Information Research - Vol. 31 No. iConf (2026)

Did I really agree to that?: making sense of policies you didn't read with models that actually did

Shikha Soneji, Mitchell Hoesing, Sujay Koujalgi, and Jonathan Dodge

DOI: https://doi.org/10.47989/ir31iConf64157

Abstract

Introduction. While Privacy Policies and Terms of Service (ToS) are intended to inform users; they often overwhelm, mislead, and confuse in practice. This work investigates automated techniques for analyzing such legal documents, with the goal of supporting user comprehension and regulatory auditing.

Method. We use expert-driven annotations from Terms of Service; Didn't Read (ToS;DR) to train classification models that assign case labels to individual sentences. We also develop a classifier to distinguish document types: Privacy Policy, ToS, or other legal text.

Analysis. Models were evaluated using F1 score, and we compared traditional fine-tuned models (RoBERTa, PrivBERT) against GPT-4 Turbo. We then applied the best-performing models to real-world policies to uncover conceptual overlaps between Privacy Policies and ToS.

Results. Our case classifier achieved a 0.73 F1 score, while the document-type classifier reached 0.79. GPT-4 performed worse on case classification (0.58 F1). We found that GDPR-relevant clauses often appear in both Privacy Policies and ToS, blurring distinctions and raising risks for user misinterpretation and regulatory non-compliance.

Conclusion. By surfacing hidden structures and overlapping clauses, our system enhances transparency and supports digital literacy by increasing accessibility of complex documents. This work lays the foundation for tools that promote user agency and platform accountability.

Introduction

Users typically agree to lengthy and complicated Terms of Service (ToS) and privacy policies without reading or comprehending them, leading to uninformed consent and risky data-sharing practices (Auxier et al., 2019; Obar and Oeldorf-Hirsch, 2020). This calls into question the rationale of expecting users to make informed decisions, given that they do not comprehend the terms. Moreover, overlapping and redundant content in these policies can further obscure key points, creating compliance risks under regulations such as the GDPR (The European Parliament and the Council of the European Union, 2018).

Automating the simplification and categorisation of popular ToS documents would be immensely beneficial, enhancing user understanding of accepted policies and facilitating the identification of concerning changes. We envision an automated system that takes the full text of a ToS or Privacy Policy and produces a user-facing summary, such as a simplified bullet list of key takeaways or a letter/number score designed to improve accessibility and comprehension. To support this vision, our work focuses on extracting key policy concepts from the expert-annotated corpus provided by Terms of Service; Didn't Read (ToS;DR) (Terms of Service; Didn't Read, 2012). This dataset includes user-friendly summaries, case-level annotations, and crowdsourced scores, serving as a foundation for automated tools that promote policy transparency. Building on this resource, we address three tasks:

1. **Case classification:** Sentence-level classification of key concepts from policy documents, with a focus on 246 cases defined in the ToS;DR taxonomy (e.g., '*The service has a no refund policy*').
2. **Document type classification:** Sentence-level classification into one of five document types (docTypes are Terms of Service, Privacy Policy, Cookie Policy, Data Policy, and Other Policy), enabling the identification of GDPR guideline (Privacy Terms, 2023) violations through analysing content overlap. One way to determine if policy document writers are following the guidelines is to measure how well classifiers can predict the document type of the source (akin to work from (Pozen et al., 2019), predicting political party affiliation from text).
3. **Concept overlap analysis:** Quantifying redundancies across document types to assess compliance with GDPR requirements for disjoint content.

We perform these tasks by triangulating results from both case and docType classification problems. This paper explores the following RQs, one per task:

- **RQ1** *Case classification* - How well do our NLP classifiers identify and categorise the important clauses defined in our policy taxonomy?
- **RQ2** *Document type classification* - To what extent can document-type classifiers distinguish the source of policy snippets, and what does their confusion reveal about semantic overlap across document categories?
- **RQ3** *Concept overlap analysis* - Which policy concepts are most frequently shared between document types, and how can identifying these overlaps support clearer, human-centric policy understanding?

In answering these RQs, we make three main contributions.

1. **Development of an automated analysis framework:** Existing policy document analysis methods rely heavily on annotated datasets and cannot generalise to new, unseen data. The case classifier from RQ1 represents a step toward building an automated analysis

framework capable of examining fresh, unexplored policy documents, reducing dependence on manual annotations and improving scalability.

2. **Empirical analysis of NLP models:** We present an empirical comparison of transformer-based models (RoBERTa, PrivBERT) and traditional machine learning models (SVM, Random Forest), evaluating their effectiveness in simplifying and classifying policy documents. Our findings from RQ1 and RQ2 demonstrate RoBERTa's potential for building practical tools to enhance privacy policy transparency and compliance.
3. **Application of concept overlap measurement in a new domain:** We operationalise two objective methods to measure and identify conceptual overlaps in policy documents. The first extends prior work (Pozen et al., 2019) to legal and regulatory texts based on the docType classifier from RQ2. The second applies the case classifier from RQ1 to documents of different types, looking for cases that appear in both Privacy Policies and Terms of Service. Finding such cases amounts to discovering non-compliance with GDPR regulations for disjoint content.

Background and related work

This section reviews prior work relevant to understanding, engaging with, and analysing legal documents. We organise the relevant literature on legal document comprehension and analysis, spanning user studies, engagement barriers, privacy risks, and automated approaches for policy understanding.

People's understanding of legal documents

The '*biggest lie on the internet*' (Obar and Oeldorf-Hirsch, 2020) is, of course, '*Yes, I have read and agree to the terms.*' Even though there are no rules requiring websites to disclose their ToS, a number of laws may call for these declarations. Therefore, before using the services provided, users must read and comprehend the ToS provided by an organisation. While it should have taken users 15–17 minutes to read a ToS document thoroughly, they only spent an average of 51 seconds doing so (Obar and Oeldorf-Hirsch, 2020), indicating that information overload was a substantial factor influencing their reading behavior. Companies that use extremely restrictive ToS may struggle to sustain their consumer base and overall profitability (Bakos et al., 2014). Despite some user's claim to have read privacy rules, only 22% of users say they have done so thoroughly, with the majority admitting to having merely skimmed or read a section (Auxier et al., 2019). If consumers lack a comprehensive understanding and providers know their policies largely go unread, the question arises as to how users can make informed decisions.

Barriers to engagement with legal documents

Some users knowingly accept the data-for-service trade-off and thus feel little need to read dense legalese; approximately half report being at least somewhat comfortable with firms using their personal data for service improvements or product development (Auxier et al., 2019; Lippi et al., 2019). To support better communication across stakeholders including legal experts, engineers, and analysts researchers have also explored diagrammatic frameworks. A notable example is the use of concept diagrams to visually represent privacy-preserving data transformations and legal constraints (Oliver et al., 2013), enabling shared understanding across disciplines. Dense, jargon-laden policies '*obfuscate rather than clarify*' (Lipton and Steinhardt, 2019) and their sheer length discourages laypersons from engaging (Robinson and Zhu, 2020). Machine learning techniques like extractive summarisation (Tesfay et al., 2018), automated annotation (Adhikari, 2020), and sentence-level classification (Zimmeck and Bellovin, 2014); attempt to cut through this noise, though concerns about the reliability and trustworthiness of these automated insights remain.

Consequences of non-engagement: security and privacy risks

Consumers' trust erodes when companies fail to safeguard personal data or misuse user information (Melicher et al., 2016; Pilton et al., 2021). Unread policies can carry severe legal consequences; for instance, a court held a user liable for undisclosed credit card fees simply because they clicked '*agree*' without reading (Davis, 2000). This is a mismatch between user expectations and system behavior, which research on user perceptions of privacy and the challenges of configuring smart home systems showed hindered effective privacy management (Kaaz et al., 2017). This work underscores the need for human-centric tools that make privacy configurations more transparent and accessible to non-expert users.

Mining and annotation of policy documents

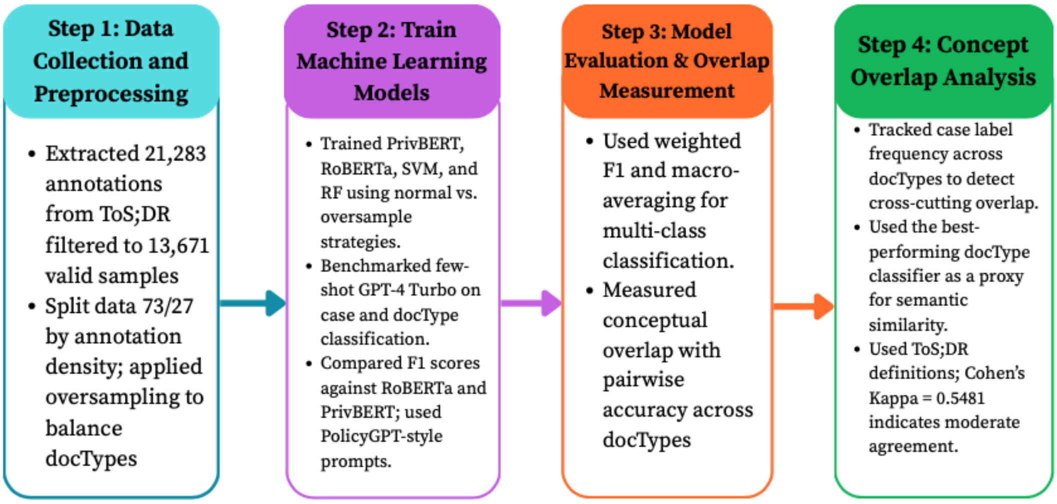
Our framework builds on a diverse range of automated annotation and mining efforts. Unsupervised extraction methods such as Contract Miner (Gao et al., 2011) and ontology-based analyses (Kost and Freytag, 2012) identify service exception clauses at scale. Crowd-sourced platforms like Privee (Zimmeck and Bellovin, 2014) and ToS;DR (Terms of Service; Didn't Read, 2012) provide human-verified sentencelevel labels, while browser extensions such as PrivacyCheck (Zaeem et al., 2018) and automated pipelines (Bui et al., 2021) generate concise summaries for end users. Large-scale studies of regulatory evolution (Amos et al., 2021) and site-wide privacy threat assessments (Alabduljabbar and Mohaisen, 2022) demonstrate the feasibility of processing millions of documents. Complementary NLP resources, including the OPP-115 corpus of expert-annotated privacy policies (Wilson et al., 2016), the PolicyQA reading comprehension dataset (Ahmad et al., 2020), and intent/slot classification approaches (Ahmad et al., 2021), advance machine understanding of legal text. We integrate these annotated datasets to train our case and document-type classifiers.

Related work

A rich ecosystem of tools has emerged for automated privacy-policy analysis (Wagner, 2023) chart in 25 years of ML-driven policy improvements, culminating in user-friendly summaries aligned with GDPR. Systems like Polisis (Harkous et al., 2018) and PrivacyCheck (Zaeem et al., 2018) leverage deep learning and crowdsourcing to generate concise policy overviews but are constrained to persite summarisation and do not explore structural overlaps across document types. Barrister (Perera and Perera, 2021) focuses on class-action waiver simplification, and (Lukose et al., 2022) introduces a hybrid extractive-abstractive ToS summariser both powerful but limited in scale or scope. Other efforts apply QA paradigms (e.g., (Ravichander et al., 2019), browser dashboards (e.g., PrivExtractor (Bolton et al., 2023), and user studies (e.g., (Golbeck and Mauriello, 2016)) to enhance understanding, yet none quantify conceptual overlap across policy corpora. By contrast, we operationalise a large, human-curated ToS;DR taxonomy for sentencelevel '*concept classification*' (Cases and docTypes) and introduce pairwise overlap metrics to reveal GDPR-relevant ambiguities. Our framework thus bridges existing summarisation and QA tools with scalable, cross-document structural analysis, enabling truly human-centric policy comprehension at internet scale.

Methodology

Figure 1 illustrates our end-to-end pipeline for analysing policy documents through a human-centered NLP lens. In Step 1, we curate a dataset of over 13,000 annotated policy clauses from ToS;DR, applying stratified splitting and oversampling to mitigate document-type imbalance. Step 2 involves training a suite of classification models including domain-specific transformers and GPT-4 Turbo to predict policy cases and document types. In Step 3, we evaluate performance using macro-averaged F1 scores and use pairwise accuracy to probe semantic separability. Finally, Step 4 leverages these models to uncover conceptual overlaps between document types, supported by manual annotation and inter-rater reliability analysis. This workflow grounds model predictions in interpretable patterns, helping us surface where policies blur boundaries that matter for user understanding.



Step 1: Data Collection and Preprocessing

- Extracted 21,283 annotations from ToS;DR filtered to 13,671 valid samples
- Split data 73/27 by annotation density; applied oversampling to balance docTypes

Step 2: Train Machine Learning Models

- Trained PrivBERT, RoBERTa, SVM, and RF using normal vs. oversample strategies.
- Benchmarked few-shot GPT-4 Turbo on case and docType classification.
- Compared F1 scores against RoBERTa and PrivBERT; used PolicyGPT-style prompts.

Step 3: Model Evaluation & Overlap Measurement

- Used weighted F1 and macro-averaging for multi-class classification.
- Measured conceptual overlap with pairwise accuracy across docTypes

Step 4: Concept Overlap Analysis

- Tracked case label frequency across docTypes to detect cross-cutting overlap.
- Used the best-performing docType classifier as a proxy for semantic similarity.
- Used ToS;DR definitions; Cohen's Kappa = 0.5481 indicates moderate agreement.

Figure 1. Four-stage workflow showing how sentence-level ToS;DR annotations are processed, modeled, evaluated and analysed to reveal conceptual overlap across policy document types.

Data collection and preprocessing

Figure 2 illustrates that we scraped 21,283 sentence-level annotations from the ToS;DR API (Terms of Service; Didn't Read, 2012) and the ToS;DR website via BeautifulSoup (Richardson, 2007), each annotation pairing a text snippet ('*Description*') with one of 246 human-curated case labels ('Case') and one of five normalised document-type labels ('*docType*'). After removing duplicates and incomplete entries, 13,671 descriptions remained. We split the documents containing these annotations by annotation density greater than 10 annotations/document in the training set (896 docs) versus less than 10 in the test (1,096 docs) so that better-covered documents drive model learning. Splitting documents following this procedure led to a 73/27-train/test split at the sentence level. In simple terms, we train on documents that contain richer information and test on documents with fewer examples, allowing us to evaluate how well the models generalise beyond heavily annotated policies. Finally, to correct the 92% concentration in ToS and Privacy Policy classes, we applied standard oversampling on the training split so that each of the five docTypes has 4,728 samples.

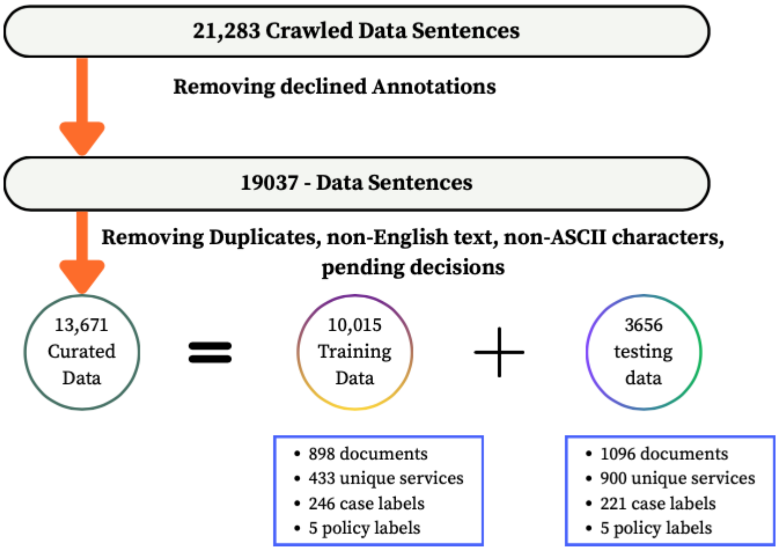


Figure 2. Data dissection pipeline depicting the transition from raw crawled annotations to curated training and test sets, including document counts, service coverage and label distributions.

Training machine learning models

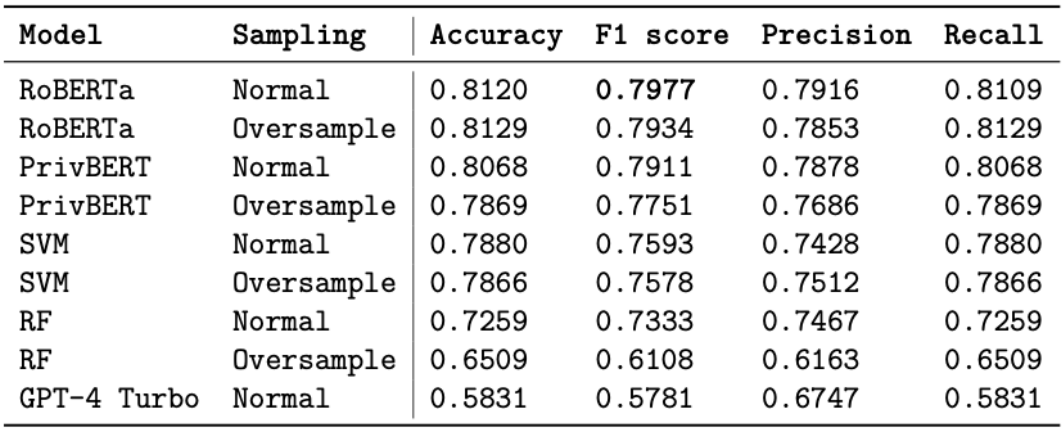
In our Case and Document Type Classification tasks, we employ four notable models: PrivBERT (Srinath, Sundareswara, et al., 2021), RoBERTa (Liu et al., 2019), Linear Support Vector Machines (SVM), and Random Forest (RF). With PrivBERT's specific tailoring for privacy rules and RoBERTa's dynamic token masking and larger training dataset, both models' transformer-based architectures effectively manage long dependencies and complex structures common to intricate policy documents. Our third model, the Linear SVM classifier, handles the highest number of classes (246) in the Case Classification task, which also offers resistance to overfitting. The RF model uses the same models for both classification problems, yielding a simplified presentation while simultaneously managing a vast feature space and possible correlations with built-in overfitting avoidance. We chose to utilise the same model version or configurations when possible to keep the tasks commensurate to simplify comparison. In essence, '*Normal*' and '*Oversample*' represent two different strategies for handling class imbalance in the dataset, each with its unique implications on the model's training and performance. In the context of this experiment, '*Normal*' refers to the original, unmodified set of training and testing data. This is the base dataset that maintains its initial distribution of classes, thereby providing a direct reflection of the real-world scenario, including any class imbalance inherent to the phenomenon and its detection.

In contrast, '*Oversample*' signifies our attempt to address the issue of an imbalanced dataset, specifically within the realm of long-tailed distribution problems. In such cases, certain classes possess significantly more samples than others, potentially causing the model to bias towards the majority class. We trained our NLP models (RoBERTa and PrivBERT) for 5 epochs with warm-up steps for the learning rate set to 500 and weight decay set to 0.01. Based on our computing resources, we set the batch sise for training to 16 and for the test to 64. We also evaluate one-shot GPT-4 Turbo to assess its ability to map policy snippets to our 246 cases and five docTypes, aligning with recent work on PolicyGPT (Tang et al., 2023). We compare GPT's weighted F1 against PrivBERT (Srinath, Wilson, and Giles, 2021) and RoBERTa (Liu et al., 2019) to assess whether LLMs can capture nuanced legal semantics beyond our dedicated classifiers. We instructed the model to map a privacy-related sentence to exactly one of 246 predefined ToS;DR 'Cases', using a strict one-shot format: a single input example and explicit constraints to choose only from the provided list without generating new labels. Full prompt templates and engineering details are provided in Supplemental Materials.

Evaluate models and measure overlap

To perform the overlap measurement, we need to select the best classifiers from those available. The F1 score summarises a model's ability to balance precision and recall, making it well-suited for evaluating performance on imbalanced datasets, which is why we choose F1 score over accuracy for evaluation. We have used weighted F1 scores for all the metrics on the Case Classification and Document Type Classification tasks. Later, Table 2 will show scores for the individual docTypes. We use macro-averaging to construct evaluation metrics because it offers equal weight to each class while doing multi-class classification. It is important to clarify that the F1 score does not measure conceptual overlap between docTypes. Instead, our approach utilises pairwise accuracy for this specific purpose. This method reflects the likelihood of correctly identifying the docType based on semantic content, similar to how political affiliation prediction can measure conceptual overlap in speeches (Pozen et al., 2019). Pairwise accuracy effectively captures the nuances in different docTypes, thereby indicating the degree of overlap better than the macro-averaged F1 score.

Once we determined the best classifiers for Case Classification and document type classification, we set about Concept Overlap Analysis between docTypes in two ways. First, we use the best performance attained on Document Type Classification as a proxy for measuring the conceptual overlap between docTypes, adapting methodology from prior work (Pozen et al., 2019). Second, we run the best performing classifier for case classification and examine the frequency at which it detects a particular case in each docType. If there is no concept overlap, then we should anticipate a high performance on docType classification and a few cases with large numbers of occurrences in each docType. Having identified cases that we considered to overlap, the next question is to determine the proper home for those cases. To assess the privacy relevance of each ToS;DR case, we conducted a binary annotation task involving 245 distinct cases (excluding a special '*abstain*' label for unresolved items). To do this, each of the four authors labeled each case as non-privacy related (0) or privacy-related (1), resulting in 980 total annotations. They performed this binary classification based on reading the definitions or one paragraph explanations given by the ToS;DR team on their website. In cases of disagreements or ties, the first author made the final determination. Due to explicit reference to privacy-related issues, certain cases, such as '*Some personal data may be kept for business interests or legal obligations*,' lend themselves to unambiguous labeling. However, given the inherent subjectivity of individual opinions, the researchers' labels differed at times, such as for cases with less specific information (e.g., '*Pseudonyms are allowed*' or '*Only necessary logs are kept by the service to ensure quality*').



Model	Sampling	Accuracy	F1 score	Precision	Recall
RoBERTa	Normal	0.8120	**0.7977**	0.7916	0.8109
RoBERTa	Oversample	0.8129	0.7934	0.7853	0.8129
PrivBERT	Normal	0.8068	0.7911	0.7878	0.8068
PrivBERT	Oversample	0.7869	0.7751	0.7686	0.7869
SVM	Normal	0.7880	0.7593	0.7428	0.7880
SVM	Oversample	0.7866	0.7578	0.7512	0.7866
RF	Normal	0.7259	0.7333	0.7467	0.7259
RF	Oversample	0.6509	0.6108	0.6163	0.6509
GPT-4 Turbo	Normal	0.5831	0.5781	0.6747	0.5831

Table 1. Performance comparison across models and sampling strategies.

Having created varied individual labeling, we needed to combine them into a final verdict in order to determine the best acceptable category for each case. To do this, the team employed the kappa score (Cohen, 1960) as a statistical measure to gauge the inter-rater reliability for this binary classification task. The kappa score serves as a robust measure of the degree of agreement between the researchers. Utilising this score, the team was able to evaluate and consolidate the independent labels. To quantify consistency among annotators, we computed Cohen's kappa, yielding a score of 0.5481, which indicates moderate agreement (Cohen, 1960). This level of agreement reflects both the inherent subjectivity of the task and the complexity of borderline cases, for example, platform bans, user-generated content, or arbitration clauses that may have indirect privacy implications. These privacy relevance annotations are integral to our later analyses: they allow us to partition the 245 cases into privacy and non-privacy buckets, compare their distribution across document types, and assess which overlaps may violate GDPR's intended separation of legal functions. In this way, our human-in-the-loop labeling provides a grounded, compliance-aware foundation for understanding how policy documents communicate user rights and responsibilities.

Empirical results

We now report empirical findings from our classification and overlap analysis. Results are organised around our three research questions, beginning with case-level classification performance, followed by document-type overlap measurement, and concluding with an analysis of overlapping legal concepts.

RQ1 - Analysing document contents via case classification

To explore how well NLP models can map policy sentences to human-defined categories, we cast the task as a 246-way classification problem using the ToS;DR '*Case*' taxonomy. Each '*Case*' label corresponds to a distinct user-centric legal concept such as '*data not sold,*' '*account deletion,*' or '*tracking via cookies.*' These represent recurring clause types found in real-world terms of service and privacy policies and serve as the core of our ontology. Table 1 shows that transformer-based models consistently outperform traditional classifiers. The RoBERTa model achieves the highest macro-averaged F1 score of 0.7367 under the natural (unbalanced) data distribution, with PrivBERT close behind. While the absolute difference between transformer variants is modest, RoBERTa consistently ranks highest across evaluation metrics, whereas the margin over classical models is substantially larger. This suggests that transformer architectures, pre-trained on large corpora and further fine-tuned on policy text, are well-suited to capture the syntactic and semantic patterns unique to legalese such as passive constructions, conditionals, and finegrained distinctions in consent or data usage.

In contrast, classical models such as SVM and Random Forests exhibit significantly lower performance (macro F1: 0.64), likely due to the sparsity and high dimensionality of the input space. These models are further handicapped by the longtail distribution of case labels: over half of the 246 classes have fewer than 50 instances, amplifying the challenge of generalisation in low data regimes. Notably, our experiments reveal that oversampling rare classes, a common mitigation for class imbalance, consistently leads to decreased performance. This counterintuitive result suggests that forcing artificial balance disrupts the semantic grounding present in organically skewed policy text. Many rare '*Cases*' are not annotation noise but rather meaningful outliers representing edge-case legal commitments (e.g., biometric data opt-outs or arbitration clauses) that occur sparsely but are still important. Our findings imply that for domains like policy analysis, preserving natural class distributions may better reflect real-world document composition and improve generalisation. We also tested the performance of GPT-4 Turbo in a strict one-shot setting, using a structured prompt (see Supplemental Materials) that listed all 246 cases and included a single example. Despite GPT-4's broad language capabilities, it underperformed all fine-tuned classifiers on this task (macro F1: 0.5781). Qualitative inspection reveals that GPT often selected semantically adjacent cases suggesting partial comprehension but insufficient specificity in clause recognition. Without direct gradient-based learning on the ToS;DR taxonomy, the model lacked the precision to resolve subtle distinctions (e.g., '*data shared with third parties*' vs. '*data shared for advertising*'), which are crucial for downstream applications like compliance checking or user alert systems.

RQ2 - Quantifying overlap via document type classification

Table 2 shows that the Document Type Classification task exhibits accuracy and F1 scores in the 0.78–0.81 range, indicating moderate concept overlap among policy docTypes. This suggests that while overlaps exist, they are not severe. A perfect separation of concepts would yield scores closer to 1, while complete overlap (random guessing across 5 classes) would result in scores around 0.2. In line with the trends observed in the Case Classification task, transformer models (RoBERTa and PrivBERT) outperform traditional approaches (RF and SVM). However, in this simpler 5-class classification task, the gap between transformers and classical models narrows significantly. This is likely due to the smaller and less complex label space, allowing classical models to remain competitive. Despite this, RoBERTa remains the top-performing model, followed closely by PrivBERT. This advantage likely stems from RoBERTa's optimised pretraining regime, larger and more diverse training data, dynamic masking, and longer training, which yields more robust and transferable contextual representations than earlier transformer architectures (Özkurt, 2024, Timoneda and Vera, 2025).

	ToS	PP	CP	DP	?P	F1 score
ToS	**1300**	133	2	2	28	0.84
PP	228	**1581**	39	6	21	0.85
CP	6	52	**42**	0	1	0.45
DP	3	12	0	**1**	0	0.05
?P	95	73	1	13	**17**	0.13

	ToS	PP	CP	DP	?P	F1 score
ToS	**1281**	145	3	3	33	0.83
PP	228	**1586**	51	5	5	0.85
CP	3	56	**40**	2	0	0.41
DP	3	56	40	**2**	0	0.00
?P	100	78	1	10	**10**	0.08

Table 2. (**Top**) Macro Averaging Confusion Matrix and F1 for the RoBERTa docType classifier with Normal sampling. (**Bottom**) Same, but with Oversampling. **ToS** = Terms of Service, **PP** = Privacy Policy, **CP** = Cookie Policy, **DP** = Data Policy, **?P** = Unknown Policy.

We expected oversampling to improve performance on underrepresented docTypes (e.g., Data Policy). However, Table 2 shows that oversampling had limited success and even degraded performance for some docTypes. For example, RoBERTa and PrivBERT models exhibited zero performance on the 'Data Policy' docType with oversampling. This indicates that our current oversampling strategy is insufficient and suggests the need for improved data augmentation or additional data collection, consistent with recommendations in prior work (Buolamwini and Gebru, 2018). Initially, we were hoping to provide the output of a statistical test comparing the distributions of case labels between the two docTypes. However, we were not able to find one that was capable of handling the categorical data we have. Thus, we have created and computed the following loss function for binary classification (here the classes are Terms of Service and Privacy Policy documents):

$$\text{Loss.} \quad = \frac{1}{2} \sum_{c} |f_c(\text{PP}) - f_c(\text{ToS})| \qquad (1)$$

Equation 1 shows $f_c(\cdot)$, which is the fraction of samples in a document type assigned

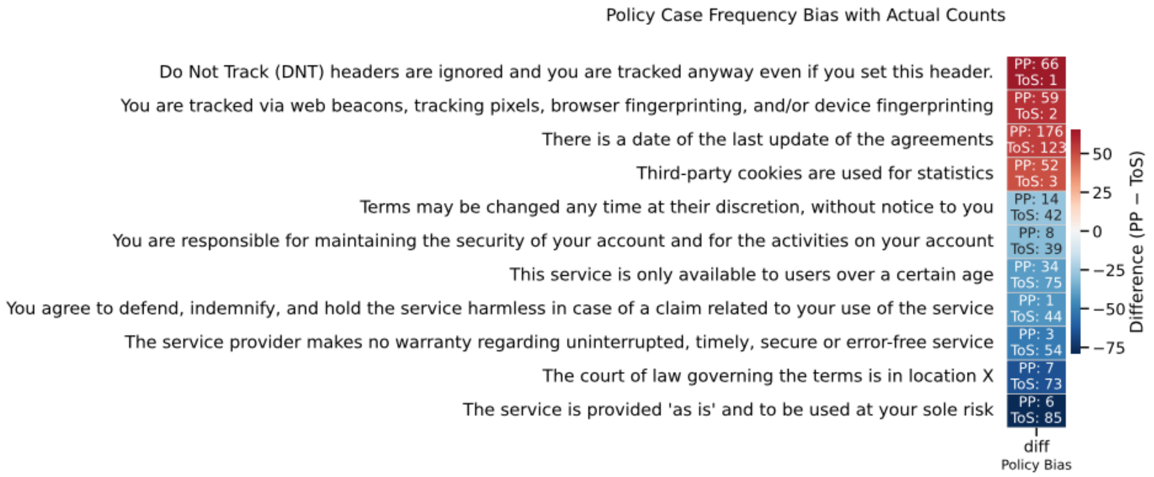


Figure 3. Visualising case-level frequency biases between Privacy Policies (PP) and Terms of Service (ToS). Each row represents a case that appears in both policy types. Cell colors encode the directional bias (red = PP-dominant, blue = ToSdominant), while the annotations show raw counts (e.g., 'PP: 12 | ToS: 7'). This dual representation highlights not only which cases are unevenly distributed, but also how prevalent they are overall helping identify misplaced or ambiguous clauses that contribute to policy overlap.

case $_c$, and the sum runs over all case labels. This loss ranges from 0 (identical distributions) to 1 (completely distinct).

To measure the extent of overlap, we devised a custom loss function that compares the distributions of case labels between Terms of Service and Privacy Policy documents. This function quantifies the difference in the fraction of data receiving each label, yielding a value from 0 (identical distributions) to 1 (completely nonoverlapping). For the best-performing model (RoBERTa), this metric results in a score of 0.6146, suggesting that the distributions of cases in these docTypes are moderately dissimilar. These results suggest that while some concept overlap exists between policy documents, the overlap is not extensive. This partial overlap has critical implications for adherence to GDPR guidelines, which require clear and disjoint content between docTypes. Moderate dissimilarity (as indicated by the loss function score of 0.6146) implies that some clauses may improperly appear in both ToS and privacy policies. This risks creating redundancies that could obscure key privacy information and breach GDPR compliance.

RQ3 - Quantifying overlap via case classification

To deepen our understanding of conceptual entanglement between policy types, we first examine case-level frequency distributions across privacy policies (PP) and Terms of Service (ToS) as shown in Figure 3. Building on this analysis, we focus on the top seven most overlapping cases i.e., those frequently occurring in both PP and ToS, as shown in Figure 4. While some of these clauses (e.g., change notifications, account eligibility) may appear generic, our manual annotations revealed that at least two of them are explicitly privacy-relevant. For instance:

- '*Instead of asking directly, this Service will assume your consent merely from your usage*' reflects inferred consent mechanisms, an explicitly privacy-focused concept under frameworks like GDPR.

- '*This service is only available to users over a certain age*' engages age-based access restrictions, often tied to privacy regulations like the Children's Online Privacy Protection Act (COPPA) and GDPR's age of digital consent.

Despite their privacy implications, both clauses appear with substantial frequency in both ToS and PP documents. This kind of conceptual encroachment undermines the principle of clear policy separation and creates friction for end users trying to locate privacy-specific information.

- The '*consent by usage*' clause appears in roughly equal proportions across PP and ToS, despite its core relevance to privacy expectations.
- The '*age restriction*' clause, although grounded in user protection, often surfaces in ToS where users may not expect to find privacy-related criteria.

This misalignment highlights a deeper structural issue: even when content is privacyrelevant, it is not reliably siloed into the privacy policy. Instead, it is interwoven with terms governing service use leading to blurred boundaries between user rights and service conditions.

From a regulatory perspective, such blending can dilute the effectiveness of mandated transparency. GDPR, for example, emphasises that privacy information should be conveyed in a clear and distinguishable manner, separate from commercial or contractual terms. Our findings reveal that many policies do not meet this standard. Furthermore, from a user-centric standpoint, the interleaving of clauses across documents:

- Increases search effort when looking for specific rights or obligations,
- Undermines mental models users may have about where to find particular types of information,
- And increases the risk of misinterpretation, especially when redundant or conflicting statements exist.

These results call for more structured authorship practices and possibly automated tools that can flag and separate privacy-relevant clauses during the drafting process. Identifying and mitigating such conceptual overlaps will not only improve document clarity but also advance compliance and user trust.

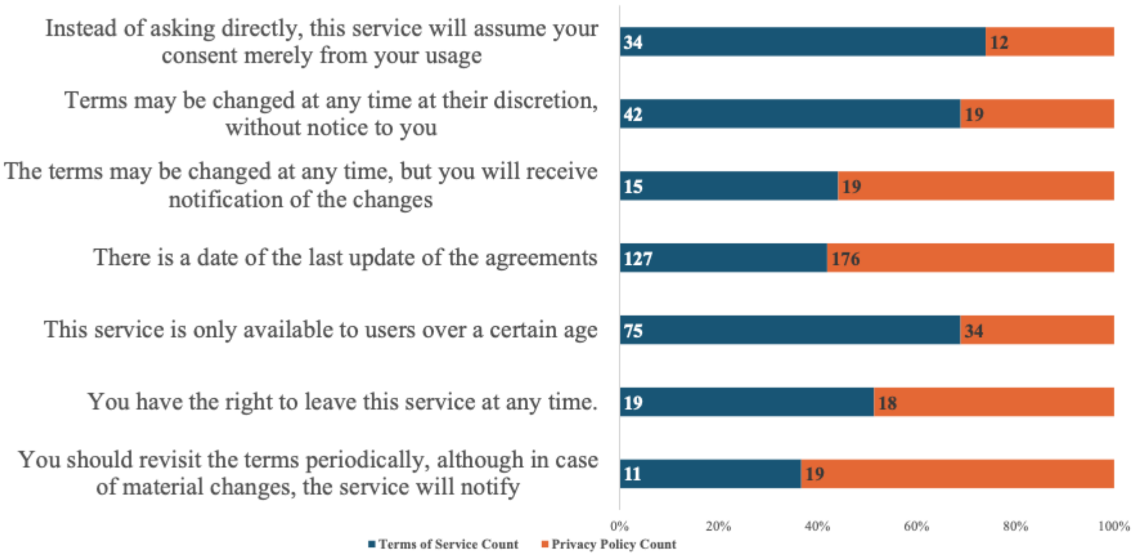


Figure 4. Top 7 overlapping policy cases where the same clause is mentioned in both Terms of Service (ToS) and Privacy Policies (PP), though often with varying frequency. Each bar shows the percentage distribution between ToS and PP, with absolute frequencies inset. The high presence of certain clauses across both documents, such as age restrictions, consent assumptions, and policy change notifications, raises concerns about redundancy and blurred document boundaries. Such overlaps can confuse users about where to locate key information, undermining the interpretability and legal distinctiveness of these policy types.

Discussion

Our study set out to explore how well NLP models can support the structured understanding of policy documents through three central research questions: clauselevel classification (RQ1), document type prediction (RQ2), and analysis of conceptual overlap across documents (RQ3). Our findings suggest both technical feasibility and the need for further human-centric evaluation.

Implications for case classification (RQ1)

We demonstrate that transformer-based models like RoBERTa and PrivBERT are highly effective at categorising policy sentences into 246 distinct human-defined labels. The RoBERTa model achieved an F1 score of 0.7367, indicating a strong capacity to recognise nuanced clause semantics in legal text. However, our results also caution against common pre-processing strategies like oversampling: despite expectations, balancing the class distribution led to worse performance. This suggests that rare clauses in ToS or Privacy Policies may carry important semantic signals that should not be masked through synthetic resampling. While these models enable advanced features like concept heatmaps and clause navigation in user interfaces, their true utility must be measured by their impact on actual users. Future work should involve both qualitative and quantitative user studies to assess whether users can more efficiently or confidently extract knowledge from legal documents using model-driven visual scaffolds. This follows methodologies found in prior work on usable transparency (Golbeck and Mauriello, 2016).

Lessons from document type prediction (RQ2)

Our experiments on docType classification reveal moderate conceptual disjointness between different policy document categories. Transformer models again performed best (F1 > 0.79), but the narrow gap between these and classical models (e.g., SVM) indicates that the task is less semantically demanding than case classification. This aligns with our hypothesis: a five-way classification over broader document categories involves less fine-grained interpretation. Interestingly, oversampling again yielded minimal benefits and sometimes degraded performance on underrepresented types like '*Data Policy*' or '*Cookie Policy*.' This suggests that overfitting to synthetic samples may outweigh any gains in class representation. These outcomes recommend caution when applying standard balancing techniques in domain-specific NLP pipelines.

Reflections on conceptual overlap (RQ3)

Our final analysis highlighted critical breakdowns in structural authorship norms across documents. Across both Privacy Policies and Terms of Service, many clauses recur in similar frequencies, indicating that conceptual boundaries between document types are often weakly enforced in practice. Cases such as '*Two-factor authentication*' and '*Providing identifiable information*' defy conventional expectations about the separation between security-focused ToS and privacy-centric Privacy Policy. Such overlap may reflect poor documentation practices or evolving industry norms, but regardless of cause, the impact on users is concerning. When the same concept appears in both documents, often with subtle differences in language, it increases the cognitive burden on users trying to understand their rights and responsibilities. From a human-centered design perspective, this violates principles of clarity, locality, and minimal redundancy. Our classifiers allow us to surface these inconsistencies systematically, opening new possibilities for visual interfaces that flag clause repetition, recommend reorganisation, or offer interactive explanations. Future systems could even incorporate live feedback from end-users or regulatory experts to iteratively improve policy presentation and alignment with frameworks like GDPR.

Threats to validity

In discussing the threats to the validity of our research, we follow the framework from (Yin, 2018).

Construct validity

We chose explainable models (e.g., Random Forest, SVM, BERT variants) over large blackbox LLMs for better interpretability and future UI integration. GPT-4 Turbo, while promising, performed poorly in few-shot settings and lacks the fine-grained alignment needed for legal classification. Our models offer strong predictive accuracy, but they are not the end goal of human comprehension. We view these classifiers as enabling infrastructure for future policy tools and plan to evaluate their real-world utility through iterative design studies and interface deployment. Throughout our experiments, we adhered to a single-instance classification framing where each sentence is mapped to exactly one case. This simplification was driven by the structure of available ToS;DR-labeled data but may not reflect the true complexity of policy text. Future work could explore multi-label or phraselevel classification to capture overlapping semantics more faithfully. Our use of classification-based simplification enables explainability and future scoring mechanisms, but it abstracts away semantic continuity across sentences or paragraphs. This may limit the expressiveness of our system compared to summarisation-based alternatives. While suitable for structured labeling and case-level visualisations, it may not capture broader rhetorical strategies used in policy writing.

Internal validity

Our work relies on the ToS;DR taxonomy and human curated annotations for both case labels and document types. While this enables high interpretability, the original labels vary in granularity, contain overlapping semantics (e.g., negations like '*data is sold*' vs. '*data is not sold*'), and were not designed for strict classification tasks. Moreover, our own binary privacy-relevance labels used in RQ3 showed only moderate inter-rater agreement (Cohen's $_{\kappa}$= 0.548), reflecting the subjectivity and nuance involved in legal language interpretation.

External validity

Our decision to benchmark both normal and oversampled distributions was informed by common NLP practices. However, oversampling did not consistently improve performance and occasionally harmed it particularly for underrepresented classes like Data Policy. This raises concerns about the stability of results under alternate sampling or augmentation strategies. Moreover, our long-tailed label distribution (246 case classes) increases the likelihood of overfitting to dominant labels or underfitting rare but semantically rich clauses. While transformer-based models like RoBERTa and PrivBERT outperformed classical baselines, these models are fine-tuned on ToS/DR-style documents and may not generalise to other policy formats or languages. Our GPT-4 Turbo experiments highlight the limitations of few-shot prompting on tasks that require domain grounding and structured taxonomies. Similarly, our evaluation of document-type overlap may reflect patterns inherent to our dataset rather than universal characteristics of privacy/legal documentation.

Reliability

Although we demonstrate how our models support policy analysis through caselevel predictions and overlap visualisations, we have not yet tested these tools in real user workflows. Our interpretations about improved document comprehension, reduced redundancy, or GDPR implications remain speculative until validated through user-centered evaluations. As future work, we plan controlled studies that measure how our interface affects users' ability to locate, understand, and compare policy content.

Conclusion

As privacy concerns become increasingly central to people's everyday lives, it is important to understand how natural language processing models can make their lives better. With the central objective of this paper being to break down the legalese in a manner that is concise and understandable to everyone, we hope that this automated summarisation of policy documents will help to simplify the verbosity of the original documents. From a service provider's perspective, we highlighted the redundancies and the overlaps in the different documents that need to be unique to certain texts. By training the model on the dataset of over 10,000 annotations, the RoBERTa model yields an F1 score of 0.74 when classifying a case. The transformer-based models invariably perform better than their traditional counterparts, with an increased efficacy coupled with higher computational time. Meanwhile, results from RQ2 and RQ3 suggest some commonalities in the policy documents, suggesting a substantial amount of overlap. These significant overlaps in the number of cases are indicative of the lack of clarity of terminology between the two types of policy documents and their contents. Our approach provides actionable insights for regulators, document authors, and compliance officers, enabling the detection and mitigation of GDPR violations related to redundant or overlapping content.

Acknowledgements

The authors confirm that this study received no external funding from a grant. The authors would like to thank the Terms of Service; Didn't Read (ToS;DR) community for maintaining the annotated corpus that made this research possible. We are also grateful to the colleagues and reviewers for their thoughtful feedback throughout the development of this manuscript. We also thank Penn State University for providing the institutional resources that supported this research.

About the authors

Shikha Soneji is a doctoral candidate in the College of Information Sciences and Technology at The Pennsylvania State University, USA. Her research focuses on privacy policy comprehension, human-centered AI, and natural language processing to improve transparency in digital consent systems. She works on ontology-driven classification of legal documents and user-centric explainability tools for privacy analysis. She can be contacted at sxs7000@psu.edu or shikhasoneji8@gmail.com

Mitchell Hoesing has completed his master's in computer science from the College of Information Sciences and Technology at The Pennsylvania State University, USA. His research interests include applied machine learning, natural language processing, and computational analysis of legal and regulatory text. His work explores scalable methods for understanding complex legal documents. He can be contacted at mdh5934@psu.edu

Sujay Koujalgi is currently working as a Software Engineer II at GE Healthcare. He graduated with a master's in computer science from the College of Information Sciences and Technology at The Pennsylvania State University, USA. He has worked on LLMs and machine-learning ranking systems, large data pipelines, and enterprise platforms used at scale. He can be reached at koujalgisujay@gmail.com

Jonathan Dodge is an Assistant Professor in the College of Information Sciences and Technology at The Pennsylvania State University, USA. His research centers on explainable AI, human-centered machine learning, and improving trust and transparency in AI systems. He investigates how interactive explanations and model interpretability techniques can support better human collaboration. He can be contacted at dodge@psu.edu

References

Adhikari, A. D. (2020). Automated change detection in privacy policies [Doctoral dissertation, University of Denver].

Ahmad, W., Chi, J., Le, T., Norton, T., Tian, Y., & Chang, K.-W. (2021). Intent classification and slot filling for privacy policies. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 4402–4417.

Ahmad, W., Chi, J., Tian, Y., & Chang, K.-W. (2020). PolicyQA: A reading comprehension dataset for privacy policies. Findings of the Association for Computational Linguistics: EMNLP 2020, 743–749.

Alabduljabbar, A., & Mohaisen, D. (2022). Measuring the privacy dimension of free content websites through automated privacy policy analysis and annotation. Companion Proceedings of the Web Conference 2022, 860–867.

Amos, R., Acar, G., Lucherini, E., Kshirsagar, M., Narayanan, A., & Mayer, J. (2021). Privacy policies over time: Curation and analysis of a million-document dataset. Proceedings of the Web Conference 2021, 2165–2176.

Auxier, B., Rainie, L., Anderson, M., Perrin, A., Kumar, M., & Turner, E. (2019). Americans' attitudes and experiences with privacy policies and laws. Pew Research Center: Internet, Science & Tech.

Bakos, Y., Marotta-Wurgler, F., & Trossen, D. R. (2014). Does anyone read the fine print? consumer attention to standard-form contracts. The Journal of Legal Studies, 43(1), 1–35.

Bolton, T., Dargahi, T., Belguith, S., & Maple, C. (2023). PrivExtractor: Toward redressing the imbalance of understanding between virtual assistant users and vendors. ACM Transactions on Privacy and Security, 26(3), 1–29.

Bui, D., Shin, K. G., Choi, J.-M., & Shin, J. (2021). Automated extraction and presentation of data practices in privacy policies. Proceedings on Privacy Enhancing Technologies.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Conference on fairness, accountability and transparency, 77–91.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and psychological measurement, 20(1), 37–46.

Davis, A. J. (2000). The american prosecutor: Independence, power, and the threat of tyranny. Iowa L. Rev., 86, 393.

Gao, X., Singh, M. P., & Mehra, P. (2011). Mining business contracts for service exceptions. IEEE Transactions on Services Computing, 5(3), 333–344.

Golbeck, J., & Mauriello, M. L. (2016). User perception of Facebook app data access: A comparison of methods and privacy concerns. Future Internet, 8(2), 9.

Harkous, H., Fawaz, K., Lebret, R., Schaub, F., Shin, K. G., & Aberer, K. (2018). Polisis: Automated analysis and presentation of privacy policies using deep learning. 27th {USENIX} security symposium ({USENIX} security 18), 531–548.

Kaaz, K. J., Hoffer, A., Saeidi, M., Sarma, A., & Bobba, R. B. (2017). Understanding user perceptions of privacy, and configuration challenges in home automation. 2017 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC), 297–301.

Kost, M., & Freytag, J. C. (2012). Privacy analysis using ontologies. Proceedings of the second ACM conference on Data and Application Security and Privacy, 205–216.

Lippi, M., Pałka, P., Contissa, G., Lagioia, F., Micklitz, H.-W., Sartor, G., & Torroni, P. (2019). Claudette: An automated detector of potentially unfair clauses in online terms of service. Artificial Intelligence and Law, 27(2), 117–139.

Lipton, Z. C., & Steinhardt, J. (2019). Troubling trends in machine learning scholarship: Some ML papers suffer from flaws that could mislead the public and stymie future research. Queue, 17(1), 45–77.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimised BERT pretraining approach. arXiv preprint arXiv:1907.11692.

Lukose, E., De, S., & Johnson, J. (2022). Privacy pitfalls of online service terms and conditions: A hybrid approach for classification and summarisation. Proceedings of the Natural Legal Language Processing Workshop 2022, 65–75.

Melicher, W., Sharif, M., Tan, J., Bauer, L., Christodorescu, M., & Leon, P. G. (2016). Preferences for web tracking. Proceedings on Privacy Enhancing Technologies, 2016(2), 1–20.

Obar, J. A., & Oeldorf-Hirsch, A. (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), 128–147.

Oliver, I., Howse, J., & Stapleton, G. (2013). Protecting privacy: Towards a visual framework for handling end-user data. 2013 IEEE Symposium on Visual Languages and Human Centric Computing, 67–74.

Özkurt, C. (2024). Comparative analysis of state-of-the-art Q&A models: BERT, RoBERTa, DistilBERT, and ALBERT on SQuAD v2 dataset. Chaos and Fractals, 1(1), 19–30.

Perera, T., & Perera, T. (2021). Barrister-processing and summarisation of terms & conditions/privacy policies. 2021 6th International Conference for Convergence in Technology (I2CT), 1–7.

Pilton, C., Faily, S., & Henriksen-Bulmer, J. (2021). Evaluating privacy-determining user privacy expectations on the web. Computers & Security, 105, 102241.

Pozen, D. E., Talley, E. L., & Nyarko, J. (2019). A computational analysis of constitutional polarisation. Cornell L. Rev., 105, 1.

Privacy Terms. (2023). Privacy policy vs terms and conditions. https://privacyterms.io/privacy/privacy-policy-vs-terms-and-conditions/

Ravichander, A., Black, A. W., Wilson, S., Norton, T., & Sadeh, N. (2019). Question answering for privacy policies: Combining computational and legal perspectives. arXiv preprint arXiv:1911.00841.

Richardson, L. (2007). Beautiful Soup documentation. https : / / beautiful - soup - 4.readthedocs.io/en/latest/

Robinson, E. P., & Zhu, Y. (2020). Beyond 'I agree': Users' understanding of web site terms of service. Social media+society, 6(1), 2056305119897321.

Srinath, M., Sundareswara, S. N., Giles, C. L., & Wilson, S. (2021). PrivaSeer: A privacy policy search engine. 21st International Conference on Web Engineering, ICWE 2021, 286–301.

Srinath, M., Wilson, S., & Giles, C. L. (2021). Privacy at scale: Introducing the PrivaSeer corpus of web privacy policies. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 6829–6839.

Tang, C., Liu, Z., Ma, C., Wu, Z., Li, Y., Liu, W., Zhu, D., Li, Q., Li, X., Liu, T., et al. (2023). PolicyGPT: Automated analysis of privacy policies with large language models. arXiv preprint arXiv:2309.10238.

Terms of Service; Didn't Read. (2012). Project website for 'Terms of Service; Didn't Read'. https://tosdr.org/

Tesfay, W. B., Hofmann, P., Nakamura, T., Kiyomoto, S., & Serna, J. (2018). PrivacyGuide: Towards an implementation of the EU GDPR on internet privacy policy evaluation. Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics, 15–21.

The European Parliament and the Council of the European Union. (2018). European Union General Data Protection Regulation (GDPR) [Accessed: 1/07/2026]. https://gdpr-info.eu/

Timoneda, J. C., & Vera, S. V. (2025). BERT, RoBERTa, or DeBERTa? comparing performance across transformers models in political science text. The Journal of Politics, 87(1), 347–364.

Wagner, I. (2023). Privacy policies across the ages: Content of privacy policies 1996– 2021. ACM Transactions on Privacy and Security, 26(3), 1–32.

Wilson, S., Schaub, F., Dara, A. A., Liu, F., Cherivirala, S., Leon, P. G., Andersen, M. S., Zimmeck, S., Sathyendra, K. M., Russell, N. C., et al. (2016). The creation and analysis of a website privacy policy corpus. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1330–1340.

Yin, R. K. (2018). Case study research and applications: Design and methods (Sixth edition). SAGE.

Zaeem, R. N., German, R. L., & Barber, K. S. (2018). Privacycheck: Automatic summarisation of privacy policies using data mining. ACM Transactions on Internet Technology (TOIT), 18(4), 1–18.

Zimmeck, S., & Bellovin, S. M. (2014). Privee: An architecture for automatically analysing web privacy policies. 23rd {USENIX} Security Symposium ({USENIX} Security 14), 1–16.