

LexAbSumm: Aspect-based Summarization of Legal Decisions

Santosh T.Y.S.S, Mahmoud Aly, Matthias Grabmair

School of Computation, Information, and Technology;
Technical University of Munich, Germany
{santosh.tokala, mahmoud.aly, matthias.grabmair}@tum.de

Abstract

Legal professionals frequently encounter long legal judgments that hold critical insights for their work. While recent advances have led to automated summarization solutions for legal documents, they typically provide generic summaries, which may not meet the diverse information needs of users. To address this gap, we introduce LexAbSumm, a novel dataset designed for aspect-based summarization of legal case decisions, sourced from the European Court of Human Rights jurisdiction. We evaluate several abstractive summarization models tailored for longer documents on LexAbSumm, revealing a challenge in conditioning these models to produce aspect-specific summaries. We release LexAbSumm to facilitate research in aspect-based summarization for legal domain.

Keywords: Aspect-based Summarization, Legal Judgements, European Court of Human Rights

1. introduction

Legal professionals, including lawyers, judges, and researchers, regularly face the challenge of sifting through lengthy legal judgments that encompass multiple critical aspects for case law interpretation and judicial reasoning. Recent strides in NLP have addressed this challenge by offering automated legal case summarization systems (eg., Agarwal et al. 2022; Bhattacharya et al. 2021, 2019 *inter alia*). These improvements owe much to the availability of datasets such as UK-Abs, IN-Abs, IN-Ext (Shukla et al., 2022), BVA (Zhong et al., 2019), Rulingbr (de Vargas Feijó and Moreira, 2018), Multilexsumm (Shen et al., 2022) facilitating research in this field.

However, existing legal case summarization datasets and systems rely on a single, generic summary, which may not meet the diverse demands of users. Legal case documents are inherently multifaceted and often contain distinct aspects relevant to different users. This complexity is further amplified when documents involve multiple allegations or claimants. This one-size-fits-all approach risks omitting critical details and failing to provide specific information that individual users require (Angelidis and Lapata, 2018; Woodsend and Lapata, 2012). Therefore, there is a need to develop legal case summarization systems capable of generating concise, aspect-specific summaries that cater to users' specific information needs more effectively.

Despite the undeniable demand for such systems, there currently exists no dataset designed explicitly for aspect-based legal case summarization. In this work, we present a novel dataset,

named **LexAbSumm**¹, tailored for evaluating single-document aspect-oriented abstractive summarization. LexAbSumm is sourced from the jurisprudence of the European Court of Human Rights (ECtHR), responsible for addressing complaints against states for alleged human rights violations under the European Convention of Human Rights. To create this dataset, we use theme-based fact sheets from the ECtHR's press service². From these sheets, we extract aspects from section titles and use the corresponding case details and brief summaries as aspect-based summaries of the case document. Our assessment of several abstractive summarization models, tailored for longer documents on LexAbSumm reveal that these models struggle to condition their summaries on the provided aspects, often producing generic summaries regardless of the aspect.

2. Dataset Construction

Judgements Collection We collect English judgments from the publicly accessible database of the ECtHR, HUDOC³, based on their metadata (Document Type: HEJUD). They follow a structured format, with sections like *Procedure*, outlining the procedural steps; *The Facts*, covering case background; *The Law*, providing legal reasoning and *Conclusion*, stating the Court's verdict on alleged violations. We use handcrafted rules and regular expressions to parse the documents due to inconsistent HTML structure and section titles. We exclude the *Procedure* section as it focuses

¹Our LexAbSumm dataset is available at <https://github.com/TUMLegalTech/LexAbSumm>

²<https://www.echr.coe.int/factsheets>

³<http://hudoc.echr.coe.int/>

Reproductive rights

Medically-assisted procreation

Evans v. United Kingdom

19 April 2007 (Grand Chamber)

The applicant, who was suffering from ovarian cancer, underwent in-vitro fertilisation (IVF) with her then partner before having her ovaries removed. Six embryos were created and placed in storage. When the couple's relationship ended, her ex-partner withdrew his consent for the embryos to be used, not wanting to be the genetic parent of the applicant's child. National law consequently required that the eggs be destroyed. The applicant complained that domestic law permitted her former partner effectively to withdraw his consent to the storage and use by her of embryos created jointly by them, preventing her from ever having a child to whom she would be genetically related.

For the reasons given by the Chamber in its [judgment of 7 March 2006](#), namely that the issue of when the right to life began came within the State's margin of appreciation, the Grand Chamber found that the embryos created by the applicant and her former partner did not have a right to life. It therefore held that there had been **no violation of Article 2** (right to life) of the Convention. The Grand Chamber further considered that, given the lack of European consensus, the fact that the domestic rules had been clear and brought to the attention of the applicant and that they had struck a fair balance between the competing interests, there had been **no violation of Article 8** (right to respect for private and family life) of the Convention. Lastly, the Grand Chamber held that there had been **no violation of Article 14** (prohibition of discrimination) **taken in conjunction with Article 8** of the Convention.

Figure 1: Example of an aspect-judgement-summary triplet from the fact sheet PDF file.

on procedural details. Throughout this work, we collectively refer to the combination of *The Law* and *Conclusion* as 'the law section' since the conclusion primarily offers a final verdict based on the legal reasoning described in the Law section.

Aspects and Summaries Collection We curate our aspect-based summarization dataset using fact sheets available on the ECHR website. These sheets, prepared by the court's press service division, provide an analysis of case law developments across various thematic areas (e.g., Reproductive Rights⁴). In total, there are 73 fact sheets available as PDF documents, organized under 16 broad themes. Each sheet is further divided into sections (e.g., "Medically-assisted procreation" in the Reproductive Rights sheet), encompassing various aspects within a theme. Within each section, multiple cases and their summaries are provided. Each case summary includes two components: one summarizing the case's facts (in black font) and the other summarizing the case's legal aspects (in blue font). Fig. 1 illustrates an example of a case name with its summary and section title as presented in the PDF file. We employ various heuristics to extract section titles, cases, and their summaries from these PDF files. Subsequently, we manually filter out aspect-summary pairs in which either the summary is missing or the section title doesn't represent an aspect of the theme (e.g., "Pending applications"). Finally, we match case names in aspect-summary pairs back to our judgment collection and exclude pairs that couldn't be linked to our judgment collection, as some might refer to non-English documents not included in our collection.

⁴https://www.echr.coe.int/documents/d/echr/FS_Reproductive_ENG

	Whole	Facts	Law
Input Length	14357.14	3929.77	10427.38
Summ. Length	251.1	81.19	169.91
Comp. Ratio	66.25	59.75	85.02
Coverage (1/2-gram)	0.95/0.73	0.87/0.46	0.96/0.74
Density (1/2-gram)	6.32/5.18	2.56/1.49	7.08/5.94
Copy Length	1.98	1.53	2.01
Novelty (1/2-gram)	0.07/0.49	0.26/0.75	0.08/0.45
Novelty (3-gram)	0.7	0.9	0.66

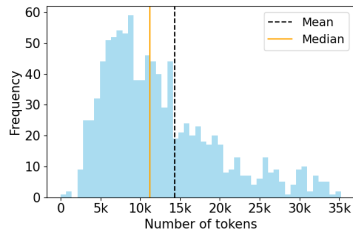
Table 1: Characteristics of LexAbSumm.

2.1. Dataset Splits & Characteristics

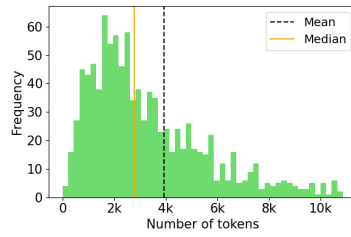
We end up with a total of 1053 aspect-judgement-summary triplets with 376 unique aspects. We randomly sample 7.5% of these unique aspects (28) to create a test set of 91 triplets, allowing us to evaluate the models' generalization to new, unseen aspects. The remaining triplets are divided into training (810), validation (95), and test (57) sets. Notably, both the judgment and summary are available at a fine-grained level, segmented into facts and law sections. Based on this, we create three variants of aspect-based summarization task: using only the facts section, the law section and the whole (both the facts and the law) case as input, along with the aspect, to produce the corresponding summaries.

Data Analysis: We report the following characteristics (i) Average number of tokens in the input and the summary. We also plot the token length distribution for the input and summary of three tasks in Fig. 2. (ii) Compression Ratio indicates the token ratio between the input and the summary. (iii) Coverage-n (Grusky et al., 2018) indicates the % of n-grams in the summary that are part of an extractive fragment within the input and quantifies the extent to which a summary is derivative of an input text. (iv) Density-n (Grusky et al., 2018) quantifies how well the n-gram sequence of a summary can be described as a series of extractions and is derived from the average length of the extractive fragment to which each n-gram in the summary belongs. (v) Copy Length (Chen et al., 2020) denotes the average length of segments in summary copied from the input. (vi) Novelty-n (See et al., 2017) denotes the ratio of new n-grams present in the summary that are not in the input. (vii) We plot the distribution of aspect length tokens in Fig. 3.

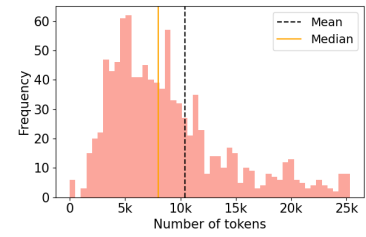
Table 1 shows that the law sections are longer than the facts sections in both the input and output, even with higher compression ratios, requiring more precise capture of crucial points from the input with respect to aspect presenting a greater challenge. We notice that the data exhibits higher coverage and density at the unigram level, suggesting higher token-level extractiveness. However, this does not necessarily translate to overall extractiveness, as it tends to rise with longer input



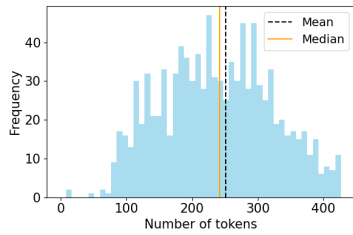
(a) Distribution of number of tokens in the input of the whole split



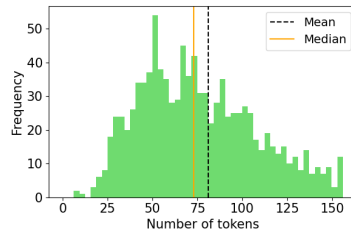
(b) Distribution of number of tokens in the input of the facts split



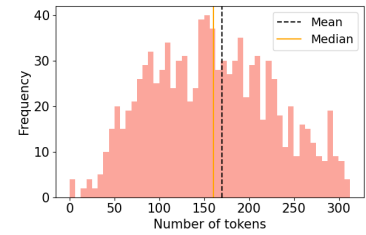
(c) Distribution of number of tokens in the input of the law split



(d) Distribution of number of tokens in the summary of the whole split



(e) Distribution of number of tokens in the summary of the facts split



(f) Distribution of number of tokens in the summary of the law split

Figure 2: Distribution of input and output tokens in three LexAbSumm splits.

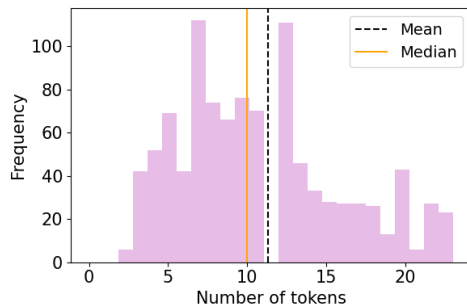


Figure 3: Distribution of tokens in the aspects of LexAbSumm.

documents that contain various summary tokens dispersed throughout. At the bigram level, both coverage and density drop, indicating token dispersion. This observation is further supported by the copy length metric. These metrics are notably higher for the law section compared to the facts, due to the specific vocabulary used in legal reasoning, tests and principles which leave little room for paraphrasing, while the progression of events in the facts are more adaptable to paraphrasing. This is also evident in higher novelty for facts compared to law. While unigram novelty may appear smaller, it increases with n-grams, indicating that the summaries are highly abstractive, requiring synthesis of information from multiple segments across the input. Average number of tokens in aspect is 10.2.

3. Experiments

3.1. Models & Metrics

We evaluate the following abstractive models designed to handle longer inputs on our three task variants. These are structured as sequence-to-sequence tasks, where the aspect is prepended to the document with a delimiter as the sequence input and the summary as output.

LED (Beltagy et al., 2020) is a longformer variant equipped with both encoder and decoder. In the encoder, it uses efficient local+global attention pattern instead of full self-attention, while the decoder utilizes full quadratic attention. It is initialized from the pre-trained BART model (Lewis et al., 2020). We use the LED-base version, which can handle input lengths of up to 16,384 tokens.

PRIMERA (Xiao et al., 2022) is initialized with the LED model and pre-trained with a summarization specific entity-based sentence masking objective and can handle upto 4096 tokens.

LongT5 (Guo et al., 2022) uses local+global attention from ETC (Ainslie et al., 2020) and adopt summarization specific pre-training from PEGASUS into the T5 model (Raffel et al., 2020) for longer sequences. We use the LongT5-base version which can handle upto 16384 tokens.

SLED (Ivgi et al., 2023) processes long sequences by partitioning the input into overlapping chunks and encode each chunk with a short-range pre-trained models encoder. It relies on decoder to fuse information across chunks attending to all input tokens, similar to fusion-in-decoder (Izacard

and Grave, 2021). SLED can be applied on top of any short-range models and thus we derive SLED-BART-base by applying on BART-base.

Unlimiformer (Bertsch et al., 2023) adopts a strategy similar to SLED, but rather than attending to all input tokens, it focuses exclusively on the top-k tokens retrieved from a k-nearest-neighbor index constructed over the hidden states of all input tokens at every decoder layer. This design helps to handle unbounded length during testing, in contrast to SLED, which is limited by memory when attending to all input tokens in the decoder. We create the Unlimiformer-BART-base model.

For both SLED and Unlimiformer, we append the aspect to every chunk of the input.

Metrics: We calculate ROUGE- $\{1,2,L\}$ (Lin, 2004) and BERTScore (BS) (Zhang et al., 2019) to compute the lexical and semantic overlap between the generated and the golden summaries for each instance and report the averaged values across all the 148 test (seen + unseen) instances.

3.2. Implementation Details

We implement our code using Huggingface Transformers library (Wolf et al., 2020). We sweep over learning rates $\{1e-5, 3e-5, 5e-5, 1e-4, 3e-4\}$ and select the best model based on the R-L score on the validation set. The model is trained for 10 epochs with Adam optimizer (Kingma and Ba, 2014). The inputs and outputs are truncated to a maximum of 16384/512 tokens for all models except PRIMERA, which is 4096/512 tokens. SLED and Unlimiformer have a chunk size of 256, and a chunk overlap of 0.5. In Unlimiformer, 16k tokens are used at training time, while unlimited tokens at evaluation.

3.3. Results

We use three extractive baselines: **Random** selects input sentences randomly as the summary. **Similarity** chooses sentences similar to the query using cosine similarity with BERT embeddings. **Extractive Oracle** is a greedy algorithm that iteratively selects sentences which maximize ROUGE-2 with the reference summary, usually regarded as the *upper bound* of the extractive method (Nallapati et al., 2017). These baselines use the same number of sentences as the reference summary.

Table 2 indicates that abstractive models outperform the Random and Similarity baselines. Among the abstractive models, LED outperforms PRIMERA, due to its longer input capacity (16k vs. 4k), evident in R-L scores for the law and whole variants with longer inputs. LongT5 surpasses them all, benefiting from its end-to-end pre-training for lengthy sequences. SLED and Unlimiformer outperform all, suggesting that adapting

short-range pre-trained models into those frameworks boost performance. Unlimiformer is comparable or better to SLED demonstrating that attending to the top-k input keys can be an accurate approximation of full attention. The Extractive Oracle, which leverages target summaries, serves as a strong baseline, emphasizing the potential of an extract-then-summarize paradigm in future.

3.4. Analysis

Generalizability to new aspects We report the models' performance separately for triplets covering seen (57 instances) and unseen aspects (91 instances) during training using the whole split in Table 3. We notice a decline in all models when handling unseen aspects. Future work should focus on enhancing model robustness to new aspects, especially in the legal domain, where the evolution of case law and legal norms requires adaptability to novel aspects.

Sensitivity to aspects We assess the model's aspect sensitivity by examining triplets in the test set where the judgment document remains the same, but the aspects differ. We calculate the BLEU score (Papineni et al., 2002) between every pair of summaries (42 pairs) with the same case and report the average in Table 3. Lower BLEU scores indicate the model's ability to generate distinct summaries for different aspects rather than a generic one, with a lower bound set by the Oracle score. We observe that long-range models like LED and LongT5 tend to produce general summaries with less aspect conditioning due to the longer positional distance between the aspect and the input. In contrast, SLED mitigates this effect by concatenating the aspect as a prefix to every chunk. However, this concatenation effect is subdued by top-k attention in Unlimiformer. Although PRIMERA exhibits better aspect sensitivity, it lags in ROUGE and BS, indicating a room for improvement in effectively conditioning on the aspect, especially for longer documents.

4. Related Work

ECtHR as data resource This has been utilized in various NLP tasks including judgement prediction (Aletras et al., 2016; Chalkidis et al., 2019, 2021; Santosh et al., 2022; Tyss et al., 2023a,b; Xu et al., 2023b), argument mining (Habernal et al., 2023; Poudyal et al., 2019, 2020), vulnerability detection (Xu et al., 2023a), event extraction (Filtz et al., 2020; Navas-Loro, 2022). While these tasks center around the judgment corpus, we utilize the press release fact sheets from the ECtHR to develop our aspect-based summarization dataset, LexAbSumm.

Models	Whole				Facts				Law			
	R-1	R-2	R-L	BS	R-1	R-2	R-L	BS	R-1	R-2	R-L	BS
Random	41.24	10.94	19.29	82.77	25.85	4.99	15.06	78.98	36.13	9.38	18.98	82.63
Similarity	44.24	14.88	22.11	83.58	31.23	7.6	18.22	83.89	41.32	14.07	22.45	84.14
Ext. Oracle	68.27	46.46	34.11	89.1	46.22	22.95	26.67	86.97	68.48	48.59	39.57	90.04
LED	49.56	25.53	30.28	87.28	40.43	15.38	26.36	88.09	47.08	26.82	33.49	88.41
PRIMERA	47.90	22.03	27.89	86.94	38.71	14.29	25.28	87.81	46.09	21.69	29.04	87.25
Long-T5	50.91	26.51	31.27	87.72	40.68	15.74	26.83	88.37	47.64	27.30	33.86	88.85
SLED-BART	52.41	28.28	32.97	88.33	40.61	15.91	26.26	87.99	49.24	29.73	35.34	88.97
Unlim.-BART	51.53	27.77	32.28	88.21	40.67	16.04	26.45	88.27	50.62	30.28	35.55	88.86

Table 2: Results on three LexAbSumm datasets. Extractive Oracle (Grey) leverages target summaries.

Models	Seen				Unseen				BLEU
	R-1	R-2	R-L	BS	R-1	R-2	R-L	BS	
Oracle									21.06
LED	49.04	26.68	31.3	88.01	48.81	24.82	29.64	87.43	46.92
PRIMERA	50.16	22.74	28.16	86.88	48.18	21.59	27.71	86.98	31.70
Long-T5	51.10	27.95	32.68	88.42	50.79	25.61	30.39	87.84	51.07
SLED-BART	54.17	29.80	34.14	88.84	51.31	27.32	32.24	88.27	42.12
Unlim.-BART	52.49	29.72	34.55	88.55	50.93	26.55	30.86	87.99	54.41

Table 3: Analysis on the Whole variant of LexAbSumm. ‘Seen’ and ‘Unseen’ column depict new aspect generalizability, ‘BLEU’ column depict aspect sensitivity.

Legal Case Summarization There exist datasets from various jurisdictions such as HOLJ (Grover et al., 2004) and UK-Abs from UK, IN-Abs and IN-Ext from India (Shukla et al., 2022), BVA from US (Zhong et al., 2019), CanLII from Canada (Zhong and Litman, 2022), Rulingbr (de Vargas Feijó and Moreira, 2018) from Brazil, Multilexsumm (Shen et al., 2022) from civil rights cases of US. Traditional approaches have leaned on extractive algorithms, known for their faithfulness (Zhong et al., 2019; Bhattacharya et al., 2019), while the recent ones explored abstractive methods (Shukla et al., 2022; Shen et al., 2022; Elaraby and Litman, 2022). These datasets provide single, generic summary, which may not meet differentiated user needs. In contrast, our work pioneers aspect-based legal case summarization.

Aspect-based Summarization Initial works were based on aspects in opinions and reviews (Woodsend and Lapata, 2012; Angelidis and Lapata, 2018; Wu et al., 2016) which got extended to news (Ahuja et al., 2022; Bahrainian et al., 2022; Frermann and Klementiev, 2019), scientific documents (Meng et al., 2021) and wikipedia (Hayashi et al., 2021; Yang et al., 2022). Closely related area is query-based summarization (Kulkarni et al., 2020; Zhong et al., 2021; Wang et al., 2022) focusing on specific fact or knowledge based questions instead of high-level aspects which mimic user intents. We introduce LexAbSumm, an aspect-based summarization data for the legal domain.

5. Conclusion

We introduce LexAbSumm, the first aspect-based summarization dataset for legal judgments, sourced from ECtHR fact sheets. Unlike traditional summarization, LexAbSumm targets differentiated summaries based on user needs (aspects). We evaluate abstractive models tailored for longer inputs, highlighting their limitations in aspect conditioning, indicating room for improvement. We hope our data and models can foster research in this area. In future, we intend to develop a model that not only conditions on the given aspect but also discourages the generation of non-aspect-related words during decoding.

6. Limitations

While our work has introduced a novel aspect-based summarization dataset for legal judgments, it is essential to acknowledge its limitations:

Language and Jurisdiction Specificity: LexAbSumm is primarily derived from legal judgments within the jurisdiction of the European Court of Human Rights (ECtHR). As such, the language, legal principles, and specific nuances in these judgments may not be directly transferable to other jurisdictions, limiting its generalizability. Further evaluation need to be carried out to understand the cross domain generalization abilities.

Evaluation Metrics: The paper primarily uses standard NLP evaluation metrics like ROUGE and BLEU. While these metrics provide useful insights, they may not fully capture the nuanced quality of legal summaries, and developing domain-specific

evaluation metrics could be beneficial.

Aspects Coverage: The dataset’s aspects are constructed based on fact sheet section titles. This approach may not comprehensively cover all the potential aspects relevant to legal professionals, as it relies on the available thematic organization within the fact sheets. Future work could explore methods for incorporating a wider range of aspects, reflecting diverse legal cases and user needs.

7. Ethics Statement

We anticipate no ethical issues in redistributing our dataset as a publicly accessible resource. This is because the legal judgment data and summaries extracted from the fact sheets are already available through HUDOC and the ECtHR website, and our utilization aligns with the ECtHR data policy. While it’s worth noting that these judgments contain real names of individuals, our approach does not introduce any harmful use beyond their existing public accessibility. Our primary goal is to encourage further research in the realm of Legal NLP, which currently suffers from a scarcity of data. Our aim is to create technology that assists legal professionals. It’s important to acknowledge that our use of pre-trained language models may inherit biases from their training data. We emphasize the need to thoroughly investigate and address any biases that may emerge, ensuring the fairness and ethical usage of the systems we develop.

8. Bibliographical References

- Abhishek Agarwal, Shanshan Xu, and Matthias Grabmair. 2022. Extractive summarization of legal decisions using multi-task learning and maximal marginal relevance. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1857–1872.
- Ojas Ahuja, Jiacheng Xu, Akshay Gupta, Kevin Horecka, and Greg Durrett. 2022. Aspectnews: Aspect-oriented summarization of news documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6494–6506.
- Joshua Ainslie, Santiago Ontanon, Chris Alberti, Vaclav Cvicek, Zachary Fisher, Philip Pham, Anirudh Ravula, Sumit Sanghai, Qifan Wang, and Li Yang. 2020. Etc: Encoding long and structured inputs in transformers. In *Proceedings of the 2020 Conference on Empirical Meth-*
- ods in Natural Language Processing (EMNLP)*, pages 268–284.
- Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel PreoŃiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ computer science*, 2:e93.
- Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. *arXiv preprint arXiv:1808.08858*.
- Seyed Ali Bahrainian, Sheridan Feucht, and Carsten Eickhoff. 2022. Newts: A corpus for news topic-focused summarization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 493–503.
- Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.
- Amanda Bertsch, Uri Alon, Graham Neubig, and Matthew R Gormley. 2023. Unlimiformer: Long-range transformers with unlimited length input. *arXiv preprint arXiv:2305.01625*.
- 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. In *Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part I 41*, pages 413–428. Springer.
- 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*, pages 22–31.
- Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in english. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323.
- 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*, pages 226–241.
- Yiran Chen, Pengfei Liu, Ming Zhong, Zi-Yi Dou, Danqing Wang, Xipeng Qiu, and Xuan-Jing Huang. 2020. Cdevalsumm: An empirical study of cross-dataset evaluation for neural summarization systems. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3679–3691.
- Diego de Vargas Feijó and Viviane Pereira Moreira. 2018. Rulingbr: A summarization dataset for legal texts. In *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13*, pages 255–264. Springer.
- Mohamed Elaraby and Diane Litman. 2022. Arglegalsumm: Improving abstractive summarization of legal documents with argument mining. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6187–6194.
- Erwin Filtz, María Navas-Loro, Cristiana Santos, Axel Polleres, and Sabrina Kirrane. 2020. Events matter: Extraction of events from court decisions. *Legal Knowledge and Information Systems*, pages 33–42.
- Lea Frermann and Alexandre Klementiev. 2019. Inducing document structure for aspect-based summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6263–6273.
- Claire Grover, Ben Hachey, and Ian Hughson. 2004. The holj corpus. supporting summarisation of legal texts. In *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora*, pages 47–54.
- 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)*, pages 708–719.
- Mandy Guo, Joshua Ainslie, David C Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. Longt5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736.
- Ivan Habernal, Daniel Faber, Nicola Recchia, Sebastian Bretthauer, Iryna Gurevych, Indra Spiecker genannt Döhmann, and Christoph Burchard. 2023. Mining legal arguments in court decisions. *Artificial Intelligence and Law*, pages 1–38.
- Hiroaki Hayashi, Prashant Budania, Peng Wang, Chris Ackerson, Raj Neervannan, and Graham Neubig. 2021. Wikiasp: A dataset for multi-domain aspect-based summarization. *Transactions of the Association for Computational Linguistics*, 9:211–225.
- Maor Ivgi, Uri Shaham, and Jonathan Berant. 2023. Efficient long-text understanding with short-text models. *Transactions of the Association for Computational Linguistics*, 11:284–299.
- Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *EACL 2021-16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 874–880. Association for Computational Linguistics.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Sayali Kulkarni, Sheide Chammas, Wan Zhu, Fei Sha, and Eugene Ie. 2020. Aquamuse: Automatically generating datasets for query-based multi-document summarization. *arXiv preprint arXiv:2010.12694*.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Rui Meng, Khushboo Thaker, Lei Zhang, Yue Dong, Xingdi Yuan, Tong Wang, and Daqing He. 2021. Bringing structure into summaries: a faceted summarization dataset for long scientific documents. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1080–1089.

- Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.
- Rodriguez-Doncel Victor Navas-Loro, Maria. 2022. Whenthefact: Extracting events from european legal decisions. In *Legal Knowledge and Information Systems*, pages 219–224. IOS Press.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Prakash Poudyal, Teresa Gonçalves, and Paulo Quaresma. 2019. Using clustering techniques to identify arguments in legal documents. *ASAIL@ ICAIL*, 2385.
- Prakash Poudyal, Jaromír Šavelka, Aagje Ieven, Marie Francine Moens, Teresa Goncalves, and Paulo Quaresma. 2020. Echr: Legal corpus for argument mining. In *Proceedings of the 7th Workshop on Argument Mining*, pages 67–75.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.
- Tyss Santosh, Shanshan Xu, Oana Ichim, and Matthias Grabmair. 2022. Deconfounding legal judgment prediction for european court of human rights cases towards better alignment with experts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1120–1138.
- Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.
- Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. Multi-lexsum: Real-world summaries of civil rights lawsuits at multiple granularities. *Advances in Neural Information Processing Systems*, 35:13158–13173.
- 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*, pages 1048–1064.
- Santosh Tyss, Oana Ichim, and Matthias Grabmair. 2023a. Zero-shot transfer of article-aware legal outcome classification for european court of human rights cases. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 593–605.
- Santosh Tyss, Marcel Perez San Blas, Phillip Kemper, and Matthias Grabmair. 2023b. Leveraging task dependency and contrastive learning for case outcome classification on european court of human rights cases. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1103–1103.
- Alex Wang, Richard Yuanzhe Pang, Angelica Chen, Jason Phang, and Samuel Bowman. 2022. Squality: Building a long-document summarization dataset the hard way. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1139–1156.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.
- Kristian Woodsend and Mirella Lapata. 2012. Multiple aspect summarization using integer linear programming. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 233–243.
- Haibing Wu, Yiwei Gu, Shangdi Sun, and Xiaodong Gu. 2016. Aspect-based opinion summarization with convolutional neural networks. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 3157–3163. IEEE.
- Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. Primera: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263.

Shanshan Xu, Leon Staufer, Santosh T.y.s.s, Oana Ichim, Corina Heri, and Matthias Grabmair. 2023a. VECHR: A dataset for explainable and robust classification of vulnerability type in the European court of human rights. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11738–11752, Singapore. Association for Computational Linguistics.

Shanshan Xu, Santosh T.y.s.s, Oana Ichim, Isabella Risini, Barbara Plank, and Matthias Grabmair. 2023b. From dissonance to insights: Dissecting disagreements in rationale construction for case outcome classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9558–9576, Singapore. Association for Computational Linguistics.

Xianjun Yang, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Xiaoman Pan, Linda Petzold, and Dong Yu. 2022. Oasum: Large-scale open domain aspect-based summarization. *arXiv preprint arXiv:2212.09233*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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*, pages 163–172.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, et al. 2021. Qmsum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921.

Yang Zhong and Diane Litman. 2022. Computing and exploiting document structure to improve unsupervised extractive summarization of legal case decisions. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 322–337.

A. Appendix

A.1. Models Performance Analysis

Tables 4 and 5 provided the detailed results on generalizability to new aspects and sensitivity to aspects for the facts and law variants, across all the models respectively.

Models	Seen				Unseen				BLEU
	R-1	R-2	R-L	BS	R-1	R-2	R-L	BS	
LED	40.73	15.78	26.87	88.04	40.57	15.13	26.69	88.12	39.4
PRIMERA	38.82	15.64	25.83	87.69	38.64	13.38	24.94	87.89	32.78
Long-T5	41.12	15.91	26.78	88.52	40.91	15.62	26.63	88.28	61.42
SLED-BART	41.73	19.38	29.81	88.61	36.67	12.6	24.04	87.6	57.1
Unlim.-BART	41.94	19.67	29.46	88.66	38.25	13.77	24.57	88.02	34.94

Table 4: Anlysis on the facts varirant of LexAbSumm. 'Seen' and 'Unseen' column depict new aspect generalizability, 'BLEU' column depict aspect sensitivity.

Models	Seen				Unseen				BLEU
	R-1	R-2	R-L	BS	R-1	R-2	R-L	BS	
LED	47.52	28.91	35.36	88.49	46.58	25.51	32.31	88.24	53.01
PRIMERA	45.28	21.65	28.71	87.09	45.98	21.72	28.71	87.35	27.01
Long-T5	47.8	29.53	36.09	88.79	46.91	25.91	32.53	88.53	86.56
SLED-BART	50.66	32.25	37.51	89.32	48.35	28.15	33.98	88.76	41.35
Unlim.-BART	51.44	31.86	37.52	89.09	50.1	29.3	34.31	88.73	61.44

Table 5: Anlysis on the Law varirant of LexAbSumm. 'Seen' and 'Unseen' column depict new aspect generalizability, 'BLEU' column depict aspect sensitivity.