendo al vissuto criminale del soggetto interessato (nei suoi risvolti pregressi ed in quelli prognostici) e, dall'altro lato, analizzando il potenziale offensivo insito nelle condotte criminose alle quali il medesimo risulti essere dedito, con una precisa direzionalità lesiva, quanto ai beni esposti a pregiudizio (Cons. St., sez. III, 20 giugno 2018, n. 3782).5.4. Queste considerazioni valgono, a maggior ragione, dopo la recente sentenza n. 24 del 27 febbraio 2019 della Corte costituzionale che, in seguito alla sentenza della Corte europea dei diritti dell'uomo del 23 febbraio 2017, De Tommaso c. Italia, e seppure con riferimento alle ipotesi di cui alle lett. a) e b) dell'art. 1, comma 1, del d. lgs. n. 159 del 2011, ha sottolineato l'esigenza generale di rispettare, anche per il diritto della prevenzione, essenziali garanzie di tassatività sostanziale, inerente alla precisione, alla determinatezza e alla prevedibilità degli elementi costitutivi della fattispecie legale, che costituisce oggetto di prova, ed altrettanto essenziali garanzie di tassatività processuale, attinente invece alle modalità di accertamento probatorio in giudizio.8.1. La semplice presenza in un picchetto di molte persone finalizzato ad ostacolare gli automezzi in entrata o in uscita dallo stabilimento industriale, non connotata da elementi fattuali che consentano di rintracciare specifici e individuali condotte di violenza o minaccia da parte di un determinato soggetto, non può integrare da sola sintomo di pericolosità sociale a carico di questo, se non si vuole trasformare il diritto della prevenzione e, in particolare, il foglio di via obbligatorio in un surrettizio, indebito, strumento di repressione della libertà sindacale e del diritto di sciopero e, in ultima analisi, in una misura antidemocratica.10.1. Merita invece correzione la motivazione della sentenza nella parte in cui sembra avere perentoriamente escluso che la misura di prevenzione possa essere adottata nei confronti di soggetti che, anche in occasione di conflitti sindacali, usino violenza o minaccia nei confronti delle forze dell'ordine, perché l'esercizio del diritto di sciopero, anche nella forma del c.d. picchettaggio, non deve mai legittimare l'impiego della violenza o della minaccia, e la commissione di reati, come quelli di cui all'art. 337 c.p. o dell'art. 610 c.p., ben può essere assunta a presupposto anche delle misure di prevenzione, ove offenda o metta a repentaglio la sicurezza o la tranquillità pubblica.

Figure 4: An example of a report for qualitative analysis of extracted holdings. On the right is the holding produced by HM-BERT with the 5 top scoring sentences highlighted with different colors.

project "Giustizia Agile/ Agile Justice" project[5], funded by the Italian Ministry of Justice.

REFERENCES

[1] Abhishek Agarwal, Shanshan Xu, and Matthias Grabmair. 2022. Extractive Summarization of Legal Decisions using Multi-task Learning and Maximal Marginal Relevance. https://doi.org/10.48550/ARXIV.2210.12437

[2] Dogu Araci. 2019. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. https://doi.org/10.48550/ARXIV.1908.10063

[3] Dennis Aumiller, Ashish Chouhan, and Michael Gertz. 2022. EUR-Lex-Sum: A Multi- and Cross-lingual Dataset for Long-form Summarization in the Legal Domain. https://doi.org/10.48550/ARXIV.2210.13448

[4] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3615–3620. https://doi.org/10.18653/v1/D19-1371

[5] Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating Evaluation in Text Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 9347–9359. https://doi.org/10.18653/v1/2020.emnlp-main.751

[6] Paheli Bhattacharya, Kaustubh Hiware, Subham Rajgaria, Nilay Pochhi, Kripabandhu Ghosh, and Saptarshi Ghosh. 2019. *A Comparative Study of Summarization Algorithms Applied to Legal Case Judgments.* 413–428. https://doi.org/10.1007/978-3-030-15712-8_27

[7] Paheli Bhattacharya, Soham Poddar, Koustav Rudra, Kripabandhu Ghosh, and Saptarshi Ghosh. 2021. Incorporating Domain Knowledge for Extractive Summarization of Legal Case Documents. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law* (São Paulo, Brazil) *(ICAIL '21)*. Association for Computing Machinery, New York, NY, USA, 22–31. https://doi.org/10.1145/3462757.3466092

[8] Jaime Carbonell and Jade Goldstein. 1998. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, Australia) *(SIGIR '98)*. Association for Computing Machinery, New York, NY, USA, 335–336. https://doi.org/10.1145/290941.291025

[9] Arun Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evalaution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 643–653. https://doi.org/10.18653/v1/P18-1060

[10] Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019. Large-Scale Multi-Label Text Classification on EU Legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 6314–6322. https://doi.org/10.18653/v1/P19-1636

[11] Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. MultiEURLEX - A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 6974–6996. https://doi.org/10.18653/v1/2021.emnlp-main.559

[12] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of Law School. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 2898–2904. https://doi.org/10.18653/v1/2020.findings-emnlp.261

[13] Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. Paragraph-level Rationale Extraction through Regularization: A case study on European Court of Human Rights Cases. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 226–241. https://doi.org/10.18653/v1/2021.naacl-main.22

[14] Jianpeng Cheng and Mirella Lapata. 2016. Neural Summarization by Extracting Sentences and Words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 484–494. https://doi.org/10.18653/v1/P16-1046

[15] Peng Cui and Le Hu. 2021. Sliding Selector Network with Dynamic Memory for Extractive Summarization of Long Documents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 5881–5891. https://doi.org/10.18653/v1/2021.naacl-main.470

[5] More information is available at https://www.unitus.it/it/unitus/mappatura-della-ricerca/articolo/giustizia-agile.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/ARXIV.1810.04805

[17] Yue Dong, Andrei Mircea, and Jackie Chi Kit Cheung. 2021. Discourse-Aware Unsupervised Summarization for Long Scientific Documents. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, 1089–1102. https://doi.org/10.18653/v1/2021.eacl-main.93

[18] G. Erkan and D. R. Radev. 2004. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research* 22 (dec 2004), 457–479. https://doi.org/10.1613/jair.1523

[19] Atefeh Farzindar and Guy Lapalme. 2004. Letsum, an automatic legal text summarizing system. *Jurix* (01 2004), 11–18.

[20] Matt Grenander, Yue Dong, Jackie Chi Kit Cheung, and Annie Louis. 2019. Countering the Effects of Lead Bias in News Summarization via Multi-Stage Training and Auxiliary Losses. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 6019–6024. https://doi.org/10.18653/v1/D19-1620

[21] Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 708–719. https://doi.org/10.18653/v1/N18-1065

[22] Peter Henderson, Mark S. Krass, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho. 2022. Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset. https://doi.org/10.48550/ARXIV.2207.00220

[23] Huan Yee Koh, Jiaxin Ju, Ming Liu, and Shirui Pan. 2022. An Empirical Survey on Long Document Summarization: Datasets, Models, and Metrics. *ACM Comput. Surv.* 55, 8, Article 154 (dec 2022), 35 pages. https://doi.org/10.1145/3545176

[24] Anastassia Kornilova and Vladimir Eidelman. 2019. BillSum: A Corpus for Automatic Summarization of US Legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*. Association for Computational Linguistics, Hong Kong, China, 48–56. https://doi.org/10.18653/v1/D19-5406

[25] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics (Oxford, England)* 36 (09 2019). https://doi.org/10.1093/bioinformatics/btz682

[26] Daniele Licari and Giovanni Comandè. 2022. ITALIAN-LEGAL-BERT: A Pre-trained Transformer Language Model for Italian Law. In *Companion Proceedings of the 23rd International Conference on Knowledge Engineering and Knowledge Management (CEUR Workshop Proceedings, Vol. 3256)*, Danai Symeonidou, Ran Yu, Davide Ceolin, María Poveda-Villalón, Davide Audrito, Luigi Di Caro, Francesca Grasso, Roberto Nai, Emilio Sulis, Fajar J. Ekaputra, Oliver Kutz, and Nicolas Troquard (Eds.). CEUR, Bozen-Bolzano, Italy. https://ceur-ws.org/Vol-3256/#km4law3 ISSN: 1613-0073.

[27] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013

[28] Chao-Lin Liu and Kuan-Chun Chen. 2019. Extracting the Gist of Chinese Judgments of the Supreme Court. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law* (Montreal, QC, Canada) *(ICAIL '19)*. Association for Computing Machinery, New York, NY, USA, 73–82. https://doi.org/10.1145/3322640.3326715

[29] Yang Liu and Mirella Lapata. 2019. Text Summarization with Pretrained Encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3730–3740. https://doi.org/10.18653/v1/D19-1387

[30] Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028* (2002).

[31] Potsawee Manakul and Mark Gales. 2021. Long-Span Summarization via Local Attention and Content Selection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 6026–6041. https://doi.org/10.18653/v1/2021.acl-long.470

[32] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 1906–1919. https://doi.org/10.18653/v1/2020.acl-main.173

[33] Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Barcelona, Spain, 404–411. https://aclanthology.org/W04-3252

[34] Derek Miller. 2019. Leveraging BERT for Extractive Text Summarization on Lectures. https://doi.org/10.48550/ARXIV.1906.04165

[35] Ankan Mullick, Abhilash Nandy, Manav Kapadnis, Sohan Patnaik, Raghav R, and Roshni Kar. 2022. An Evaluation Framework for Legal Document Summarization. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 4747–4753. https://aclanthology.org/2022.lrec-1.508

[36] Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-Judgment-Prediction: A Multilingual Legal Judgment Prediction Benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 19–35. https://doi.org/10.18653/v1/2021.nllp-1.3

[37] Seth Polsley, Pooja Jhunjhunwala, and Ruihong Huang. 2016. CaseSummarizer: A System for Automated Summarization of Legal Texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*. The COLING 2016 Organizing Committee, Osaka, Japan, 258–262. https://aclanthology.org/C16-2054

[38] Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 4512–4525. https://doi.org/10.18653/v1/2020.emnlp-main.365

[39] Abhay Shukla, Paheli Bhattacharya, Soham Poddar, Rajdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal, and Saptarshi Ghosh. 2022. Legal Case Document Summarization: Extractive and Abstractive Methods and their Evaluation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online only, 1048–1064. https://aclanthology.org/2022.aacl-main.77

[40] Andrea Tagarelli and Andrea Simeri. 2021. Unsupervised law article mining based on deep pre-trained language representation models with application to the Italian civil code. *Artificial Intelligence and Law* 30, 3 (sep 2021), 417–473. https://doi.org/10.1007/s10506-021-09301-8

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[42] Ruifeng Yuan, Zili Wang, and Wenjie Li. 2020. Fact-level Extractive Summarization with Hierarchical Graph Mask on BERT. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 5629–5639. https://doi.org/10.18653/v1/2020.coling-main.493

[43] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big Bird: Transformers for Longer Sequences. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 17283–17297. https://proceedings.neurips.cc/paper/2020/file/c8512d142a2d849725f31a9a7a361ab9-Paper.pdf

[44] Hao Zheng and Mirella Lapata. 2019. Sentence Centrality Revisited for Unsupervised Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 6236–6247. https://doi.org/10.18653/v1/P19-1628

[45] Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset of 53,000+ Legal Holdings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law* (São Paulo, Brazil) *(ICAIL '21)*. Association for Computing Machinery, New York, NY, USA, 159–168. https://doi.org/10.1145/3462757.3466088

[46] Linwu Zhong, Ziyi Zhong, Zinian Zhao, Siyuan Wang, Kevin D. Ashley, and Matthias Grabmair. 2019. Automatic Summarization of Legal Decisions Using Iterative Masking of Predictive Sentences. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law* (Montreal, QC, Canada) *(ICAIL '19)*. Association for Computing Machinery, New York, NY, USA, 163–172. https://doi.org/10.1145/3322640.3326728

[47] Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A Robustly Optimized BERT Pre-training Approach with Post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*. Chinese Information Processing Society of China, Huhhot, China, 1218–1227. https://aclanthology.org/2021.ccl-1.108

A DATASET, CODE AND EXPERIMENTS

Where can I find the data set and code? The data set is publicly available on Hugging face at ITA-CASEHOLD dataset card. To

download the data set from hugging face, simply install hugging face data sets and then write the following lines:

```
from datasets import load_dataset

ds = load_dataset('itacasehold/itacasehold')
```

How to run the experiments? To reproduce the experiments,

(1) The first step is cloning the repository from our GitHub ITA-CASEHOLD repo with the following command:
> *git clone https://github.com/dlicari/ITA-CASEHOLD.git*
(2) Install the dependencies from the requirements.txt file with
> *pip install -r requirements.txt*
(3) To run the model use run_model.py file, giving the model name as argument.
> *python run_model.py –model modelname*

The generated results will be stored in the outputs folder in JSON files.

Received 12 February 2023; revised April 2023; accepted