ECE1786: Unfair ToS Final Report

Jianing Zhang
Jiazhou Liang

word count: 1987
penalty: 0

Dec 9th, 2023

Permission

	Jianing Zhang	Jiazhou Liang
Permission to post video	Yes	Yes
Permission to post final report	Yes	Yes
Permission to post source code	Yes	Yes

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. Such clauses may violate consumer laws [2], compromise users' rights, and raise privacy concerns [3].

The Large Language Model's (LLM) proven ability to efficiently extract summaries from complex texts [4] makes it ideal for addressing ToS complexities. The Unfair-ToS project employs a GPT-based framework to highlight crucial ToS sentences, offer simplified explanations, and evaluate their fairness, also providing reasons for any unfair terms.

Model Structure Overview

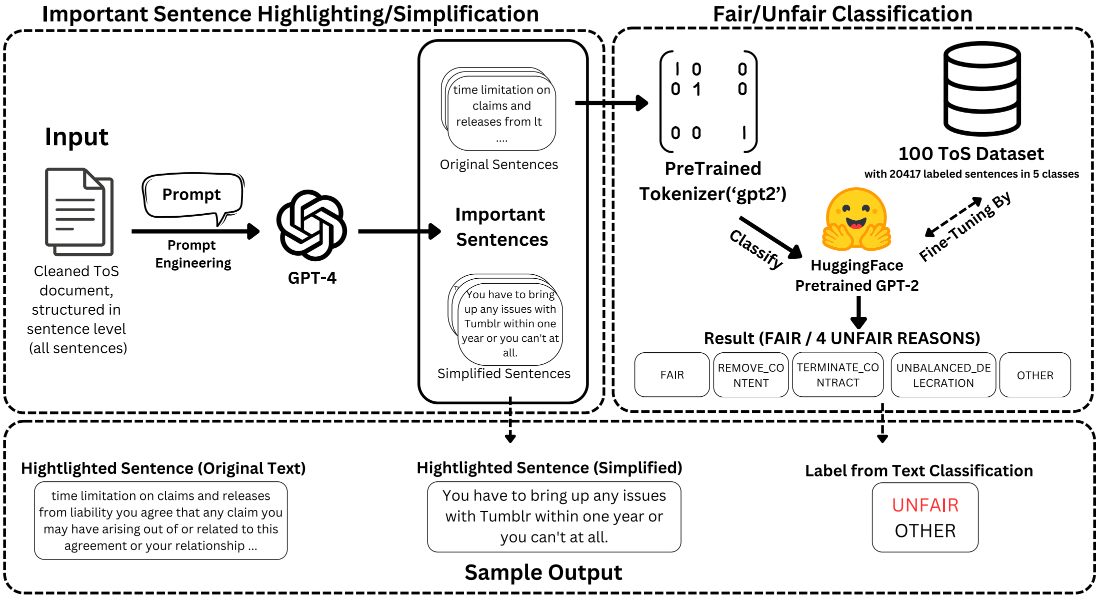
Important Sentence Highlighting/Simplification
Input
Cleaned ToS document, structured in sentence level (all sentences)
Prompt
Prompt Engineering
GPT-4
time limitation on claims and releases from lt ....
Original Sentences
Important Sentences
You have to bring up any issues with Tumblr within one year or you can't at all.
Simplified Sentences
Fair/Unfair Classification
1 0 0
0 1 0
0 0 1
PreTrained Tokenizer('gpt2')
100 ToS Dataset
with 20417 labeled sentences in 5 classes
Classify
Fine-Tuning By
HuggingFace Pretrained GPT-2
Result (FAIR / 4 UNFAIR REASONS)
FAIR
REMOVE_CONTENT
TERMINATE_CONTRACT
UNBALANCED_DELECRATION
OTHER
Hightlighted Sentence (Original Text)
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 ...
Hightlighted Sentence (Simplified)
You have to bring up any issues with Tumblr within one year or you can't at all.
Label from Text Classification
UNFAIR
OTHER
Sample Output

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

In their paper 'CLAUDETTE: an Automated Detector of Potentially Unfair Clauses in Online Terms of Service' [5], Lippi et al. explored using machine learning models like SVM, CNN, and hybrids to detect unfair clauses in ToS. While the results show promise, there is potential for further improvement.

This work's significant contribution is a dataset of 50 ToS from different online platforms, annotated by legal experts. Sentences were labeled as fair or unfair in eight categories (Figure 2).

Adapting to advances in machine learning, a recent paper [6] expanded the dataset to 100 ToS for broader coverage. Using the same annotation method, it achieved improved classification results with Memory-Augmented Neural Networks.

Existing approaches often label crucial terms as fair if they don't violate laws, yet users could benefit from awareness of such terms. Furthermore, the use of fine-tuned LLMs in legal studies [7] hasn't been

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]

explored for Unfair Term classification. These issues have prompted our attention to design a more comprehensive model.

Data and Data Processing

Text Highlighting

The first challenge was defining 'importance' for highlighting in ToS. We utilized a dataset from TOS;DR [8], where contributors highlighted sentences based on 200 predefined cases [9], covering user protection, neutrality, or rights violations. These cases define importance in our project, with highlighted sentences as ground truth and original ToS documents for training in prompt engineering.

For feasibility, we focused on ToS documents with 40-50 highlighted sentences, resulting in 11 documents, each around 300 sentences. Cleaning involved using NLTK's 'English.pickle' sentence detector [10], converting text to lowercase, and removing special characters, HTML tags, and duplicates. Sentences were labeled '1' if highlighted or '0' otherwise. 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

For the fair/unfair classification model, we utilized the same 100 ToS dataset as in related work [6]. Legal experts divided this into 20,417 sentences, labeled 'fair' or as one of eight unfair clause types (Figure 2).

Of the 20,417 sentences, 18,235 are marked fair and 2,182 as unfair, with fewer in specific classes (Table 2). This class imbalance is expected, if too many unfair terms exist, users directly detecting unfair terms in raw documents would render the detector meaningless. We addressed this by aggregating the unfair sentences into four types, as outlined in Table 2.

For the 89 samples labelled with multiple classes, we assigned them to the class with the fewest samples to reduce imbalance. We employed oversampling, duplicating samples to match the 'unfair' classes with the 'fair' class in total samples. This maintained the original distribution without adding human bias in synthetic samples, thus avoiding noise in the data. The dataset was split 80% for training and 20% for testing, as shown 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

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

Architecture

Text Highlighting and Simplification

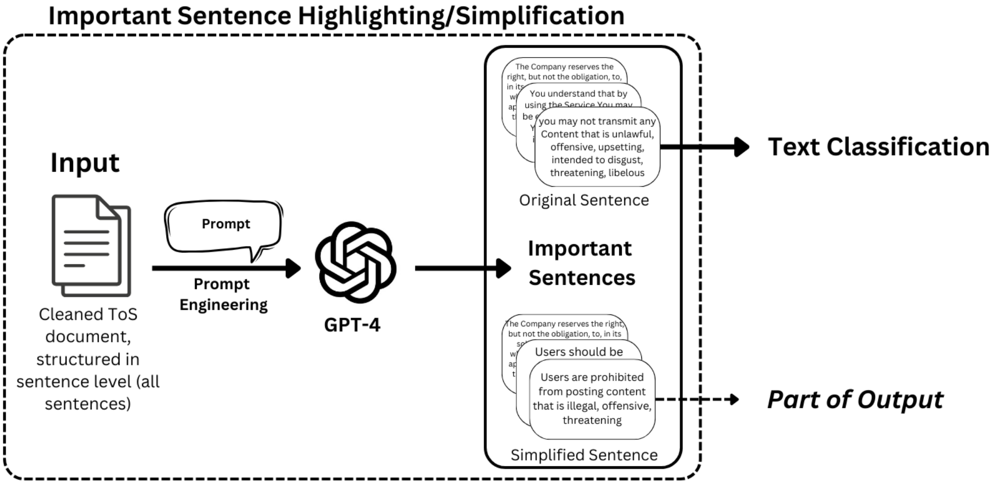
Important Sentence Highlighting/Simplification
Input
Cleaned ToS document, structured in sentence level (all sentences)
Prompt
Prompt Engineering
GPT-4
The Company reserves the right, but not the obligation, to, in its
You understand that by using the Service You may
you may not transmit any Content that is unlawful, offensive, upsetting, intended to disgust, threatening, libelous
Original Sentence
Important Sentences
The Company reserves the right, but not the obligation, to, in its
Users should be
Users are prohibited from posting content that is illegal, offensive, threatening
Simplified Sentence
Text Classification
Part of Output

Figure 3: Model of Text Highlighting/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 in our model was fine-tuned with 5/11 ToS documents (about 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?

To assist GPT-4 in selecting sentences to highlight, we condensed the 200 pre-defined cases of importance into 57 categories (few shots). Although we initially sought to incorporate several highlighted sentences as examples, doing so resulted in a decrease in recall without improvement in precision. Including noisy examples may have constrained GPT-4's selection, leading to decreased 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 added requirements to ensure documents with lengths beyond capacity can be handled, and the output in 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 manually added an index for each sentence to align GPT-4's output with the original index.

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

The original highlighted sentences are the input of the 'Text Classification' model, and simplification will be part of the final output.

Text Classification

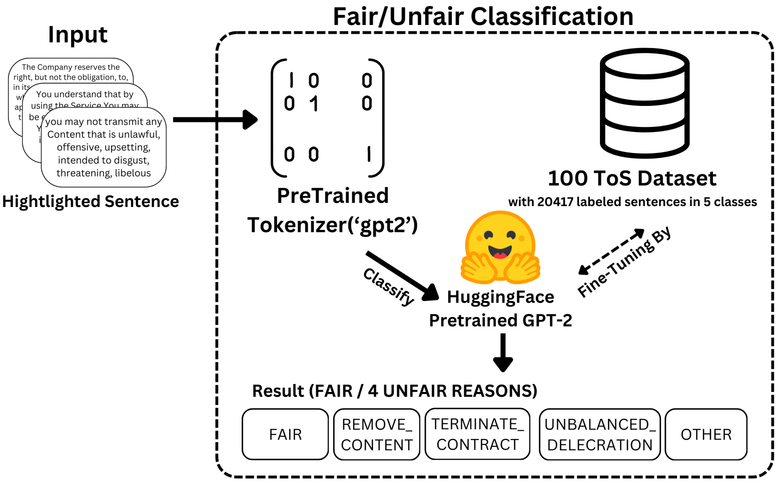
Input
The Company reserves the right, but not the obligation, to, in its
You understand that by using the Service You may be
you may not transmit any Content that is unlawful, offensive, upsetting, intended to disgust, threatening, libelous
Hightlighted Sentence
Fair/Unfair Classification
PreTrained Tokenizer('gpt2')
100 ToS Dataset
with 20417 labeled sentences in 5 classes
Fine-Tuning By
Classify
HuggingFace Pretrained GPT-2
Result (FAIR / 4 UNFAIR REASONS)
FAIR
REMOVE_ CONTENT
TERMINATE_ CONTRACT
UNBALANCED_ DELECRATION
OTHER

Figure 4: Text Classification

Important sentences identified by GPT-4 are inputted into a text classification model (Figure 4). They are tokenized using HuggingFace's 'GPT-2' tokenizer and inputted into the 'GPT-2' model [11], fine-tuned on the 100 ToS dataset. The output falls into one of five classes in Table 2, labeling each highlighted sentence to help users identify 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 2. 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.

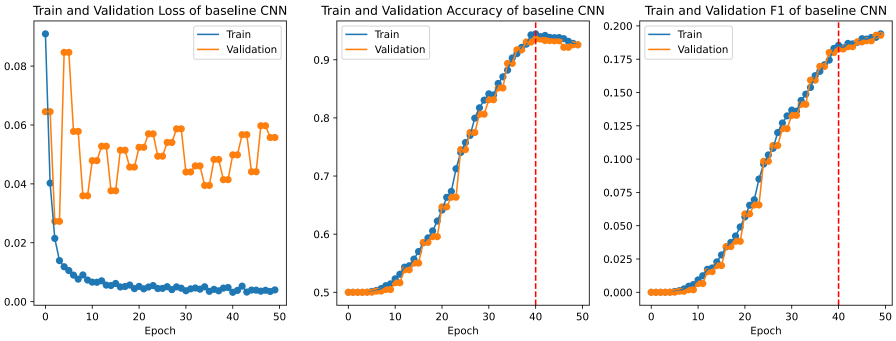

Figure 5: Training and Validation Result of CNN model

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

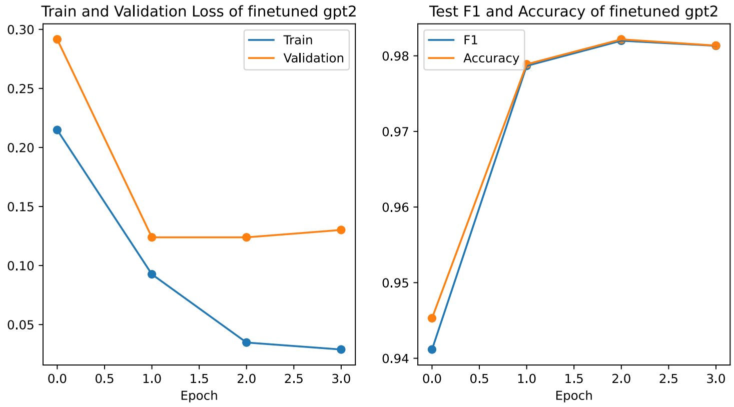

Figure 6: **Left**: The training and validation loss **Right**: Test F1 and Accuracy through Epoch

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

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

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.

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.

Individual Contributions

Jianing's Contribution

- Crawled the ToS and the highlighted sentences from ToS;DR website
- Manually selected the 57 categories of importance from ToS;DR
- implemented the GPT-4 model
- Engineered GPT-4 prompt
- Manually identified key sentences missed by human reviewers on the website
- Wrote evaluation script for GPT-4 model

Jiazhou's Contribution

- Cleaned dataset for Text Highlight and Simplification
- Engineered GPT-4 prompt and found the best parameters
- Cleaned ToS-100 dataset
- Implemented the baseline CNN model
- Implemented and fine-tuned GPT-2 model for text classification
- Wrote evaluation script for both GPT-4 and GPT-2 models

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