Unfair ToS: LLM-Based Unfair Contract Term Detector

Jianing Zhang
Jiazhou Liang

Introduction

Terms of service (ToS) are agreements between service providers and their users. These ToS documents often contain complex legal language that users struggle to understand. A Deloitte survey discovered that 91% of individuals consent to ToS without reading them thoroughly [1], potentially leading to inadvertent acceptance of unfair terms. These unfair clauses may violate consumer laws [2], compromise users' rights, and raise privacy concerns [3].

The proven advantage of the Large Language Model (LLM) in accurately and efficiently sifting through extensive text and extracting summaries from complex sections [4] makes them one of the best candidates to resolve this issue. The Unfair-ToS project uses a GPT-based framework to highlight key ToS sentences, create simplified explanations, and assess their fairness. In cases where the terms are unfair, the system will also provide a reason to explain the cause.

Model Structure Overview

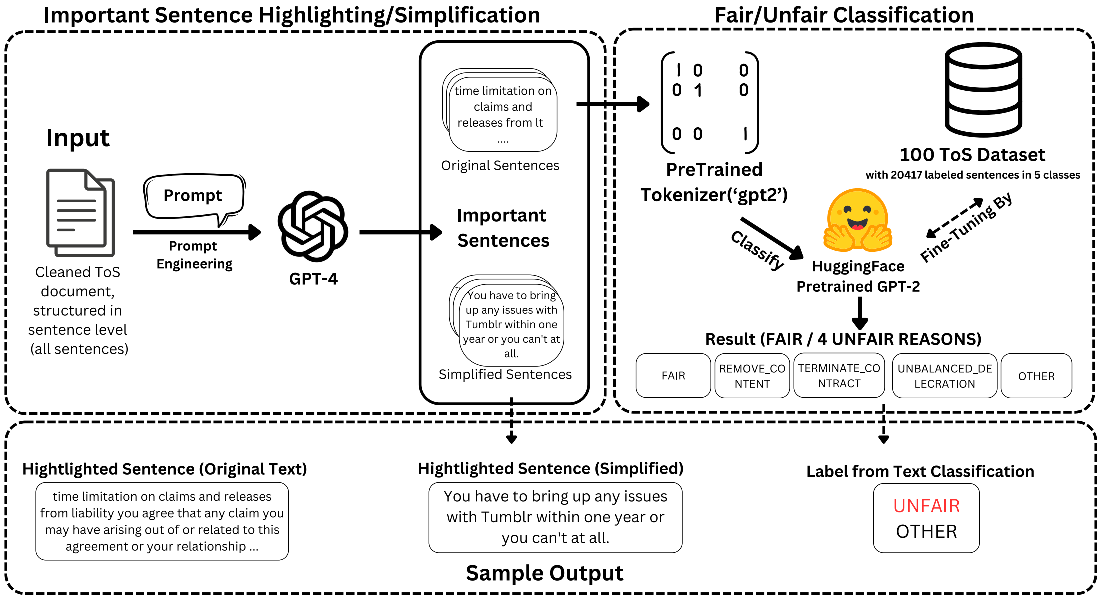

Figure 1: The overview model structure, the input is the cleaned ToS in sentence level, through GPT-4 to highlight and simplify sentences and fine-tuned GPT-2 to classify fair/unfair reasons

Background & Related Work

Lippi et al. addressed the presence of unfair terms in infrequently read contracts in their paper 'CLAUDETTE: an Automated Detector of Potentially Unfair Clauses in Online Terms of Service' [5]. The authors introduced the idea of utilizing machine learning models, including SVM, CNN, and several hybrid models, to identify potential unfair clauses in ToS. While the classification results from this paper are deemed acceptable, there is room for improvement.

One of the noteworthy contributions of this work is a dataset comprising 50 ToS from various online platforms. Legal professionals annotated this dataset at the sentence level, categorizing sentences as either fair or unfair, across eight distinct categories (Figure 2).

Type of clause	Symbol
Arbitration	`<a>`
Unilateral change	`<ch>`
Content removal	`<cr>`
Jurisdiction	`<j>`
Choice of law	`<law>`
Limitation of liability	`<ltd>`
Unilateral termination	`<ter>`
Contract by using	`<use>`

Figure 2: The eight categories of unfair clauses in Lippi et al.'s paper [5]

Building upon this foundation to adapt to the rapid evolution in the machine learning community, a more recent paper [6] expanded the dataset to include 100 ToS, providing a more comprehensive coverage of service providers. The same annotated mechanism used in the previous paper was applied, and the study achieved improved classification results using Memory-Augmented Neural Networks.

However, in existing approaches, many crucial terms are annotated as fair simply because they do not violate any laws. Users could benefit from being made aware of such terms without having to read through all the fair terms. Furthermore, the utilization of fine-tuned LLMs, a popular approach in legal studies [7], remains unexplored in the context of Unfair Term classification. These issues have prompted our attention to design a more comprehensive model.

Data and Data Processing

Text Highlighting

The first challenge involved determining which sentences are 'Important' for highlighting. To tackle this, a dataset from the TOS;DR website [8]. This dataset comprises a comprehensive collection of over 1000 Terms of Service (ToS) documents. Contributors on this platform highlighted sentences based on predefined 200 cases [9], which include instances of protecting users, neutrality, or violating users' rights. These 200 cases will also be the definition of importance in our project. The highlighted sentences serve as the ground truth for evaluation, and the original ToS documents are used as training samples during the prompt engineering process.

To manage the voluminous data, we focused on ToS documents with an average of 40-50 highlighted sentences among samples. This led to a curated set of 11 documents, each containing approximately 300 sentences. The cleaning process involved splitting text into sentences using the NLTK 'English.pickle' sentence detector [10], converting to lowercase, and removing special characters, HTML tags, and duplicates. Each sentence was then labelled '1' if highlighted for importance or '0' if not. A few samples are provided in Table 1.

Sentence	Label
You may terminate your Crunchyroll account at any time and for any reason.	1
At 444 Bush Street, San Francisco, CA 94108, phone: (415) 796-3560	0
The site and services may be used and accessed for lawful purposes only.	1

Table 1: Samples in the TOS;DR Dataset

Text Classification

The second challenge is the definition of 'unfair' to fine-tuning GPT-2 in classification. We utilize the same 100 ToS dataset mentioned in the previous paper [6]. Legal experts segmented this dataset into 20,417 sentences, labelled as either 'fair' or falling into one of eight types of unfairness categories outlined in Figure 2. Of the total sentences, 18,235 are marked as fair, while 2,182 are classified as unfair, with fewer samples when considering each specific reason category. This class imbalance is expected, as the majority of sentences in ToS documents are typically considered fair. Directly detecting unfair terms in raw documents would render the detector meaningless.

To address the class imbalance in unfair categories, we aggregated them into four types of unfairness, as outlined in Table 3.

Index	**Class Name**	**Description**	**Original Sample**	**Samples after over-sampling**
0	FAIR	The sentence does not have any unfair clauses.	18235	14554
1	REMOVE _CONTENT	The provider removes consumer content from the service	216	1454
2	TERMINATE _CONTRACT	The provider terminates or modifies the contract	653	4703
3	UNBALANCED _DECLARATION	Limitations affect the balance between the parties' rights	705	4362
4	OTHER	Other reasons, such as choice of law	608	4103

Table 2: Class, description, and sample before and after oversampling

The 89 samples with more than one class label were assigned to the class with the least total samples, further alleviating class imbalance. We employed oversampling by duplicating existing samples to match the number of total samples in the four 'unfair' classes with the 'fair' class. This approach aims to reduce imbalances while maintaining the original class distribution. To avoid introducing human bias, such as including terms that do not exist in the original text, synthetic samples were not created. Preventing unwanted noise in the dataset. In the end, we divided the dataset into 80% training and 20% testing. Examples in the dataset are shown in Table 3.

Sentence	Label
You understand and agree that Mozilla reserves the right, at its discretion, to remove any submission that it deems violates these terms.	REMOVE_CONTENT
our websites include multiple domains such as mozilla.org , mozillians.org , firefox.com , mozillafestival.org , openstandard.com , openbadges.org and webmaker.org .	FAIR
we reserve the right to modify any provision hereof from time to time, in our sole discretion, and such modification shall be effective immediately upon its posting on the website.	TERMINATE_CONTRACT

Table 3: Samples in the 100 ToS Dataset

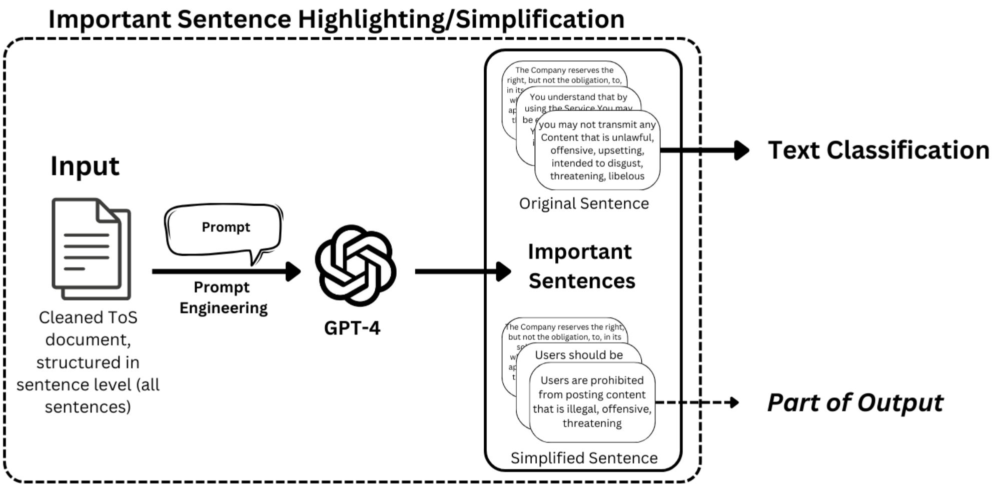

Figure 3: Model of Text Highlighting/Simplification

Architecture

Text Highlighting and Simplification

Figure 3 utilizes GPT-4 through prompts to highlight sentences and generate simplifications for each highlighted sentence from the input, cleaned ToS. The prompt used in our model has been fine-tuned using 5/11 ToS documents mentioned above, comprising approximately 1500 sentences, through prompt engineering.

The initial segment of the prompt offers a summary of the tasks:

Detail Instruction:
Using the content in the ’term of service document’ given below, accomplish the following two tasks:
***Hightlight:** Hightlight max 50 indexes that are most important for users to carefully read before accepting the term, using the ’definition of important’ given below. You can also use your knowledge to identify ’important’ within the Terms of Service documents. Keep a good balance between precision and recall. Obtain the index of highlighted sentences*
***Simplification:** For each highlighted sentence, based on the document’s content, craft a plain-language simplification that is easily understandable for a general audience. Aim for a Gunning Fog index below 9.*

Then specifying the desired output format, GPT-4 will only output the index of highlight sentences to reduce the length of total output tokens.

Please provide the output in a text file format with ’Highlight’ and ’Simplification’ as header (first line). Each line contains the ”index” and its corresponding ”Simplification” enclosed in quotation marks and separated by a comma.
Output Format
Highlight,Summary
”index”,”Simplification 1”
”index”,”Simplification 2”
...

We utilized a chain of thought to guide GPT-4 in text highlighting, ensuring that the output aligns with the user group of specific service providers.

Steps
think step by step,
1. who is the service provider? who is its user population?
2. what should be considered as important sentences for users to read? using the definition given below
3. How can you quantify the importance of a sentence using this definition?
4. What are the 50 most important sentences?

Subsequently, we condensed the 200 cases from the TOS;DR website into 57 categories, as the definition of importance. This aids GPT-4 in deciding which sentences to highlight based on these examples (few shots). Although we initially sought to incorporate several highlighted sentences as examples as well, doing so resulted in a decrease in recall without a corresponding improvement in precision. The inclusion of noisy examples might constrain GPT-4's selection, potentially leading to a reduction in recall.

Definition of Important###
A sentence is important if it relates to one or more of the following 57 practices:
*1. **Retention of User-Generated Content**: Keeping user content even after the user closes their account.*
*2. **User Tracking**: Monitoring users on other websites.*
(Truncated due to word limit)

The completed 57 categories in the prompt can be found here. We also added a few requirements to ensure the document with a length beyond GPT-4 capacity can be properly handled, and the output adheres to the desired format.

Additional requirements
1. The output should only contain the format listed above, turn off any warning or error message.
2. If the text is too long, please only the most important part of the text.

The last part consists of the input ToS. We also manually incorporated an index for each sentence in the ToS to guarantee that GPT-4's output aligns with the original index.

service provider
[service provider]
term of service document
index 1 [input sentence 1]
index 2 [input sentence 2]
...

The original texts of highlighted sentences will be input into the 'Text Classification' model, and their simplifications will form part of the final output.

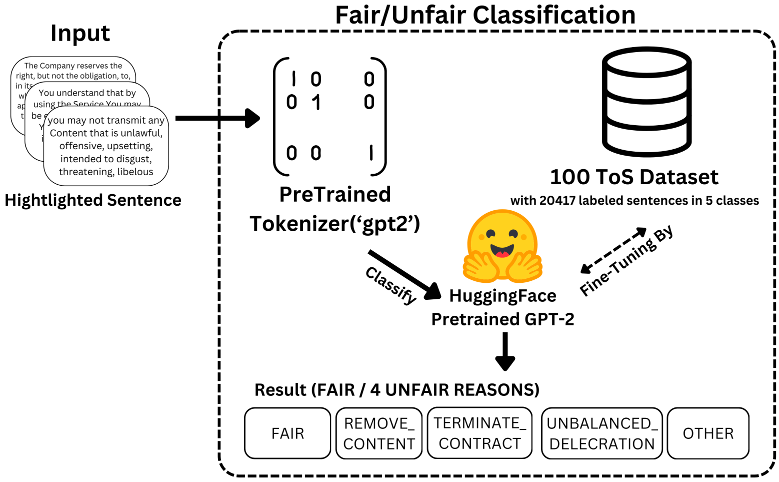

Figure 4: Text Classification

Text Classification

Important sentences identified by GPT-4 will serve as the input to the text classification model (shown in Figure 4). Each sentence will be tokenized using HuggingFace's predefined 'GPT2' tokenizer. And feed them into the pre-trained 'gpt2' model [11], which is fine-tuned on the 100 ToS dataset with the same tokenizer and padded to the maximum length. The model's output will fall into one of the five classes outlined in Table 3. This model assigns a label to each highlighted sentence, aiding users in identifying the fairness of sentences.

Baseline Model

Text Highlighting

The baseline model for text highlighting uses Text Rank, an unsupervised graph-based ranking algorithm [12]. It identifies important sentences based on their similarity to all others, a method popular for its efficiency in sentence extraction [13].

Text Simplification

Simplified sentences will be compared with originals using the Gunning Fog Index [14], measuring the required years of education for first-read comprehension, as shown in Table 4.

Gunning Fog Index	Years of Education
17	College Graduate
15	College Junior
13	College Freshman
11	High School Junior
9	High School Freshman
7	Seventh Grade

Table 4: Gunning Fog Index

Text Classification

We constructed a baseline CNN classification model Figure 5, inspired by Yoon Kim's paper [15], with k1 = 4 and k2 = 4. The model's objective, the same as our fine-tuned model, is to classify sentences into five distinct categories as outlined in Table 3. It employs the same training and testing data from the 100 ToS dataset and was assessed using accuracy and F1-Score, allowing a fair comparison with our GPT-2 model while accounting for class imbalance.

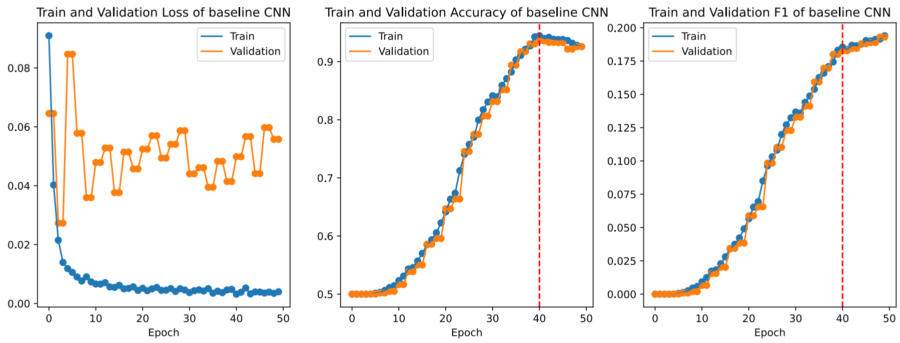

Figure 5: Training and Validation Result of CNN model

Text Highlighting Evaluation

Quantitative Results

The quantitative evaluation of our text highlighting model used Precision, Recall, and F1 Score across 11 ToS documents. The ground truth was established from the highlighted sentences in the ToS;DR dataset. To account for minor differences between model output and original sentences, we used cosine similarity with 'all-mpnet-base-v2' embedding for matching model results with ground truth.

All Docs	Precision	Recall	F1
Text Rank	0.390	0.339	0.353
GPT-4	0.433	0.642	0.512

Table 5: Quantitative Result of Text Highlight Model

Table 5 shows GPT-4 outperforming the baseline Text Rank model in all metrics. Table 6 reveals significant F1 score differences between the two models, which proved GPT-4's efficacy and also suggested a solution for key terms extraction in various documentation.

Documents	Text Rank F1	GPT-4 F1
LBRY	0.199	0.474
Google	0.296	0.531
Pure Dating	0.298	0.529
IDrive	0.211	0.481
HuffPost	0.215	0.418

Table 6: F1 score of two models' result in selected documents

Qualitative Results

While GPT-4's precision with TOS;DR isn't high, TOS;DR's human biases contribute to imperfections. Reviewing GPT-4's outputs, we identified a subgroup of important sentences based on the 200 pre-defined cases that were only identified by GPT-4.

Example	Reason
Crunchyroll: we reserve the right to adjust pricing for our service or any components thereof in any manner and at any time as we may determine in our sole and absolute discretion.	Freely adjusting price without notices
Pinterest: if you choose to submit comments, ideas or feedback, you agree that we are free to use them without any restriction or compensation to you.	Using user's content without notices

Table 7: Important Sentences only highlighted by GPT-4

In Table 7, sentences regarding the right to adjust pricing and use of user content without notice and compensation are only highlighted in GPT-4. Human contributors missed them. These cases where the model compensates for human oversight add context to our findings and show the model's potential in parsing and simplifying complex ToS documents.

Text Simplification Evaluation

Quantitative Results

The original texts had a high Gunning Fog Index of 24.821, indicating college-level complexity, whereas the simplified texts scored 9.484, aligning with a ninth-grade level. This shows the model's success in simplifying complex legal language. Additionally, the BERT Score [16], assessing similarity between simplified and original sentences, indicated a high F1 score of 0.815, confirming the model's effectiveness in retaining essential information.

Qualitative Results

Example 1:	
Original Sentence:	*Without limiting the foregoing, to the full extent permitted by law, tumblr disclaims all warranties, express or implied, of merchantability, fitness for a particular purpose, or non-infringement.*
Simplified Sentence:	*Tumblr doesn't promise that the service will meet your needs or that there won't be errors.*
Example 2:	
Original Sentence:	*time limitation on claims and releases from liability you agree that any claim you may have arising out of or related to this agreement or your relationship with tumblr must be filed within one year after such claim arose.*
Simplified Sentence:	*You have to bring up any issues with Tumblr within one year or you can't at all.*

Table 8: Original and Simplified Text Comparison

The original text in Table 8 uses complex vocabulary like 'merchantability' and 'infringement', requiring multiple readings. In contrast, our model significantly reduced this complexity, producing shorter, simpler texts.

However, the model had limitations in including a clear definition of key terms in the sentences. For example, 'service' is defined in the original text but not in the simplified samples, so users might need to refer back to the original for clarification. Nevertheless, the various advantages of simplified text can assist users in quickly understanding.

Text Classification Evaluation

Quantitative Results

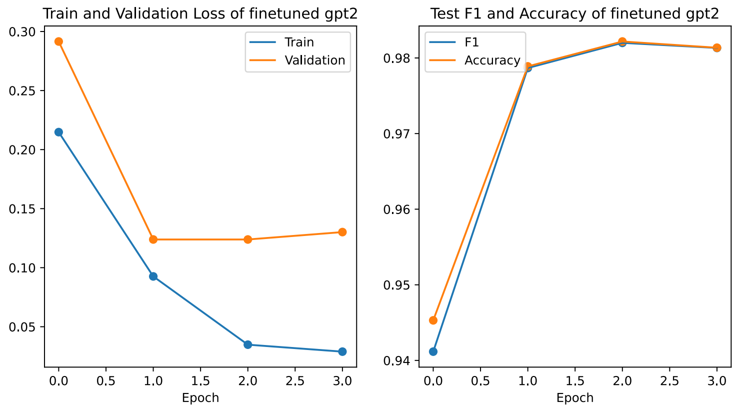

Figure 6: **Left**: The training and validation loss **Right**: Test F1 and Accuracy through Epoch

From Table 9, the baseline CNN model's F1 score is only 0.193, indicating a tendency to classify samples into the class with the majority of samples, which is the 'fair' class. As a result, the baseline CNN struggles to properly identify 'unfair' sentences from the samples.

Model	Accuracy	F1
Baseline CNN	0.962	0.193
Fine Tuned GPT-2	0.982	0.981

Table 9: The Accuracy and F1 of text classification from two models

In contrast, our fine-tuned GPT-2 model (Figure 6) outperforms the baseline in both metrics. These results suggest that the GPT-2 model is more resistant to class imbalance, especially when fair sentences dominate the corpus, as is often the case.

Qualitative Results

The first example in Table 10 presents explicit evidence of unfair terms, specifically suspending users' access to the services. The classification model correctly categorizes it into the 'Terminate Contract' class, underscoring the effectiveness of the classification model.

Although the second example is highlighted, the classification model categorizes it as a fair term. Upon reviewing the text, it is evident that the sentence pertains to protecting users' rights, which should be considered fair. However, it remains beneficial for users to be aware of such sentences.

This example justifies the inclusion of both highlighting and classification models. Relying solely on text highlighting model results will make customers difficult to discern whether a highlighted sentence is violating or protecting their rights. Conversely, relying just on text classification might obscure important but fair sentences within the 'fair' category, contradicting the project's overarching goal.

Example 1:	
Hightlight Sentence (Original):	*suspending or terminating your access to Google services google reserves the right to suspend or terminate your access to the services or delete your Google account if any of these things happen: ...*
Highlight Sentence (Simplified):	*Google can suspend or terminate your access to their services if you seriously or repeatedly violate the terms or policies, or if required by law.*
Label:	*Unfair - Terminate Contract*
Example 2:	
Hightlight Sentence (Simplified):	*You can discontinue using Google services whenever you want.*
Label:	*Fair*

Table 10: The examples of completed Results from all three tasks

Discussion and Learnings

In our project, the text simplification and classification model demonstrated strong performance, effectively parsing and categorizing the Terms of Service. Indicating LLMs can perform well when the ground truth is clearly defined. However, we faced challenges in evaluating GPT-4's performance, especially in quantitative analysis, due to the lack of a ground truth to construct a proper metric.

To improve our approach in future projects, we recognize the potential benefits of consulting with legal experts. Their specialized insights could greatly refine our evaluation process and lead to more precise qualitative outcomes.

Bibliography

[1] C. Cakebread, "You're not alone, no one reads terms of service agreements," Business Insider, 11 2017. [Online]. Available: https://www.businessinsider.com/deloitte-study-91-percent-agree-terms-of-service-without-reading-2017-11#:~:text=A%20Deloitte%20survey%20of%202%2C000

[2] G. Pearson, "Regarding unfair terms in financial services contracts," *University of Western Australia Law Review*, Jan. 2013. [Online]. Available: https://search.informit.org/doi/abs/10.3316/ielapa.301269270309298

[3] G. S. Hans, "Privacy Policies, Terms of Service, and FTC Enforcement: Broadening Unfairness Regulation for a New Era," *University of Michigan Law School Scholarship Repository*, vol. 19, no. 1, pp. 163–200, 2012. [Online]. Available: https://repository.law.umich.edu/mttlr/vol19/iss1/5

[4] Y. A. Arbel and S. Becher, "How Smart are Smart Readers? LLMs and the Future of the No-Reading Problem," Jun. 2023, [Online; accessed 9. Dec. 2023].

[5] M. Lippi, P. Palka, G. Contissa, F. Lagioia, H.-W. Micklitz, G. Sartor, and P. Torroni, "CLAUDETTE: an Automated Detector of Potentially Unfair Clauses in Online Terms of Service," *arXiv*, May 2018.

[6] F. Ruggeri, F. Lagioia, M. Lippi, and P. Torroni, "Detecting and explaining unfairness in consumer contracts through memory networks," *Artif. Intell. Law*, vol. 30, no. 1, pp. 59–92, Mar. 2022.

[7] D. Liga and L. Robaldo, "Fine-tuning GPT-3 for legal rule classification," *Computer Law & Security Review*, vol. 51, p. 105864, Nov. 2023.

[8] T. Team, "Terms of service; didn't read," Tosdr.org, 2019. [Online]. Available: https://tosdr.org/

[9] "Terms of Service; Didn't Read - Cases," Dec. 2023, [Online; accessed 9. Dec. 2023]. [Online]. Available: https://edit.tosdr.org/cases

[10] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, 2009. [Online]. Available: https://books.google.ca/books?id=KGIbfiiP1i4C

[11] HF Canonical Model Maintainers, "gpt2 (revision 909a290)," 2022. [Online]. Available: https://huggingface.co/gpt2

[12] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004, pp. 404–411.

[13] C. Mallick, A. K. Das, M. Dutta, A. K. Das, and A. Sarkar, "Graph-based text summarization using modified textrank," in *Soft Computing in Data Analytics: Proceedings of International Conference on SCDA 2018*. Springer, 2019, pp. 137–146.

[14] R. Gunning, “The fog index after twenty years,” *Journal of Business Communication*, vol. 6, no. 2, pp. 3–13, 1969. [Online]. Available: https://doi.org/10.1177/002194366900600202

[15] Y. Kim, “Convolutional neural networks for sentence classification,” 2014.

[16] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, and Y. Artzi, “Bertscore: Evaluating text generation with bert,” in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=SkeHuCVFDr