Harmful Terms and Where to Find Them: Measuring and Modeling Unfavorable Financial Terms and Conditions in Shopping Websites at Scale

Anonymous Author(s)

Abstract

Terms and conditions for online shopping websites often contain terms that can have significant financial consequences for customers. Despite their impact, there is currently no comprehensive understanding of the types and potential risks associated with unfavorable financial terms. Furthermore, there are no publicly available detection systems or datasets to systematically identify or mitigate these terms. In this paper, we take the first steps toward solving this problem with three key contributions.

First, we introduce *TermMiner*, an automated data collection and topic modeling pipeline to understand the landscape of unfavorable financial terms. *Second*, we create *ShopTC-100K*, a dataset of terms and conditions from shopping websites in the Tranco top 100K list, comprising 1.8 million terms from 8,251 websites. Consequently, we develop a taxonomy of 22 types from 4 categories of unfavorable financial terms—spanning purchase, post-purchase, account termination, and legal aspects. *Third*, we build *TermLens*, an automated detector that uses Large Language Models (LLMs) to identify unfavorable financial terms.

Fine-tuned on an annotated dataset, *TermLens* achieves an F1 score of 94.6% and a false positive rate of 2.3% using GPT-4o. When applied to shopping websites from the Tranco top 100K, we find that 47.21% of these sites contain at least one unfavorable financial term, with such terms being more prevalent on less popular websites. Case studies further highlight the financial risks and customer dissatisfaction associated with unfavorable financial terms, as well as the limitations of existing ecosystem defenses.

CCS Concepts

• **Information systems** → **Web mining**; • **Security and privacy** → *Social engineering attacks*; • **Social and professional topics** → *Commerce policy*.

Keywords

Topic modeling, unfavorable terms, consumer protection, terms and conditions dataset, deceptive content

1 Introduction

In 2023, U.S. e-commerce sales reached $1.12 trillion [1], with users frequently engaging with websites that impose terms and conditions on financial transactions. While these terms are often benign, they can also facilitate scams or impose unfair financial consequences on unsuspecting users. This risk is heightened by the fact that most users rarely read these lengthy, jargon-filled terms [4, 57, 69], and are often not required to do so before completing a purchase.

In this work, we define unfavorable financial terms as those that are one-sided, imbalanced, unfair, or malicious, thereby disadvantaging users. Figure 1 shows a real-world example of harmful financial terms on a website selling earbuds at seemingly attractive prices. When users make a purchase, the terms and conditions obligate the users to a fitness app subscription with a recurring $86 monthly fee. *This obligation is not disclosed at all during the purchase process*.

Unfair or harmful financial terms can also exist on legitimate websites—unlike traditional social engineering scams, these terms may not be inherently deceptive but can still cause substantial losses. Figure 5 in the Appendix presents the T&Cs from Celsius [14], a cryptocurrency company bankrupt in 2022. These terms stipulate that if Celsius goes bankrupt, users could lose digital investments since they would be treated as unsecured creditors. A judge later ruled that Celsius owned its users' cryptocurrency deposits based on these terms [34], highlighting the real financial risks such terms pose to users.

It is worth noting that the website in Figure 1 operated for at least a year without being flagged by major browsers before its shutdown in June 2024, showing the current defense ecosystem's lack of understanding and mitigation strategies for unfavorable financial terms. Likewise, Celsius's unfair terms only gained attention during bankruptcy proceedings. Despite their impact, few mitigation methods exist for unfavorable financial terms.

A possible approach to addressing the concern is to extend the current methods for detecting social engineering scams and dark patterns [6, 10, 20, 41, 48, 49, 81]. Unfortunately, such an extension is not straightforward. Many of these methods are not designed to detect unfavorable financial terms, as they typically focus on content-based features like word patterns, images, website structures, or external indicators like link length and certificates [5, 7, 22, 39, 63, 82], which are unrelated to the detection of unfavorable financial terms. Similarly, work on online agreement analysis focuses on privacy policies [13, 32, 78, 83] or on terms deemed invalid under EU law [11, 25, 26, 40, 42, 43].

The advent of large language models (LLMs) provides a more effective approach to analyzing policies like T&Cs [24, 61], but their potential in this area has yet to be fully explored. This paper aims to fill this gap. Several challenges need to be addressed for the approach to be effective: (1) *a lack of understanding of unfavorable financial terms in the wild*, (2) *no publicly available datasets for studying these terms*, and (3) *the absence of detection systems to identify and mitigate them*. This work seeks to address these challenges through a large-scale measurement study. To the best of our knowledge, this is the first systematic effort to categorize and detect unfavorable financial terms in real-world online shopping websites. Our contributions are as follows:

- ***Data collection and topic modeling pipeline*:** We present *TermMiner*, a scalable pipeline for collecting and analyzing terms and conditions from shopping websites. The pipeline includes (1) a data collection module to collect from the

Tranco top websites and fraudulent e-commerce datasets, (2) an LLM-based classification module of terms, and (3) an interactive topic modeling module. We identify unfavorable financial terms using the FTC's definitions of unfair practices [15].

- ***ShopTC-100K dataset*: :** We create the *ShopTC-100K* dataset, consisting of terms and conditions from shopping websites in the Tranco top 100K list, which will be open-sourced. This dataset includes 1.8 million terms extracted from 8,251 shopping websites.
- ***Unfavorable financial term taxonomy*: :** We develop a comprehensive taxonomy for unfavorable financial terms, covering 4 categories and 22 types, including terms related to purchase, billing, post-purchase activities, account termination and recovery, and legal conditions.
- ***Unfavorable financial terms detection*: :** We develop *TermLens*, a Chrome plugin-based detection framework using LLMs to automatically identify unfavorable financial terms. With a fine-tuned GPT-4o model, *TermLens* achieves a 94.6% F1 score and a 2.4% FPR on an annotated evaluation dataset.
- ***Measurement study*: :** We analyze 1.9 million terms from 8,979 online shopping websites, finding that 47.21% of English websites in the Tranco top 100k contain unfavorable financial terms. Our analysis reveals these terms are more common on less popular websites, with case studies highlighting the potential financial and legal harm to consumers.

2 Related Work

***Scam and fake e-commerce website detection*:** Detection methods for scam and fake e-commerce websites (FCW) typically rely on two types of features: external (e.g., URLs, certificates, logos, redirect mechanisms) [6, 9, 22, 51, 62, 63, 76, 85] and content-based (e.g., visual and HTML structures, images, scripts, hyperlinks) [6, 36, 39, 79–81]. These models are either rule-based or machine learning-based, with feature selection grounded in domain knowledge (e.g., indicative images, third-party scripts). However, no prior work in this line has considered terms and conditions and their financial impacts on users.

We consider social engineering scams to overlap with our detection target. The unfavorable financial terms in Figure 1 function similarly by deceiving users into signing up for additional subscriptions. However, as discussed in §3.3 and §5.2, unfavorable financial terms are not exclusive to scam websites. However, it's important for users to stay alert on websites with such terms. We view our work as the first to measure and detect unfavorable financial terms at scale.

***Dark patterns*:** Dark patterns are deceptive user interface designs intended to manipulate users into actions against their best interests [48]. Recent research has examined their psychological impact and influence on user decision-making [49, 53, 56, 77], while also exploring legal frameworks and strategies for intervention [29, 47].

Although terms and conditions are not part of the user interface design, we consider the unfavorable financial terms we identify to be closely related to dark patterns. The unilateral nature of these terms and their potential to hide uncommon or unexpected conditions

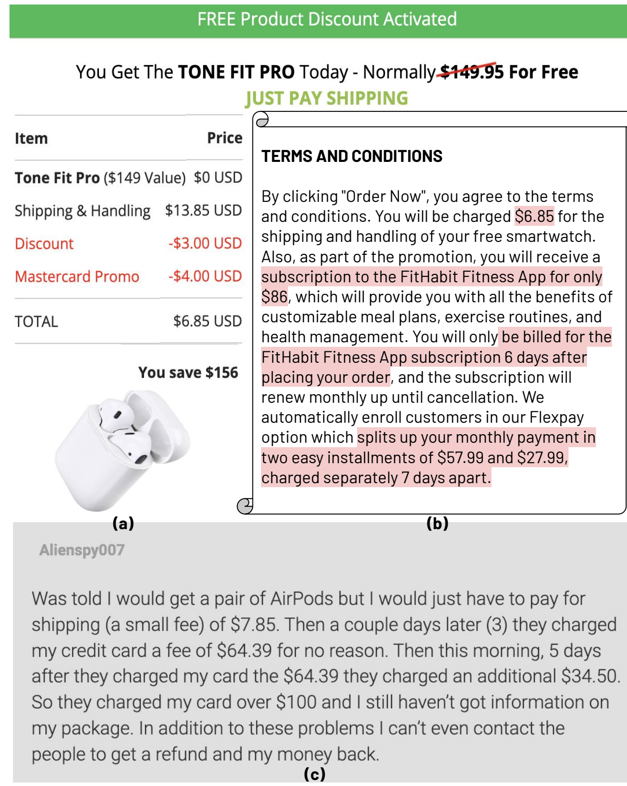


Figure 1: Unfavorable financial term example — (a) shows the payment page for Tone Fit Pro, a now-defunct website, with no mention of the subscription service on the payment page. (b) displays the T&Cs for Tone Fit Pro, which state that customers will be automatically enrolled in an $86 per month FitHabit Fitness App subscription with automatic renewal. (c) shows a screenshot of real-life victim complaints.

make them closely align with the characteristics of dark patterns: asymmetric, covert, deceptive, hiding information, and restrictive.

***Terms and conditions legal analysis*:** There is limited NLP-based analysis of legal documents like online contracts and terms of service [11, 26, 35, 40, 42, 43]. Prior studies, such as Lippi et al. [43] and Galassi et al. [26], typically focus on small datasets of terms and conditions (25 to 200 documents). However, their focus is mainly on assessing unfairness under the European Union's Unfair Contract Terms Directive [75] (i.e., clauses invalid in court). In contrast, our work specifically targets terms with direct financial impacts on users.

In this paper, we focus on the financial terms in the large-scale measurement of terms and conditions from English shopping websites, assessed using the definition of unfair acts or practices as provided by the Federal Trade Commission (FTC)'s Policy Statement on Deception [16]. A detailed comparison of our proposed term taxonomy with prior work is provided in Appendix E.

***Privacy policy analysis*:** A significant body of work investigates the viability of NLP-based analysis for privacy policies. One significant line of such research focuses on detecting contradicting policy statements (e.g., via ontologies [3] and knowledge graphs [17]) or

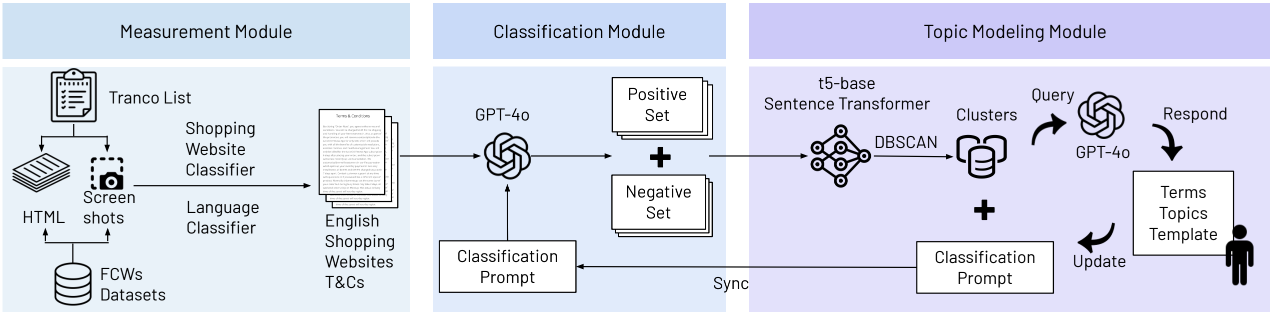


Figure 2: *TermMiner* (data collection and topic modeling pipeline)—(1) Measurement module: collects shopping websites from the Tranco list and fake e-commerce website datasets, filtering out English terms and conditions from shopping websites. (2) Term classification module: classify the terms into binary categories based on a given prompt. (3) Topic modeling module: leverages t5-base Sentence Transformer and DBSCAN for clustering. Topics are derived from the clusters using a combination of manual inspection and GPT-4o, employing a snowball sampling method [27] to iteratively develop a topic template of terms.

ambiguities [68]. Other areas include improving user comprehension [32], topic-modeling, and summarization of policies [2, 64].

In this work, we focus on financial terms and conditions which are distinct from privacy policies. While we also perform topic modeling, we are the first to apply such a pipeline to construct a taxonomy for unfavorable financial terms. Furthermore, detecting contradictions and ambiguities is orthogonal to the detection of malicious financial terms, making it difficult to apply similar techniques directly.

3 Understanding Unfavorable Terms

In this section, we outline our detection goal and present *TermMiner*, a pipeline for collecting, clustering, and topic modeling unfavorable financial terms from English shopping websites in the Tranco top 100,000 [73] and datasets of fraudulent e-commerce sites [6, 37]. As shown in Figure 2, the pipeline categorizes two types of terms: (1) financial terms that may pose future financial risks, and (2) unfavorable financial terms, identified as unfair, unfavorable, or concerning for customers. We then summarize the taxonomy of unfavorable financial terms, which fall into four broad categories: (1) purchase and billing, (2) post-purchase, (3) termination and account recovery, and (4) legal terms.

3.1 Threat Model

We aim to detect one-sided, imbalanced, unfair, or malicious *financial* terms in online shopping websites' terms and conditions, which pose significant risks to users, potentially leading to unexpected financial losses. These risks can arise from website operators seeking to limit liability or from intentional malfeasance.

To assess whether a financial term is unfavorable, we refer to Section 5 of the Federal Trade Commission (FTC) Act [15], which defines an act as unfair if it meets the following criteria:

- **C1: Substantial Injury**. It causes or is likely to cause substantial injury to consumers;
- **C2: Unavoidable Harm**. Consumers cannot reasonably avoid it; and
- **C3: Insufficient Benefits**. It is not outweighed by countervailing benefits to consumers or competition.

During the topic modeling of term clusters, we judge the topic representing each cluster by three criteria to evaluate their fairness.

***C1*:** Since we focus on financial terms with potentially detrimental impacts, all terms inherently satisfy this criterion.

***C2*:** Terms and conditions are often hidden or difficult to avoid. Fair financial terms must be clearly displayed at critical points, like the payment page. However, terms related to cancellation, refunds, and returns are rarely shown upfront. We evaluate terms for unexpected fees (e.g., cancellation charges, non-refundable items, costly returns) that place an undue burden on consumers.

***C3*:** We classify terms as benign if they serve legitimate user or business protection, such as terms prohibiting fraud or abuse, protecting intellectual property, or ensuring legal compliance

3.2 Data Collection and Topic Modeling

As shown in Figure 2, we introduce *TermMiner*, a data collection and topic modeling pipeline for identifying unfavorable financial term at scale. By integrating LLMs like GPT-4o, *TermMiner* significantly reduces the extensive manual effort required in previous web content mining studies, such as those focused on detecting dark patterns [48]. *TermMiner* will be open-sourced and can be repurposed for various web-based text analysis tasks or longitudinal studies. Researchers can use our tools to explore different aspects of terms and conditions, such as readability, accessibility, or fairness.

***A Two-Pass Method*:** In the data collection and topic modeling steps, we employ a two-pass method. The first pass focuses on modeling and detecting *financial terms* to develop a corresponding topic template. In the second pass, we use the detected financial terms to re-conduct the classification and topic modeling modules. This time, the goal is to detect unfavorable financial terms within the financial terms identified. This approach is necessary because, to the best of our knowledge, there are no established templates or annotation schemes for (1) financial terms or (2) unfavorable financial terms in online shopping agreements. This two-pass process ensures comprehensive detection and accurate categorization of both financial and unfavorable financial terms.

***(1) Measurement Module*:** The measurement module collects terms and conditions from shopping websites to build a large, diverse dataset for analysis. For our large-scale measurement, we collect English shopping websites from two sources: the Tranco list [73], a ranking of top global websites, and two fraudulent e-commerce datasets (FCWs [6] and the Fraudulent and Legitimate Online Shops Dataset [37]). We filter out non-English content using Python's langdetect library [18]. To classify shopping websites, we evaluate several configurations: (1) GPT-3.5-Turbo [12] with URL, (2) GPT-3.5-Turbo with URL and HTML content, (3) GPT-4o [59] with URL, and (4) GPT-4o with URL and website screenshot. To evaluate our classification methods, we manually annotated a sample of 500 websites from the Tranco list, categorizing them into "shopping" and "non-shopping." GPT-4o, when prompted with URLs and screenshots, achieved an accuracy of 92%, comparable to commercial website classification services [48] (see Appendix B for details). Therefore, we use this configuration throughout our work.

We subsequently crawl the shopping websites to collect terms and conditions pages. A snowballed regex matching method detects terms and any nested policy pages, refined through positive and negative regex patterns to improve accuracy. Starting with common anchor texts, we iteratively refine the regex patterns by analyzing T&C links, which can be found in Appendix F. As shown in Table 1, we collected 1.9 million terms from 8,979 websites in total.

***(2) Classfication Module*:** The classification module categorizes terms from shopping websites' terms and conditions into binary categories: positive or negative. The categorization is based on the detection goal (such as identifying financial terms or identifying unfavorable financial terms) using corresponding prompts with the GPT-4o model [59].

We choose prompt engineering over fine-tuning the LLM during the term classification process due to cost considerations. Prior research [58, 70], along with our empirical observation (see §5.1), indicates that clear task descriptions and relevant examples (taxonomy) significantly enhance LLM performance in text classification. Therefore, for a given set of terms and conditions, we begin with zero-shot term classification. This process outputs sets of positive and negative terms, which are then used for clustering, inspection, and topic modeling. The resulting template generated from this analysis will, in turn, enhance the classification accuracy, creating a feedback loop that continuously improves our detection capabilities.

***(3) Topic Modeling Module*:** The topic modeling module uses LLMs and manual inspection to organize terms into meaningful topics. We generate sentence embeddings with the T5 model [60] and apply the DBSCAN clustering algorithm [23] to group terms by semantic similarity. The DBSCAN hyperparameters are decided through manual inspection.

To extract high-frequency topics, we leverage GPT-4o [59], building on recent findings that show LLMs outperform traditional topic modeling methods like Latent Dirichlet Allocation (LDA) [8] and BERTopic [31] in topic analysis [52, 67].

We develop an iterative topic modeling approach assisted by GPT-4o proceeds as follows:

(1) We analyze DBSCAN clusters and create an initial topic template for financial terms.

Table 1: Dataset and detection statistics—We source data from the Tranco top 100k list, the FCWs dataset [6], and the FLOS (Fraudulent and Legitimate Online Shops Dataset) [37], resulting in a total of 8,979 English shopping websites with terms and conditions. We report the statistics of detected unfavorable financial terms within them.

	Source	ShopTC-100K (Ours)	FCW	FLOS	Total
Datasets	Website to Query	100,000	6,127	1,040	27,167
	Accessible	61,466	1,378	542	63,386
	English	38,674	1,157	317	40,148
	Website with T&C	8,251	463	265	8,979
	Total Term Count	1,825,231	56,921	27,604	1,909,756
	Website with Unfavorable Terms	3,895	185	171	4,251
Number of Detected Terms	Purchase and Billing	2,920	75	53	3,048
	Post-Purchase	7,075	423	404	7,902
	Termination and Account Recovery	780	2	4	786
	Legal	1,108	1	1	1,110
	Others	81	4	4	89
	Total Unfavorable Financial Terms	11,079	519	466	12,064

(2) GPT-4o performs topic modeling on random samples from each cluster, assigning them to existing topics or suggesting new ones.
(3) We review and refine new topic suggestions through manual inspection, and updating the template.
(4) This process iterates until all clusters are assigned to a meaningful and satisfactory topic.

This iterative workflow, combining clustering, human-guided template creation, and GPT-4o's advanced topic modeling, enables efficient and comprehensive extraction of the topic template. We analyze 22,112 clusters in total, creating the unfavorable financial term taxonomy below.

***ShopTC-100K Dataset.*:** In the data collection stage, we filter 8,251 shopping websites from the Tranco top 100K, resulting in 1.8 million terms. This dataset will be made public on Huggingface. We present its statistics alongside two fake e-commerce datasets in Table 3, and report the statistics of unfavorable financial terms identified in later large-scale measurement (§5.2).

3.3 Unfavorable Financial Term Taxonomy

We report the 22 types of unfavorable financial terms, grouped into 4 categories as detailed below and in Table 2. Each type is exemplified by a unfavorable financial term taken directly from real-world shopping websites and analyzed for its alignment with the three criteria proposed by the FTC Act [15]. A detailed taxonomy of unfavorable financial terms can be found in Appendix C. While these unfavorable financial terms are not inherently deceptive, they often entail future financial obligations that customers should be aware of. We do not claim this list is exhaustive; however, it represents the most

Table 2: Categories, types, and examples of unfavorable financial terms are clustered, extracted, and topic-modeled from 1.9 million terms across 8,979 websites. All examples are extracted as-is from real-world shopping websites. The criteria are as follows: C1 = "Substantial Injury" (the term causes or is likely to cause substantial injury to consumers), C2 = "Unavoidable Harm" (consumers cannot reasonably avoid it), C3 = "Insufficient Benefits" (it is not outweighed by countervailing benefits to consumers or competition). The symbols represent the likelihood of satisfaction of a given criterion: ● = Always, ◒ = Sometimes.

Category	Type	Example	C1	C2	C3
Purchase and Billing Terms	Immediate Automatic Subscription	Also, as part of the promotion, you will receive a subscription to the FitHabit Fitness App for only $86, and the subscription will renew monthly up until cancellation.	●	◒	●
	Automatic Subscription after Free Trial	After the Promotion period has ended, unless you cancel the service before the end of the free trial period, you will automatically be subscribed onto the regular paid 1-year plan at the price of $275.40, which will automatically renew for successive 12-month periods, until cancelled.	●	◒	◒
	Unilateral Unauthorized Account Upgrades	Brevo reserves the right to automatically increase the contacts limit in the User account and upgrade the User's plan without prior notice.	●	●	●
	Late or Unsuccessful Payment Penalty	In addition, if any payment is not received within 30 days after the due date, then we may charge a late fee of $10 and we may assess interest at the rate of 1.5% of the outstanding balance per month (18% per year), or the maximum rate permitted by law.	●	◒	◒
	Overuse Penalty	If the Company establishes limits on the frequency with which you may access the Site, or terminates your access to or use of the Site, you agree to pay the Company one hundred dollars ($100) for each message posted in excess of such limits or for each day on which you access the Site in excess of such limits, whichever is higher.	●	◒	◒
	Retroactive Application of Price Change	When an applicable exchange rate is updated or when a change of price is notified to Brevo by its suppliers or WhatsApp, Brevo might immediately apply with retroactive effect the new Ratio and price increase to the User.	●	●	◒
Post-Purchase Terms	Non-Refundable Subscription Fee	If you or we cancel your subscription, you are not entitled to a refund of any subscription fees that were already charged for a subscription period that has already begun.	◒	◒	●
	No Refund For Purchase	Unless a refund is required by law, there are No Refund For Purchases for POS terminals and all transactions are final.	◒	◒	◒
	Strict No Cancellation Policy	As Research and Markets starts processing your order once it is submitted, we operate a strict no cancellation policy.	◒	◒	◒
	Cancellation Fee or Penalty	Some Bookings can't be canceled for free, while others can only be canceled for free before a deadline.	◒	◒	◒
	Non-Refundable Additional Fee	For this service, National Park Reservations charges a 10% non-refundable reservation fee based on the total dollar amount of reservations made.	◒	◒	◒
	Non-Monetary Refund Alternatives	Refund Policy: Refunds are not in cash but in the form of a "coupon".	◒	◒	●
	No Responsibility for Delivery Delays	We will not be held responsible if there are delays in delivery due to out-of-stock products.	◒	◒	●
	Customers Responsible for Shipping Issues	If the parcel is on hold by the Customs department of the shipping country, the customer is liable to provide all relevant and required documentation on to the authorities. Asim Jofa is not liable to refund the amount in case of non-clearance of the parcel.	◒	◒	●
	Customers Pay Return Shipping	All shipping costs will have to be borne by the customer.	◒	◒	●
	Restocking Fee	An 8% restocking fee and shipping fees for both ways will be borne by the buyer if returned without defects within 30 days from the purchase date or 7 days from delivery date, whichever is later.	◒	◒	●
Termination and Account Recovery Terms	Account Recovery Fee	To recover an archived or locked account, the legitimate creator of the account shall provide verifiable information about one's identity and will be charged a 10% administrative fee for the additional work caused by the account recovery process.	◒	◒	◒
	Digital Currency, Reward, Money Seizure on Inactivity	Please be noted that if your account is dormant for a period of 12 consecutive calendar months or longer, ..., any amounts in your account's balance, including any outstanding fees owed to you, shall be considered as forfeited and shall be fully deducted to Appnext.	◒	◒	●
	Digital Currency, Reward, Money Seizure on Termination or Account Closure	All Currency and/or Virtual Goods shall be cancelled if Your account is terminated or suspended for any reason or if We discontinue providing the Games and we will not compensate you for this loss or make any refund to you.	◒	◒	◒
Legal Terms	Exorbitant Legal Document Request Fee	Responding to requests for production of documents, and other matters requiring more than mere ministerial activities on our part, will incur a fee of two hundred dollars ($200) per hour.	◒	◒	◒
	Forced Waiver of Legal Protections	You hereby waive California Civil Code Section 1542. You hereby waive any similar provision in law, regulation, or code.	◒	◒	◒
	Forced Waiver of Class Action Rights	This agreement includes a class action waiver and an arbitration provision that governs any disputes between you and Sendinblue.	◒	◒	◒
	Other Legal Unfavorable Financial Term	...			

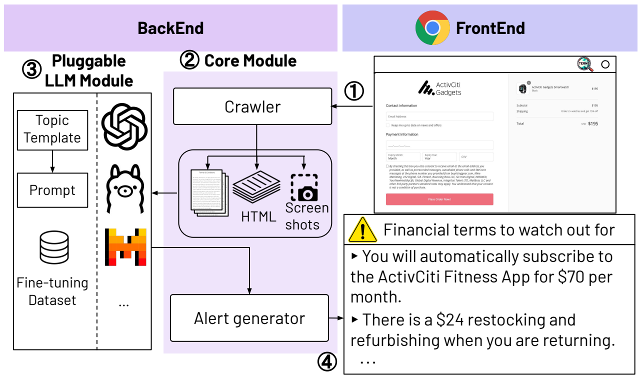


Figure 3: *TermLens* Design— (1) When the user activates the plugin, the current page URL is sent to the backend. (2) The terms and conditions are crawled and combined with the page information. (3) The pluggable LLM module analyzes this data, and if unfavorable financial terms are detected. (4) Alerts are generated and displayed on the front end to warn users of potentially unfair financial terms.

prominent types among the 1.9 million terms from 8,979 websites. We also report the financial term template in Appendix A.

It is important to note that this paper *does not* aim to analyze the fairness of terms from the legal perspective. We consider our work to be a complementary addition to the AI & Law datasets, by focusing on the natural phrasing found in online shopping websites' terms and conditions. A comparison between our unfavorable financial term template and previous work on online agreement fairness can be found in Appendix E.

4 Unfavorable Financial Term Detection System

In this section, we introduce *TermLens*, a Chrome plugin designed to detect unfavorable financial terms on e-commerce websites. Built upon the insights gained from the unfavorable financial term template and topic modeling analysis, *TermLens* enables efficient identification of potentially harmful financial terms, providing users with real-time protection against unfair or unfavorable conditions.

4.1 System Overview

Our detection system is illustrated in Figure 3. When a user activates *TermLens*, the URL of the current page is sent to the backend. Upon receipt, the backend crawler collects the terms and conditions pages. These term pages, along with the HTML content of the current page (and a screenshot if paired with a multimodal LLM), are preprocessed and sent to the pluggable LLM module for further analysis. If the LLM module flags any terms as unfavorable financial terms, the alert generator sends the identified terms back to the frontend, where they are displayed to the user.

***Pluggable LLM Module*:** We parse and preprocess the current page to determine if it is a payment page, improving alert accuracy by cross-checking unfavorable financial terms with payment details. For example, in Figure 1, the term "You will be charged $6.85 for the shipping and handling of your free smartwatch" aligns with the payment page, making it less concerning than the Immediate Automatic Subscription term, "you will receive a subscription to the FitHabit Fitness App for only $86," which is not shown on the payment page.

The Pluggable LLM Module, a key part of our system, analyzes both terms and conditions pages and the current webpage. By keeping the LLM decoupled from the backend, we allow flexibility in integrating different models. This enables multimodal models like GPT-4, GPT-4o [59], or LLaMA 3.2 [33] to process screenshots and terms, or text-based models such as GPT-3.5 [12], LLaMA [72], Mistral [38], or Gemma [71] to analyze HTML and terms.

***Backend Core Module*:** The alert generator receives flagged unfavorable financial terms and checks if the user is on a payment page. If so, it only flags terms not displayed on that page. GPT-4o analyzes page screenshots to mimic the user's experience and guard against adversarial text-based evasion. When the page is not a payment page, all flagged financial terms are shown. Since returns and refunds are rarely disclosed on payment pages, our evaluation in §5.1 focuses on scenarios where the user is not on a payment page and seeks to assess financial risks in advance.

Table 3: Statistics of annotated datasets for fine-tuning and validation for each term category.

Type	Fine-tuning	Validation
Post-Purchase	51	48
Legal	30	30
Termination and Account Recovery	15	16
Purchase and Billing	32	32
Unfavorable Terms Combined	128	126
Benign	116	119
Total Count	244	245

5 Evaluation and Large-Scale Measurement

We implement and evaluate *TermLens* using a manually annotated dataset. Our evaluation focuses on two key aspects: (1) assessing detection performance to determine how effectively LLMs, including both zero-shot and fine-tuned models, identify unfavorable financial terms (§5.1), and (2) analyzing findings from large-scale measurements using *TermLens* (§5.2).

5.1 Evaluation on an Annotated Dataset

***Dataset*:** We created an annotated dataset by randomly selecting 500 terms from clusters of both unfavorable financial terms and negative clusters (i.e., benign financial or non-financial terms). This yielded 250 potential unfavorable financial terms and 250 benign terms. Three researchers independently labeled the terms using the unfavorable financial template, without knowledge of the clusters. Disagreements were resolved in a second pass, and duplicates were removed, resulting in a final corpus of 489 terms. The dataset was split into 244 terms for fine-tuning and 245 terms for validation, as shown in Table 3.

***Baselines*:** To our knowledge, no prior work has directly addressed the detection of unfavorable financial terms. Recent advances in large language models (LLMs) demonstrate superior performance in common sense reasoning, complex text classification, and contextual understanding [12, 59, 72], outperforming older models like

Table 4: Performance of LLMs in detecting unfavorable financial terms evaluated using zero-shot classification with GPT-3.5-Turbo, GPT-4-Turbo, GPT-4o, Llama 3, and Gemma, along with fine-tuned GPT-3.5-Turbo and GPT-4o models.

Configuration	Model	Simple Prompt				Unfavorable Term Taxonomy Prompt			
		FPR (%) (↓ better)	TPR (%) (↑ better)	F1 (%) (↑ better)	AUC (%) (↑ better)	FPR (%) (↓ better)	TPR (%) (↑ better)	F1 (%) (↑ better)	AUC (%) (↑ better)
Zero-Shot	GPT-3.5-Turbo	59.5	71.9	59.9	56.2	74.0	96.5	68.5	61.2
	GPT-4o	58.6	72.6	61.4	56.2	**34.4**	96.6	**82.5**	**80.3**
	GPT-4-Turbo	58.6	72.6	61.4	56.2	64.8	**100.0**	73.8	67.2
	Llama 3	73.5	82.3	61.4	54.4	48.5	86.7	71.3	69.1
	Gemma	56.8	80.5	65.2	61.9	74.2	100.0	69.8	62.8
Fine-Tuned	GPT-3.5-Turbo	-	-	-	-	5.5	91.5	92.7	**91.8**
	GPT-4o	-	-	-	-	**2.3**	**92.1**	**94.6**	**91.8**

BERT [19] and RoBERTa [45]. Therefore, we evaluate state-of-the-art LLMs: (1) GPT-3.5-Turbo [12], (2) GPT-4-Turbo [59], and (3) GPT-4o, along with two open-source LLMs: (1) LLaMA 3 8B [72] and (2) Gemma 8B [71].

***Evaluation Configurations*:** We evaluate two configurations: (1) Zero-shot classification with a simple binary prompt describing the unfavorable financial term and a multi-class taxonomy prompt explaining term types, and (2) Fine-tuning the LLM using the taxonomy prompt to improve detection accuracy.

Simple Prompt

```
Classify the following term as 'malicious' or 'benign'.
A term is 'malicious' if it is a financial term that is
one-sided, unbalanced, unfair, or harmful to users.
Respond only with 'malicious' or 'benign'.
```

Unfavorable Term Taxonomy Prompt

```
You will be provided with a paragraph extracted from the
terms and conditions. Your task is to classify them into
one of the topics below or 'benign':

- Automatic Subscription after Free Trial: Automatically
subscribing users after free trials.
[...]

If the term is reasonable based on common sense, reply
'benign'.
If the term is malicious and financial, reply with a
topic from the template above.
```

***Evaluation Metrics*:** We evaluate the models in terms of false positive rate, true positive rate (recall), F1 score, and AUC (Area Under the Curve). AUC represents the area under the ROC (Receiver Operating Characteristic) curve, measuring the model's ability to distinguish between classes.

***Zero-shot Classification Performance*:** As a baseline for unfavorable financial term detection, we evaluated zero-shot classification with two prompts: (1) a simple prompt defining unfavorable financial terms and (2) a taxonomy prompt explaining term types. Using the taxonomy improved True Positive Rate (TPR) by 4.4% to 27.4% and boosted the F1 score by 4.5% to 21.1%, showing a better balance of precision and recall. However, the False Positive Rate (FPR) increased in most cases, except for GPT-4o, where it dropped by 24.2%. GPT-4o achieved the best overall performance with a TPR of 96.6% and an F1 score of 82.5%, demonstrating the importance of a unfavorable financial term taxonomy for more accurate detection.

***Fine-tuned LLM Classification Performance*:** We fine-tune GPT-3.5-Turbo and GPT-4o for 10 epochs with a batch size of 1. As shown in Table 1, the dataset maintained a similar distribution of terms across categories. Fine-tuning resulted in significant performance improvements, with GPT-4o achieving a True Positive Rate (TPR) of 92.1% and an F1 score of 94.6%. The fine-tuned GPT-4o model outperforms other LLMs in distinguishing true positives from false positives. These results demonstrate that fine-tuning, even with a limited dataset, can substantially enhance detection performance.

5.2 Large-Scale Measurement

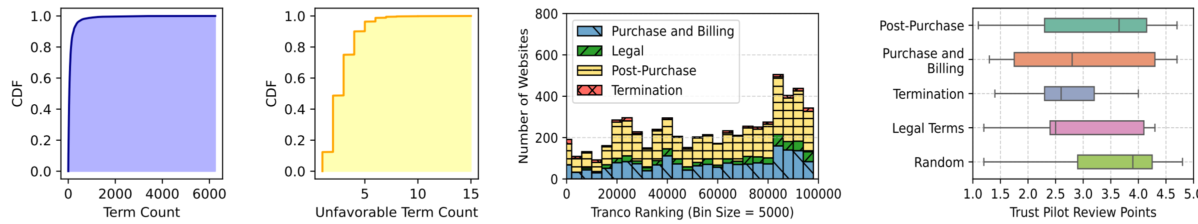
To understand the prevalence of unfavorable financial terms, we deploy the fine-tuned GPT-4o model with *TermLens* for detection. The backend detection system was applied to English shopping websites filtered from the Tranco list's top 100,000 sites, along with two fake e-commerce website datasets: the FCWs dataset [6] and the FLOS dataset [37]. This large-scale measurement serves as a qualitative study on the prevalence of unfavorable financial terms in popular shopping websites. We present our findings below.

***Categorizing Websites with Unfavorable Financial Terms*:** As shown earlier in Table 1, we collect terms and conditions from 8,979 English shopping websites, resulting in 1.9 million terms. Using a GPT-4o model with the unfavorable financial term taxonomy, 12,064 terms (approximately 0.6%) were flagged as unfavorable financial terms. Notably, 47.21% (3,895 out of 8,251) of the English shopping websites from the top 100,000 Tranco-ranked sites contain at least one type of unfavorable financial term. Figure 4(a) and (b) show the number of terms and unfavorable financial terms across 8,251 websites, underscoring how difficult it is for consumers to review lengthy T&Cs and pinpoint questionable financial terms thoroughly. This emphasizes the importance of automated detection systems to protect users from unfavorable terms.

***Trend Analysis*:** Figure 4(c) shows the distribution of unfavorable financial terms across various categories among the top 100K Tranco-ranked websites [73]. Post-purchase terms (in yellow) are the most common across all ranking levels, with a higher concentration in lower-ranked sites, suggesting these terms are more frequent on less popular websites. Purchase and billing terms (in blue) also have significant representation. Termination and account recovery terms (in red) and legal terms (in green) are less frequent but more evenly spread across the rankings. This trend highlights the widespread presence of unfavorable financial terms, especially on lower-ranked sites, underscoring the need for greater regulation to protect consumers from harmful practices, particularly on less reputable websites.

***Comparing Tranco with Fake E-commerce Datasets*:** Interestingly, the percentage of websites with unfavorable financial terms

Figure 4: Statistics from Large-scale measurement of unfavorable financial term detection on Tranco top 100K websites.



(a) CDF of the number of terms per website.

(b) CDF of the number of unfavorable financial terms per website.

(c) Distribution of unfavorable financial terms in each category across Tranco-ranked websites.

(d) Trustpilot ratings comparison between the top 10 websites with the most unfavorable financial terms and a random sample of 40 websites.

from the Tranco list (47.21%) is similar to that of fraudulent e-commerce websites (48.90%). This suggests that unfavorable financial terms are not limited to fraudulent sites but are also prevalent among high-ranking websites, pointing to a broader issue in consumer protection. Websites from *ShopTC-100K* have more unfavorable financial legal terms, indicating that legitimate websites are more inclined to shift liability onto customers than fraudulent ones.

***Qualitative Study on User rating*:** From the English shopping websites in the top 100k Tranco list, we select those with the highest frequency of unfavorable financial terms across categories. We analyze Trustpilot [74] reviews for the top 10 websites in each unfavorable financial term category with the highest presence, alongside 40 randomly selected websites. As shown in Figure 4(d), websites with unfavorable financial terms tend to have lower Trustpilot ratings, particularly those with “Post-Purchase Terms” and “Purchase and Billing Terms,” indicating negative customer satisfaction. “Termination and Account Recovery” and “Legal Terms” also correlated with lower ratings, though with more variation, suggesting mixed experiences. This suggests a link between unfavorable financial terms and consumer dissatisfaction.

***Qualitative Study on Current Ecosystem Defense*:** We examine whether the top 10 websites with the highest frequency of unfavorable financial terms are flagged by ScamAdviser [66], Google Safe Browsing [28], and Microsoft Defender SmartScreen [50]. Out of 40 websites, only 6 (15%) have a ScamAdviser score below 90, and 5 (12.5%) scored below 10, while the majority receive a perfect score of 100. None of the websites are flagged by Google Safe Browsing or Microsoft Defender, which is expected since unfavorable financial terms are not inherently indicative of scams.

***Qualitative Study on User Perception*:** To illustrate the potential harm of unfavorable financial terms, we present four case studies on user perception and financial harm in each category in Appendix D. This underscores the urgent need for automated systems to detect unfavorable financial terms effectively.

6 Discussion

We introduce *TermMiner*, an automated pipeline for collecting and modeling unfavorable financial terms in shopping websites with minimal human input. Researchers can utilize our tools to examine various aspects of web-based text, such as readability or accessibility, and to conduct longitudinal studies, as discussed in Appendix E.

TermLens assumes that the financial terms in question are not *adversarially perturbed*. Recent studies have highlighted LLM vulnerabilities to jailbreak and prompt-injection attacks [30, 44, 84]. These attacks can result in incorrect or overridden outputs. However, for T&Cs, such adversarial perturbations are likely to subject to manual scrutiny, particularly in post-complaint scenarios, such as legal disputes [34]. We leave the exploration of adversarial robustness in LLM-based unfavorable financial term detection for future work.

Ethics

We query and crawl terms and conditions from online shopping websites, collecting data from each site only once. Since terms and conditions are usually limited to a few subpages, this process does not overburden the servers hosting these websites. In this paper, we report some terms and conditions along with the associated companies, all of which are publicly available information. No personal data is collected during the measurement process.

7 Conclusion

This paper is one of the first attempts to understand, categorize, and detect unfavorable financial terms and conditions on shopping websites. These terms, which can significantly impact consumer trust and satisfaction, have not been extensively studied. By highlighting the prevalence and types of unfavorable financial terms, we hope to pave the way for increased awareness and further research in this area. We develop an automated data collection and topic modeling pipeline, analyzing 1.9 million terms from 8,979 websites to create a taxonomy for unfavorable financial terms. This taxonomy includes 22 types across 4 categories, covering purchase and billing, post-purchase activities, account termination, and legal aspects.

TermLens is the first study to evaluate the effectiveness of LLMs in identifying unfavorable financial terms. Using a fine-tuned GPT-4o model on a manually annotated dataset, *TermLens* achieves an F1 score of 94.6% with a false positive rate of 2.3%. In large-scale deployment, we find that approximately 47.21% of shopping websites in the Tranco top 100,000 contain at least one category of unfavorable financial terms. Our qualitative analysis shows that current ecosystem defenses are inadequate to protect users from these terms, that less popular websites are more likely to include unfavorable financial terms, and that there is a correlation between unfavorable financial terms and user dissatisfaction.

References

[1] Digital Commerce 360. 2023. U.S. e-commerce sales reached $1.119 trillion in 2023. https://www.digitalcommerce360.com/article/us-ecommerce-sales/.

[2] Abdulrahman Alabduljabbar, Ahmed Abusnaina, Ülkü Meteriz-Yildiran, and David Mohaisen. 2021. Automated privacy policy annotation with information highlighting made practical using deep representations. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. Association for Computing Machinery, New York, NY, USA, 2378–2380.

[3] Benjamin Andow, Samin Yaseer Mahmud, Wenyu Wang, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Tao Xie. 2019. {PolicyLint}: investigating internal privacy policy contradictions on google play. In *28th USENIX security symposium (USENIX security 19)*. USENIX Association, Berkeley, CA, USA, 585–602.

[4] Yannis Bakos, Florencia Marotta-Wurgler, and David R Trossen. 2014. Does anyone read the fine print? Consumer attention to standard-form contracts. *The Journal of Legal Studies* 43, 1 (2014), 1–35.

[5] Leyla Bilge, Engin Kirda, Christopher Kruegel, and Marco Balduzzi. 2011. Exposure: Finding malicious domains using passive DNS analysis.. In *Ndss*. Internet Society, Reston, VA, USA, 1–17.

[6] Marzieh Bitaab, Haehyun Cho, Adam Oest, Zhuoer Lyu, Wei Wang, Jorij Abraham, Ruoyu Wang, Tiffany Bao, Yan Shoshitaishvili, and Adam Doupé. 2023. Beyond Phish: Toward Detecting Fraudulent e-Commerce Websites at Scale. In *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, Online, 2566–2583.

[7] Marzieh Bitaab, Haehyun Cho, Adam Oest, Penghui Zhang, Zhibo Sun, Rana Pourmohamad, Doowon Kim, Tiffany Bao, Ruoyu Wang, Yan Shoshitaishvili, et al. 2020. Scam pandemic: How attackers exploit public fear through phishing. In *2020 APWG Symposium on Electronic Crime Research (eCrime)*. IEEE, Online, 1–10.

[8] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[9] Aaron Blum, Brad Wardman, Thamar Solorio, and Gary Warner. 2010. Lexical feature based phishing URL detection using online learning. In *Proceedings of the 3rd ACM Workshop on Artificial Intelligence and Security*. Association for Computing Machinery, New York, NY, USA, 54–60.

[10] Kerstin Bongard-Blanchy, Arianna Rossi, Salvador Rivas, Sophie Doublet, Vincent Koenig, and Gabriele Lenzini. 2021. " I am Definitely Manipulated, Even When I am Aware of it. It's Ridiculous!"-Dark Patterns from the End-User Perspective. In *Designing Interactive Systems Conference 2021*. Association for Computing Machinery, New York, NY, USA, 763–776.

[11] Daniel Braun and Florian Matthes. 2021. NLP for consumer protection: Battling illegal clauses in German terms and conditions in online shopping. In *Proceedings of the 1st Workshop on NLP for Positive Impact*. Association for Computational Linguistics, Online, 93–99.

[12] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[13] Duc Bui, Brian Tang, and Kang G Shin. 2023. Detection of inconsistencies in privacy practices of browser extensions. In *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, Online, 2780–2798.

[14] Celsius. 2023. 1.7 million people call Celsius their home for crypto. https://celsius.network/.

[15] Federal Trade Commission. 1914. Federal Trade Commission Act Section 5: Unfair or Deceptive Acts or Practices. https://www.federalreserve.gov/boarddocs/supmanual/cch/200806/ftca.pdf.

[16] Federal Trade Commission. 1983. FTC Policy Statement on Deception. https://www.ftc.gov/system/files/documents/public_statements/410531/831014deceptionstmt.pdf.

[17] Hao Cui, Rahmadi Trimananda, Athina Markopoulou, and Scott Jordan. 2023. {PoliGraph}: Automated privacy policy analysis using knowledge graphs. In *32nd USENIX Security Symposium (USENIX Security 23)*. USENIX Association, Berkeley, CA, USA, 1037–1054.

[18] Michal Danilk. 2024. langdetect. https://pypi.org/project/langdetect/.

[19] Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

[20] Linda Di Geronimo, Larissa Braz, Enrico Fregnan, Fabio Palomba, and Alberto Bacchelli. 2020. UI dark patterns and where to find them: a study on mobile applications and user perception. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 1–14.

[21] Kasper Drazewski, Andrea Galassi, Agnieszka JABONOWSKA, Francesca Lagioia, Marco Lippi, Hans-Wolfgang Micklitz, Giovanni Sartor, Giacomo Tagiuri, and Paolo Torroni. 2021. A corpus for multilingual analysis of online terms of service. In *Proceedings of the Natural Legal Language Processing Workshop 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 1–8.

[22] Vincent Drury and Ulrike Meyer. 2019. Certified phishing: taking a look at public key certificates of phishing websites. In *15th Symposium on Usable Privacy and Security (SOUPS'19)*. USENIX Association, Berkeley, CA, USA, 211–223.

[23] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD)*. AAAI Press, Online, 226–231.

[24] Mirgita Frasheri, Arian Bakhtiarnia, Lukas Esterle, and Aleksandros Iosifidis. 2024. Are LLM-based methods good enough for detecting unfair terms of service?

[25] Andrea Galassi, Kasper Drazewski, Marco Lippi, and Paolo Torroni. 2020. Cross-lingual annotation projection in legal texts. In *Proceedings of the 28th international conference on computational linguistics*. International Committee on Computational Linguistics (ICCL), Barcelona, Spain, 915–926.

[26] Andrea Galassi, Francesca Lagioia, Agnieszka Jabłonowska, and Marco Lippi. 2024. Unfair clause detection in terms of service across multiple languages. *Artificial Intelligence and Law* (2024), 1–49.

[27] Leo A Goodman. 1961. Snowball sampling. , 148–170 pages.

[28] Google. 2024. Google Safe Browsing. https://safebrowsing.google.com/.

[29] Colin M Gray, Cristiana Santos, Nataliia Bielova, Michael Toth, and Damian Clifford. 2021. Dark patterns and the legal requirements of consent banners: An interaction criticism perspective. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–18.

[30] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*. Association for Computing Machinery, New York, NY, USA, 79–90.

[31] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure.

[32] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G Shin, and Karl Aberer. 2018. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27th USENIX Security Symposium (USENIX Security 18)*. USENIX Association, Berkeley, CA, USA, 531–548.

[33] Hugging Face. 2024. Llama-3.2-90B-Vision Model. https://huggingface.co/meta-llama/Llama-3.2-90B-Vision.

[34] Investopedia. 2023. Judge Rules Celsius Networks Owns Customers' Crypto Depositss. https://www.investopedia.com/judge-s-ruling-on-celsius-7092044.

[35] Agnieszka Jablonowska, Francesca Lagioia, Marco Lippi, Hans-Wolfgang Micklitz, Giovanni Sartor, and Giacomo Tagiuri. 2021. Assessing the cross-market generalization capability of the claudette system. In *Legal knowledge and information systems*. IOS Press, Online, 62–67.

[36] Ankit Kumar Jain and Brij B Gupta. 2017. Phishing detection: analysis of visual similarity based approaches.

[37] Audrone Janaviciute and Agnius Liutkevicius. 2023. Fraudulent and Legitimate Online Shops Dataset. https://doi.org/10.17632/m7xtkx7g5m.1

[38] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and et al. 2023. Mistral 7B.

[39] Amin Kharraz, William Robertson, and Engin Kirda. 2018. Surveylance: Automatically detecting online survey scams. In *2018 IEEE Symposium on Security and Privacy (SP)*. IEEE, Online, 70–86.

[40] Francesca LAGIOIA, Hans-Wolfgang MICKLITZ, Yannis PANAGIS, Giovanni SARTOR, and Paolo TORRONI. 2017. Automated detection of unfair clauses in online consumer contracts. In *Legal Knowledge and Information Systems: JURIX 2017: The Thirtieth Annual Conference*, Vol. 302. IOS Press, Amsterdam, The Netherlands, 145.

[41] Xigao Li, Anurag Yepuri, and Nick Nikiforakis. 2023. Double and nothing: Understanding and detecting cryptocurrency giveaway scams. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*. Internet Society, Reston, VA, USA.

[42] Nut Limsopatham. 2021. Effectively leveraging BERT for legal document classification. In *Proceedings of the Natural Legal Language Processing Workshop 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 210–216.

[43] Marco Lippi, Przemysław Pałka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. 2019. CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law* 27 (2019), 117–139.

[44] Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2023. Autodan: Generating stealthy jailbreak prompts on aligned large language models.

[45] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

[46] Marco Loos and Joasia Luzak. 2016. Wanted: a bigger stick. On unfair terms in consumer contracts with online service providers. *Journal of consumer policy* 39 (2016), 63–90.

[47] Jamie Luguri and Lior Jacob Strahilevitz. 2021. Shining a light on dark patterns. *Journal of Legal Analysis* 13, 1 (2021), 43–109.

[48] Arunesh Mathur, Gunes Acar, Michael J Friedman, Eli Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. 2019. Dark patterns at scale: Findings from a crawl of 11K shopping websites. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–32.

[49] Arunesh Mathur, Mihir Kshirsagar, and Jonathan Mayer. 2021. What makes a dark pattern... dark? Design attributes, normative considerations, and measurement methods. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 1–18.

[50] Microsoft. 2024. Microsoft Defender SmartScreen. https://learn.microsoft.com/en-us/windows/security/operating-system-security/virus-and-threat-protection/microsoft-defender-smartscreen/.

[51] Mahmood Moghimi and Ali Yazdian Varjani. 2016. New rule-based phishing detection method. *Expert systems with applications* 53 (2016), 231–242.

[52] Yida Mu, Chun Dong, Kalina Bontcheva, and Xingyi Song. 2024. Large Language Models Offer an Alternative to the Traditional Approach of Topic Modelling.

[53] Arvind Narayanan, Arunesh Mathur, Marshini Chetty, and Mihir Kshirsagar. 2020. Dark Patterns: Past, Present, and Future: The evolution of tricky user interfaces. *Queue* 18, 2 (2020), 67–92.

[54] National Park Reservations. 2024. National Park Reservations. https://www.nationalparkreservations.com/.

[55] Neteller. 2024. Neteller: Online Payment Account. https://www.neteller.com

[56] Midas Nouwens, Ilaria Liccardi, Michael Veale, David Karger, and Lalana Kagal. 2020. Dark patterns after the GDPR: Scraping consent pop-ups and demonstrating their influence. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 1–13.

[57] Jonathan A Obar and Anne Oeldorf-Hirsch. 2020. The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society* 23, 1 (2020), 128–147.

[58] OpenAI. 2023. Best Practices for Prompt Engineering with the OpenAI API. https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-the-openai-api.

[59] OpenAI. 2023. GPT-4 Technical Report. https://platform.openai.com/docs/guides/gpt.

[60] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* 21, 140 (2020), 1–67.

[61] David Rodriguez, Ian Yang, Jose M Del Alamo, and Norman Sadeh. 2024. Large language models: a new approach for privacy policy analysis at scale. , 25 pages.

[62] Ozgur Koray Sahingoz, Ebubekir Buber, Onder Demir, and Banu Diri. 2019. Machine learning based phishing detection from URLs. *Expert Systems with Applications* 117 (2019), 345–357.

[63] Yuji Sakurai, Takuya Watanabe, Tetsuya Okuda, Mitsuaki Akiyama, and Tatsuya Mori. 2020. Discovering HTTPSified phishing websites using the TLS certificates footprints. In *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE, Online, 522–531.

[64] David Sarne, Jonathan Schler, Alon Singer, Ayelet Sela, and Ittai Bar Siman Tov. 2019. Unsupervised topic extraction from privacy policies. In *Companion Proceedings of The 2019 World Wide Web Conference*. Association for Computing Machinery, New York, NY, USA, 563–568.

[65] ScamAdviser. 2019. ScamAdviser Algorithm Explainer. https://www.scamadviser.com/articles/scamadviser-algorithm-explainer.

[66] ScamAdviser. 2024. ScamAdviser. https://www.scamadviser.com.

[67] Kabir Manandhar Shrestha, Katie Wood, D Goodman, and M Mistica. 2023. Do We Need Subject Matter Experts? A Case Study of Measuring Up GPT-4 Against Scholars in Topic Evaluation. In *Proceedings of the Seventh Workshop on Natural Language for Artificial Intelligence (NL4AI 2023) co-located with 22th International Conference of the Italian Association for Artificial Intelligence (AI* IA 2023)*. CEUR-WS, Rome, Italy.

[68] Yan Shvartzshnaider, Noah Apthorpe, Nick Feamster, and Helen Nissenbaum. 2019. Going against the (appropriate) flow: A contextual integrity approach to privacy policy analysis. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. AAAI Press, Online, 162–170.

[69] Nili Steinfeld. 2016. "I agree to the terms and conditions":(How) do users read privacy policies online? An eye-tracking experiment. *Computers in human behavior* 55 (2016), 992–1000.

[70] Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models.

[71] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology.

[72] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models.

[73] Tranco. 2024. Tranco, A Research-Oriented Top Sites Ranking Hardened Against Manipulation. https://tranco-list.eu/.

[74] Trustpilot. 2024. Trustpilot Reviews: Experience the power of customer reviews. https://www.trustpilot.com

[75] European Union. 1993. Council Directive 93/13/EEC of 5 April 1993 on unfair terms in consumer contracts. *Official Journal of the European Communities*. , 29–34 pages.

[76] Thijs van Den Hout, Thymen Wabeke, Giovane CM Moura, and Cristian Hesselman. 2022. LogoMotive: detecting logos on websites to identify online scams-a TLD case study. In *Passive and Active Measurement: 23rd International Conference, PAM 2022, Virtual Event, March 28–30, 2022, Proceedings*. Springer, Cham, Switzerland, 3–29.

[77] Ari Ezra Waldman. 2020. Cognitive biases, dark patterns, and the 'privacy paradox'. *Current opinion in psychology* 31 (2020), 105–109.

[78] Anhao Xiang, Weiping Pei, and Chuan Yue. 2023. PolicyChecker: Analyzing the GDPR Completeness of Mobile Apps' Privacy Policies. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*. Association for Computing Machinery, New York, NY, USA, 3373–3387.

[79] Guang Xiang, Jason Hong, Carolyn P Rose, and Lorrie Cranor. 2011. Cantina+ a feature-rich machine learning framework for detecting phishing web sites. *ACM Transactions on Information and System Security (TISSEC)* 14, 2 (2011), 1–28.

[80] Peng Yang, Guangzhen Zhao, and Peng Zeng. 2019. Phishing website detection based on multidimensional features driven by deep learning. *IEEE access* 7 (2019), 15196–15209.

[81] Zheng Yang, Joey Allen, Matthew Landen, Roberto Perdisci, and Wenke Lee. 2023. {TRIDENT}: Towards Detecting and Mitigating Web-based Social Engineering Attacks. In *32nd USENIX Security Symposium (USENIX Security 23)*. USENIX Association, Berkeley, CA, USA, 6701–6718.

[82] Haijun Zhang, Gang Liu, Tommy WS Chow, and Wenyin Liu. 2011. Textual and visual content-based anti-phishing: a Bayesian approach. *IEEE transactions on neural networks* 22, 10 (2011), 1532–1546.

[83] Lu Zhou, Chengyongxiao Wei, Tong Zhu, Guoxing Chen, Xiaokuan Zhang, Suguo Du, Hui Cao, and Haojin Zhu. 2023. {POLICYCOMP}: Counterpart Comparison of Privacy Policies Uncovers Overbroad Personal Data Collection Practices. In *32nd USENIX Security Symposium (USENIX Security 23)*. USENIX Association, Berkeley, CA, USA, 1073–1090.

[84] Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models.

[85] Mouad Zouina and Benaceur Outtaj. 2017. A novel lightweight URL phishing detection system using SVM and similarity index. *Human-centric Computing and Information Sciences* 7, 1 (2017), 1–13.

Appendix

A Financial Terms Topic Template

The financial term taxonomy consists of 12 categories:

(1) *Subscription/Product Terms*: Terms related to subscription fees, billing, and automatic renewals.
(2) *Service Termination Policy*: Terms outlining the financial implications of service termination.
(3) *Payment and Purchase Term*: Terms about payments, processing fees, and currency transactions.
(4) *Return and Refund Policy*: Terms governing product returns and service cancellations.
(5) *Insurance and Warranty Term*: Terms related to coverage, claims, limitations, and premiums.
(6) *Promotions and Rewards*: Terms about offers, discounts, loyalty programs, and rewards.
(7) *Shipping and Handling Terms*: Terms on product shipping costs and policies.
(8) *Dispute Resolution Policy*: Terms outlining dispute resolution methods and processes.
(9) *Investment and Trading Terms*: Terms specific to investment/trading platforms.
(10) *Intellectual Property Terms*: Terms on rights and restrictions for intellectual property use.
(11) *Financial Glossary*: Financial terminology definition.

Table 5: Performance and Cost Analysis for Zero-Shot Shopping Website Classification Using GPT-3.5-Turbo [12] and GPT-4o [59] Models on 500 randomly selected websites.

Model	Configuration	ACC (%)	TP (%)	FP (%)	TN (%)	FN (%)
GPT-3.5-Turbo	URL	73	80.9	36.5	63.5	19.1
	URL+HTML	24	31.5	5.5	94.5	68.5
GPT-4o	URL	63	40.9	10.1	89.9	59.1
	URL+Image	86	90.7	20.6	79.4	9.3

(12) *Others*: Includes less frequently mentioned financial terms.

B Shopping website classification Evaluation

To evaluate our classification methods, we randomly sampled 500 websites from the top 10,000 on the Tranco list for manual annotation to identify shopping sites (offering products or services for sale). Of these, 257 were categorized as "shopping", 219 as "non-shopping", and 24 were inaccessible, due to either server issues or geo-blocking IPs in the US. As indicated in Table 5, GPT-4o's accuracy reached 86% when analyzing URLs with corresponding screenshots. Further examination of a focused group, specifically English-language websites with available T&Cs (115 out of 500), reveals that accuracy improved to 81% for GPT-3 using only URLs and to 92% for GPT-4o with screenshots, approaching the commercial-grade classification benchmarks reported in previous studies (89%-93%) [48]. Based on these statistics, GPT-4o paired with image data is selected for the broader scale measurement of shopping websites.

C Unfavorable Financial Terms Taxonomy

We discover unfavorable financial term types falling under 4 broader categories: 1) purchase and billing terms; 2) post-purchase terms; 3) termination and account recovery terms; and 4) legal terms. We describe the taxonomy with an explanation for each type below:

Unfavorable Purchase and Billing Terms. This category includes subscription, purchase, and billing terms that are unfavorable or concerning for customers:

- **Immediate Automatic Subscription.** Additional subscriptions are automatically added when purchasing an item or during promotions without clear consent from the user.
- **Automatic Subscription after Free Trial.** Users are automatically enrolled in a paid subscription after a trial period ends unless they actively cancel.
- **Unilateral Unauthorized Account Upgrades.** Accounts may be upgraded and charged at higher rates without providing prior notice to the user.
- **Late or Unsuccessful Payment Penalty.** Penalties or interest charges are applied for late or unsuccessful payments.
- **Overuse Penalty.** Charging extra fees if usage limits are exceeded. Typically found in subscription-based services such as data plans, cloud storage, and streaming services.
- **Retroactive Application of Price Change.** Price (of subscription-based services) increases can be applied retroactively without prior notice to the user.

Unfavorable Post-Purchase Terms . This category includes cancellation, shipping, return, and refund terms that are unfavorable or concerning for customers:

- **Non-Refundable Subscription Fee.** Subscription fees that have already been charged are not refunded.
- **No Refund for Purchase.** Purchases of individual items are final and non-refundable.
- **Strict No Cancellation Policy.** Orders cannot be canceled once they have been processed.
- **Cancellation Fee or Penalty.** Fees are applied for canceling certain bookings, services, or online purchase orders.
- **Non-Refundable Additional Fee.** Charging non-refundable additional fees under various labels such as service fees, transfer fees, pre-authorization fees, administrative fees, subscription upgrade fees, handling product fees, etc.
- **Non-Monetary Refund Alternatives.** Refunds are provided in the form of coupons, reward points, or store credit rather than money.
- **No Responsibility for Delivery Delays.** Companies are not held liable for delays in product delivery.
- **Customers Responsible for Shipping Issues.** Customers are responsible for handling customs issues, additional shipping charges, and any shipping-related complications that do not involve delays.
- **Customers Pay Return Shipping.** Customers bear the cost of return shipping for products.
- **Restocking Fee.** A fee is charged for restocking returned items.

Unfavorable Termination and Account Recovery Terms. This category includes account or service termination, deactivation, and reactivation terms that are unfavorable or concerning for customers:

- **Account Recovery Fee.** A Fee is charged to recover locked or archived accounts.
- **Digital Currency, Reward, Money Seizure on Inactivity.** Digital assets, such as rewards, points, and virtual currencies, are forfeited or otherwise taken away if accounts remain inactive for extended periods.
- **Digital Currency, Reward, Money Seizure on Termination or Account Closure.** Digital assets, such as rewards, points, and virtual currencies, are forfeited or otherwise taken away upon service termination or account closure.

Unfavorable Legal Terms. This category includes legal terms that are unfavorable or concerning for customers:

- **Exorbitant Legal Document Request Fee.** High fees are charged for requesting legal documents.
- **Forced Waiver of Legal Protections.** Customers are required to waive certain legal protections.
- **Forced Waiver of Class Action Rights.** Customers waive their rights to participate in class action lawsuits.
- **Other Legal Unfavorable Financial Term.** Additional legal terms that impose financial burdens or limit legal recourse for the customer.

Many terms and conditions for online shopping websites include strong legal language, such as waivers of class action rights, arbitration clauses, and limitations of liability. This study does not

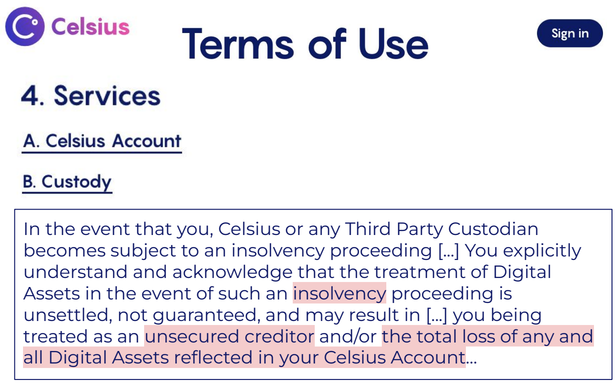


Figure 5: Extracted from the T&C of Celsius Network LLC, a now bankrupt cryptocurrency company.

specifically focus on the legal aspects for two main reasons: (1) Although legal terms can impact users financially, they differ from other categories we report. These terms, despite their potential future implications, are not the primary concern when customers make a purchase or sign up for a service. (2) There is another line of work (see Appendix E) that focuses on terms deemed invalid in court. We consider our work complementary to these studies. By not intensively focusing on legal terms, we maintain a focus on terms with immediate and direct financial implications for users.

D Case Studies

We present four case studies illustrating the potential harm of unfavorable financial terms in each unfavorable financial term category.

Unfavorable post-purchase terms case study: National Park Reservations [54], a company providing national park hotel and lodging reservation service with a 1-star review on Yelp, includes the following terms and conditions:

> For this service, National Park Reservations charges a 10% non-refundable reservation fee based on the total dollar amount of reservations made. This reservation fee will be billed separately to your credit card and will be billed under the memo "National Park Reservations". By using National Park Reservations, the customer authorizes National Park Reservations to charge their credit card the 10% non-refundable fee.

The above terms fall under the "Non-Refundable Additional Fee" category. Figure 7 shows a word cloud that displays the most frequent words from the top 50 Yelp reviews (2021-2024), excluding the company name. Despite the non-refundable additional fee being clearly stated in the terms and conditions, many customers still find it deceptive. "Scam" is among the most frequent words in the reviews. This shows the potential harm caused by unfavorable financial terms and perceived deceptive practices, significantly impacting customer trust and satisfaction.

Unfavorable termination and account recovery terms case study: Compared to other categories of unfavorable financial terms, those related to unfair, unfavorable, or concerning service termination and

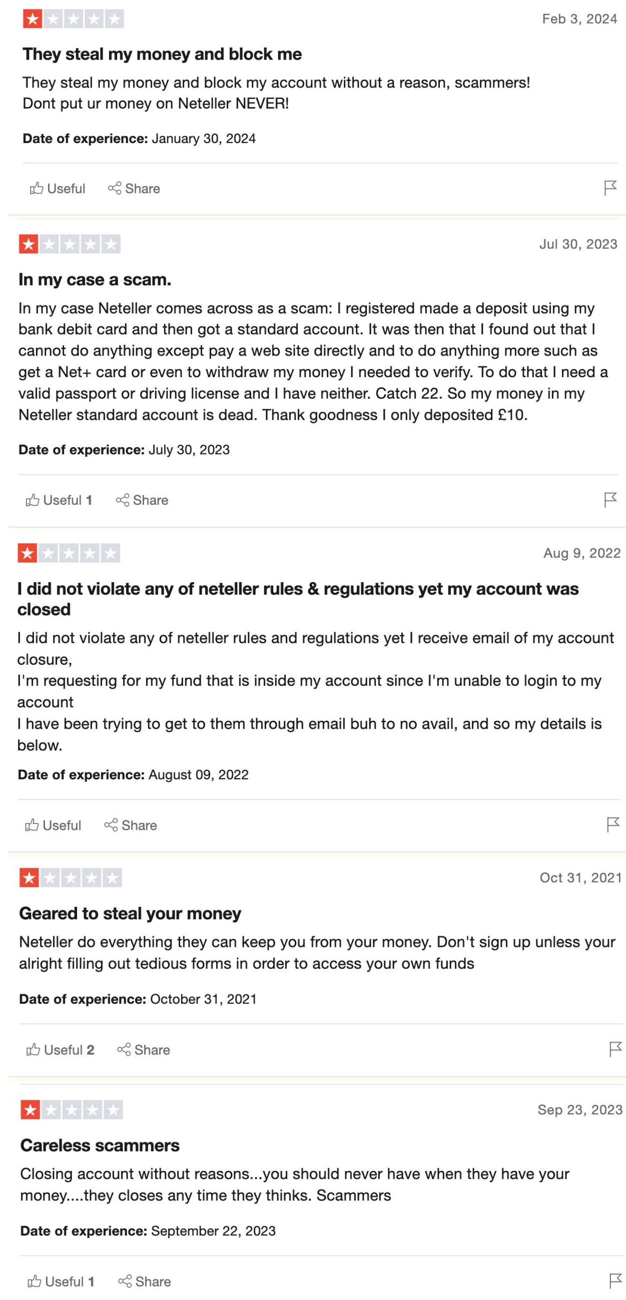


Figure 6: Complied reviews for Neteller [55] from Trust Pilot [74] regarding account closure.

account management are significantly less prevalent in our measurements, as shown in Figure 4. However, these terms can still impose substantial costs on customers. For example, the Terms of Use from Neteller [55]—a digital wallet with a TrustScore of 10 out of 100 from Scamadviser [65]—include such terms:

> If an Account has been closed, [...]. Fees relating to ongoing management of inactive accounts will

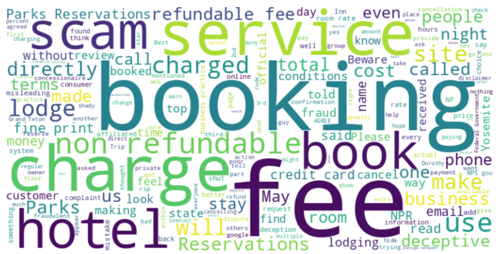


Figure 7: Word cloud based on the top 50 Yelp reviews of National Park Reservations, whose T&Cs specify a 10% non-refundable booking fee. "Scam" and "non-refundable" are frequently mentioned words in the reviews.

> also continue to be charged following the closure of your Account. This provision shall survive the termination of the relationship between you and us.

This term means that even after an account is closed, fees for managing inactive accounts will continue to be charged, potentially resulting in unexpected costs for customers. See Figure 6 in the Appendix for a compilation of customer complaints about account closures and issues with retrieving deposited funds. This demonstrates the potential harm of unfavorable termination and account recovery terms.

***Unfavorable legal terms case study*:** Another concerning unfavorable financial term is the inclusion of a waiver for California Civil Code Section 1542, found in 2.1% (152 out of 7,225) of the top 80,000 Tranco-ranked websites. An example of such a term states:

> If you are a California resident, you shall and hereby do waive California Civil Code Section 1542, which says: "A general release does not extend to claims which the creditor does not know or suspect to exist in his favor at the time of executing the release, which, if known by him must have materially affected his settlement with the debtor."

California Civil Code Section 1542 protects individuals from unknowingly giving up their right to make claims for issues they were not aware of at the time they signed a release. By including a waiver for this code, websites are essentially asking customers to give up this important protection. This indicates that these sites are aware that most people do not thoroughly read the terms and conditions. This oversight can be leveraged to disable significant legal protection, which can make co-existing unfavorable financial terms harder to dispute. In fact, 60.5% (92 out of 152) of the websites with the California Civil Code Section 1542 waiver also have at least one other category of unfavorable financial terms.

E Comparison with Other Online Agreement Annotation Scheme

In this section, we introduce the annotation templates proposed under the European Union (EU) framework for identifying unfair contract terms [21, 25, 26, 43, 46]. While these studies emphasize legal categories and jurisdictional issues, our research specifically targets financial terms in online service agreements.

Loos et al. [46] analyze the unfair contract terms of online service providers in light of the Unfair Contract Terms Directive (UCTD) [75] of the European Union. The authors examine various types of contractual terms from international online service providers, identifying those that are unlikely to pass the Directive's fairness test in the following five categories:

- **Unilateral Changes**: Our analysis also considers unilateral changes made by online service providers, particularly regarding financial aspects such as unilateral price changes, plan upgrades, and various penalties.
- **Termination Clauses**: We examine termination clauses focusing on their financial consequences, including the seizure of digital currency, reward points, or money upon termination.
- **Liability Exclusions and Limitations**: Both our paper and the authors' findings highlight the problematic nature of liability exclusions and limitations, which often unjustly limit the providers' responsibility for service failures, thereby creating a significant imbalance in the parties' rights and obligations.
- **International Jurisdiction and Choice-of-Law Clauses**: Although we have a category for unfavorable legal terms, the detailed categorization of unfair legal terms is deferred to future work. This is because, compared to other unfavorable financial terms, legal terms typically have a more indirect impact on users.
- **Transparency**: While the readability and accessibility of terms and conditions are not the main focus of this paper, our data collection and topic modeling pipeline can be readily adapted for future research in these areas.

Another worth-noting line of work in unfair online agreements [21, 25, 26, 43]. The CLAUDETTE [43] system evaluates the fairness of terms within the jurisdiction of the European Union by leveraging legal standards and principles established within the EU framework. The annotation scheme is as follows:

- Jurisdiction for disputes in a country different from the consumer's residence.
- Choice of a foreign law governing the contract.
- Limitation of liability.
- The provider's right to unilaterally terminate the contract/access to the service.
- The provider's right to unilaterally modify the contract/service.
- Requiring a consumer to undertake arbitration before court proceedings can commence.
- The provider retaining the right to unilaterally remove consumer content from the service, including in-app purchases.
- Having a consumer accept the agreement simply by using the service, without having to click on "I agree/I accept"
- The scope of consent granted to the ToS also takes in the privacy policy, forming part of the "General Agreement"

We consider our work to be a complementary addition to the AI & Law database, with our template being more aligned with the natural phrasing found in terms and conditions of online shopping websites.

We hope that future research will incorporate both the legal templates and our proposed template to provide a more comprehensive understanding of the landscape of unfair (financial) terms.

F Term Page Regex

Below are the positive and negative regex matching pattern we deploy for this work:

Positive

```
positive_regex = [
    "terms.*?conditions",
    "terms.*?of.*?use",
    "terms.*?of.*?service",
    "terms.*?of.*?sale",
    "terms.*?of.*?conditions",
    "terms.*?and.*?conditions",
    "terms.*?&.*?conditions",
    "conditions.*?of.*?use",
    "intellectual.*property.*policy",
    "return[s]?.*?policy",
    "refund[s]?.*?policy",
    "return.*?and.*?refund.*?policy",
    "cancellation.*?and.*?returns",
    "cancellation.*?returns",
    "prohibited.*conduct",
    "electronic.*communication.*policy",
    "safety.*guideline",
    "requests.*from.*law.*enforcement",
    "bonus.*terms.*apply",
    "community.*rules",
    "gift.*card.*policy",
    "contact.*us.*here",
    "shipping.*policy",
    "warranty",
    "end.*user.*license",
    "user.*?agreement",
    "payment.*terms",
    "content.*policy",
    "terms"
]
```

The negative regex list is as follows:

Positive

```
negative_regex = [
    "privacy.*?policy",
    "cookie.*?policy",
    "privacy.*?notice",
    "sale.*?tax.*?policy",
    "prohibited.*?items",
    "1099.*?k.*?form",
    "dmca.*copyright.*notification",
]
```