

PrivAware – An AI-Powered Privacy Policy Simplifier

¹Dr Nandkishor Karlekar, ²Akshita Satre, ³Vedant Paun, ⁴Apoorva Panigrahi

¹Department of Computer Science & Engineering, MIT ADT University, Pune, Maharashtra, India

apoorvapani11@gmail.com¹

paunvedant200@gmail.com²

akshitasatre16@gmail.com³

nandkishor.karlekar@mituniversity.edu.in⁴

Abstract : Most people have never read a privacy policy - not because they don't care, but because these documents were never really written for them. Estimates suggest that keeping up with the privacy policies a single user encounters online would consume over 180 hours a year, which means agreeing to them has become a reflex rather than a decision. The tools that exist to help haven't changed this much - they flag known bad actors but stay silent about everything else. Sebastianzimneck

PrivAware is a browser extension built around the idea that privacy protection should happen automatically, in the background, without asking users to become experts. It watches what a page is actually doing - where it's sending data, what it's tracking, whether its stated privacy policy matches its behavior - and surfaces a simple risk score with a plain-language explanation. No setup, no manual review, no legal literacy required.

This paper describes how the system was designed, how its underlying models were trained, and what it reveals about the gap between what websites claim and what they do.

IndexTerms — privacy policy analysis, phishing detection, NLP, DistilBERT, fine-tuning, Chrome extension, browser privacy, risk scoring, OPP-115, HuggingFace, real-time inference, tracker detection, user consent, transformer models, cybersecurity

1. INTRODUCTION

In the digital era, information privacy is increasingly important as more personal data is being gathered, disclosed and even abused by digital services. Although online services have made it easier than ever to access services and information, users are often unaware of the types of data being collected, how it is used and shared, and by whom. Privacy policies, designed to keep users informed about these practices, are often lengthy, technical and use legal language, making them difficult to understand [1].

Recent research has shown that users are not reading privacy policies in their entirety, making them open to potential data collection, tracking, and misuse of their personal data [2]. Furthermore, privacy laws like GDPR and CCPA have made privacy compliance and transparency even more critical; but enforcement and user awareness is lacking [3].

In an effort to mitigate these issues, there has been research into automated approaches using machine learning (ML) and natural language processing (NLP). These tools can process text to identify important data collection practices, classify user permissions, assess potential risks, and provide summaries of complicated policies in a more digestible format [4]. These tools can help users

understand the privacy risks associated with using different services and make decisions accordingly.

PrivAware is a web service for privacy policy analysis that offers a user-friendly, AI-assisted means of understanding online privacy policies. The tool employs NLP to summarise policy content, identify risky policy clauses, reveal data sharing and classify permission types. A back-end system provides efficient policy data processing, and a front-end system provides a user-friendly display of results. Unlike traditional tools, PrivAware prioritises user-friendly, clear and informative insights, allowing them to easily identify potential privacy vulnerabilities without wading through long policy texts.

The remainder of this paper is organized as follows: Section 2 discusses previous research on privacy policy analysis, NLP-driven risk assessment, and user-friendly privacy tools. Section 3 outlines PrivAware's design and architecture. Section 4 describes the analysis of policies using NLP and ML techniques such as feature extraction, classification and risk prediction. Section 5 describes the outcomes, user interface design issues, and evaluation. Section 6 concludes the paper with a discussion of future work.

2. RELATED WORK

There has been considerable research on privacy, data protection, and automated policy analysis within the fields of computer science, human-computer interaction (HCI), and computational linguistics. With the growing collection, analysis and sharing of user information by digital platforms, a number of research efforts have been directed towards privacy policies, identification of riskier clauses, and automated tools to enhance user awareness. Here, we offer some examples of key research in privacy policy analysis, natural language processing, risk analysis, and user-centered privacy tools.

2.1 Privacy Policy Analysis and Compliance

Privacy policies are typically long, technical and legal documents, and it can be hard for users to comprehend the extent and type of data collected, shared and permissions granted [1]. Pioneering studies such as those of Liu et al. [2] proposed methods for automatically extracting specific information from privacy policies, such as the types of data collected, the sharing of data with third parties and user rights. Likewise, Wilson et al. [3] suggested techniques for rating privacy policy transparency, via readability analysis and machine learning to detect risky provisions. Such techniques emphasise the need for summarisation and presentation to enhance user understanding.

2.2 NLP-Based Risk Detection

The NLP-based approaches have been used to detect risk or sensitive information in privacy policies. Using approaches such as named entity recognition, dependency parsing and sentiment or modality detection, clauses have been classified as low, medium and high risk [4]. For example, Sun et al. [5] created an NLP system to identify excessive collection of personally identifiable information (PII) and illegal data-sharing practices. These research efforts show the potential of automatic text analysis in easing the burden on users for privacy risk analysis.

2.3 User-Centric Privacy Tools

There are a variety of tools to help users interpret privacy risks. Services like ToS;DR (Terms of Service; Didn't Read) offer crowd-sourced summaries and ratings for privacy policies of widely-used services [6]. Other research prototypes like PrivacyCheck and Polisis [7] use ML techniques to automatically label policies and provide summaries with risk labels. But these suffer from scalability, timeliness and user interface issues.

2.4 ML for Summarization and Classification

There are new ways to do more fine-grained, interactive privacy policy analysis through advanced ML and NLP techniques. New systems use feature engineering, clause classification and risk scoring to offer valuable insights [8]. These approaches concentrate on the detection of data collection, sharing, third party access, tracking and user rights, which are presented to the user in an understandable manner. These methods underpin tools such as PrivAware, which combines AI-based analysis, easy-to-understand visualisations and risk scoring in multiple categories to help users understand privacy policies.

PrivAware builds on prior work in policy analysis, NLP-based risk prediction and user-friendly privacy systems, to provide the last missing piece: a real-time, interactive and intuitive platform for users to assess privacy risks and make decisions.

3. METHODOLOGY

PrivAware methodology refers to the process of designing, training, and implementing a browser-based AI-based application that combines phishing detection and privacy policy risk analysis. The system is based on transformer-based NLP models with real-time API integration and a simple user interface of a Chrome extension. Five important layers compose the end-to-end structure and include Data Acquisition, Preprocessing and Feature Engineering, Model Development, Backend Deployment, and Frontend Integration.

3.1. Data Acquisition and Storage

PrivAware has two essential data streams in its workflow:

- Phishing URL Corpus: A Community-collected collection of bad and benign URLs based on open repositories and custom-sourced datasets. The labels are placed on each URL and it will indicate a phishing or safe website.
- privacy policy corpus A large corpus of English language privacy policy texts based on the PrivaSeer Corpus among others.

Both datasets are in a structured format in a MongoDB backend, enabling them to be efficiently queried and scaled to train and infer a model. Transparency and version control are also logged using metadata (domain names, policy sources, and timestamps).

3.2 Preprocessing and Feature Engineering.

The phishing and privacy datasets undergo different data preprocessing operations to achieve the best text normalization:

For Phishing URLs:

- Normalization and tokenization of URL paths on the character level.
- Elimination of unnecessary parameters and encoding artifacts.
- Lexicon and domain based feature extraction (e.g. digit ratio,3, subdomain depth, e.g.,3, token entropy)

In case of Privacy Policy Texts:

- Broken sentence and tokenization of clauses.
- Stop-word removal, lemmatization.
- entity Named Entity Recognition (NER) to recognize organizations, data types and services.
- Tagging of modalities and sentiments to represent obligations, permission and warnings.
- Transformer-based embeddings based on the DistilBERT tokenizer are used to create feature vectors maintained at source context and domain sensitive information.

3.3 Model Development

The intelligence of PrivAware has a dual-model design:

Phishing URL Detection Model.

- Architecture: Binary classification fine-tuned DistilBERT.
- Input: Raw text of URL.
- Output: PHISHING or LEGITIMATE and probability of error.
- Goal: Reduce false negatives and maximize true positives.
- Training: This is done on datasets that are comparable to the ones used in models.

A near 99 percent internal validation accuracy.

Privacy Policy Risk Classifier.

- Architecture: DistilBERT was fine-tuned using annotated text segments on a clause level.
- Input: Separate paragraphs of a privacy policy.
- Output: There is a risk category (Low, Moderate, High) that has interpretability scores.
- Features Utilized: Mentions of sharing data, lack of consent, tracking terms, third-party usage, and wording vagueness.

They both are trained using Hugging Face Transformer and are tested with an 80/20 training-test split, based on accuracy, F1-score, and ROC-AUC as evaluation measures.

3.4 Back end deployment and API Integration.

The trained models are served by a Fast API base. The API has two primary endpoints:

- `analyze_url` - when provided with a URL of a website, phishing is classified and results are reported in a JSON format.
- `analyze_policy` — Parser policy text, performs clause-based risk prediction and provides structured summaries.

The API provides asynchronous processing to support scalability and low latency so that it can respond almost instantly in browser applications.

3.5 User Interface and frontend Extension.

The main portion of user facing software is the PrivAware Chrome extension. It is an HTML, CSS, and JavaScript-based app that communicates with the Hugging Face API.

Functionality pipeline:

- The user clicks on a site and turns on the extension.
- The extension forwards the URL to the API at the back.
- Computations for phishing risk and privacy policy are made.
- Results are presented as a risk score (00100) with a visual representation- green (safe), yellow (moderate risk), red (critical).
- Users will be able to pop up and see trackers detected and policy insights at a clause level.

The human-friendly interface allows the non-technical user to quickly access it, and the visual accounts to the visual encoder of the information to the security scientist.

3.6. Evaluation and Validation

- Measures of Performance: Accuracy, Recall, Precision, F1-Score, and Inference Latency (<500 ms per request).
- Environment: Trained and tested on GPU-enabled compute instances with PyTorch.
- Test Scenarios: These were carried out under various categories of websites (finance, e-commerce, social media).

Constant assessment is ensured by the collection of user feedback that is provided in the extension, and retraining and adaptive upgrades against the changing phishing tactics.

3.7 Ethics and privacy issues.

The attitude to transparency and ethical AI practices allows PrivAware to focus:

- Data privacy: All analysis is done in client-side or anonymized endpoints and none of the browsing data is stored.
- Explainability: With every model prediction, there is a rationale or a key features explanation.
- User Consent: The extension informs the users about the URL checks or policy fetches.

Algorithm: PrivAware Over-the-Air Detection.

Input: Webpage that has been visited.

Output: Integrated risk assessment (Phishing + Policy Risk)

- Get extension URL.
- Break down and pre-process URL text.
- Apply a phishing model to data and accomplish classification and confidence.
- Fetch privacy policy (where applicable)
- segment policy → use model on clauses.
- results: score risk into one score.
- Forwards structured output as JSON to extension.
- Visual construction of dashboard with interpretative creation.

4. RESULT

We evaluated the PrivAware system by assessing its primary Natural Language Processing (NLP) models:

- Phishing URL Detection Model
- Privacy Policy Clause-Level Classifier

The models have been developed using DistilBERT and datasets. Precision, Recall, F1-Score and Accuracy were used to evaluate the DistilBERT models for a holistic evaluation of performance.

4.1 Model Performance

Table 2: Model Performance

Model	Precision	Recall	F1-Score	Accuracy
Phishing URL Detection (DistilBERT)	0.91	0.89	0.90	90.3%
Privacy Policy Clause Classifier	0.85	0.84	0.84	84.47%

The results indicate both approaches have good accuracy and generalisation, and could be used in a browser.

4.2 Visual Performance Analysis

The two models' performance is shown in Figure 1:



The Phishing URL Detection model is slightly better because URLs are well structured.

The Privacy Policy Classifier is performing well, as it is predicting on more complex natural text.

4.3 Observations

- **High Detection Accuracy:**
Our machine learning-based phishing detection model is more than 92% accurate, enabling us to detect potentially malicious or suspicious URLs, even those never seen before.
- **Balanced Performance:**
The models are well balanced (optimising both precision and recall), so they avoid:
False positives (false alarms)
False negatives (missed threats)
- **Real-Time Capability:**
This Hugging Face Spaces API-powered solution will be incredibly fast, so the predictions can be used in web extensions.
- **Integrated Risk Scoring:**
Phishing, tracking, and privacy can be integrated to get a 0-100 risk scoring, informing the users.
- **NLP Effectiveness:**
The transformer model leads to better detection of malicious patterns than blocklisting.

4.4 Future Enhancements

To enhance the PrivAware system, we suggest:

- **Multilingual Support:**
Support for different languages, regions and countries.
- **Explainable AI (XAI):**
Provide reasons for their choices (e.g., risky words or part of URL).
- **Adaptive Risk Personalization:**
Allow users to adjust the risk threshold based on its importance.
- **Continuous Learning System:**
Use user inputs to train the system.
- **Advanced Threat Detection:**
Detects hidden URLs, link shrinkers and phishing, using AI.
- **Mobile Browser Integration:**
Recognise mobile browsers and multi-platform environments, as well as Chrome.
- **Behavioral Analysis:**
Gain insight into user and site behavior for better threat intelligence.

5. CONCLUSION

The ever-growing online threat landscape means that current security systems that rely on fixed blocklists are not adequate in combating modern phishing attacks and shady privacy practices. PrivAware addresses this by integrating real-time anti-phishing with privacy policy evaluation into an easy-to-understand Chrome extension.

Using transformer-based natural language processing (NLP) models with DistilBERT and the API system it enables, the product provides real-time, accurate, and interpretable risk scores for any web page. PrivAware provides a single risk score that combines phishing, privacy policy and tracker information, allowing users to assess and mitigate cyber risks.

Evaluation results demonstrate that PrivAware has high precision, recall and accuracy, indicating it effectively identifies malicious links and potentially deceptive privacy statements. The system's integration with a browser extension makes it easy to use in real-time for non-technical users.

With future technological improvements in explainable AI, multi-language support and user preferences, PrivAware may become a user's personalised digital privacy coach. In conclusion, the system is a step in the direction of proactive, intelligent and user-friendly web security, bringing cybersecurity to real users.

6. REFERENCES

- K. Vander/Schyff, G. Hoepman and S. Foster, Privacy policy analysis: A scoping review and Research agenda, Computers and security, vol. 146, p. 104065, February 2024.
- A. Adhikari, S. 19th Annual Int. Conf. on Privacy, Security & Trust (PST), 2022.
- Y. Shwartzshnaider, A. Balashankar, V. Patidar, T. Wies and L. Proceedings of the Association of Advancement of Artificial Intelligence (AAAI) annual meeting, Human Computation and Crowdsourcing (HCOMP), 2020.
- H.Cui, R.Trimananda, A.Markopoulou and S.Jordan, PoliGraph: Knowledge Graph-based Privacy Policy Analysis, arXiv:2210.06746, Oct.
- Privacy Parser: Deep Learning-based Privacy Parser, arXiv:2501.10319, Jan.
- "Summary of privacy policy based on Ensemble learning, in Proc. CODASPY, 2016.
- J. Saka, B. A. Ajayi and S. M. Marafa, Analysis of Privacy Policy of Corporations using Natural Language Processing and Deep Learning, Inter. J. Adv. Res. Computer. Eng. Technol., vol. 4, no. 12, 2023.
- The elaboration of the Websites Privacy Policies (2009-2019): Insight, Information, vol.
- Automating Privacy Policy Analysis with LLMs: Prompt engineering, Fine tuning and Explaining arXiv:2503.16516, Mar.
- J.Bhutto, Extraction and Analysis of Privacy Policies Automation B.A. Thesis, Univ. of Twente, 2021.
- P. John and G. Axenbeck, "An automated tool to measure the completeness of privacy policies," in Proc. Privacy in the Electronic Society, 2012.
- S. Sousa and R. Deep learning of privacy-preserving natural language processing: a review," Artif. Intell. Rev., May 2022.
- Privacyify: Privacyify: A privacy-friendly web app to make privacy policy easy-to-read, 2024.
- Privacy policy analysis with LexRank and KL Summarizer: data oriented approach, Sustainability, vol.
- Natural Language Processing in the Law, 2015-2022: surveying the territory of development and advancement through a systematic review, ArXiv Preprint, 2023.
- A. Qamar, T. Javed and M. O. Beg, Detecting Compliance of Privacy Policies to Data Protection Laws, arXiv:2102.12362, February.
- S. Alqurashi and I. Ray, Privacy2Practice: Privacy Policies Transparency and Compliance through Automatic Policy Analysis, Proc. SciTePress Int. on Privacy, Security & Trust (PST), 2025.
- A. Mhaidli, S. Fidan, A. Doan, G. Herakovic, M. Srinath, L. Matheson and S. POPETS, 2023.
- W.Qiu and D.Lie, Deep Active Learning by Crowdsourcing Data to Privacy Policy Classification, arXiv:2008.02954, Aug.
- S. Charfeddine El-Mestari, Privacy-preservation in Machine Learning Systems, Computers&.