Check for updates

ACM DIGITAL LIBRARY | Association for Computing Machinery | acm open

Latest updates: https://dl.acm.org/doi/10.1145/3769126.3769206

SHORT-PAPER

Exploring Metric Correlations for Legal Text Summarization Evaluation

PDF Download
3769126.3769206.pdf
15 January 2026
Total Citations: 0
Total Downloads: 0

Published: 16 June 2025

Citation in BibTeX format

ICAIL 2025: 20th International Conference on Artificial Intelligence and Law
June 16 - 20, 2025
IL, Chicago, USA

ICAIL '25: Proceedings of the Twentieth International Conference on Artificial Intelligence and Law (June 2025)
https://doi.org/10.1145/3769126.3769206
ISBN: 9798400719394

Exploring Metric Correlations for Legal Text Summarization Evaluation

Martim Zanatti
INESC-ID
Lisboa, Portugal
Instituto Superior Técnico
Lisboa, Portugal
martim.zanatti@tecnico.ulisboa.pt

Ricardo Ribeiro
Iscte
Lisboa, Portugal
INESC-ID
Lisboa, Portugal
ricardo.ribeiro@inesc-id.pt

H. Sofia Pinto
INESC-ID
Lisboa, Portugal
Instituto Superior Técnico
Lisboa, Portugal
sofia@inesc-id.pt

Abstract

The rapid advancements in legal text summarization have not been matched by equivalent progress in evaluation metrics capable of assessing the quality of legal summaries. Traditional evaluation approaches, such as **ROUGE**, remain widely used despite their inability to capture semantic fidelity. While more recent metrics focus on semantic evaluation, their applicability to legal summarization has not been thoroughly tested, and their performance is highly dependent on embedding models and computational resources, particularly for long and complex legal texts. Furthermore, the absence of publicly available datasets with expert annotations hinders the development and validation of domain-specific evaluation methods. In this paper, we address these challenges by introducing the first publicly available dataset of Portuguese legal summaries, annotated by legal experts across multiple dimensions such as Coherence and Relevance. We use this dataset to systematically evaluate several recent evaluation metrics, comparing their performance against **ROUGE**, the standard metric for summarization tasks. Our analysis, based on Spearman correlation with human judgments, reveals that **ROUGE-2** maintains the highest correlation across almost every evaluated dimension, outperforming more recent metrics, including semantic-based approaches. These results emphasize the challenges of adapting new evaluation frameworks to the legal domain and underscore the need for further research into metrics that can better capture domain-specific requirements.

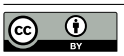
CCS Concepts

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

Keywords

Legal Automatic Evaluation, Semantic similarity, Lexical similarity, Spearman Correlation, Expert Evaluation Dataset

ACM Reference Format:
Martim Zanatti, Ricardo Ribeiro, and H. Sofia Pinto. 2025. Exploring Metric Correlations for Legal Text Summarization Evaluation. In *20th International Conference on Artificial Intelligence and Law (ICAIL 2025), June 16–20, 2025, Chicago, IL, USA*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3769126.3769206



This work is licensed under a Creative Commons Attribution 4.0 International License.
ICAIL 2025, Chicago, IL, USA
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1939-4/25/06
https://doi.org/10.1145/3769126.3769206

1 Introduction and Motivation

Despite the significant growth in legal text summarization, initially driven by the advent of transformers [16] and more recently accelerated by the development of Large Language Models (LLMs) [1], there has not been a corresponding advancement in evaluation metrics capable of accurately assessing the inherent quality of legal summaries. ROUGE [7], a straightforward metric that measures overlap by matching n-grams between candidate and reference summaries, remains the most commonly used metric for evaluating legal text summarization.

In other areas of natural language processing, such as machine translation, significant progress has been made in evaluation metrics. For example, COMET [11], a neural evaluation model, has been proposed as a replacement for traditional metrics like BLEU [10], providing a more nuanced and context-aware assessment of translation quality. In this context, Krubi'nski and Pecina [6] explored adapting the principles underlying COMET to develop a similar framework, COMES, for summarization evaluation. However, the absence of a standardized framework for collecting human evaluations, coupled with the scarcity of annotated datasets, significantly complicates this task.

The challenges are even more acute in the context of legal texts, which are often extensive [15]. This length presents substantial processing demands for models akin to those in the COMET framework, as these models require considerable computational resources to process entire documents. Moreover, evaluation models in machine translation tend to generalize more effectively across domains, as they are less dependent on subject-specific nuances. This underscores an additional challenge in creating universal evaluation metrics for summarization, particularly for complex and specialized domains like legal texts.

Although ROUGE remains the most widely used evaluation metric for text summarization, several alternative evaluation methods have been proposed in recent years to address its limitations. To the best of our knowledge none of these approaches have been applied to the field of legal summarization. In fact, they all rely on the CNN/DailyMail News dataset[1] or its variants, which is a common choice due to its size and extensive annotations across dimensions such as Coherence, Consistency, and other quality indicators.

One of the primary obstacles to advancing evaluation methods for legal summarization is the absence of annotated datasets across languages and legal systems. This deficiency not only hampers the development of reliable, domain-specific evaluation metrics but

[1] https://www.kaggle.com/datasets/gowrishankarp/newspaper-text-summarization-cnn-dailymail

also reduces the effectiveness of general-purpose methods when applied to the legal domain. Creating such datasets is inherently challenging and resource-intensive, requiring the expertise of legal professionals with a comprehensive understanding of the specific legal system under consideration. Their involvement is crucial to ensure that annotations capture the intricacies of legal reasoning and adhere to the judicial standards and practices of the respective jurisdiction.

This work introduces the first publicly available dataset of Portuguese legal summaries, annotated by experts across multiple dimensions such as Coherence and Consistency (see Section 3). As recent evaluation metrics remain largely untested in legal summarization, we aim to assess their suitability for this domain. Our main goal was to evaluate the quality of Portuguese Supreme Court summaries using various metrics and analyze their correlation with expert annotations. Notably, despite the rise of semantics-based metrics, **ROUGE-2** showed the strongest correlation with human evaluations across all dimensions.

The remainder of the article is organized as follows: Section 2 reviews the recent methods for evaluating text summarization. Section 3 introduces our dataset and its characteristics. Section 4 explains the necessary adaptations to use the evaluated evaluation metrics. Section 5 shows the correlation results of the several metrics tested in the Legal Summarization dataset and finally, Section 6 presents our conclusions and future work.

2 Related Work

The goal of this paper is to understand how well evaluation metrics align with expert evaluations of Portuguese legal summaries. To the best of our knowledge, this is the first work to provide this kind of analysis for the legal domain.

BERTScore is an unsupervised metric introduced by Zhang and et al. [20], designed to measure semantic equivalence between a reference and a hypothesis by leveraging token-level matching functions in a distributed representation space. Further highlighting the need for enhanced metrics in abstractive summarization, Zhang and et al. [19] developed ROUGE-SEM, a hybrid evaluation metric combining ROUGE with semantic information. ROUGE-SEM measures the lexical aspects of a summary using ROUGE, while the semantic component is assessed with a Siamese-BERT model to evaluate the similarity between the candidate summary and the reference summary.

Taking a novel approach, Yuan et al. [17] framed text evaluation as a text generation problem, using BART, an encoder-decoder pre-trained model. Their method evaluates four dimensions: Faithfulness ($s \rightarrow h$), which measures the likelihood of generating the hypothesis from the source text to assess alignment with the original content; Precision ($r \rightarrow h$), evaluating the likelihood of generating the hypothesis from the reference summary for precision-focused scenarios; Recall ($h \rightarrow r$), which assesses the ease of generating the reference from the hypothesis; and F1-score ($r \leftrightarrow h$), which quantifies the semantic similarity between reference and generated texts.

Finally, Krubi'nski and Pecina [6] adapted the COMET framework to develop COMES, a model that uses a stack of feed-forward layers to output scores across four quality dimensions: Coherence, Consistency, Fluency, and Relevance. COMES takes as input the source document, reference summary, and hypothesis summary to comprehensively evaluate summary quality.

Summing up, existing evaluation metrics for summarization exhibit varying strengths in capturing lexical, semantic, and contextual nuances. While metrics such as **ROUGE** and **BERTScore** provide valuable insights into lexical and semantic equivalence, advancements like **ROUGE-SEM** integrate semantic understanding more effectively to address the limitations of traditional approaches. On the other hand, methods like **BARTScore** and **COMES** represent a paradigm shift by leveraging pre-trained models and domain-specific architectures to assess multiple quality dimensions.

Despite these advances, to the best of our knowledge, no prior work has systematically analyzed how well these metrics align with expert evaluations in the legal domain. This gap underscores the importance of our study, which aims to fill this void by evaluating and contextualizing these metrics within the specific requirements of Portuguese legal summaries.

3 Dataset

The dataset[2] used in this study was developed as part of the AI & Law Master's program,[3] involving 5 experts who evaluated multiple summaries of 12 Portuguese Supreme Court of Justice (STJ) judgments. These judgments ranged from 696 to 13,397 tokens in length. Each expert read all judgments and rated the corresponding summaries from 1 (lowest) to 5 (highest) across six dimensions: Coherence (Ch), assessing sentence quality and flow; Consistency (Cs), checking factual alignment; Relevance (Rv), evaluating content coverage; Semantic Correspondence (SC), measuring semantic similarity; Accuracy (A), reflecting overall fidelity; and Replacement (Rp), judging whether the summary could substitute the original summary.

For the expert annotation process, a total of 69 distinct summaries were used, categorized as follows: 10 summaries authored by judges, 29 extractive summaries generated using the LexRank algorithm [3], and 30 abstractive summaries produced by the Llama 3.1 8B model [2].

Each judgment in the dataset is paired with one to three extractive summaries, all generated using the LexRank algorithm. Variations arise from three factors: the **embedding model** used to represent paragraphs, the **legal sections** selected as **context** for similarity computation, and the **sections** chosen for **extraction**. These differences result in distinct summaries, though in some cases, variations yield identical outputs, leading to a total of 29 summaries across the 12 judgments.

The abstractive summaries were derived from two judgments. For these summaries, each expert designed three distinct prompts and evaluated the generated summary corresponding to each prompt. Furthermore, for these two judgments, only one expert assessed an extractive summary for each judgment.

As a result of the conducted study, we obtained 37 summaries (10 human-generated and 27 extractive summaries) that were evaluated by all 5 experts, 2 extractive summaries that were evaluated by a single expert, and 6 abstractive summaries evaluated by each expert.

[2] https://huggingface.co/datasets/MartimZanatti/LegalSumm
[3] https://ai-and-law-lisbon.com/llm

In Section 1, we highlighted that the length of legal texts presents significant challenges for recent evaluation approaches, such as BERTScore, which imposes a limit of 512 tokens each for the source text and the hypothesis text. Legal documents, such as judgments and summaries, are often substantially longer than news articles [4, 5, 15]. This makes it challenging to encapsulate all the necessary text to effectively evaluate the quality of the hypothesis summary.

As previously noted, varying the judgment sections used as context and for extraction leads to different summaries for the same case. Judgments from the Portuguese Supreme Court of Justice are generally divided into four main sections: Report, Facts, Law, and Decision. Discussions with STJ judges revealed that the Law section is the most relevant for generating extractive summaries (see [18] for details on the segmentation of these judgments).

For comparison, the widely used CNN-DailyMail dataset has a mean of 781 tokens per article and 56 per summary.[4] In contrast, the Portuguese legal judgments in this study average 4069 tokens per judgment and 226.7 per summary. This substantial difference underscores the greater complexity and length of legal texts, posing challenges for summary evaluation using standard approaches.

4 Adaptations for Evaluation Metrics

This section details the adaptation of **BERTScore**, **ROUGE-SEM**, and **BARTScore** to meet the specific needs of our study. **BERTScore** and **BARTScore** were chosen for their widespread use as baselines and state-of-the-art metrics in evaluation research [6, 19], particularly for assessing semantic alignment and faithfulness. In contrast, **ROUGE-SEM** is a more recent contribution that, despite limited validation across datasets, introduces semantic awareness and holds promise for evaluating summaries in the Portuguese legal domain. Its inclusion allows exploration of its potential in a novel and complex context.

Since there are three types of summaries — **human-written**, **extractive**, and **abstractive** — we compute the evaluation metrics using different combinations of hypotheses and references. For **human-written summaries**, we compare them against the Law section of the corresponding judgments, as it is the most relevant for summary production and helps reduce the overall token count. For **extractive summaries**, we use the human-written summaries as reference. Finally, for **abstractive summaries**, we compute the metrics against both the human-written summaries and the Law section of the respective judgments.

All these metrics depend on the embedding model employed to capture the semantic properties of the summaries. To ensure a fair comparison, we use the same embedding model across all evaluation metrics. Specifically, we employ legal-BERTImbau [9] as the embedding model,[5] a specialized variant of BERTImbau large [12], which has been fine-tuned for the Portuguese legal domain.

4.1 BERTScore

One of the evaluation metrics employed in this study is BERTScore. BERTScore provides three outcomes: **Precision**, which quantifies the degree of semantic alignment from the hypothesis to the reference text; **Recall**, which measures the extent to which the reference text aligns semantically with the hypothesis; and **F1-Score**, the harmonic mean of Precision and Recall. As stated in Section 3, BERTScore imposes a limit of 512 tokens each for the source text and the hypothesis text. To address this limitation, we divide the text into chunks when necessary, calculate the BERTScore for each chunk, computing the final result as the mean across all chunks.

4.2 ROUGE-SEM

The ROUGE-SEM approach consists of three components: lexical evaluation using ROUGE, semantic evaluation via a Siamese-BERT model trained on English news data, and a rewriting phase using Google Translate.[6] The semantic model, based on contrastive learning, is ill-suited for European Portuguese legal texts. To address this, we adopt BERTScore and our embedding model (Section 4), applying the same strategy as in Section 4.1 to handle documents exceeding 512 tokens. Summaries are categorized as **Good**, **Pearl**, **Glass**, or **Bad** based on lexical (threshold = 0.2839) and semantic (threshold = 0.55) scores. Pearl and glass summaries are rewritten: we translate PT–EN using Google Translate and back EN–PT using DeepL.[7] Google Translate sometimes returns Brazilian variants or near-identical output, limiting diversity. Rewritten summaries are evaluated with ROUGE. For pearls, the version with the higher ROUGE score is selected; for glass, the lower. Rewriting usually yields lower ROUGE scores, suggesting this step needs improvement. Final scores are scaled: $1.5 \times ROUGE$ for pearls and $0.5 \times ROUGE$ for glass.

4.3 BARTScore

As discussed in Section 2, BARTScore is an evaluation metric that frames text evaluation as a text generation problem. Following the explanation in Section 4, we assess *Faithfulness* (fa) as the evaluation between **human-written summaries** and **abstractive summaries** with respect to the **document source**. For the remaining dimensions — Precision, Recall, and F1 Score — we evaluate the **extractive summaries** and **abstractive summaries** against the **human-written summaries**.

Tang and et al. [14] extended the mBART model [8] to support 50 languages,[8] including Portuguese. For Portuguese, the TED58 dataset was used, consisting of a training set with 49446 sentences. This BART model enforces a limit of 1024 tokens for both the source text and the hypothesis text. To address this constraint, the same approach applied to other evaluation metrics was adopted. Specifically, the text is divided into chunks when necessary, the BARTScore is calculated for each chunk, and the final score is obtained by averaging the scores across all chunks.

5 Results

To assess the correlation between the expert evaluations of all dimensions, as outlined in Section 3, and the several evaluation metrics scores presented in Section 4, we employ the Spearman Correlation [13].

[4] https://www.kaggle.com/datasets/gowrishankarp/newspaper-text-summarization-cnn-dailymail
[5] https://huggingface.co/stjiris/bert-large-portuguese-cased-legal-tsdae-gpl-nli-sts-v1
[6] https://translate.google.com/?sl=en&tl=pt-PT&op=translate
[7] https://www.deepl.com/
[8] https://huggingface.co/facebook/mbart-large-50

Table 1 present the correlation between expert evaluations and various evaluation metrics across all dimensions, considering all summaries. The three highest scores for each dimension are highlighted in bold, while scores displayed in italics correspond to cases where the p-value exceeds a threshold of 0.05.

Metric	Ch	Cs	Rv	SC	A	Rp	Mean Score
ROUGE-1 r	**0.37**	**0.39**	**0.42**	**0.45**	**0.5**	**0.46**	**0.7**
ROUGE-1 p	*-0.01*	*-0.04*	*-0.05*	*-0.05*	*-0.12*	*-0.07*	0.26
ROUGE-1 f1	*0.05*	*0.03*	*0.03*	*0.02*	*-0.05*	*0.004*	0.32
ROUGE-2 r	**0.46**	**0.48**	**0.48**	**0.52**	**0.55**	**0.55**	0.42
ROUGE-2 p	0.23	0.21	0.21	0.21	0.17	0.24	0.14
ROUGE-2 f1	0.34	0.34	0.34	0.34	0.32	0.38	0.18
ROUGE-L r	0.34	0.4	0.37	0.41	0.45	**0.46**	**0.68**
ROUGE-L p	*-0.02*	*-0.04*	-0.06	*-0.13*	*-0.06*	*0.04*	0.29
ROUGE-L f1	*0.04*	*0.05*	*0.02*	*0.03*	*-0.05*	*0.02*	0.36
BERTScore p	0.29	0.29	0.31	0.33	0.27	0.26	0.61
BERTScore r	0.27	0.27	0.23	0.22	0.16	0.24	0.56
BERTScore f1	0.28	0.28	0.27	0.26	0.19	0.24	0.59
ROUGE-SEM-1 r	0.35	0.37	0.41	0.4	0.45	0.39	**0.74**
ROUGE-SEM-1 p	*0.05*	*0.04*	*0.03*	*0.003*	*-0.05*	*-0.005*	0.28
ROUGE-SEM-1 f1	*0.12*	*0.13*	*0.13*	*0.08*	*0.04*	*0.08*	0.34
ROUGE-SEM-2 r	**0.43**	**0.46**	**0.47**	**0.45**	**0.51**	**0.5**	0.47
ROUGE-SEM-2 p	0.25	0.28	0.26	0.22	0.21	0.28	0.15
ROUGE-SEM-2 f1	0.36	**0.39**	0.37	0.34	0.35	0.4	0.2
ROUGE-SEM-L r	0.35	0.38	0.4	0.39	0.44	0.39	0.73
ROUGE-SEM-L p	*0.06*	*0.07*	*0.05*	*0.02*	*-0.03*	*0.02*	0.31
ROUGE-SEM-L f1	*0.13*	*0.17*	*0.14*	*0.09*	*0.06*	*0.1*	0.39
BARTScore p	*-0.007*	*-0.1*	*-0.2*	*-0.1*	*-0.17*	*-0.11*	5.76×10^{-9}
BARTScore r	*0.12*	*0.14*	*0.12*	*0.09*	*0.11*	*0.08*	1.04×10^{-8}
BARTScore f1	*-0.02*	*-0.09*	*-0.2*	*-0.11*	*-0.17*	*-0.13*	5.28×10^{-9}
BARTScore fa	0.31	0.31	0.37	0.32	*0.23*	*0.29*	5.02×10^{-9}

Table 1: Spearman Correlation between evaluation metrics and human evaluation. The Mean Score presented in the table represents the average score of the corresponding evaluation metric.

ROUGE-2 recall demonstrated a stronger correlation across all dimensions, significantly outperforming BERTScore and BARTScore, which evaluates the semantic aspects of summaries. ROUGE-SEM-2 recall, which combines a semantic score with ROUGE-2, achieved results that were highly comparable to those obtained with ROUGE-2 recall. In the legal domain, the use of technical language and the repetition of specific terms is common, leading to high overlap of words in summaries. ROUGE-1, which measures the overlap of unigrams (individual words), might overestimate the summary's quality, as the repetition of terms does not necessarily reflect semantic adequacy. In contrast, ROUGE-2, which considers the overlap of bigrams (pairs of consecutive words), is more sensitive to the combination and context of words. As a result, ROUGE-2 is more strongly correlated with summary quality, as it better captures the structure and meaning of sentences, offering a more accurate reflection of the relevance and cohesion of content in the legal domain. The ROUGE-2 results are consistent with those reported by Zhang and et al. [19] for the CNN/Daily dataset.

When examining the correlation across the different dimensions, the results reveal unexpected patterns. Given that ROUGE is a lexical overlap-based metric, it was anticipated to exhibit stronger correlations with expert evaluations in **Ch**, **Cs**, **Rv**, and **SC** compared to **A** and **Rp**. Specifically, moderate correlations were expected for **Ch** and **Cs**, as maintaining logical flow and factual consistency often relies on preserving the wording of the source text. For **Rv** and **SC**, higher correlations were anticipated, particularly with ROUGE Recall, as these dimensions emphasize capturing key information, aligning with ROUGE's recall-focused approach. However, the highest correlations were observed in the **A** and **Rp** dimensions, which theoretically rely more on deeper semantic fidelity rather than strict lexical matches, making these results particularly surprising. ROUGE precision is expected to correlate better with **Ch** and **A**, as it rewards more concise and well-structured summaries. Higher precision reflects the ability to preserve key details while minimizing extraneous information, which aligns with the requirements for maintaining coherence and accuracy in summaries. However, this expected correlation does not occur.

BERTScore, being a semantic similarity metric, is expected to correlate more strongly with dimensions requiring deeper semantic alignment rather than surface-level similarity. However, this does not occur; BERTScore's correlation with **A** and **Rp** is lower compared to almost all other dimensions across all BERTScore metrics. **SC** should also exhibit a strong correlation due to BERTScore's ability to capture nuanced meanings beyond word overlap; however, this is only observed for BERTScore precision.

ROUGE-SEM-2 demonstrated stronger correlations than ROUGE-SEM-1 and ROUGE-SEM-L, while still falling short of ROUGE-2 across nearly all evaluated dimensions, consistent with the findings reported by its authors. However, all metrics exhibited higher correlations when applied to the CNN/Daily dataset compared to our dataset. This discrepancy underscores the increased complexity inherent in evaluating legal summarization tasks.

Although the model used in the BARTScore evaluation metric for the Portuguese language was trained on a limited number of instances and was not specifically adapted to the Portuguese legal domain, the results remain promising. For BARTScore faithfulness scores, certain dimensions, such as **Cs**, **Rv** and **Rp**, demonstrated a higher correlation with human experts compared to BERTScore, which leverages a BERT model specifically adapted to the Portuguese legal domain. We anticipate that with a BART model tailored to the Portuguese legal domain, the correlation could be significantly improved, as such a model would better capture the semantic intricacies of the legal domain.

In general, the results indicate that there is still significant room for improvement, as the highest correlation scores range only from 0.46 to 0.55 across all dimensions. Based on the findings, ROUGE remains the metric with the strongest correlation with human evaluations for this dataset. However, ROUGE-SEM-2 achieved comparable results, albeit at a higher computational cost.

6 Conclusion and Future Work

In this paper, we present a novel dataset annotated by experts across multiple dimensions for Portuguese legal summarization, which, to the best of our knowledge, is the first publicly available dataset dedicated to this task. Using Spearman correlation, we evaluate several recent evaluation metrics and compare their performance with the standard metric, **ROUGE**. The results reveal significant room for improvement. Despite recent advances in evaluation metrics designed to capture semantic information, **ROUGE** still achieves better correlation with expert evaluations. The performance of newer metrics heavily relies on the ability of embedding models to

effectively capture semantic nuances, a challenge exacerbated by the considerable length of legal documents.

For future work, we aim to explore different embedding models, which are crucial for modern evaluation metrics such as **ROUGE-SEM** and **BERTScore**. Given that **ROUGE-2** demonstrated stronger correlations compared to **ROUGE-1** due to its focus on higher-order n-grams, it would be worthwhile to explore correlations using **ROUGE-SU4**. This metric incorporates 4-words skips, enabling it to better account for sentence reordering, while its inclusion of unigrams maintains a balance between higher-level semantic matching and basic lexical overlap.

COMES has shown promising results for English news summarization, but its effectiveness in non-English contexts remains uncertain. Its adaptation relies on automatic translation of annotated English data, which may hinder semantic evaluation in complex domains like legal texts. Moreover, its multilingual assessment was limited to only two dimensions — *Focus* and *Coverage* — neglecting key aspects such as *Coherence* and *Consistency*, which are central to our dataset. **COMES** also underperformed compared to established metrics like **ROUGE** and **BERTScore**, and its evaluation was based on small datasets (270 instances per language), limiting its generalizability to more specialized domains.

Given these limitations, we chose to focus on more established metrics – **BERTScore**, and **BARTScore** – and propose, as future work, the development of a **COMES**-based architecture for evaluating Portuguese legal summaries. Additionally, we propose testing the BARTScore evaluation metric with a BART model specifically adapted to the Portuguese legal domain as part of future work.

This will allow us to gain a broader perspective on recent evaluation approaches. We hope that this comprehensive evaluation will provide a deeper understanding of the strengths and weaknesses of existing metrics, enabling the development of an evaluation metric capable of overcoming these limitations.

In parallel, we have also conducted preliminary experiments using GPT-based evaluation and plan to present these results in future work. Building on this, we intend to explore the use of other LLMs and prompt engineering strategies to better understand their potential for evaluating legal summaries in Portuguese.

Acknowledgments

This research was supported by Fundação para a Ciência e Tecnologia (FCT), through the INESC-ID multi-annual funding with reference DOI:10.54499/UIDB/50021/2020. This research was also supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for Responsible AI).

We would also like to thank the five students of the AI & Law Master's program[9] for their invaluable contribution to the creation of the dataset.

References

[1] Lochan Basyal and Mihir Sanghvi. 2023. Text Summarization Using Large Language Models: A Comparative Study of MPT-7b-instruct, Falcon-7b-instruct, and OpenAI Chat-GPT Models. *CoRR* abs/2310.10449 (2023). arXiv:2310.10449

[2] Abhimanyu Dubey and et al. 2024. The Llama 3 Herd of Models. *CoRR* abs/2407.21783 (2024). arXiv:2407.21783

[3] Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research* 22 (2004), 457–479.

[4] Ambedkar Kanapala, Sukomal Pal, and Rajendra Pamula. 2019. Text summarization from legal documents: a survey. *Artif. Intell. Rev.* 51, 3 (2019), 371–402.

[5] Marios Koniaris, Dimitris Galanis, and et al. 2023. Evaluation of Automatic Legal Text Summarization Techniques for Greek Case Law. *Inf.* 14, 4 (2023), 250.

[6] Mateusz Krubi'nski and Pavel Pecina. 2022. From COMET to COMES - Can Summary Evaluation Benefit from Translation Evaluation?. In *Proceedings of the 3rd Workshop on Evaluation and Comparison of NLP Systems, Eval4NLP 2022, Online, November 20, 2022*, Daniel Deutsch and et al. (Eds.). Association for Computational Linguistics, 21–31.

[7] 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.

[8] Yinhan Liu and et al. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Trans. Assoc. Comput. Linguistics* 8 (2020), 726–742.

[9] Rui Melo, Pedro A. Santos, and João Dias. 2023. A Semantic Search System for the Supremo Tribunal de Justiça. In *Progress in Artificial Intelligence - 22nd EPIA Conference on Artificial Intelligence, EPIA 2023, Faial Island, Azores, September 5-8, 2023, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 14116)*, Nuno Moniz and et al. (Eds.). Springer, 142–154.

[10] 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, July 6-12, 2002, Philadelphia, PA, USA*. ACL, 311–318.

[11] Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber and et al. (Eds.). Association for Computational Linguistics, 2685–2702.

[12] Fábio Souza, Rodrigo Frassetto Nogueira, and Roberto de Alencar Lotufo. 2020. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Intelligent Systems - 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20-23, 2020, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 12319)*, Ricardo Cerri and Ronaldo C. Prati (Eds.). Springer, 403–417.

[13] C. Spearman. 1904. The Proof and Measurement of Association Between Two Things. *American Journal of Psychology* 15 (1904), 88–103.

[14] Yuqing Tang and et al. 2020. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning. *CoRR* abs/2008.00401 (2020). arXiv:2008.00401

[15] Howard R. Turtle. 1995. Text Retrieval in the Legal World. *Artif. Intell. Law* 3, 1-2 (1995), 5–54.

[16] Ashish Vaswani and et al. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon and et al. (Eds.). 5998–6008.

[17] Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating Generated Text as Text Generation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021*, Marc'Aurelio Ranzato and et al. (Eds.). 27263–27277.

[18] Martim Zanatti, Ricardo Ribeiro, and H. Sofia Pinto. 2024. Segmenting Model for Portuguese Judgments. In *Progress in Artificial Intelligence - 23nd EPIA Conference on Artificial Intelligence, EPIA 2024, Viana do Castelo September 3-6, 2024.*

[19] Ming Zhang and et al. 2024. ROUGE-SEM: Better evaluation of summarization using ROUGE combined with semantics. *Expert Syst. Appl.* 237, Part A (2024), 121364.

[20] Tianyi Zhang and et al. 2019. BERTScore: Evaluating Text Generation with BERT. *CoRR* abs/1904.09675 (2019). arXiv:1904.09675

[9] https://ai-and-law-lisbon.com/llm