Clear & Simple Languages: Understanding Your Terms of Service

Transforming Hazardous ToS into user-friendly language for easy understanding

Team 10
Doam Lee (20234290)
Yokyung Lee (20200775)
Hyeonjun Boo (20210294)

Link To Repository

Introduction

In the rapidly changing realm of digital interactions and online services, the significance of Terms of Service (ToS) agreements cannot be overstated. Serving as foundational documents that regulate user behavior and responsibilities in their interactions with platforms and services, ToS agreements not only establish legal boundaries but also wield considerable influence over how users engage with these platforms. However, the formidable challenge lies in the intricate legal language and comprehensive scope of these agreements, making them largely complex and difficult for the general public to comprehend.

Do not read ToS

A common tendency emerges where users agree to ToS without a comprehensive understanding. This behavior was vividly demonstrated in a research scenario where a fictitious social networking site, "Name Drop," incorporated absurd clauses, such as users surrendering their firstborn child as payment. As a result, a staggering 98% of participants agreed to these outrageous terms. This is a shocking experimental result indicating how many people are not reading the ToS.[1]

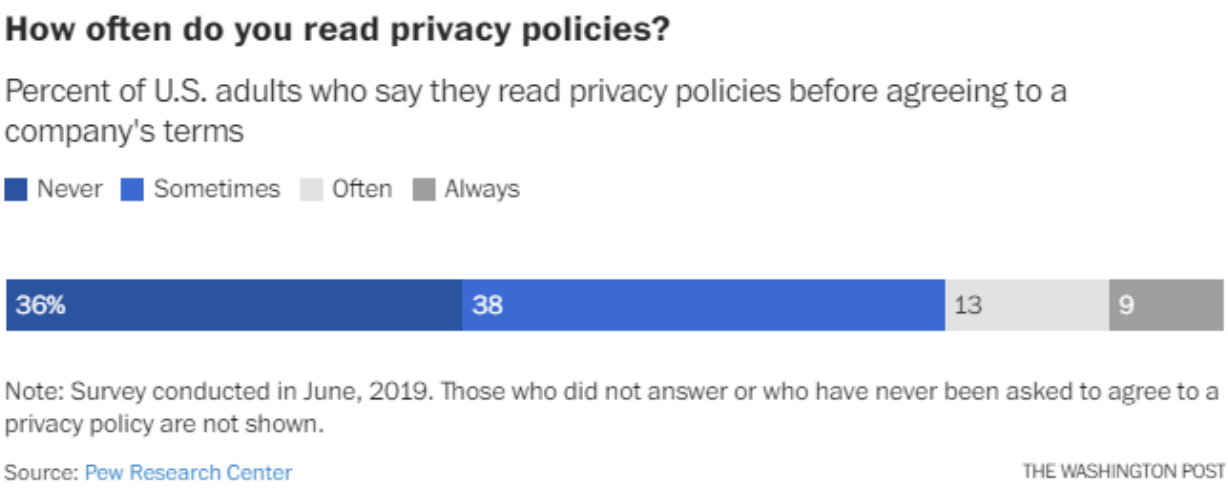
Deloitte's survey of 2,000 U.S. consumers revealed that 91% of individuals do not meticulously read legal terms and conditions before agreeing. Notably, among the younger age group of 18-34, this percentage rises to 97%, indicating a higher proportion of individuals in this demographic who do not read ToS. It is evident that younger individuals may not fully grasp the significance of ToS and are less inclined to engage with lengthy texts. Consequently, they are more prone to facing difficulties as a result.[2]

---

[1] Obar, J. A., & Oeldorf-Hirsch, A. (2016). *The biggest lie on the Internet: Ignoring the privacy policies and terms of service policies of social networking services*. Proceedings of the 44th Research Conference on Communication, Information & Internet Policy. Arlington, VA. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2757465

[2] Deloitter *2017 Global Mobile Consumer Survey: US edition p.12* https://www2.deloitte.com/content/dam/Deloitte/us/Documents/technology-media-telecommunications/us-tmt-2017-global-mobile-consumer-survey-executive-summary.pdf

In another survey, results revealed a pattern similar to the aforementioned findings. Here, the proportion of individuals consistently reading ToS stood at a mere 9%, while those who either "Never" or "Sometimes" read the ToS constituted a significant majority at 74%.[3]



The length of the ToS

Through this figure (https://www.visualcapitalist.com/terms-of-service-visualizing-the-length-of-internet-agreements/), the result of examining ToS from notable U.S. companies have revealed that reading times extend beyond 20 minutes, with Microsoft's ToS, the lengthiest among them, requiring over an hour. This intricate and time-consuming nature contributes significantly to user reluctance in engaging with ToS agreements thoroughly.

The trend in the changing of ToS

Google(https://policies.google.com/terms?hl=en-US) and Meta(https://www.facebook.com/legal/terms/) have transformed their ToS to be more reader-friendly. They are aware that users tend to avoid difficult ToS due to the prevalence of complex legal terminology and the lack of categorization, making it challenging for users to comprehend. Consequently, various leading companies have made changes to their ToS, aiming to make them more accessible and easy to read.

Purpose

Our program focuses on extracting potentially hazardous content from ToS agreements and translating it into language that is not only easy to read but specifically tailored for secondary school students. By addressing the accessibility gap and promoting user

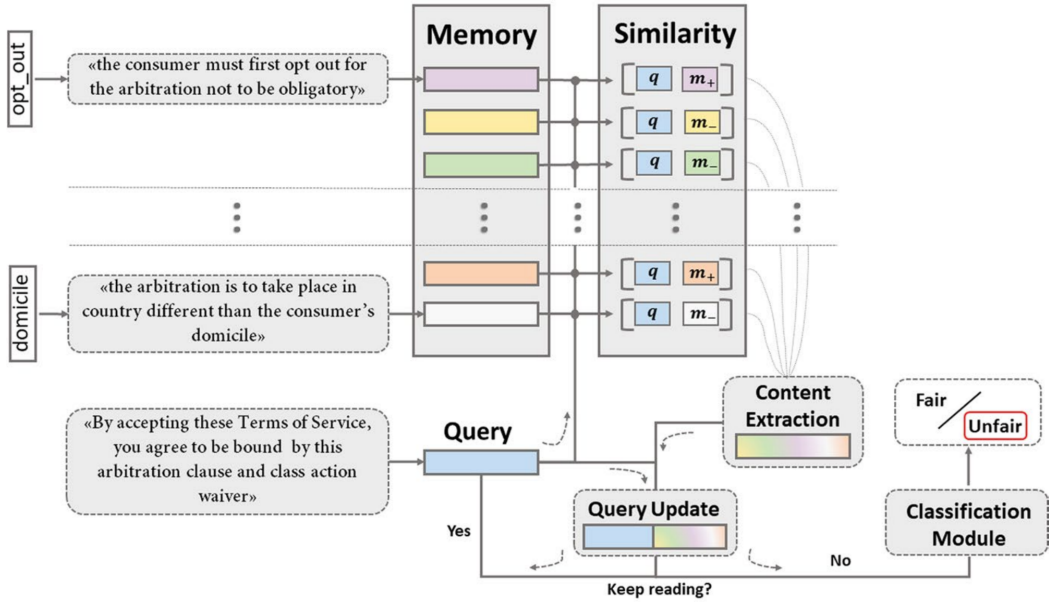
[3] Geoffrey A. Fowler *I tried to read all my app privacy policies. It was 1 million words.* https://www.washingtonpost.com/technology/2022/05/31/abolish-privacy-policies/

understanding, we envisions mitigating issues stemming from the pervasive lack of ToS comprehension.

MANN (Memory Augmented Neural Network)

Model summarization

In the first stage of the model, we have to identify potentially hazardous clauses from given ToS. To achieve this, we have applied the MANN model from the paper we've cited"[4]. The basic idea of this model is to provide external knowledge (so called memory) that helps the model to identify hazardous clauses.



The detailed workflow of the model is given as a figure. We provide the memories to the model that are used as standards for judging the hazardousness of the clause. The sample memories given on the figure are "The consumer must first opt out for the arbitration not to be obligatory" and "The arbitration is to take place in a country different from the consumer's domicile". Like these in this model, the memories are sentences explaining some cases of hazardous ToS.

So when the ToS we want to analyze is given, we parse it into clauses. Then we embed the clauses and the predefined memories, and evaluate the similarity between them. After that, concatenate the embedded clauses with the similarity vectors and run it through a classifier to judge if it's hazardous or not.

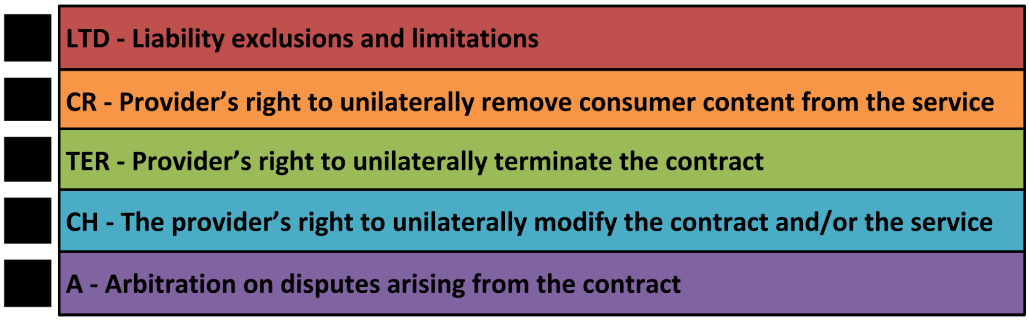
[4] Ruggeri, F., Lagioia, F., Lippi, M., & Torroni, P. (2022). Detecting and explaining unfairness in consumer contracts through memory networks. *Artificial Intelligence and Law*, *30*(1), 59-92.

And additionally to avoid overfitting, we applied 10-fold cross validation. So after finishing training, we have 10 models trained on different datasets and sum up the results of them to get a final prediction.

Dataset

For training and testing, we have used the same data from the original paper. It contains 100 ToS documents and 21,063 sentences. Among the data, 2346 sentences were marked as clearly or potentially unfair. Since 11.1% of the sentences are marked to be unfair, it emphasizes the importance of our work once again.

In detail, we are dealing with 5 types of unfairness that are most challenging to detect. Those types are described on the table below. Arbitration (A) clauses were the most uncommon cases in which they appeared on only 43 of 100 documents, and other types of clauses appeared on at least 83 of 100 documents.



Results

Type of unfairness	Average result (Accuracy, F1, Recall, Precision)
A	98.0%, 34.8%, 63.4%, 25.2%
CH	96.8%, 47.3%, 81.7%, 33.7%
CR	98.1%, 47.0%, 79.8%, 33.9%
TER	96.4%, 48.7%, 76.1%, 36.0%
LTD	95.8%, 51.7%, 81.7%, 38.1%

We have trained our model for 20 epochs for every 10 fold models, and the average of the results of the models are on the table. It was promising that all of them have shown accuracy over 95%. But the precision metric is really low, showing that nearly two of

three clauses we have marked as unfair are false positive cases for every type of unfairness.

This happens because as we have mentioned earlier, the data is biased (Only 11.1% of sentences are marked as unfair). So to solve this problem, we can consider several solutions like applying imbalanced classification training techniques or making the data more balanced (more unfair clauses) by considering every type of unfairness as one, not only specific kinds of them. The code, dataset, models and the results are on github: Link To Repository

Large Language Models

Up-Goer Five Challenge

During our need finding process, we identified the need to reduce the user's burden of understanding terms and service. The lengthy terms of service document can be reduced with the use of the MANN model by selecting the terms that are potentially hazardous with their hazardous categories. However, still, the identified terms are difficult to understand with their unnecessarily complex vocabulary and lengthy descriptions. Because of the weight and tone of these documents, we understand the intention of the writer in needing to convey this information in a complex manner, but for the users, reading through the term and understanding in first-glance is very important.

With this in mind, we were inspired by the 'Up-Goer Five Challenge,'[5] which was originally intended to raise the value of research scientists do. It challenges scientists to communicate complex concepts in an understandable way such that a broader audience can understand the concept without any complex phrases or jargons. It aims to identify and articulate the fundamental aspects of their work, getting to the heart of their research. Scientists could showcase works in an understandable and appreciable manner. The same context applies for our approach. The terms of service writers need to reduce their jargon so that the point can be presented in an 'understandable' and 'appreciable' manner. The method is simple, rephrase whatever description text using only the most used 1000 words in English. However, 1000 words covers a very small scope of words when trying to portray such complex ideas forward.

In this project, we use LLMs to simplify the terms of service, evaluate the proportion of words in the simplified version that belong to the most used 1000 words[6]. If the

[5] *The up-goer five text editor*. The Up-Goer Five Text Editor. (n.d.). https://splasho.com/upgoer5/
[6] Norman, D. (n.d.). *1,000 most common US English words*. Discover gists · github. https://gist.githubusercontent.com/deekayen/4148741/raw/98d35708fa344717d8eee15d11987de6c8e26d7d/1-1000.txt

proportion of words does not exceed a minimum threshold, we iterate the simplification after passing the words that do not belong in the frequent word list.

Prompt Engineering

We first attempted the simplification just to see the performance of the large language models. We used the Meta Llama 2's Llama 2 7B Chat[7] considering the scope of this project and because it was free and open source. Thankfully the results were satisfactory so we proceeded with building the structure of the prompts. Before we proceed, we have to mention that the accuracy of simplification is very subjective due to the nature of evaluation of Large Language Models. We set the threshold proportion the frequent words have to cover as 75% and set the maximum number of interactions as 3 based on several experimental trials. These values showed reasonable results and processing time. The final prompt we settled upon is as follows:

```
INPUT:
Can you please rephrase the following term of service using simpler
language that is easy to understand? <INPUT TERM>
Give me an output in the form: 'Rephrased Version: <rephrased term
of service>'. Your response must have the string 'Rephrased
Version:' in it! You must have a simplified version that follows
the string 'Rephrased Version"

OUTPUT:
You have requested analysis on the term: <INPUT TERM>

The given term may be dangerous because it belongs to type
category:<HAZARD CATEGORY> -> <HAZARD CATEGORY DESCRIPTION>
To better understand the term, this is a simplified version of it:
<SIMPLIFIED TERM>
```

A sample runthrough of the KakaoTalk Term of Service is as follows:

```
Input:
The personal information may be used or provided to a third
party, without consent from the user and in accordance with
related laws, to the extent that the information is collected
and used for reasons related to the purpose of collection.

Output:
```

[7] *TheBloke/llama-2-7B-chat-GGML · hugging face*. TheBloke/Llama-2-7B-Chat-GGML · Hugging Face. (n.d.). https://huggingface.co/TheBloke/Llama-2-7B-Chat-GGML

```
The personal information may be shared with third parties
without asking for permission from the user, as long as it is
used for reasons related to its original collection purpose.
```

DISCUSSION

Through this project, we have identified and addressed complex and hazardous terms in Terms of Service agreements by leveraging a model to pinpoint such terms and employing Language Models to simplify and enhance user comprehension.

During our user study we asked 13 participants how they thought of the whole process by asking them to look through the original terms of service and the simplified version. Every participant was satisfied with the simplified version of the ToS clauses. They agreed that the simple version is easy to read and understand while preserving the original meaning of it and stressed the complexity of the original versions. Some commented that these simplifications would definitely make them more relieved and willing to read ToS clauses with more attention.

However 10 (76.9%) of the participants argued that it was hard to understand the reason why the clauses are marked as a certain type of unfairness. This is because we have chosen the types of unfairness in interest which are challenging to detect. So since we are using memories, if we select the memory with highest similarity it can explain the reason why the clause is marked as unfair.

Furthermore, 4 (30.8%) of the participants argued that they couldn't understand why the clauses are unfair. It's because we have lots of false positive (false unfair) cases, so some of the clauses marked as unfair are sometimes fair. To minimize this, we can apply some solutions that we have discussed at the results section of MANN.

LIMITATIONS AND FUTURE WORK

Our project has demonstrated promising outcomes in simplifying terms of service using a small-scale large language model. However, several limitations and areas for improvement have been identified. Our initial choice of a smaller scale language model was driven by considerations of time and cost. However, when we compared our model's performance to ChatGPT from OpenAI using the same prompt, we observed that the latter provided faster results with a consistent output format. Our language models encountered challenges such as refusing to generate an output or not adhering to the structured "Rephrased Version: <Simplified Term> " format, making parsing difficult. To enhance accuracy and satisfaction, employing larger models with OpenAPI

keys might have yielded better outcomes.

Prompt engineering is another primary area for potential development in our project. The restriction on prompt length limited our creativity in approaching prompts, impacting prompt quality. Exploring different prompt engineering strategies could be a promising avenue for future steps. For instance, integrating Retrieval based Augmentation to present a frequent list of words and constraining the language model to rephrase within that scope, or employing few-shot engineering by providing examples of rephrased terms, could be explored. Extracting a subset of indispensable words, particularly those with specific purposes like 'privacy,' could further enhance output quality. Also, based on the role of the LLM, we could try to set different personas that would make the simplified output more relatable and effectively conveyed (having some humorous personality that makes the content more engaging and lightweight). We could also use LLMs to make up some real-life examples in which these terms of service, when misunderstood or neglected, can cause detrimental problems. Not only that, we could expand further to use generative AI to use various modalities to illustrate such scenarios.

Despite investing significant time in model selection, environment setup, and determining our direction for aiding users in understanding terms of service, we did not reach the stage of deploying the system with a user-friendly interface. Recognizing the importance of accessibility in meeting identified needs and maximizing the system's merit, developing an appropriate user interface stands out as a crucial milestone for the long-term application of our system. We wholly acknowledge that the success of the system is not solely dependent on the underlying technology but also on how well users can interact with and benefit from the system's capabilities.

Furthermore, our evaluation process, as highlighted in the assessment of the language model and user study results, points out that subjectivity is a significant limitation. A potential approach is exploring NLP approaches to compare semantic and syntactic similarity between original and simplified versions of our terms of service. However, we acknowledge the impracticality of entirely assessing response quality in this manner.

Another potential approach is conducting more diverse user studies involving individuals with varying levels of understanding of terms of service, including experts like terms of service writers themselves and novices such as middle school students. Additionally, leveraging crowdsourcing platforms like MTurk to collect data on model performance and implementing reinforcement learning with a reward model could provide a more comprehensive evaluation framework.

Addressing these limitations and pursuing these avenues will contribute to the continued development and refinement of our project. We believe there are various thought provoking areas we can touch upon to make a more impactful deployment of our system.

References

Obar, J. A., & Oeldorf-Hirsch, A. (2016). *The biggest lie on the Internet: Ignoring the privacy policies and terms of service policies of social networking services*. Proceedings of the 44th Research Conference on Communication, Information & Internet Policy. Arlington, VA.

Deloitter *2017 Global Mobile Consumer Survey: US edition p.12* https://www2.deloitte.com/content/dam/Deloitte/us/Documents/technology-media-telecommunications/us-tmt-2017-global-mobile-consumer-survey-executive-summary.pdf

Geoffrey A. Fowler *I tried to read all my app privacy policies. It was 1 million words.* https://www.washingtonpost.com/technology/2022/05/31/abolish-privacy-policies/

Norman, D. (n.d.). *1,000 most common US English words*. Discover gists · github. https://gist.githubusercontent.com/deekayen/4148741/raw/98d35708fa344717d8eee15d11987de6c8e26d7d/1-1000.txt

Ruggeri, F., Lagioia, F., Lippi, M., & Torroni, P. (2022). Detecting and explaining unfairness in consumer contracts through memory networks. *Artificial Intelligence and Law*, *30*(1), 59-92.

TheBloke/llama-2-7B-chat-GGML · hugging face. TheBloke/Llama-2-7B-Chat-GGML · Hugging Face. (n.d.). https://huggingface.co/TheBloke/Llama-2-7B-Chat-GGML

The up-goer five text editor. The Up-Goer Five Text Editor. (n.d.). https://splasho.com/upgoer5/