

Using Sentence Embedding Techniques for Enhancing Terms-of-Service Text Summarization

Harry Peach¹, Nicolay Rusnachenko¹, Mayank Baraskar¹, and Huizhi Liang¹

School of Computing, Newcastle University, Newcastle, United Kingdom
harrypeach1999@gmail.com, {nicolay.rusnachenko, m.p.baraskar2,
huizhi.liang}@newcastle.ac.uk

Abstract. Summarization is useful for extracting salient information from linguistically complex texts. This is especially relevant in the legal domain, where it can be used to make content more accessible to layman readers. A simplified representation can help foster transparency and trust between an organization and individuals. We examine the background of the latest advances in extractive and abstractive summarization approaches. The recent appearance of transformer architecture with a self-attention mechanism has a huge impact on abstractive summarization performance. However, a major limitation of abstractive summarization pertains to constraints on input size. To address these shortcomings, in this paper, we propose a target-oriented sentence embedding classification (SEC) architecture. It is designed specifically for Terms-of-Service (ToS) document summarization and is intended to serve the preliminary text processing for abstractive summarization. The results of experiments conducted under a collection of ToS documents from the service TOS;DR show that SEC model results in average 11% increment across all ROUGE metrics (F-measure) in comparison with other extractive summarizers for significantly short summaries. The application of SEC in general-purpose abstractive summarizers results in models that illustrate increment in ROUGE-2 by 11-12% and equal or better ROUGE-L. We accompany the proposed architecture with the annotation service and complex word simplification modules, formed into a publicly available system¹.

Keywords:

1 Introduction

By their nature, legal texts follow a strict and thorough writing style. This is one of the many reasons that can cause comprehension challenges for ordinary readers. *Terms-of-Service* (ToS) documents are the type of legal documents which represent and describe an agreement between two parties: a service and a user. In order to avoid ambiguity within the agreement, ToS texts are usually long and complex.

¹ <https://github.com/HarryPeach/simplifying-legal-content>

Architecturally, there are several common approaches to reducing text complexity: *text summarization* [1], and *text simplification* from complex words [2]. In terms of the result summary, automatic text summarizations are either *extractive*, *abstractive*, or *hybrid* [1]. Extractive approaches create a summary that uses *verbatim* sentences from the input and drops sentences that are deemed as not salient to the user by relying on semantic analysis of contents [3], similarity metrics [4] (LexRank), etc. Conversely, abstractive approaches process the given text and reword it by keeping the original meaning intact while shortening it. The hybrid approach represents a combination of the two above approaches. *Text simplification* aims to improve text comprehension, especially for layman readers who are not experienced with the intricacies of the particular domain [5].

The recent advances in text-translation and language modeling illustrate a significant research interest towards the *transformer* architecture [6] application for abstractive summarization [7, 8]. Despite the advances, such models have the following limitations: (i) input size constraints, i.e. requiring a shortened representation, (ii) domain-specific fine-tuning, which is highly demanding on computational resources.

In this paper, we propose a text summarization system tailored for the legal domain. Our legal document processing system first applies extractive summarization, which then serves as the foundation for (i) annotation, (ii) abstractive summarization, (iii) text simplification, thereby enabling a more effective context reading. We propose a target-oriented extractive summarization architecture dubbed as *sentence embedding classifier* (SEC) as a preliminary component for abstractive summarizers. We conduct extensive experiments on ToS legal documents and demonstrate higher recall in comparison with LSA and LexRank models, with an average 11% increment across all ROUGE metrics by F-measure for the summaries of 10% volume of the original texts. We show the benefit of the proposed SEC in a hybrid approach by improving the results of the general-purpose language model-based summarization models.

2 Background

Firstly, we provide an overview of related systems, followed by the text summarization approaches that they use. CLAUDETTE [9] is a system for identifying “unfair clauses” in contracts, using conventional machine learning techniques, performed on sentence classification format. The output of the system for a given text is a list of specific unfair reasons, as well as a more general ‘unfairness’ rating. In [10] the authors propose the HECTOR system, which comes to the hybrid approach from the perspective of text transformations in the following order: syntax, discourse, and lexical simplification. Authors dedicate their system to raw texts written in French.

One of the most important considerations when creating an *extractive* model, is that the saved content is presented as-is. In the case of extractive methods, unsupervised learning methods are used to complete summarization. One of the initial approaches used was *Latent Semantic Analysis* (LSA) [3, 11]. This

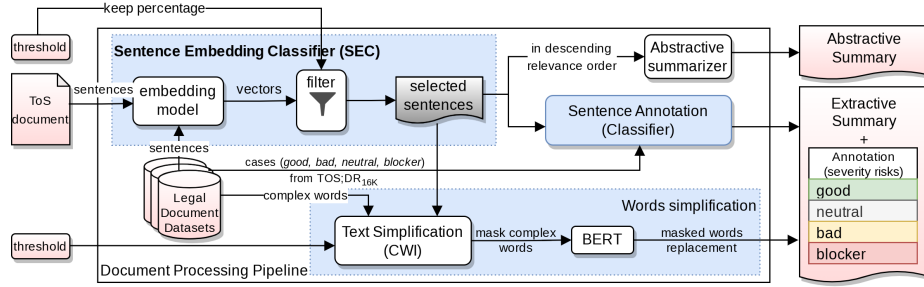


Fig. 1. Term-of-Service text summarization system workflow with Sentence Embedding Classifier (SEC) module for initial salient sentence filtering stage; *bottom curved rectangle* corresponds to the input and result data (documents); *rectangles* are related to text processing components of the workflow; grayshaded filled components corresponds to a data; *blue backgrounded boxes* correspond to target-oriented system components

approach yielded several limitations. For instance, it is only partially able to capture words with multiple meanings (polysemy), due to the fact that every occurrence of a word is assumed to have the same meaning. Extractive summarization methods can also be treated as *classifiers* for the most salient sentences: a summary is treated as a set of sentences, where each sentence could be classified (ranked) by its relevance in the summary [12, 13]. In [4], authors propose the **LexRank** model, with salient sentence selection based on their intra-sentence cosine similarity with other sentences in the text.

As for the abstractive methods, the appearance of the encoder-decoder architecture with self-attention mechanism or *transformer* [6], originally dedicated for the machine translation problem, causes a significant effect with its further applications in a variety of NLP tasks. Being pre-trained on a large amount of data, transformers serve the knowledge about one or multiple languages and are therefore considered as *language models*. The same is applicable to the decoder part of the transformer, with the appearance of **GPT** [14] and further successors such as **Longformer** [15], **T5** [7], and **LongT5** [8]. A combination of the generative component of the GPT architecture and BERT results in **BART** architecture [16]. The latter affects further studies in the so-called domain of *target-oriented transformers*. In [17] authors propose a metric and algorithm of salient sentence selection and masking scheme of the transformers, formed into **PEGASUS** architecture.

To evaluate the model’s effectiveness in summarisation contexts, the researchers tested its efficacy on a variety datasets, including CNN/DailyMail [18], BigPatent [19]. Comparing the LARGE sized transformer models, PEGASUS illustrates the best results CNN/DailyMail PRIMERA for BigPatent, while for the rest of the mentioned datasets, the highest results were achieved² by LongT5 [8].

² According to the related analysis of LARGE-sized models

We propose a ToS text summarization system, in which salient sentence selection is driven by analogies from similar summarized documents via embeddings, rather than existing metrics grounded solely on the content of the document itself [17]. This offers a unique and more contextualized method of content reduction, tailored specifically for abstractive summarization.

3 Terms-Of-Service text summarization system

We present a system for extracting and annotating salient information from Terms of Service (ToS) documents. This system employs a hybrid text summarization approach, leveraging third-party abstractive summarization models. Figure 1 depicts the text processing workflow of the system. When a document is input into this system, two outputs are provided: (i) abstractive summary, (ii) annotated extractive summarization.

Table 1. TOS;DR_{summ} results of legal text summarization models experiment tasks of performance assessment for: (1) extractive summarizers, (2) abstractive summarizers in hybrid mode; best scores across all the methods in each block are **bolded**

Task Model	Type	ROUGE-1(p/r/f)	ROUGE-2(p/r/f)	ROUGE-L(p/r/f)
1 LSA _{10%} limit	extractive	0.324/0.434/0.321	0.080/0.098/0.076	0.165/0.201/0.156
LexRank _{10%} limit	extractive	0.372/0.431/0.321	0.101/0.100/0.080	0.179/0.195/0.149
SEC _{mpnet 10% limit} (our)	extractive	0.395/0.408/0.356	0.111/0.090/0.089	0.181/0.179/0.160
T5 _{base}	abstractive	0.457/0.155/0.218	0.092/0.032/0.045	0.250/0.083/0.116
T5 _{base} over LexRank _{50%} limit	hybrid	0.508/0.175/0.248	0.122/0.039/0.057	0.261/0.088/0.125
2 T5 _{base} over SEC _{mpnet 50% limit} (our)	hybrid	0.480/0.173/0.239	0.129/0.045/0.064	0.276/0.097/0.135
LongT5 _{base}	abstractive	0.491/0.242/0.308	0.108/0.050/0.066	0.239/0.119/0.151
LongT5 _{base} over LexRank _{50%} limit	hybrid	0.496/0.245/0.312	0.110/0.051/0.066	0.244/0.121/0.153
LongT5 _{base} over SEC _{mpnet 50% limit} (our)	hybrid	0.498/0.242/0.311	0.113/0.050/0.067	0.244/0.118/0.152
LEGAL-PEGASUS _{base}	abstractive	0.502/0.259/0.325	0.127/0.069/0.086	0.250/0.129/0.162
ChatGPT-4 _{25.09.23}	abstractive	0.435/0.260/0.310	0.092/0.060/0.070	0.214/0.133/0.156

Extractive Summarizer. In this paper, we consider the Sentence Embedding Classifier (SEC) architecture, illustrated in Figure 1 (SEC block). This architecture involves an embedding model, which aims to provide vectorised representation for: (i) document sentences, (ii) sentences of the legal document dataset collection. The *similarity* represents a cosine distance of the input sentence to a list of embeddings of sentences that are of known importance, created from legal document datasets. The similarity of the vectorised input sentences to the known embeddings is gated against a user-configurable *threshold* which represents a percentage of the relevant sentences to be considered.

Sentence Annotation. We apply a neural network classification model for risk classification assessment by relying on annotated cases to signify the level of risk for the user³: *good* (poses minimal risk, likely beneficial to user), *bad* (poses

³ We consider TOS;DR resource and the related annotation, Section 4

some level of risk or misunderstanding for the user), *blocker* (poses a significant risk), and *neutral* (neither beneficial nor harmful).

Words Simplification. Represents the following components adopted sequentially: complex word annotator, and language model for masked words prediction [20] (Figure 1). For annotation of complex words, we consider resources of *complex word identification* (CWI) [21, 22]. We define complexity level via the predefined *threshold* [23].

Abstractive Summarizer. According to the background analysis, provided in Section 2 we consider transformer-based text summarization models for the related component of the proposed system. The input of the related models represents textual information in the form of a limited sequence of tokens. We trim the original document by a maximum number of available tokens in input, retaining only the initial part of the ToS content. In this paper, we consider pre-trained and publicly available models, described in Section 5.

4 Datasets and Services

In this section, we cover the resources that were used in our experiments. Table 2 illustrates the statistics of the datasets and resources described below. LexGLUE [24] is a benchmark dataset for legal language understanding in English. As part of this benchmark, it provides legal multi-label classification tasks alongside the related sub-datasets: UNFAIR-ToS and LEDGAR.

Table 2. Statistics of the datasets utilized in experiments.

Dataset	#docs	Sentences
TOS;DR _{16K}	16K	13 358
UNFAIR-ToS	50	12 000
LEDGAR	–	80 000
CWIG3G2 [20]	42	500
TOS;DR _{summ} (Raw Texts)	10	1263
TOS;DR _{summ} (Summaries)	10	283

UNFAIR-ToS [9] sentences are only ranked by whether or not they are any unfair terms and how unfair they are. Texts are annotated by legal experts, so the quality of the data is ensured. The dataset contains 50 documents (contracts), with over 12K sentences in total. **LEDGAR** [25] consists of contracts, sourced from the publicly available US Securities and Exchange Commission filings. The dataset represents 80K sentences, separated into 6:1:1 proportions of train, validation, and test folds respectively.

TOS;DR⁴ is an online service that provides annotated Terms of Service and Privacy Policy documents from companies across the web. The annotations classify the severity risks into four classes: *good* (22%), *bad* (37%), *blocker* ($\approx 1\%$),

⁴ <https://tosdr.org>

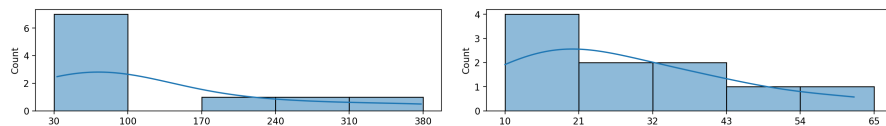


Fig. 2. Histogram and distribution of the TOS;DR_{summ} document lengths in sentences for the *raw texts* (left) and their *summaries* (right); for the most (7/10) raw texts, the overall length does not exceed 100 sentences, while the remaining 3 documents exceed this amount due to the presence of the Privacy Policy (PP)

and *neutral* (40%). We contacted the service administration for the whole list of organizations to create the following datasets: (i) **TOS;DR_{summ}**, which involves five organizations and manual summaries of their terms, (ii) **TOS;DR_{16K}**⁵, which represents ToS texts for 16000 organizations. To establish convention with UNFAIR-ToS with a binary annotation towards every sentence (fair/unfair), we consider TOS;DR document sentence importance by treating the following severity classes⁶: “neutral and good” (fair), “bad” (unfair). We consider a split onto train and test folds with 6:4 proportion. **CWIG3G2** [20] comprises of texts from WikiNews (42 articles), and Wikipedia articles (500 sentences).

The TOS;DR_{summ} consists of 10 documents alongside their summaries, that represent a concatenated list of the most significant cases manually extracted from the texts by experts of TOS;DR platform. The original documents may represent a single paragraph or may represent a couple of the following paragraphs: Terms-of-Use and Privacy Policy. The presence of a Privacy Policy affects the lengths for some documents in TOS;DR_{summ}. The length of summaries in sentences⁷ represent $\approx 22.4\%$ out of the original document (see Table 2). Figure 2 illustrates the 5-bin histogram and distribution plots of the document’s lengths in sentences, separately for the raw texts and their summaries.

5 Experimental Setup

We use TOS;DR and UNFAIR-ToS for relevance sentence assessment in extractive/abstractive modes and their combinations in hybrid models. For SEC results we consider all sentences with similarity values, exceeding the threshold of p .

Text summarization. For extractive approaches, we experiment with LSA, LexRank, and sentence embedding classifier (dubbed as SEC_{mpnet} model from here) based on embeddings, obtained by **all-mpnet-base-v2**⁸ model. We consider NLTK library [26] (**punkt/english** model) for splitting text into sentences.

⁵ <https://github.com/HarryPeach/simplifying-legal-content/blob/main/data/cases.csv>

⁶ *blocker* class of TOS;DR was not considered due to the relatively small amount of the related cases ($\approx 1\%$)

⁷ Details on sentence tokenization in Section 5

⁸ <https://huggingface.co/sentence-transformers/all-mpnet-base-v2>

The sentence embedding model was trained on UNFAIR-ToS and TOS;DR_{16K} (train fold). For the abstractive summarization, the following general-purpose models were selected: T5_{base}⁹ (512 tokens), LongT5_{base}¹⁰ (1K tokens). For comparison reasons, we provide evaluation results of the target-oriented pre-trained LEGAL-PEGASUS¹¹ transformer, alongside with application results for dialogue assistant ChatGPT-4 by OpenAI¹².

Text Annotation. We consider TOS;DR_{16K} dataset for training a fully connected layer four class classification model (set of classes in TOS;DR_{16K}), with vocabulary based on the training dataset. The input size for sentences was limited to 64 tokens. The model was trained for 10 epochs.

Text simplification. The LEDGAR dataset was adopted for *masked language modelling* (MLM) of the BERT_{base-uncased}¹³ model. Due to the resource limits, the size of the tokenized representation of the input sentences was limited to 128 tokens. Since the MLM variant of BERT is finetuned, MLM Probability of 15% is used which is generally used in BERT. The result BERT_{base-lexglue} model¹⁴ is published. For the complex word, the annotation of the CWIG3G2 was used as a source of complex words with their simplified alternatives (1191 examples). For the latter the threshold as non-complex at all and that with the lowest threshold 80% of words filtered, with a threshold of the word selection equal to 0.1. In the resulting system, we consider the related threshold as 0.5, and therefore the result vocabulary of 84 complex words was used.

6 Results and Discussion

We organize experiments in two tasks to assess and compare the effectiveness of: (1) extractive summarization methods, (2) general purpose abstractive methods and their integration into hybrid mode. We use the ROUGE metric to evaluate the proposed methods on TOS;DR_{summ} documents.

Table 1 presents the results of text summarization models separately for each task. In the case of Task 1 and extractive models (LexRank, SEC_{mpnet}), we considered to output sentences in order of their decreasing relevance for inclusion in the summary. Starting with the extractive model and the threshold of 10% of most salient content (10% limit, Table 1), we investigate the similar behavior of LexRank and LSA. Using SEC_{mpnet} results in a better selection of salient sentences across all ROUGE metrics, with average 11% increment by F-measure. Switching to Task 2 (Table 1), we conducted an experiment with a higher threshold (50% of the document content) to complete the language model input. The application of SEC_{mpnet} in hybrid mode for T5 results in increased performance by ROUGE-2 and ROUGE-L measures, with an average of 11-12%

⁹ <https://huggingface.co/t5-base>

¹⁰ <https://huggingface.co/pszemraj/long-t5-tglobal-base-16384-book-summary>

¹¹ <https://huggingface.co/ksi319/legal-pegasus>

¹² <https://platform.openai.com/docs/models/gpt-4>

¹³ <https://huggingface.co/bert-base-uncased>

¹⁴ <https://huggingface.co/mayankb96/bert-base-uncased-finetuned-lexglue>

increment by F-measure. In the case of LongT5, the application of extractive summarizers (LexRank, SEC_{mpnet}) slightly increases the result of the hybrid systems by ROUGE-1 and ROUGE-L. In fact, the application of SEC_{mpnet} allows us to achieve results comparable with the results of the transformer models in the legal domain (see LEGAL-PEGASUS results, Table 1). The application of SEC_{mpnet} in hybrid mode with general-purpose abstract summarizers results in a ToS summarization solution, which does not require extra resources for fine-tuning transformer models.

For most documents whose length does not exceed 100 sentences (Figure 2), we analyse the T5 and LongT5 tokenizers. Both models yield an average of 2665 tokens per document. The document’s content coverage is 20% in the case of T5 model and 39% for the LongT5.

For the comparison reason, the last block in Table 1 illustrates an application of the domain-oriented LEGAL-PEGASUS model and ChatGPT-4. In the case of ChatGPT-4, the pretrained state¹⁵ dated by 2023/11/25¹⁶ was used. We consider a default temperature parameter and utilize the following prompt message for each document of the TOS;DR_{summ} collection: “*Provide the raw text summary in a single paragraph for the following document: [tos-text]*”. According to the F-measure results (f) and as LEGAL-PEGASUS, it slightly outperforms other systems of Task 2. Similar to the related experiments in the legal case summarization domain [27], the fine-tuned version of the transformer (LEGAL-PEGASUS) illustrates better performance. Despite the absence of publicly available information on the tokenization procedure and vocabulary for ChatGPT-4, we evaluate the number of words included in the input. The prompts were completely incorporated into the input for TOS;DR_{summ} texts consisting of no more than 100 sentences. For longer documents, only 35% words on average were considered.

7 Conclusions

In this paper we propose a text simplification system for the Terms-of-Service texts. We consider automatic legal text simplification task as salient sentence selection problem accompanied by abstractive summarization of the selected sentences. We propose sentence embedding extractive classification model (SEC_{mpnet}) and show the importance of its application for general-purpose abstractive summarization models. According to experiments with texts from TOS;DR resource, it was shown that the SEC_{mpnet} results in average 11% increment across all ROUGE metrics (F-measure) in comparison with other extraction methods for the relatively short summary representation of the original texts. Our integration of the SEC_{mpnet} model in a hybrid mode with general-purpose abstractive summarization models effectively leads to performance enhancement. Specifically, we report improvements in ROUGE-2 scores and at least maintained ROUGE-L scores by F-measure.

¹⁵ `text-davinci-002-render-paid`

¹⁶ <https://help.openai.com/en/articles/6825453-chatgpt-release-notes>

References

1. El-Kassas, W.S., Salama, C.R., Rafea, A.A., Mohamed, H.K.: Automatic text summarization: A comprehensive survey. *Expert Systems with Applications* 165, 113679 (2021), <https://www.sciencedirect.com/science/article/pii/S0957417420305030>
2. Garimella, A., Sancheti, A., Aggarwal, V., Ganesh, A., Chhaya, N., Kambhatla, N.: Text simplification for legal domain: Insights and challenges. In: *Proceedings of the Natural Legal Language Processing Workshop 2022*. pp. 296–304. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid) (Dec 2022), <https://aclanthology.org/2022.nllp-1.28>
3. Steinberger, J., Jezek, K.: Using latent semantic analysis in text summarization and summary evaluation. In: *Proc. ISIM 2004*, 4, 8 (2004)
4. Erkan, G., Radev, D.R.: Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research* 22, 457–479 (2004)
5. Siddharthan, A.: A survey of research on text simplification. *ITL-International Journal of Applied Linguistics* 165(2), 259–298 (2014)
6. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* 30 (2017)
7. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21(140), 1–67 (2020)
8. Guo, M., Ainslie, J., Uthus, D., Ontanon, S., Ni, J., Sung, Y.H., Yang, Y.: LongT5: Efficient text-to-text transformer for long sequences. *arXiv preprint arXiv:2112.07916* (2021)
9. Lippi, M., Palka, P., Contissa, G., Lagioia, F., Micklitz, H., Sartor, G., Torroni, P.: CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. *CoRR abs/1805.01217* (2018), <http://arxiv.org/abs/1805.01217>
10. Todirascu, A., Wilkens, R., Rolin, E., François, T., Bernhard, D., Gala, N.: HECTOR: A hybrid TExt SimplifiCation TOol for raw texts in French. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. pp. 4620–4630. European Language Resources Association, Marseille, France (Jun 2022), <https://aclanthology.org/2022.lrec-1.493>
11. Evangelopoulos, N., Zhang, X., Prybutok, V.R.: Latent semantic analysis: five methodological recommendations. *European Journal of Information Systems* 21(1), 70–86 (2012)
12. Mihalcea, R., Tarau, P.: Textrank: Bringing order into text. In: *Proceedings of the 2004 conference on empirical methods in natural language processing*. pp. 404–411 (2004)
13. Nenkova, A., Vanderwende, L.: The impact of frequency on summarization. Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005 101 (2005)
14. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training. *OpenAI Blog* (2018)
15. Beltagy, I., Peters, M.E., Cohan, A.: Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150* (2020)
16. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR abs/1910.13461* (2019), <http://arxiv.org/abs/1910.13461>

17. Zhang, J., Zhao, Y., Saleh, M., Liu, P.J.: PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. *CoRR* abs/1912.08777 (2019), <http://arxiv.org/abs/1912.08777>
18. Hermann, K.M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., Blunsom, P.: Teaching machines to read and comprehend. *Advances in neural information processing systems* 28 (2015)
19. Sharma, E., Li, C., Wang, L.: Bigpatent: A large-scale dataset for abstractive and coherent summarization. *arXiv preprint arXiv:1906.03741* (2019)
20. Yimam, S.M., Štajner, S., Riedl, M., Biemann, C.: CWIG3G2 - complex word identification task across three text genres and two user groups. In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. pp. 401–407. Asian Federation of Natural Language Processing, Taipei, Taiwan (Nov 2017), <https://aclanthology.org/I17-2068>
21. Paetzold, G., Specia, L.: SemEval 2016 task 11: Complex word identification. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. pp. 560–569. Association for Computational Linguistics, San Diego, California (Jun 2016), <https://aclanthology.org/S16-1085>
22. Yimam, S.M., Biemann, C., Malmasi, S., Paetzold, G., Specia, L., Štajner, S., Tack, A., Zampieri, M.: A report on the complex word identification shared task 2018. In: *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*. pp. 66–78. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018), <https://aclanthology.org/W18-0507>
23. Biran, O., Brody, S., Elhadad, N.: Putting it simply: a context-aware approach to lexical simplification. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. pp. 496–501. Association for Computational Linguistics, Portland, Oregon, USA (Jun 2011), <https://aclanthology.org/P11-2087>
24. Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D.M., Aletras, N.: Lexglue: A benchmark dataset for legal language understanding in english. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Dublin, Ireland (2022)
25. Tuggeener, D., Von Däniken, P., Peetz, T., Cieliebak, M.: Ledger: a large-scale multi-label corpus for text classification of legal provisions in contracts. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. pp. 1235–1241 (2020)
26. Bird, S., Loper, E.: NLTK: The natural language toolkit. In: *Proceedings of the ACL Interactive Poster and Demonstration Sessions*. pp. 214–217. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), <https://aclanthology.org/P04-3031>
27. Deroy, A., Ghosh, K., Ghosh, S.: How ready are pre-trained abstractive models and llms for legal case judgement summarization? *CEUR Workshop* 3423 (2023)