Information Research - Vol. 30 No. iConf (2025)

Taking disagreements into consideration: human annotation variability in privacy policy analysis

Tian Wang, Yuanye Ma, Catherine Blake, Masooda Bashir, and Ryan Wang
DOI: https://doi.org/10.47989/ir30iConf47581

Abstract

Introduction. Privacy policies inform users about data practices but are often complex and difficult to interpret. Human annotation plays a key role in understanding privacy policies, yet annotation disagreements highlight the complexity of these texts. Traditional machine learning models prioritize consensus, overlooking annotation variability and its impact on accuracy.

Method. This study examines how annotation disagreements affect machine learning performance using the OPP-115 corpus. It compares majority vote and union methods with alternative strategies to assess their impact on policy classification.

Analysis. The study evaluates whether increasing annotator consensus improves model effectiveness and if disagreement-aware approaches yield more reliable results.

Results. Higher agreement levels improve model performance across most categories. Complete agreement yields the best F1-scores, especially for First Party Collection/Use and Third-Party Sharing/Collection. Annotation disagreements significantly impact classification outcomes, underscoring the need for understanding annotation disagreements.

Conclusion. Ignoring annotation disagreements can misrepresent model accuracy. This study proposes new evaluation strategies that account for annotation variability, offering a more realistic approach to privacy policy analysis. Future work should explore the causes of annotation disagreements to improve machine learning transparency and reliability.

Introduction

Privacy policies are crucial for digital privacy, explaining how businesses handle personal data to help users make informed decisions. However, the FTC and other studies find most are '*incomprehensible*', with users rarely reading or understanding them (Braun, 2024; Hallgren, 2012; Tang et al., 2021). Technical language and inconsistent terminology further complicate comprehension (Azhagusundari & Thanamani, 2013), and ambiguities lead to varying interpretations of terms (Pedregosa et al., 2011). Users' understanding also depends on their education and cultural background (Hossin & Sulaiman, 2015; Krippendorff, 2018).

Researchers have used natural language processing (NLP) to improve the readability of privacy policies (Li, et al;, 2022; Wilson et al., 2016) Efforts include summarising lengthy policies (LingPipe Alias-i., 2008), using topic models to extract data practices (Harkous et al., 2018), classifying content (Grosman et al., 2020), identifying specific information (Gray, 2011; Mysore Sathyendra et al., 2017) and checking compliance automatically (Liu, et al., 2016). Using machine learning (ML) and natural language processing (NLP) to improve privacy policy readability offers benefits, such as processing large data quickly and outperforming human annotators in speed. However, human annotation, often used to create gold-standard datasets, is time-consuming and relies on inter-rater reliability, leaving disagreements unexplored. ML models, which optimise probability-based classification, prioritise data quantity over annotation variety (Plank, 2022). This overlooks data inconsistencies and diverse interpretations. A study found that even privacy experts rarely agree on policy interpretations, suggesting automated tools may struggle to accurately interpret policies, just like users (Braun, 2024).

Human annotation and interpretation of policy and legal texts are often inconsistent and prone to disagreement. For instance, annotators assigned three different labels—Other, User Choice/Control, and First Party Collection/Use—to the same text, while another sentence received two distinct labels: first and third party, and User Choice/Control. Similarly, texts (4)-(7) were inconsistently labelled as First Party Collection/Use by some annotators but not by others. This variation stems from the deliberate ambiguity and complexity of such texts, reflecting their inherently controversial and evolving nature.

Texts to be annotated/ labelled:

1. If you do not wish to share your PIN, you always have the option to not provide the information or use the MediaNews Websites that require it.
2. By use of our websites and games that have dynamic in-game advertising, you signify your assent to SCEA's privacy policy.
3. You may register or enhance your profile by linking your Facebook or Google accounts on NYTimes.com.
4. Sharing Your Information with Other Companies
5. You can delete cookies using your browser settings.
6. What Choices Do I Have?
7. You can visit our Web Sites without sharing personally identifiable information.

Annotation scheme/label choices:

OPP-115's annotation scheme consists of ten data practice categories:

1. *First party collection/use*: how and why a service provider collects user information.
2. *Third party sharing/collection*: how user information may be shared with or collected by third parties.
3. *User choice/control*: choices and control options available to users.
4. *User access, edit, & deletion*: if and how users may access, edit, or delete their information.
5. *Data retention*: how long user information is stored.
6. *Data security*: how user information is protected.

7. *Policy change*: if and how users will be informed about changes to the privacy policy.
8. *Do not track*: if and how Do Not Track signals for online tracking and advertising are honoured.
9. *International and specific audiences*: practices that pertain only to a specific group of users (e.g., children, Europeans, or California residents).
10. *Other*: additional sub-labels for introductory or general text, contact information, and practices not covered by the other categories.

We argue that revealing variations in understanding privacy policies is an underexplored area of research. This gap partly arises from the bias in using machine learning as the primary method and from the tendency to treat privacy as a uniform concept, overlooking cultural, educational, and gender differences. To address this, we conduct an empirical study using the OPP-115 corpus, which includes annotations from three annotators. We expand the characterization from three (individual, pairwise, and complete agreement) to seven: three individual, three pairwise, and one gold standard with full agreement. Unlike majority voting, pairwise agreement requires two specific annotators to agree. Complete agreement, while yielding fewer instances in the target class, may offer higher-quality annotations. Our study provides tangible measures for understanding how annotation variations impact machine learning performance. Rather than simply optimising metrics, we aim to reflect human disagreements through realistic reporting. We highlight two approaches to constructing gold standards—one ignoring disagreement and one considering it—without advocating for either, acknowledging that practical constraints often dictate these choices. We emphasise that disagreements in interpreting privacy statements are common, even among experts, and recognizing these differences is essential for user-centred privacy research.

Related work

High-quality labelled data are essential for supervised machine learning. Some argue that multiple annotators reduce human bias in evaluation (Artstein, 2017), but the number of annotators needed for high-quality annotation is unclear and often limited by budget (Mysore Sathyendra et al., 2017). Inter-rater reliability (IRR), or annotator agreement, is commonly used to measure annotation quality (Moallem, 2018). Typically, studies use either text annotated by any rater (union) or by the majority (majority vote), ignoring genuine annotator differences.

The assumption that annotators are interchangeable is also flawed, as differences exist between annotator populations and individuals (Hershcovich et al., 2022). Agreement studies show that variation can arise from heterogeneous data, complex labels, and annotator differences. Global agreement coefficients may mask these variations, while more detailed studies provide insights (Stevens et al., 2020).

A review of legal machine learning datasets found that disagreement during annotation is typically removed, with final corpora containing only '*gold standard*' annotations. Common strategies like majority vote, forced agreement, expert review, or arbitration do not account for disagreements (Prabhakaran et al., 2021). Issues of interpretation and non-transparency in reporting machine learning results, especially with legal documents, have also been flagged (Plank, 2022). Accurate predictions can build trust, but reproducibility depends on dataset validity and the discovery process (Herbert et al., 2023). Thus, documenting data preparation processes and their impact on model performance is crucial for trust and transparency in machine learning results (Bai et al., 2022).

Method

Towards this goal of using machine learning to provide privacy policy statement analysis that documents and considers annotation disagreements, our study provides ways to demonstrate how

to directly measure the impact of human (dis)agreement on machine learning model performance, by responding to the following research questions:

> RQ1: to what extent does reaching consensus amongst annotators impact the classification performance of traditional machine learning and deep learning models, respectively?
>
> RQ2: how do some alternative strategies used to create gold standards compare with the typical union and majority vote strategy?
>
> RQ3: what metrics can be used to better manage the trade-off between more annotated texts and more annotations for the same text?

To address these research questions, we conducted experiments using the OPP-115 corpora (Anaraky et al, 2019) with annotated text prepared through majority vote, union methods, and alternative strategies (individual, pairwise, and complete agreement). We tested how these strategies impact model performance using two traditional supervised learning algorithms—support vector machines (SVM) and Naïve Bayes (NB)—and two deep learning models, bidirectional long short-term memory (BiLSTM) and bidirectional encoder representations from transformers (BERT).

Dataset

The OPP-115 Corpus comprises 13,209 sentences and the number of sentences in each target class are highly imbalanced (as shown in Figure 1). For example, there are only 6, 39, and 63 sentences with complete agreement in the Do Not Track, Data retention, and Policy Change categories respectively. The OPP-115 corpora were prepared by ten independent annotators who coded privacy policy text segments using several predefined categories. Each privacy segment was annotated by three independent annotators. We arranged the unique annotator IDs in each statement in ascending order and replaced them with annotators A, B, and C respectively. By our proposed alternative strategies, the gold standards used individual annotators (A or B or C), where 2 annotators reached consensus (A and B, or A and C or B and C) and where there was complete agreement (A and B and C agree on the annotations). We also replicated the strategy used in the original work where a text is deemed relevant when at least 2 out of 3 annotators (majority vote) reached agreement, or all the 3 annotators (union).

	Orig.	Independent			Pairwise			Complete	Majority	Union
Category	Kappa	A	B	C	AB	AC	BC	ABC		
Do not track	0.91	22	27	20	10	11	11	6	20	43
International and specific audiences	0.87	643	556	484	482	406	367	338	588	778
First party collection/use	0.76	2249	2297	2297	1512	1579	1575	1266	2156	3489
Third party sharing/collection	0.76	1614	1702	1646	1100	1098	1109	873	1600	2590
User access, edit and deletion	0.74	198	215	226	129	132	134	109	178	355
Policy change	0.73	189	160	185	84	99	91	63	150	326
Data security	0.67	291	373	289	175	163	186	128	271	563
User choice/control	0.61	580	649	619	307	300	325	224	485	1154
Data retention	0.55	105	118	135	48	61	55	39	86	233
Other	0.49	2183	1393	1863	1013	1248	882	758	1646	3087

Figure 1. Number of sentences in each gold standard

Annotations created in the original OPP-115 data were segmented text into paragraphs (Anaraky et al, 2019). In contrast, the unit of analysis in our experiments is a sentence, thus both the original text and the annotations were converted into sentences using version 4.1.2 of LingPipe

(Alabduljabbar et al., 2021) and the index position of each sentence was maintained and subsequently aligned with the index position of the manual annotations. Sentences were pre-processed using the NLTK Python package (Prabhakaran et al., 2021), including: converting words to lowercases, and removing punctuation and stop words. Terms appearing infrequently (less than 5 sentences) and very frequently (more than 95% of the sentences) were removed because their presence would contribute little to the classification performance (Amos et al., 2021). Annotation categories are not mutually exclusive, a sentence can be annotated as belonging to more than one multiple categories. Lastly, we framed the problem as a binary text classification task for each of the annotation categories.

Text classification

The classification experiments used two algorithm families: traditional models (SVM, NB) and deep learning models (BiLSTM, BERT). Ten-fold cross-validation evaluated each model, splitting the dataset into ten equal parts, with nine used for training and one for testing. The data was stratified to maintain a balanced proportion of positive and negative labels in each fold. Model performance was measured using standard metrics: precision, recall, F-1, and accuracy (Hamdani et al., 2021). For traditional models, feature selection was crucial. We used version 1.0.2 of Scikit-learn (Gordon, et al., 2022) and entropy-based selection, calculating information gain to choose the top 2000 features (Amos et al., 2021; Pedregosa et al., 2011). These features were then used to construct the sentence-term matrix for the test set. TF-IDF was considered but not used due to its inability to account for target class distribution (Srinath et al., 2021).

Results

We report our findings regarding how much annotator agreement impacts the performance of automated approaches.

Alternative methods: independent, pairwise, complete agreement

We found that increasing the level of agreement from independent to pairwise to complete improved F-1, accuracy, precision, and recall across nearly all categories and classifiers (Figure 2). Complete agreement yielded the best F1 scores for several categories, including First Party Collection/Use and Third-Party Sharing/Collection. For Third Party Sharing/Collection, precision improved by 5% and recall by 4%. First Party Collection/Use saw a 7% improvement in precision and 8% in recall. The '*Other*' category showed significant gains, with F1 improving from 0.82 to 0.91.

In some categories, like User Choice/Control and International and Specific Audiences, pairwise agreement performed as well as complete agreement, with recall and accuracy reaching 0.97. Categories with fewer examples, such as Data Retention, had high metrics but raised concerns about generalizability.

Certain classifiers handled data inconsistencies better. For Third Party Sharing/Collection, precision improved from 0.85 (independent) to 0.91 (complete agreement) across all classifiers. The difference in model performance was minor (within 0.02), but the impact of different gold standards was more significant, ranging from 0.03 to 0.06 (Figures 2 and 3). Third Party Sharing/Collection also showed notable improvements in F-1, precision, recall, and accuracy, with precision and F1 for First Party Collection/Use improving by at least 0.07.

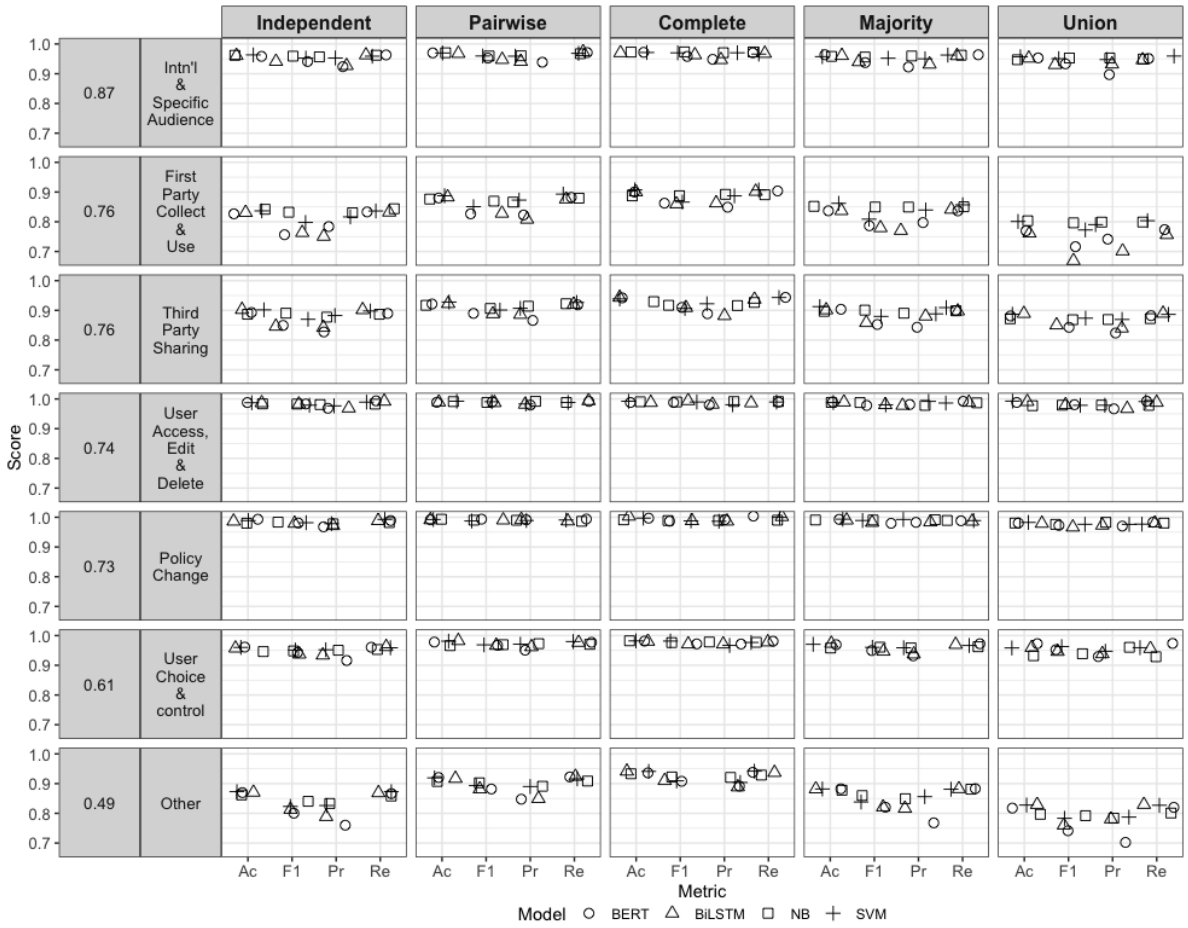

Figure 2. Average model performance for increasing levels of consensus (independent, pairwise, and complete agreement) and increasing levels of disagreement (majority vote, union) gold standards *(Ac=accuracy, F1, Pr=precision, and Re=recall)

Category	Sents	Model	Accuracy			Recall			Precision			F1		
			Complete	Union	Δ	Complete	Union	Δ	Complete	Union	Δ	Complete	Union	Δ
International and specific audiences		BERT	0.97	0.95	0.02	0.97	0.95	0.02	0.95	0.90	0.05	0.96	0.93	0.03
	Complete 338	BiLSTM	0.97	0.95	0.02	0.97	0.95	0.02	0.95	0.93	0.02	0.96	0.93	0.03
	Union 778	NB	0.97	0.95	0.02	0.97	0.95	0.02	0.97	0.95	0.02	0.97	0.95	0.02
		SVM	0.97	0.96	0.01	0.97	0.96	0.01	0.97	0.95	0.02	0.97	0.95	0.02
First party collection/use		BERT	0.90	0.77	0.13	0.90	0.77	0.13	0.85	0.74	0.11	0.86	0.72	0.14
	Complete 1266	BiLSTM	0.90	0.76	0.14	0.90	0.76	0.14	0.86	0.70	0.16	0.86	0.67	0.19
	Union 3489	NB	0.89	0.80	0.09	0.89	0.80	0.09	0.89	0.80	0.09	0.89	0.80	0.09
		SVM	0.91	0.80	0.11	0.91	0.80	0.11	0.89	0.79	0.10	0.87	0.77	0.10
Third party sharing/collection		BERT	0.94	0.88	0.06	0.94	0.88	0.06	0.89	0.82	0.07	0.91	0.84	0.07
	Complete 873	BiLSTM	0.94	0.89	0.05	0.94	0.89	0.05	0.88	0.84	0.04	0.91	0.85	0.06
	Union 2590	NB	0.93	0.87	0.06	0.93	0.87	0.06	0.92	0.87	0.05	0.92	0.87	0.05
		SVM	0.94	0.89	0.05	0.94	0.89	0.05	0.92	0.87	0.05	0.91	0.87	0.04
User access, edit and deletion		BERT	0.99	0.99	0.00	0.99	0.99	0.00	0.98	0.97	0.01	0.99	0.98	0.01
	Complete 109	BiLSTM	0.99	0.99	0.00	0.99	0.99	0.00	0.98	0.97	0.01	0.99	0.98	0.01
	Union 355	NB	0.99	0.98	0.01	0.99	0.98	0.01	0.99	0.98	0.01	0.99	0.98	0.01
		SVM	0.99	0.99	0.00	0.99	0.99	0.00	0.98	0.98	0.00	0.99	0.98	0.01
Policy change		BERT	1.00	0.98	0.02	1.00	0.98	0.02	0.99	0.97	0.02	0.99	0.97	0.02
	Complete 63	BiLSTM	1.00	0.98	0.02	1.00	0.98	0.02	0.99	0.97	0.02	0.99	0.97	0.02
	Union 326	NB	0.99	0.98	0.01	0.99	0.98	0.01	0.99	0.98	0.01	0.99	0.98	0.01
		SVM	1.00	0.98	0.02	1.00	0.98	0.02	0.99	0.98	0.01	0.99	0.98	0.01
User choice/control		BERT	0.98	0.97	0.01	0.98	0.97	0.01	0.97	0.93	0.04	0.97	0.95	0.02
	Complete 224	BiLSTM	0.98	0.96	0.02	0.98	0.96	0.02	0.97	0.94	0.03	0.97	0.95	0.02
	Union 1154	NB	0.98	0.93	0.05	0.98	0.93	0.05	0.98	0.96	0.02	0.98	0.94	0.04
		SVM	0.98	0.96	0.02	0.98	0.96	0.02	0.97	0.95	0.02	0.98	0.96	0.02
Other		BERT	0.94	0.82	0.12	0.94	0.82	0.12	0.89	0.70	0.19	0.91	0.74	0.17
	Complete 758	BiLSTM	0.94	0.83	0.11	0.94	0.83	0.11	0.89	0.78	0.11	0.91	0.76	0.15
	Union 3087	NB	0.93	0.80	0.13	0.93	0.80	0.13	0.92	0.78	0.14	0.92	0.79	0.13
		SVM	0.94	0.83	0.11	0.94	0.83	0.11	0.90	0.79	0.11	0.91	0.78	0.13

Figure 3. Average model performance for complete and union gold standards

Typical gold standard methods: union and complete

In general, we also found increasing the level of agreement improves model performance: the majority vote results outperform the union across all metrics and all classifiers, in all categories. Figure 4 shows the original Fleiss' Kappa statistic against the F1 score for complete and union gold standards. The range of Kappa in the OPP-115 collection ranges from moderate to very good, so it is possible that larger variations might show a correlation with F1, but these results suggest that: (a) Kappa is not a good substitute for F1 scores produced using different gold standards; (b) the difference between the standard metrics produced from complete and union experiments might provide a more realistic way to convey the impact of disagreement.

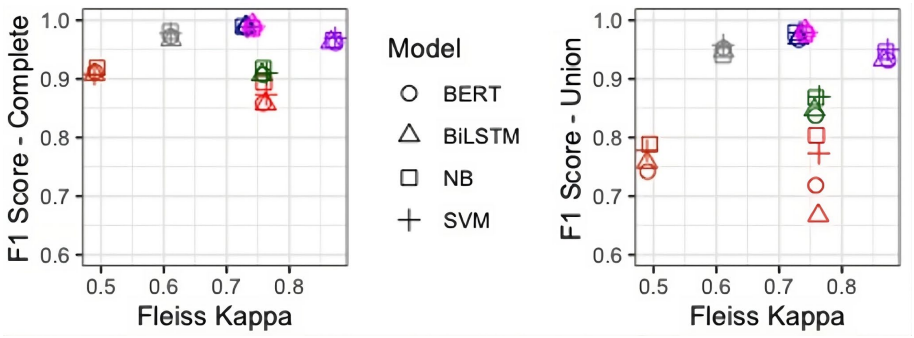

Figure 4. Fleiss Kappa versus F1-Score for complete and union gold standards

Discussion

Common methods for reporting machine learning results often focus on the model and its tuning parameters, neglecting variations in human-labelled training data. In privacy documentation,

diverse terminology and interpretations make it essential to consider human discrepancies in machine learning analysis. Traditional metrics don't always reflect the Fleiss Kappa statistic or the impact of human consensus. Based on our research, we recommend:

1. **Use multiple annotators:** for at least a subset of the corpus, multiple annotators should be employed to capture human variability. Reducing the number of categories per annotator could enhance annotation quality and efficiency.
2. **Iterative annotation:** apply an iterative approach, directing annotators to categories with less agreement. This can optimise human effort and maintain annotation quality, using complete and union gold standards as metrics.
3. **Differentiate gold standards:** use the difference between complete and union gold standards to assess how well machine learning results align with human judgments. This can provide insight into the model's accuracy in mirroring human interpretations.
4. **Avoid aggregation:** report results separately for each category rather than aggregating them. Specific categories may have varying expectations, and detailed performance information is crucial for understanding model effectiveness in real-world applications.

Conclusion

Privacy policy statements are vital for regulatory compliance and user data decisions, yet many are unreadable and often ignored. Machine learning could help by automating information extraction, but current reporting practices that don't align with human judgement undermine trust. Unlike previous methods using Fleiss' Kappa to measure disagreement, we propose an approach that uses independent, pairwise, and complete agreement in gold standards. We acknowledge that our data set is limited as it used only the OPP-115 corpus which may limit the generalizability of this study. Given this limitation, our preliminary results show that higher agreement improves precision, recall, F1, and accuracy, while more disagreement reduces these metrics.

Disagreements in privacy statement interpretation are more complex than fact-based tasks, and inter-rater reliability alone may not suffice to measure model performance. Traditional metrics like Cohen's or Fleiss Kappa are inadequate for skewed data. We suggest using precision, recall, F1, and accuracy to evaluate how different gold standards affect performance, which is crucial given the evolving nature of privacy content.

With new collections of privacy statements surpassing a million entries (Bannihatti Kumar et al., 2020; Thorleiksdóttir et al., 2022), investing in annotation adjudication is urgent. Quality assurance involves decisions such as storing multiple annotators' data and measuring their agreement (Mousavi Nejad et al., 2020). Our study highlights the need for multiple annotators on subsets of texts to assess the impact of human judgement on metrics. This may conflict with current practices aimed at maximising annotated data but is essential for realistic metric representation. An iterative approach can help allocate resources effectively, and text classification results should be reported by category rather than aggregated.

Our study is the first to explicitly raise the question of disagreements of annotating privacy policies documents and highlight the value and significance of studying such disagreements. How exactly disagreement in understanding and/or interpreting privacy policies can be leveraged remains unexplored, and future research needs to understand where and why disagreements occur when it comes to privacy policy interpretation, as there can be potentially disagreements for completely different reasons and hence require different treatment or solution. For example, there may be disagreement that originates from a lack of knowledge, linguistic ambiguity, or underlying differences in preference, each would require completely different solutions.

When marking up raw text, annotators need the flexibility to decide the appropriate text boundaries that capture the target category. Before the initial annotations can be used to construct a classifier, the unit of analysis, such as a paragraph, sentence (used in this analysis) or some other predefined '*span*' must be established. This choice impacts the predictive performance of any model constructed, more work is needed to establish what span is optimal for a given task and to quantify the impact of this decision. We have introduced a new performance metric – the difference between complete and union gold standards - that directly measures the impact of human agreement using the same metrics that are commonly used to measure an automated system. However, situated empirical user studies are needed to establish if this new metric is successful in making machine-learning models more transparent.

Acknowledgements

We would like to thank all the reviewers for the feedback on this paper. There is no funding for this research to report.

About the authors

Tian Wang is Postdoctoral Associate at the CyLab Security and Privacy Institute of Carnegie Mellon University. She received their Ph.D. from University of Illinois at Urbana Champaign, and her research interests are in mobile and app security and privacy. She can be contacted at tianwan2@andrew.cmu.edu

Yuanye Ma is Senior Research Associate at University of Illinois Discovery Partners Institute. She received their Ph.D. from University of North Carolina at Chapel Hill, and her research interests are in user-centered privacy, natural language processing and information ethics. She can be contacted at yuanyem@uillinois.edu

Catherine Blake is Professor and Associate Dean for Academic Affairs in School of Information Sciences, University of Illinois at Urbana Champaign. She received her PhD from University of California, Irvine. Her research interests are in biomedical informatics, natural language processing, evidence-based discovery, learning health systems, socio-technical systems, data analytics, literature-based discovery. She can be contacted at clblake@illinois.edu

Masooda Bashir is Associate Professor in School of Information Sciences, University of Illinois at Urbana Champaign. She received their Ph.D. from Purdue University, and her research interests are in the interface of information technology, human psychology, and society; especially how privacy, security, and trust intersect from a psychological point of view with information systems. She can be contacted at mnb@illinois.edu

Ryan Wang is a PhD student in the School of Information Sciences at the University of Illinois Urbana-Champaign. He is interested in natural language processing, machine learning, and bioinformatics. He can be reached at hywang3@illinois.edu.

References

Alabduljabbar, A., Abusnaina, A., Meteriz-Yildiran, Ü., & Mohaisen, D. (2021). TLDR: Deep Learning-Based Automated Privacy Policy Annotation with Key Policy Highlights. Proceedings of the 20th Workshop on Workshop on Privacy in the Electronic Society, 103–118. https://doi.org/10.1145/3463676.3485608

Anaraky, R. G., Cherry, D., Jarrell, M., & Knijnenburg, B. (2019). Testing a comic-based privacy policy. In *The 15th Symp. on Usable Privacy and Security*.

Artstein, R. (2017). Inter-annotator agreement. *Handbook of linguistic annotation*, 297-313.

Azhagusundari, B., & Thanamani, A. S. (2013). Feature selection based on information gain. *International Journal of Innovative Technology and Exploring Engineering* (IJITEE), 2(2), 18-21.

Bai, F., Ritter, A., & Xu, W. (2021). Pre-train or annotate? domain adaptation with a constrained budget. *arXiv preprint arXiv*:2109.04711.

Bannihatti Kumar, V., Iyengar, R., Nisal, N., Feng, Y., Habib, H., Story, P., Cherivirala, S., Hagan, M., Cranor, L., Wilson, S., Schaub, F., & Sadeh, N. (2020). Finding a Choice in a Haystack: Automatic Extraction of Opt-Out Statements from Privacy Policy Text. *Proceedings of The Web Conference 2020*, 1943–1954. https://doi.org/10.1145/3366423.3380262

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analysing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Braun, D. (2024). I beg to differ: How disagreement is handled in the annotation of legal machine learning data sets. *Artificial Intelligence and Law*, 32(3), 839–862. https://doi.org/10.1007/s10506-023-09369-4

Chen, R., Fang, F., Norton, T., McDonald, A. M., & Sadeh, N. (2021). Fighting the Fog: Evaluating the Clarity of Privacy Disclosures in the Age of CCPA. *Proceedings of the 20th Workshop on Workshop on Privacy in the Electronic Society*, 73–102. https://doi.org/10.1145/3463676.3485601

Gordon, M. L., Lam, M. S., Park, J. S., Patel, K., Hancock, J., Hashimoto, T., & Bernstein, M. S. (2022, April). Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1-19).

Gray, R. M. (2011). *Entropy and information theory*. Springer Science & Business Media.

Grosman, J. S., Furtado, P. H. T., Rodrigues, A. M. B., Schardong, G. G., Barbosa, S. D. J., & Lopes, H. C. V. (2020). Eras: Improving the quality control in the annotation process for Natural Language Processing tasks. *Information Systems*, 93, 101553. https://doi.org/10.1016/j.is.2020.101553

Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology*, 8(1), 23–34.

Hamdani, R. E., Mustapha, M., Amariles, D. R., Troussel, A., Meeùs, S., & Krasnashchok, K. (2021). A combined rule-based and machine learning approach for automated GDPR compliance checking. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, 40–49. https://doi.org/10.1145/3462757.3466081

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. *Proceedings of the 27th USENIX Conference on Security Symposium*, 531–548.

Herbert, F., Becker, S., Schaewitz, L., Hielscher, J., Kowalewski, M., Sasse, A., Acar, Y., & Dürmuth, M. (2023). A World Full of Privacy and Security (Mis)conceptions? Findings of a Representative Survey in 12 Countries. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–23. https://doi.org/10.1145/3544548.3581410

Hershcovich, D., Frank, S., Lent, H., de Lhoneux, M., Abdou, M., Brandl, S., Bugliarello, E., Cabello Piqueras, L., Chalkidis, I., Cui, R., Fierro, C., Margatina, K., Rust, P., & Søgaard, A. (2022).

Challenges and Strategies in Cross-Cultural NLP. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 6997–7013). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.482

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2), 1.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

Li, J., Liu, L., Le, T. D., & Liu, J. (2020). Accurate data-driven prediction does not mean high reproducibility. *Nature Machine Intelligence*, 2(1), 13–15. https://doi.org/10.1038/s42256-019-0140-2

LingPipe Alias-i. (2008). 4.1. 0. URL http://alias-i.com/lingpipe (2008).

Liu, F., Wilson, S., Schaub, F., & Sadeh, N. (2016). Analyzing vocabulary intersections of expert annotations and topic models for data practices in privacy policies. In 2016 AAAI Fall Symposium Series.

Moallem, A. (2018). Do You Really Trust "Privacy Policy" or "Terms of Use" Agreements Without Reading Them? In D. Nicholson (Ed.), *Advances in Human Factors in Cybersecurity* (pp. 290–295). Springer International Publishing. https://doi.org/10.1007/978-3-319-60585-2_27

Mousavi Nejad, N., Jabat, P., Nedelchev, R., Scerri, S., & Graux, D. (2020). Establishing a Strong Baseline for Privacy Policy Classification. In M. Hölbl, K. Rannenberg, & T. Welzer (Eds.), *ICT Systems Security and Privacy Protection* (pp. 370–383). Springer International Publishing. https://doi.org/10.1007/978-3-030-58201-2_25

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in python journal of machine learning research. *Journal of machine learning research*, 12, 2825-2830.

Pepperberg, I. M. (1988). An interactive modeling technique for acquisition of communication skills: Separation of "labeling" and "requesting" in a psittacine subject. *Applied Psycholinguistics*, 9(1), 59–76. https://doi.org/10.1017/S014271640000045X

Plank, B. (2022). *The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation* (arXiv:2211.02570). arXiv. https://doi.org/10.48550/arXiv.2211.02570

Prabhakaran, V., Davani, A. M., & Díaz, M. (2021). *On Releasing Annotator-Level Labels and Information in Datasets* (arXiv:2110.05699). arXiv. https://doi.org/10.48550/arXiv.2110.05699

Reidenberg, J. R., Breaux, T., Cranor, L. F., French, B., Grannis, A., Graves, J., Liu, F., McDonald, A., Norton, T., Ramanath, R., Russell, N. C., Sadeh, N., & Schaub, F. (2014). *Disagreeable Privacy Policies: Mismatches between Meaning and Users' Understanding* (SSRN Scholarly Paper 2418297). https://doi.org/10.2139/ssrn.2418297

Mysore Sathyendra, K., Wilson, S., Schaub, F., Zimmeck, S., & Sadeh, N. (2017). Identifying the Provision of Choices in Privacy Policy Text. In M. Palmer, R. Hwa, & S. Riedel (Eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2774–2779). Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1294

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. https://doi.org/10.18653/v1/2021.acl-long.532

Stevens, L. M., Mortazavi, B. J., Deo, R. C., Curtis, L., & Kao, D. P. (2020). Recommendations for Reporting Machine Learning Analyses in Clinical Research. *Circulation. Cardiovascular Quality and Outcomes*, *13*(10), e006556. https://doi.org/10.1161/CIRCOUTCOMES.120.006556

Tang, J., Shoemaker, H., Lerner, A., & Birrell, E. (2021). Defining privacy: How users interpret technical terms in privacy policies. *Proceedings on Privacy Enhancing Technologies*.

Thorleiksdóttir, T., Renggli, C., Hollenstein, N., & Zhang, C. (2022). Dynamic Human Evaluation for Relative Model Comparisons. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 5946–5955). European Language Resources Association. https://aclanthology.org/2022.lrec-1.639

Wilson, S., Schaub, F., Dara, A. A., Liu, F., Cherivirala, S., Giovanni Leon, P., Schaarup Andersen, M., Zimmeck, S., Sathyendra, K. M., Russell, N. C., Norton, T. B., Hovy, E., Reidenberg, J., & Sadeh, N. (2016). The Creation and Analysis of a Website Privacy Policy Corpus. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1330–1340). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1126

Yang, Y., & Pedersen, J. O. (1997, July). A comparative study on feature selection in text categorization. In *Icml* (Vol. 97, No. 412-420, p. 35).

Zaeem, R. N., German, R. L., & Barber, K. S. (2018). PrivacyCheck: Automatic Summarization of Privacy Policies Using Data Mining. *ACM Trans. Internet Technol.*, *18*(4), 53:1-53:18. https://doi.org/10.1145/3127519