ECE1786 Project Presentation
Unfair ToS – LLM Unfair Contract Term Detector

Jianing Zhang

Jiazhou Liang

UNIVERSITY OF TORONTO

DEFY GRAVITY

Goal & Motivation

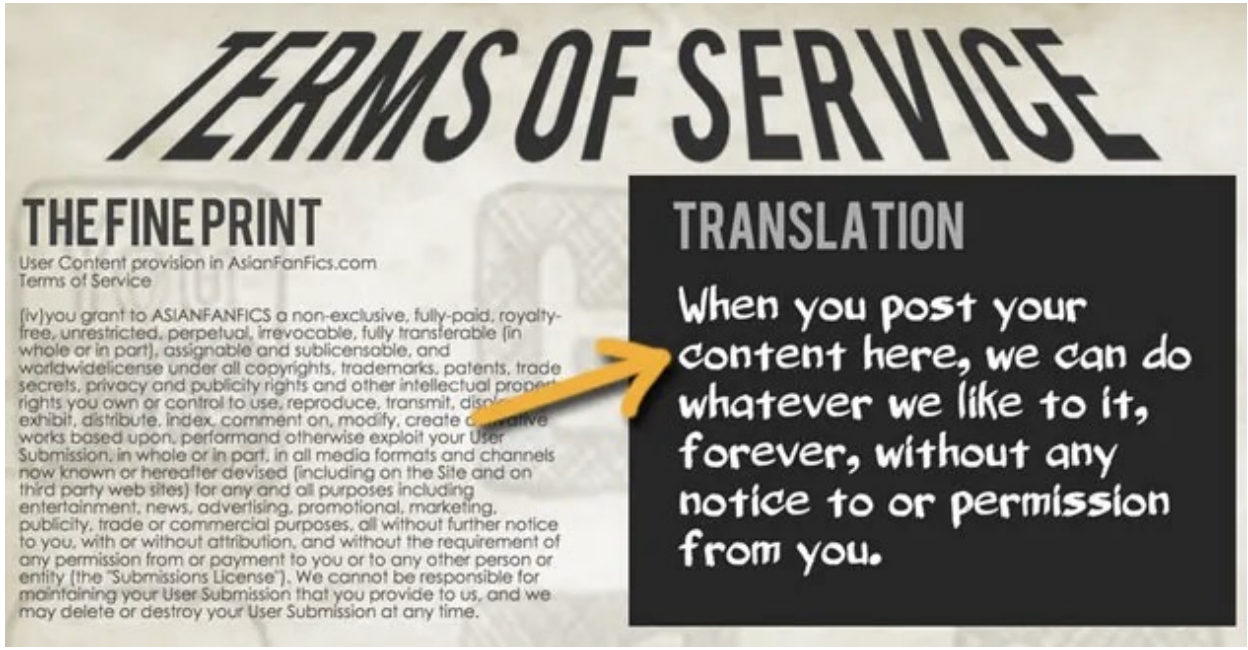
Terms of service (ToS)

- Legal agreements between service providers and its users
- Complex and lengthy, with legal words
- General users struggle to understand

Using LLMs to solve this problem!

Potential Problems

1. User accepted unfair terms (lose their rights)
2. Disadvantage in potential conflict
3. Most important: **Privacy Issue**



UNIVERSITY OF TORONTO

LLM based Contract Term Detector

1. Text Highlight

Highlighting Sentences in the ToS documents that Required User Attention (can be either fair or unfair)

2. Text Simplification

Simplified the highlighted sentences to make the general user population easy to read and comprehend

3. Text Classification

Identify whether each highlighted sentences is fair and unfair. If unfair, what is the reason?

```
Pinterest specifically disclaims
any and all warranties and
conditions of merchantability,
fitness for a particular purpose,
and non-infringement, and any
warranties arising out of course
of dealing or usage of trade
```

```
Pinterest doesn't promise services
will meet requirements or violate
any rights

Label: Unfair
```

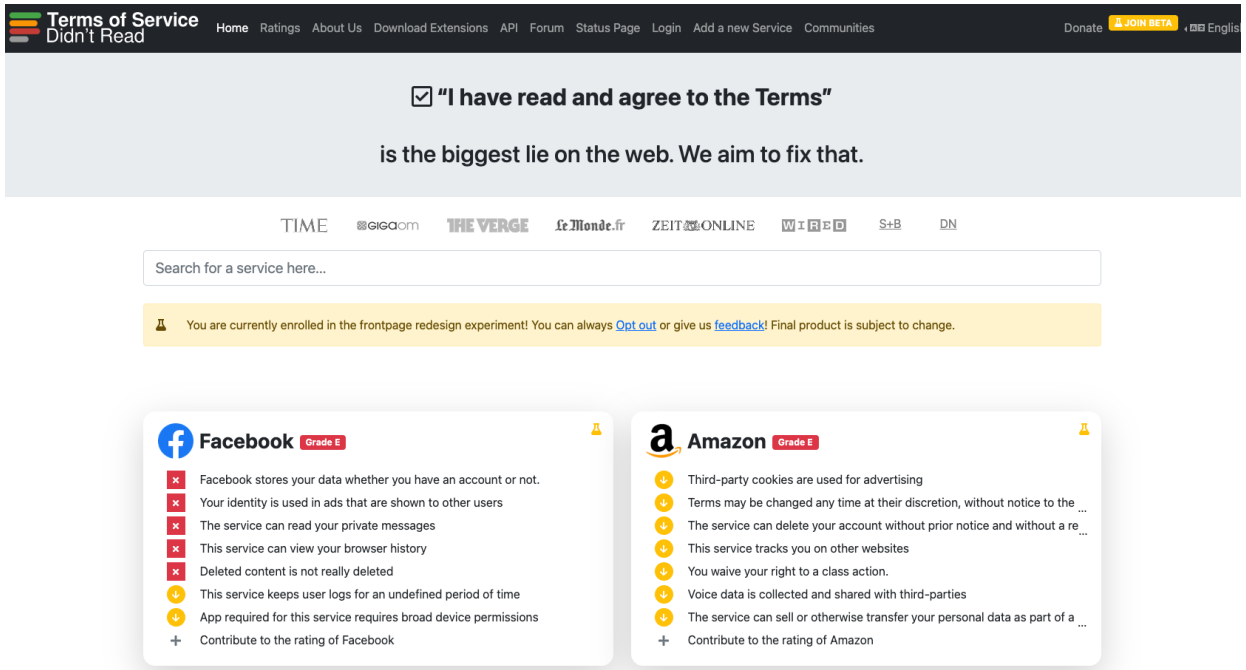
UNIVERSITY OF TORONTO

Data Collection and Processing

Text Highlight: What should be consider as ‘Important’ to highlight?

Terms of Service; Didn't Read
Contain over 1000 ToS documents

500 Contributors highlight important sentences with approvers validation



How do we use it

- Original Documents as input data for prompt engineering
- Highlight sentences as the ground truth for evaluating GPT's performance

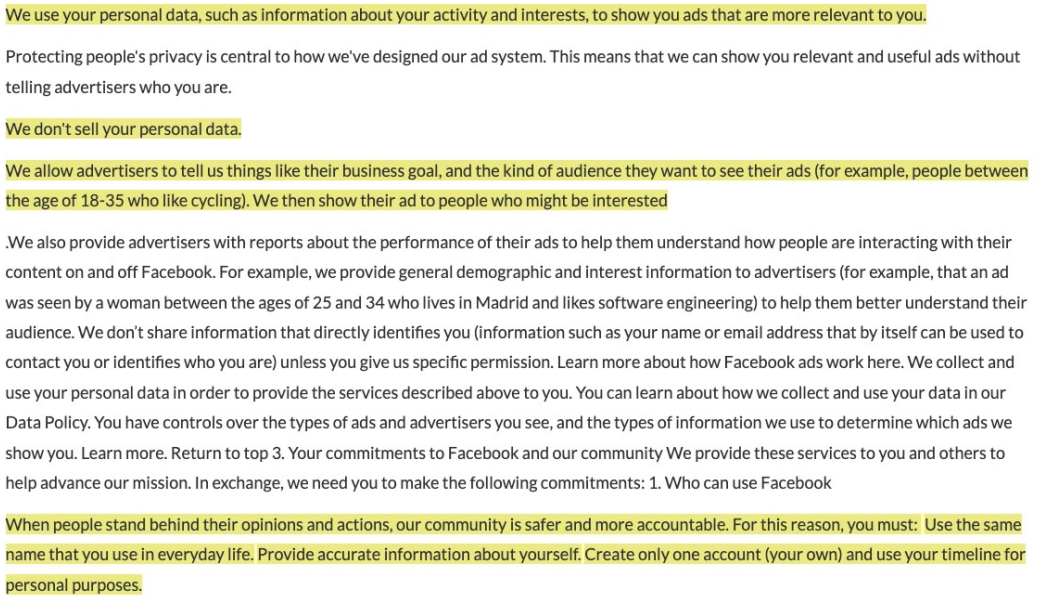
UNIVERSITY OF TORONTO

"Terms of Service; Didn't Read"

The Community that Highlight Important Sentences

- **Problem 1: Too many ToS**
  - Keep 40 - 50 highlighted sentences
  - Result in 11 documents, around 300 sentences in each documents
- **Problem 2: Too many noises**
  - Tokenized into sentences using the NLTK 'english.pickle'
  - Cleaning: Lowercase, special characters, HTML tags, duplicate sentences

- Label each sentence as 1 if Highlighted or 0 if Not



We use your personal data, such as information about your activity and interests, to show you ads that are more relevant to you.

Protecting people's privacy is central to how we've designed our ad system. This means that we can show you relevant and useful ads without telling advertisers who you are.

We don't sell your personal data.

We allow advertisers to tell us things like their business goal, and the kind of audience they want to see their ads (for example, people between the age of 18-35 who like cycling). We then show their ad to people who might be interested

.We also provide advertisers with reports about the performance of their ads to help them understand how people are interacting with their content on and off Facebook. For example, we provide general demographic and interest information to advertisers (for example, that an ad was seen by a woman between the ages of 25 and 34 who lives in Madrid and likes software engineering) to help them better understand their audience. We don't share information that directly identifies you (information such as your name or email address that by itself can be used to contact you or identifies who you are) unless you give us specific permission. Learn more about how Facebook ads work here. We collect and use your personal data in order to provide the services described above to you. You can learn about how we collect and use your data in our Data Policy. You have controls over the types of ads and advertisers you see, and the types of information we use to determine which ads we show you. Learn more. Return to top 3. Your commitments to Facebook and our community We provide these services to you and others to help advance our mission. In exchange, we need you to make the following commitments: 1. Who can use Facebook

When people stand behind their opinions and actions, our community is safer and more accountable. For this reason, you must: Use the same name that you use in everyday life. Provide accurate information about yourself. Create only one account (your own) and use your timeline for personal purposes.

Text (From Facebook)	**Label**
We use your personal data, such as information about your activity and interest, to show you ads that are more relevant to you	1
Protecting people privacy is central to how we ve designed our ad system.	0
We do not sell your personal data	1

UNIVERSITY OF TORONTO

Data Collection and Processing

Text Classification: What is fair / unfair?

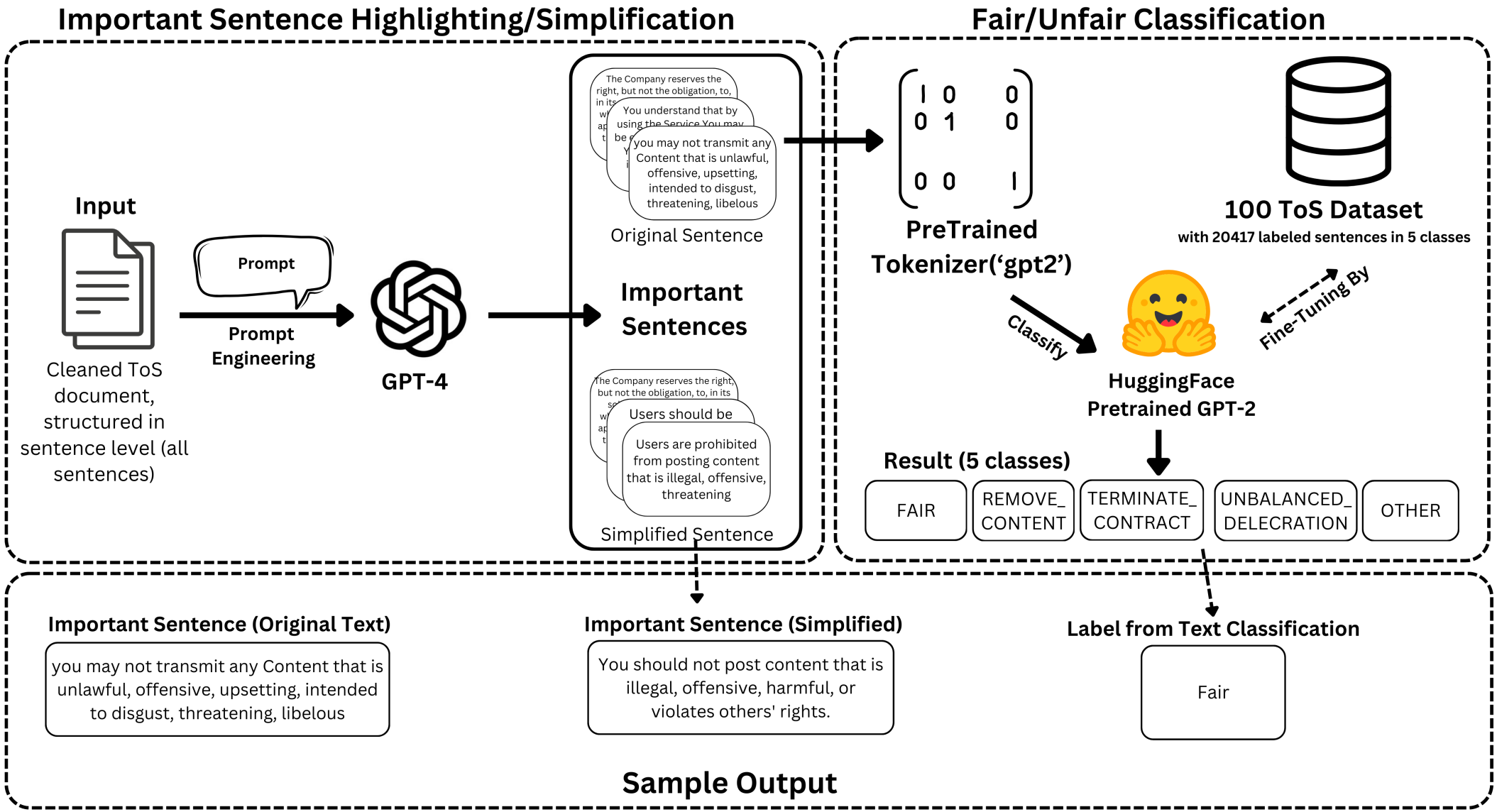
- From Paper 'Detecting and explaining unfairness in consumer contracts through memory networks'
- Analyzed 100 ToS documents, 20,471 sentences
- Labeled each sentences into
  - 'Fair' (18,235)
  - 8 Different 'Unfair' Reasons (2,182)
- Significant Imbalanced (Expected)
  - Grouped samples into 4 reasons, oversampling
- 80% training, 20% testing

Index	Class Name	Description	Original Sample	Samples after oversampling
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

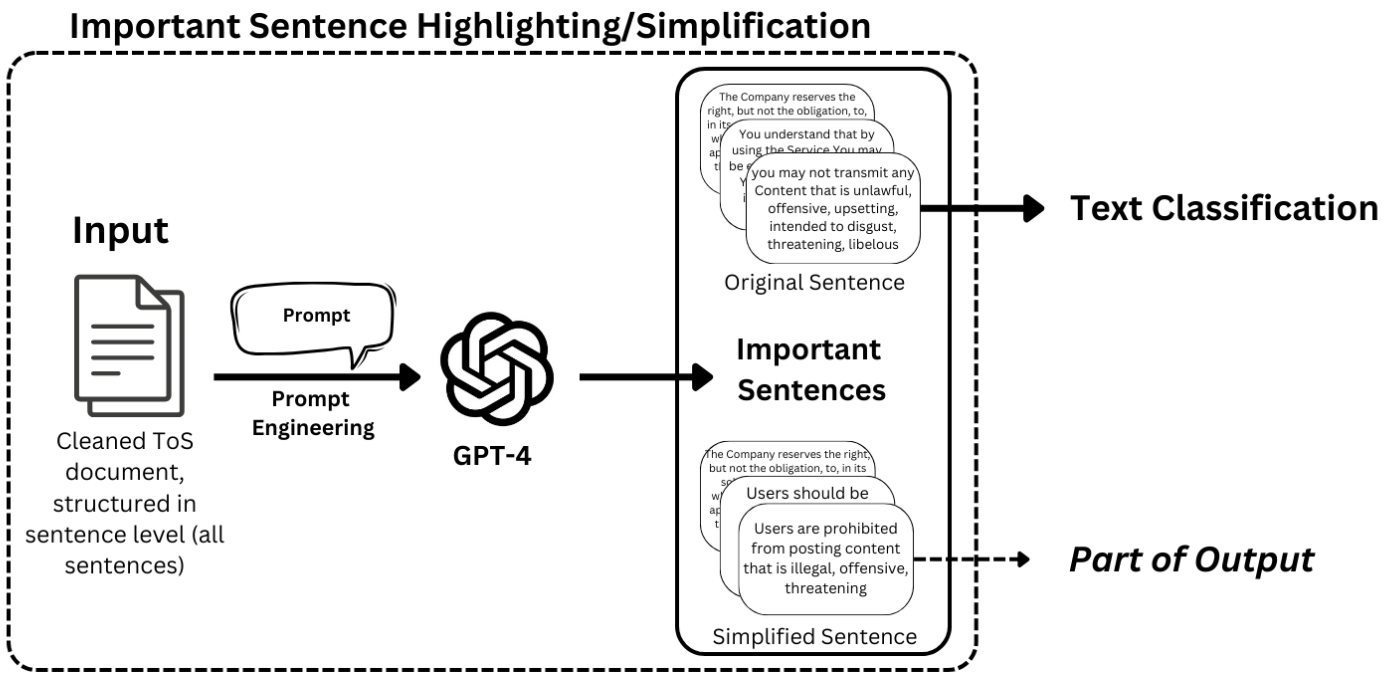
Sample: we may remove your dna results and/or dna reports and/or terminate your membership at any time , without notice.

Label: 1 Remove Content

UNIVERSITY OF TORONTO


Important Sentence Highlighting/Simplification
Fair/Unfair Classification
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,
You understand that by
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
PreTrained Tokenizer('gpt2')
100 ToS Dataset
with 20417 labeled sentences in 5 classes
Classify
Fine-Tuning By
HuggingFace Pretrained GPT-2
Result (5 classes)
FAIR
REMOVE_ CONTENT
TERMINATE_ CONTRACT
UNBALANCED_ DELECRATION
OTHER
Important Sentence (Original Text)
you may not transmit any Content that is unlawful, offensive, upsetting, intended to disgust, threatening, libelous
Important Sentence (Simplified)
You should not post content that is illegal, offensive, harmful, or violates others' rights.
Label from Text Classification
Fair
Sample Output

Text Highlight & Simplification



Use prompt to ask GPT-4 to perform following two tasks:

1. **Highlight:** 50 sentences deemed 'important' (same definition as ToS;DR website)

2. **Simplification:** For each *highlighted* sentence, generated an easier comprehensive version.

Readability is more important. BERT score is only served as a reference.

UNIVERSITY OF TORONTO

Prompt Engineering

Chain of Thought

Steps to find ‘important’

think step by step,

1. Who is the service provider? What is its user population?

2. what is important for this user group? 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 using this metrics?

Few-Shots

Examples of ‘important’

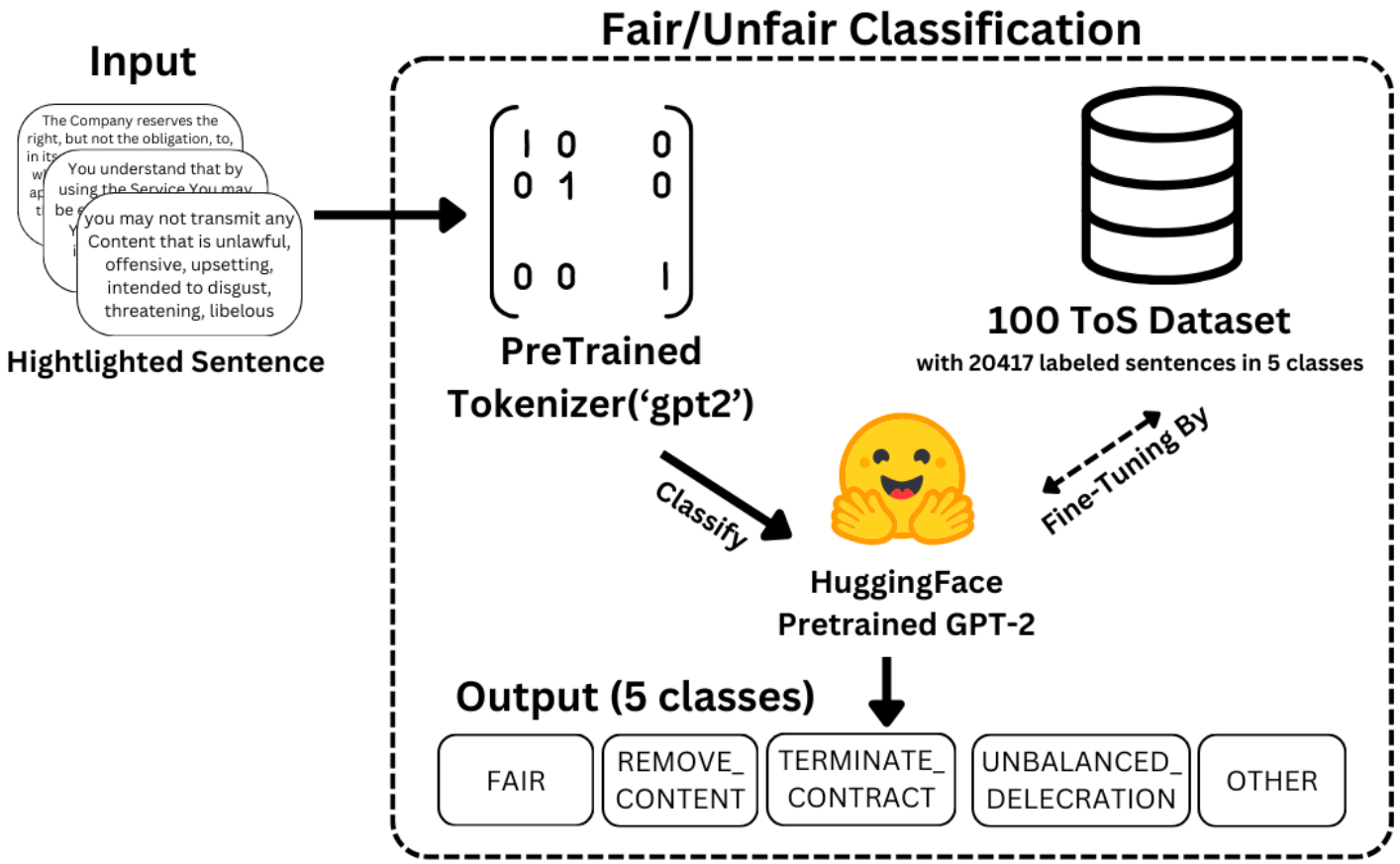
A sentence is important if it relates to one or more of the following 57 practices:

*1. **Retention of User-Generated Content:** Keeping user content even after the user closes their account.*

*2. **User Tracking**: Monitoring users on other websites.*

…

Text Classification



Input: Highlighted sentences

Fine-Tune the GPT-2 Model with the 100-ToS dataset

Output: One of five classes:

- Fair
- Remove Content
- Terminate Contract,
- Unbalanced Rights
- Others

UNIVERSITY OF TORONTO

Text Highlight Results

Quantitative

- **Ground Truth:** The highlight Sentences in ToS;DR (matched with cosine similarity)
- **Baseline:** Text Rank, an unsupervised graph-based ranking algorithm
- **Metric:** Precision, Recall and F1-Score of 11 ToS documents

Outperform baseline model in all Precision, Recall and F1 score

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

Document	Text Rank F1	GPT-4 F1
LBRY	0.199	0.474
Google	0.296	0.531
Crunchyroll	0.415	0.494
Pure Dating	0.298	0.529
IDrive	0.211	0.481
FileFactory	0.380	0.545
HuffPost	0.215	0.418
…		

Text Highlight Results

Qualitative

GPT-4 does not have a very good Precision when using TOS;DR as ground truth

- But ToS;DR is not 100% ground truth, human has bias

We manually review each results, and

- Some important sentences only highlighted by LLM

Human contributors missed them! LLM does not

Examples only highlighted by GPT-4

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.

- Crunchyroll

Reason: freely adjusting price without notices

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.

- Pinterest

Reason: using user's content without notices

UNIVERSITY OF TORONTO

Text Simplification Results

Quantitative

- Readability score metric: Gunning fog index
  - Estimates the years of formal education a person needs to understand the text on the first reading
- Summary score metric: BERT Score (to check whether the information is covered)

	Gunning Fog index
Original Text	24.821
Simplified Text	9.484

	BERT Score
Precision	0.840
Recall	0.80
F1	0.815

Original Text > 17

Fog Index	Reading level by grade
17	College graduate
16	College senior
15	College junior
14	College sophomore
13	College freshman
12	High school senior
11	High school junior
10	High school sophomore
9	High school freshman
8	Eighth grade
7	Seventh grade
6	Sixth grade

Simplified Text

UNIVERSITY OF TORONTO

Text Simplification Results

Qualitative

Original text:

- difficult vocabulary: liability, infringement,
- complex grammar, wordy
- Required read several times

Simplified Text:

- much shorter and no difficult vocabulary
- Cover most meaning in the sentences

Things simplified text cannot resolve

- Ambiguous in the original text
  - For example, what is ‘claim’?

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. (Gunning Fog: 19.70)

You have to bring up any issues with Tumblr within one year or you can't at all. (6.8)

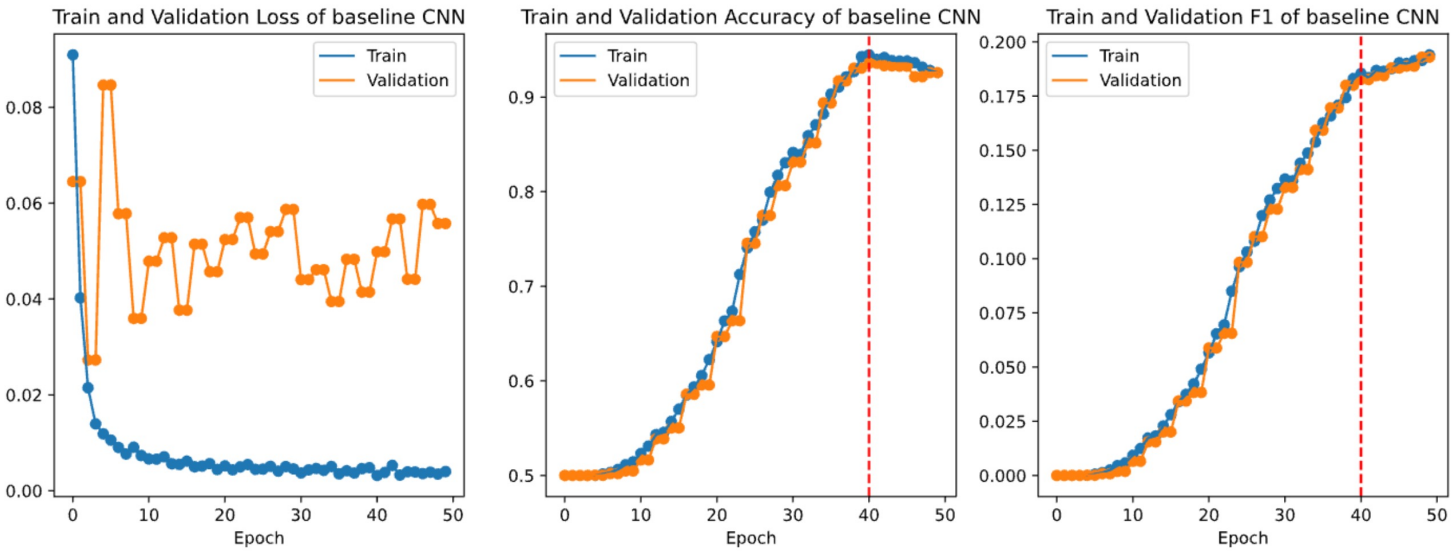
without limiting the foregoing, to the full extent permitted by law, tumblr disclaims all warranties, express or implied, of merchantability, fitness for a particular purpose, or non-infringement. (19.69)

Tumblr doesn't promise that the service will meet your needs or that there won't be errors. (6.40)

UNIVERSITY OF TORONTO

Text Classification Results

Quantitative



Training and Validation Result of baseline CNN model

Baseline CNN:

•Trained and tested on 100 ToS documents

• 92.6% accuracy, 19.3% F1 score

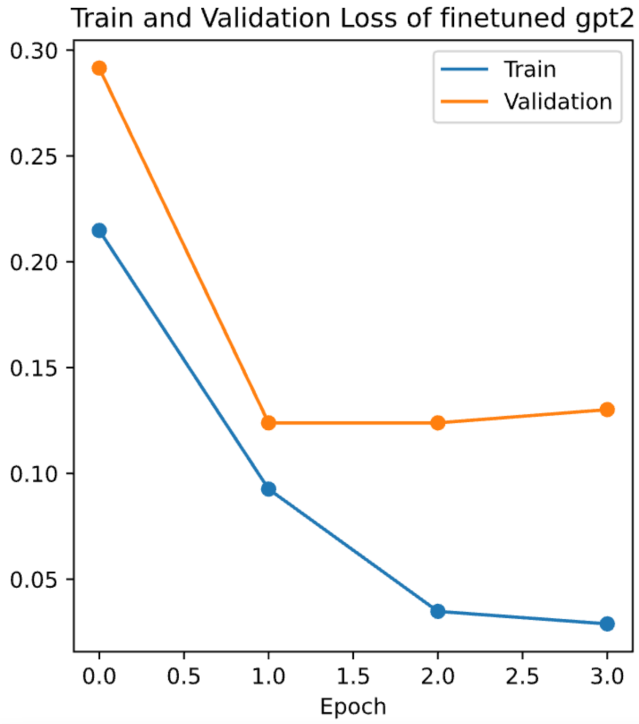
• Classify into the majority class (fair)

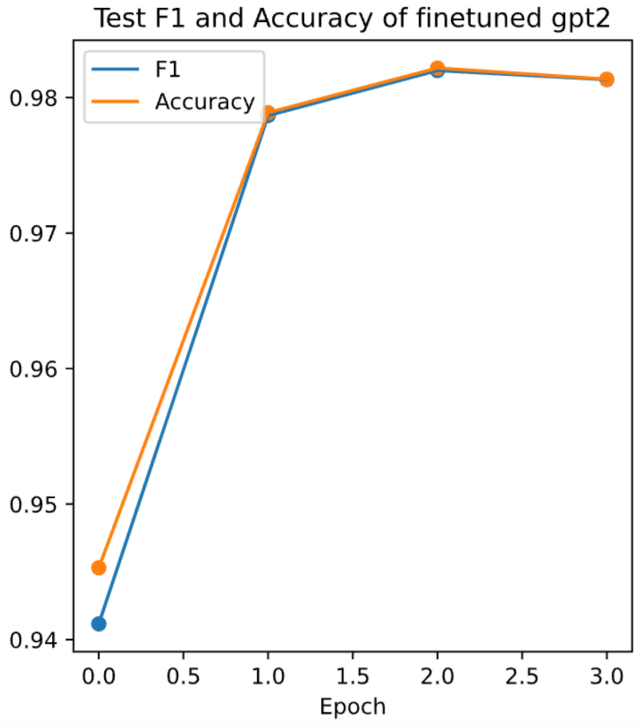
Text Classification Results

Fine-tuned GPT-2:

- 98.21% accuracy, 98.19% F1 score
- Outperforming the baseline in both metrics

- High F1 means more robust to imbalanced fair and unfair sentences
  - Most real-world cases





Left: The training and validation loss

Right: Test F1 and Accuracy through Epoch

Example

Goodreads – Around 500 lines in its ToS document, 76 highlighted / fair: 57 Unfair: 3 ,7 ,5, 4

…if these laws apply to you, some or all of the above disclaimers, exclusions, or limitations may not apply to you, and you might have additional rights.

you acknowledge that goodreads has the perpetual and irrevocable right to delete any or all of your content and data from goodreadss servers and from the service, whether intentionally or unintentionally, and for any reason or no reason, without any liability of any kind to you or any other party.

in order to protect our members from unsolicited advertising or solicitation, goodreads reserves the right to restrict the number of …

Simplification	Label
Your content can be deleted by Goodreads anytime	Unfair Remove Content

Example

Google – Around 200 lines in its ToS document

	Simplification	Label
… including google search, will continue to find and display your content as part of their search results.		
for example, to promote a google app, we might quote a review you wrote.		
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: you materially or repeatedly breach these terms, service-specific additional terms or policies we re required to do so to comply with a legal requirement or a court order we reasonably believe that your conduct causes harm or liability to a user, third party, or google for example, by hacking, phishing, harassing, spamming, misleading others, or scraping content that doesn t belong to you if you believe your google account has been suspended or terminated in error, you can appeal	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.	Unfair Terminate Contract
of course, you re always free to stop using our services at any time to use our services on behalf of an organization: an authorized representative…	You can discontinue using Google services whenever you want.	Fair

Discussion and Learnings

Model Performance

The text simplification and classification functionalities exhibited strong performance.

Evaluation Challenge

Evaluating the text highlighting function posed challenges, especially in assessing qualitative aspects.

Professional Consultation

Consult legal experts for deeper insights and improve evaluation methods.