The University of Texas at Austin
Center for Identity

PrivacyCheck v2: A Tool that Recaps Privacy Policies for You

Razieh Nokhbeh Zaeem
Safa Anya
Alex Issa
Jake Nimergood
Isabelle Rogers
Vinay Shah
Ayush Srivastava
K. Suzanne Barber

UTCID Report #20-08

June 2020

PrivacyCheck v2: A Tool that Recaps Privacy Policies for You

Razieh Nokhbeh Zaeem
Center for Identity
The University of Texas at Austin
razieh@identity.utexas.edu

Safa Anya*
Alex Issa*
Jake Nimergood*
Isabelle Rogers*
Vinay Shah*
Ayush Srivastava*
The University of Texas at Austin

K. Suzanne Barber
Center for Identity
The University of Texas at Austin
sbarber@identity.utexas.edu

ABSTRACT

Despite the efforts to regulate privacy policies to protect user privacy, these policies remain lengthy and hard to comprehend. Powered by machine learning, our publicly available browser extension, PrivacyCheck v2, automatically summarizes any privacy policy by answering 20 questions based upon User Control and the General Data Protection Regulation. Furthermore, PrivacyCheck v2 incorporates a competitor analysis tool that highlights the top competitors with the best privacy policies in the same market sector. PrivacyCheck v2 enhances the users' understanding of privacy policies and empowers them to make informed decisions when it comes to selecting services with better privacy policies.

CCS CONCEPTS

• **Security and privacy → Usability in security and privacy**; • **Computing methodologies → Supervised learning by classification**; • **Social and professional topics → Privacy policies**.

KEYWORDS

privacy policy; machine learning; GDPR; PrivacyCheck

ACM Reference Format:
Razieh Nokhbeh Zaeem, Safa Anya, Alex Issa, Jake Nimergood, Isabelle Rogers, Vinay Shah, Ayush Srivastava, and K. Suzanne Barber. 2020. PrivacyCheck v2: A Tool that Recaps Privacy Policies for You. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3340531.3417469

1 INTRODUCTION

The collection, sharing, and usage of Personally Identifiable Information (PII) over the Internet has become a major privacy concern. It is so essential to ensure lawfulness, transparency, and fairness of PII-related practices that new laws, such as the General Data Protection Regulation (GDPR)[1], are enforced around the globe to regulate privacy policies—the documents that govern these practices.

Privacy policies are legal documents that share how an organization collects, discloses, and uses a client's PII. Online privacy polices have grown into the de facto means of communicating privacy practices and are virtually ubiquitous on the Internet. Yet, we have long known that privacy policies are too long and hard to comprehend for their typical users [2, 4, 7].

To address the lack of readability in privacy policies, researchers have developed Privacy-Enhancing Technologies (PETs) that leverage Machine Learning (ML) and data mining to automatically summarize conventional privacy policies for the users. Among these PETs, however, very few are available as tools to the public.

We present PrivacyCheck v2—a novel, publicly available tool that uses ML to automatically recap a privacy policy. PrivacyCheck v2 is developed as an extension of the Google Chrome web browser and answers 20 questions concerting the privacy and security of users' PII according to the privacy policy: Ten *User Control* questions are based on the work of the Organization for Economic Cooperation and Development [9], and the Federal Trade Commission (FTC) Fair Information Practices (FIP) [3]; Another ten questions cover the most essential concerns addressed by the *GDPR*.

PrivacyCheck v2 is the second version of PrivacyCheck, our own machine learning framework to summarize privacy policies [18]. PrivacyCheck v2 adds to PrivacyCheck (1) a consumer-facing tool with higher performance, (2) new interface, (3) ten new questions aimed at the heart of the GDPR, and (4) the ability to find the top three competitors with better privacy policies (according to the User Control or GDPR standards) in the same market sector as of the privacy policy under evaluation. Finally, (5) we experimented with 13 new ML models to select the best combination of accuracy, precision, and recall for PrivacyCheck v2.

2 PRIVACYCHECK V2: THE TOOL

PrivacyCheck v2 is available online[2] and can also be found by searching for "PrivacyCheck" on the Google Chrome Web Store[3]. It currently has 850+ users and an average rating of 4.3/5 based on 12 reviews. A promotional video for our tool is available online too[4].

Figure 1 shows PrivacyCheck v2 when ran on a sample privacy policy. The user first navigates to a privacy policy using the Chrome

*These authors contributed equally to this research.

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3417469

[1] https://gdpr-info.eu
[2] https://tinyurl.com/ydf7h7dr
[3] https://chrome.google.com/webstore
[4] https://www.youtube.com/watch?v=QtQGMI7gSM4

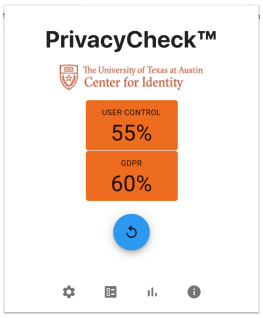

Figure 1: PrivacyCheck v2 run panel: User Control and GDPR scores of a sample privacy policy.

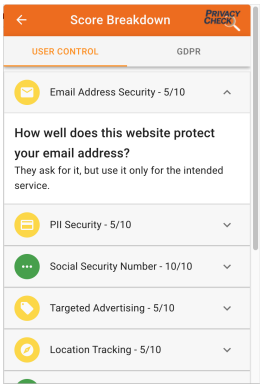

Figure 2: User Control score breakdown panel.

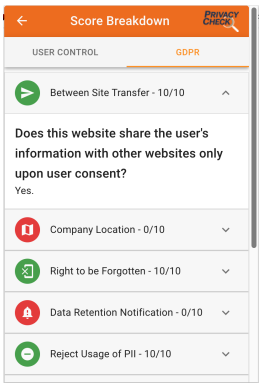

Figure 3: GDPR score breakdown panel.

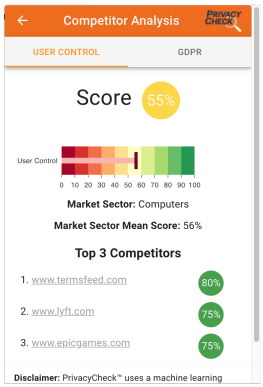

Figure 4: User Control CAT panel.

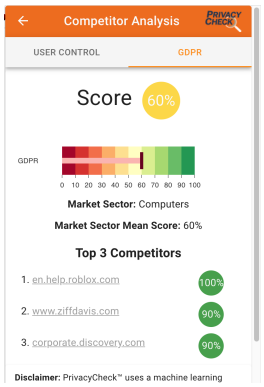

Figure 5: GDPR CAT panel.

browser and then opens PrivacyCheck v2 and clicks the run button. PrivacyCheck's machine learning models digest the privacy policy and assign two scores to it, one for the User Control and one for the GDPR standards. Clicking on each of the scores takes the user to score breakdowns explaining why the privacy policy received this score, based on the ten questions and their corresponding answers according to this privacy policy (Figures 2 and 3). Table 1 lists the User Control and GDPR questions. The interested reader can learn about how they were designed from our previous work [16, 18].

The goal of PrivacyCheck is to educate users on how their personal data is used on the Internet and to empower them to choose companies that better protect their data. Therefore, PrivacyCheck v2 adds the Competitor Analysis Tool (CAT): For each of the User Control and GDPR standards, PrivacyCheck v2 finds three other companies in the same market sector as of the policy under evaluation that have received the best scores from PrivacyCheck. PrivacyCheck v2 maintains a new back-end database of all policies it has ever been executed on and their market sectors, improving this database over time. Further, to put the score of the current policy in the context of its market sector, the CAT displays the market sector name, and the mean (average) score of privacy policies in that sector. Clicking on the graph icon at the bottom of the run panel (Figure 1) takes the user to the CAT panels of Figures 4 and 5.

3 PRIVACYCHECK V2: THE DESIGN

Figure 6 shows the high-level architecture of PrivacyCheck v2. The front-end client is the browser extension that sends the privacy policy URL to the back-end server through the REST API. The server runs (1) the machine learning classification models to calculate User Control and GDPR scores and (2) the CAT. It sends the results of these calculations back to the client to display.

3.1 PrivacyCheck v2 Front-End Client

We developed the front-end client of PrivacyCheck as a browser extension in ReactJS (a JavaScript framework). The final UI product consists of six components. Figure 7 depicts the data flow of the front-end client including references to the panels of Figures 1 to 5. The panel manager which controls the entire GUI provides data such as a global theme to all components and more specific data

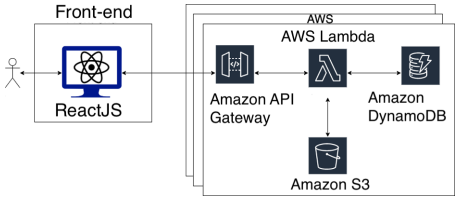

Figure 6: PrivacyCheck v2 system block diagram.

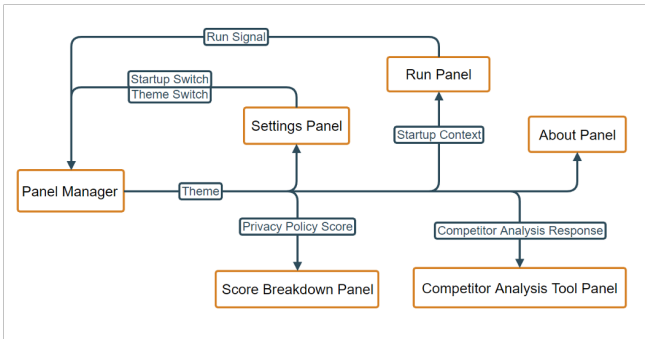

Figure 7: Data flow of the PrivacyCheck v2 client.

like the score and competitor analysis to the relevant components. Certain components like the run panel and settings panel callback to the PanelManager with settings or the signal to call the server.

ReactJS is a component-based framework that allows modularity and improves efficiency. With ReactJS, we used the Node Package Manager (NPM) that allows for easy dependency management of packages. One such package is the Material-UI we used for GUI design, an open-source project that features React components implementing Google's Material Design.

3.2 PrivacyCheck v2 Back-End Server

The server performs (1) the main task of scoring policies and (2) the competitor analysis. We host the server on Amazon Web Services (AWS) and manage it using its Serverless Framework—AWS Lambda.

Table 1: PrivacyCheck v2 questions.

	User Control	GDPR
1	How well does this website protect your email address?	Does this website share the user's information with other websites only upon user consent?
2	How well does this website protect your credit card information and address?	Does this website disclose where the company is based/user's PII will be processed & transferred?
3	How well does this website handle your social security number?	Does this website support the right to be forgotten?
4	Does this website use or share your PII for marketing purposes?	If they retain PII for legal purposes after the user's request to be forgotten, will they inform the user?
5	Does this website track or share your location?	Does this website allow the user the ability to reject usage of user's PII?
6	Does this website collect PII from children under 13?	Does this website restrict the use of PII of children under the age of 16?
7	Does this website share your information with law enforcement?	Does this website advise the user that their data is encrypted even while at rest?
8	Does this website notify or allow you to opt-out after changing their privacy policy?	Does this website ask for the user's informed consent to perform data processing?
9	Does this website allow you to edit or delete your information from its records?	Does this website implement all of the principles of data protection by design and by default?
10	Does this website collect or share aggregated data related to your identity or behavior?	Does this website notify the user of security breaches without undue delay?

Specifically, we utilized AWS Lambda to allow for parallel calls between the front-end and back-end. When the front-end sends a request to the server, the API Gateway invokes the Lambda.

There are two APIs that the front-end can call: database_get for the scores, and competitor_analysis for the CAT. The database_get API scores policies. The front-end calls database_get through the API Gateway with the privacy policy URL as a parameter. The server first checks the database to see if the scores for this URL have been calculated in the past 30 days. If a recent entry is found, the server immediately returns the entry for that privacy policy to the client, including all the User Control and GDPR scores, and the market sector. The 30 days is our indicator of freshness, as we recognize that privacy policies do not change often. If it has been more than 30 days since the run, or it is a brand new privacy policy, we decide to run our models. We have created functions to pre-process data, load in the respective ML model (GDPR, User Control, predict market sector), and run it. All of these models are called in parallel to improve efficiency. The database is then updated with the latest scores, and the entry is returned to the user. If PrivacyCheck v2 has analyzed a policy recently, it only takes 0.3s to return the data to the client, a 60x speedup from the old version of PrivacyCheck. If it has not, it takes 9.2s, which is a 1.9x speedup.

The competitor analysis API takes in the market sector as a parameter, and checks the database for all entries with that market sector. We created the database for the CAT (which stores company name, market sector, and policy scores, but no PII form PrivacyCheck v2 users) as a DynamoDB database. To improve search efficiency, we implemented the Global Secondary Index feature in DynamoDB. After getting all the entries, this API choses the top three for each of the GDPR and User Control, calculates the mean score for the market sector, and returns the results.

3.3 PrivacyCheck v2 Machine Learning Models

In this section we explain four functionalities performed at the server: (1) determining whether a URL is indeed a privacy policy, (2) finding its market sector for the CAT, and calculating (3) the User Control and (4) GDPR scores. We employed a regular expression to parse the URL in order to determine whether a website's text was a privacy policy, without actually training a ML model for it.

For the CAT API, we created a machine learning classifier that determines market sectors. We applied tf-idf pre-processing to the URL itself, and then trained a logistic regression model on the DMOZ database. DMOZ [10] is a publicly available dataset consisting of 1.5 million URL entries. The logistic regression model classifies a URL into 1 of 15 different market sectors. We had to compromise some of the accuracy of this classifier for it to fit in the AWS storage space, achieving an accuracy of 55%, based on using 70% of the dataset for training and 30% for testing.

Table 2: LightGBM metrics, averaged over the ten questions.

Standard	Accuracy	F1 Weighted	Recall Weighted	Precision Weighted
GDPR	0.60	0.55	0.60	0.53
User Control	0.60	0.57	0.60	0.58

Finally, to obtain User Control and GDPR scores, we pre-process the policy text, and apply ten trained classifier models for the User Control and ten trained classifiers for the GDPR questions.

The text pre-processing includes: (1) removing unnecessary whitespace, punctuation, and stop-words, (2) making all characters lowercase, and (3) counting words and vectorizing word frequencies as model input. No further feature extraction was performed.

The training dataset for all of the 20 questions contained 400 privacy policies, a random 10% sample of the NYSE, Nasdaq, and AMEX stock markets. More details can be found in our previous work for the User Control [18] and GDPR [16] datasets.

In order to select the best ML models, we tested 13 different classifier models on each of the ten GDPR questions. These models were (1) XGBoost, (2) Light GBM, (3) Cat Boost, and the following models from *sklearn*[5]: (4) Gradient Boost, (5) Bagging, (6) AdaBoost, (7) Random Forest, (8) Decision Tree, (9) Extra Trees, (10) KNN, (11) Stochastic Gradient Descent, (12) Naive Bayes, and (13) SVM. We measured the accuracy and F1-score (incorporating precision and recall) of correctly answering the GDPR questions (by dividing to 70% training and 30% test datasets), and we came to two clear winners: LightGBM and CatBoost.

We selected LightGBM for two practical reasons: the final model size was smaller, and it was significantly quicker to train. Table 2 shows the metrics of the final models (with 70% training and 30% testing sets), averaged over the ten questions of each standard. At the end, the models were trained on the entire dataset for the production version of PrivacyCheck v2. These models also replace the Multinomial Naive Bayes models of the legacy version of PrivacyCheck for User Control.

4 RELATED WORK

Researchers have utilized various machine learning methodologies to automatically summarize privacy policies. Few *tools*, however,

[5] https://scikit-learn.org

WWW.IDENTITY.UTEXAS.EDU

are build on such research and made publicly available. We briefly cover closely related work, with an emphasis on the available tools.

To our knowledge, Privee [22] is the first tool to automatically analyze privacy policies. Building on the crowd sourcing privacy analysis framework ToS;DR [12], Privee uses machine learning to classify privacy policies that are not already rated in ToS;DR. As opposed to PrivacyCheck v2, Privee predates the GDPR and as a result does not target it.

Polisis [5] is a browser extension that utilizes deep learning to evaluate the PII collected and shared according to a privacy policy. Pribots [6] is a chat-bot from the same authors that answers free-form questions about privacy policies. The Usable Privacy Project takes advantage of natural language processing and machine learning to semi-automatically annotate privacy policies. This project annotates [13, 14] OPP-115 (a corpus of 115 policies with attributes and data practices), which is the corpus that Polisis uses. PrivacyCheck v2 is distinct from these projects—particularly with ten new questions aimed at the heart of the GDPR, and its novel CAT.

MAPS [23] analyzes privacy policies of more than one million mobile applications. PolicyLint [1] is a natural language processing tool that identifies potential contradictions that may arise inside the same privacy policy. Close to our work is PrivacyGuide [11], a machine learning and natural language processing tool inspired by the GDPR. However, PrivacyGuide is not publicly available.

At the Center for Identity at the University of Texas at Austin, we target many aspects of identity management and privacy [8, 17, 19–21]. We developed PrivacyCheck [18] and used it to study privacy policies across industries [15] and to quantify the effect of the GDPR on the landscape of privacy policies [16].

5 CONCLUSION

We presented our publicly available browser extension, PrivacyCheck v2, which summarizes privacy policies through machine learning. It answers 20 questions based on the User Control and GDPR standards. We experimented with 13 ML classifiers to pick the best combination of accuracy, precision, and recall. In addition, PrivacyCheck v2 adds a competitor analysis tool to provide context for the current privacy policy and empowers users to select services with stronger privacy policies. PrivacyCheck v2 provides a new sleek GUI and improves the response time significantly, thanks to expanded functionality like the new AWS Dynamo database, parallelized calls, and computationally inexpensive processes using AWS Lambda and API Gateway. AWS guarantees an almost 100% uptime, and its Serverless functions allow PrivacyCheck v2 to scale to a virtually unlimited number of users. A prominent future work is to gather a more comprehensive training dataset, in order to train more accurate ML models. We further envision a finer grade market sector classifier and the implementation of PrivacyCheck v2 as extensions to other browsers.

ACKNOWLEDGMENTS

This work was in part funded by the Center for Identity's Strategic Partners. The complete list of Partners can be found at https://identity.utexas.edu/strategic-partners.

REFERENCES

[1] Benjamin Andow, Samin Yaseer Mahmud, Wenyu Wang, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Tao Xie. 2019. Policylint: investigating internal privacy policy contradictions on Google play. In *28th USENIX Security Symposium*. 585–602.

[2] Benjamin Fabian, Tatiana Ermakova, and Tino Lentz. 2017. Large-scale readability analysis of privacy policies. In *Proceedings of the International Conference on Web Intelligence*. 18–25.

[3] FTC. 2000. Privacy Online: Fair Information Practices in the Electronic Marketplace: A Federal Trade Commission Report to Congress. https://www.ftc.gov/reports/privacy-online-fair-information-practices-electronic-marketplace-federal-trade-commission

[4] Mark A Graber, Donna M D Alessandro, and Jill Johnson-West. 2002. Reading level of privacy policies on internet health web sites. *Journal of Family Practice* 51, 7 (2002), 642–642.

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

[6] Hamza Harkous, Kassem Fawaz, Kang G Shin, and Karl Aberer. 2016. Pribots: Conversational privacy with chatbots. In *Twelfth Symposium on Usable Privacy and Security (SOUPS) 2016*.

[7] Aleecia M McDonald and Lorrie Faith Cranor. 2008. the Cost of reading privacy policies. *I/S: A Journal of Law and Policy for the Information Society* 4 (2008), 543.

[8] Rima Rana, Razieh Nokhbeh Zaeem, and K Suzanne Barber. 2019. An Assessment of Blockchain Identity Solutions: Minimizing Risk and Liability of Authentication. In *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. 26–33.

[9] Having Regard. 1980. Recommendation of the Council concerning Guidelines governing the Protection of Privacy and Transborder Flows of Personal Data. (1980).

[10] Ashadullah Shawon, Syed Tauhid Zuhori, Firoz Mahmud, and Md Jamil-Ur Rahman. 2018. Website Classification Using Word Based Multiple N-Gram Models and Random Search Oriented Feature Parameters. In *2018 21st International Conference of Computer and Information Technology (ICCIT)*. IEEE, 1–6.

[11] Welderufael B Tesfay, Peter Hofmann, Toru Nakamura, Shinsaku Kiyomoto, and Jetzabel Serna. 2018. PrivacyGuide: towards an implementation of the EU GDPR on internet privacy policy evaluation. In *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics*. ACM, 15–21.

[12] ToS;DR. 2012. Terms of Service; Didn't Read. https://tosdr.org

[13] Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N Cameron Russell, et al. 2016. The creation and analysis of a website privacy policy corpus. In *Annual Meeting of the Association for Computational Linguistics*. 1330–13340.

[14] Shomir Wilson, Florian Schaub, Rohan Ramanath, Norman Sadeh, Fei Liu, Noah A Smith, and Frederick Liu. 2016. Crowdsourcing Annotations for Websites' Privacy Policies: Can It Really Work?. In *Proceedings of the 25th International Conference on World Wide Web*. 133–143.

[15] Razieh Nokhbeh Zaeem and K Suzanne Barber. 2017. A study of web privacy policies across industries. *Journal of Information Privacy and Security* 13, 4 (2017), 169–185.

[16] Razieh Nokhbeh Zaeem and K Suzanne Barber. 2020. The Effect of the GDPR on Privacy Policies: Recent Progress and Future Promise. *ACM Transactions on Management of Information Systems* (2020).

[17] Razieh Nokhbeh Zaeem, Suratna Budalakoti, K Suzanne Barber, Muhibur Rasheed, and Chandrajit Bajaj. 2016. Predicting and explaining identity risk, exposure and cost using the ecosystem of identity attributes. In *2016 IEEE International Carnahan Conference on Security Technology (ICCST)*. IEEE, 1–8.

[18] Razieh Nokhbeh Zaeem, Rachel L German, and K Suzanne Barber. 2018. PrivacyCheck: Automatic Summarization of Privacy Policies Using Data Mining. *ACM Transactions on Internet Technology (TOIT)* 18, 4 (2018), 53.

[19] Razieh Nokhbeh Zaeem, Monisha Manoharan, and K Suzanne Barber. 2016. Risk kit: Highlighting vulnerable identity assets for specific age groups. In *2016 European Intelligence and Security Informatics Conference (EISIC)*. IEEE, 32–38.

[20] Razieh Nokhbeh Zaeem, Monisha Manoharan, Yongpeng Yang, and K Suzanne Barber. 2017. Modeling and analysis of identity threat behaviors through text mining of identity theft stories. *Computers & Security* 65 (2017), 50–63.

[21] Jim Zaiss, Razieh Nokhbeh Zaeem, and K Suzanne Barber. 2019. Identity Threat Assessment and Prediction. *Journal of Consumer Affairs* 53, 1 (2019), 58–70.

[22] Sebastian Zimmeck and Steven M. Bellovin. 2014. Privee: An Architecture for Automatically Analyzing Web Privacy Policies. In *23rd USENIX Security Symposium*. USENIX Association, San Diego, CA, 1–16.

[23] Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel Reidenberg, N Cameron Russell, and Norman Sadeh. 2019. MAPS: Scaling privacy compliance analysis to a million apps. *Proceedings on Privacy Enhancing Technologies* 2019, 3 (2019), 66–86.