The University of Texas at Austin
Center for Identity

PrivacyCheck: Automatic Summarization of Privacy Policies Using Data Mining

Razieh Nokhbeh Zaeem
Rachel L. German
K. Suzanne Barber

UTCID Report #1811

MAY 2018

PrivacyCheck: Automatic Summarization of Privacy Policies Using Data Mining

RAZIEH NOKHBEH ZAEEM, University of Texas at Austin
RACHEL L. GERMAN, University of Texas at Austin
K. SUZANNE BARBER, University of Texas at Austin

Prior research shows that only a tiny percentage of users actually read the online privacy policies they implicitly agree to while using a website. Prior research also suggests that users ignore privacy policies because these policies are lengthy and, on average, require two years of college education to comprehend. We propose a novel technique that tackles this problem by automatically extracting summaries of online privacy policies. We use data mining models to analyze the text of privacy policies and answer ten basic questions concerning the privacy and security of user data, what information is gathered from them, and how this information is used. In order to train the data mining models, we thoroughly study privacy policies of 400 companies (considering 10% of all listings on NYSE, Nasdaq, and AMEX stock markets) across industries. Our free Chrome browser extension, PrivacyCheck, utilizes the data mining models to summarize any HTML page that contains a privacy policy. PrivacyCheck stands out from currently available counterparts because it is readily applicable on *any* online privacy policy. Cross validation results show that PrivacyCheck summaries are accurate 40% to 73% of the time. Over 400 independent Chrome users are currently using PrivacyCheck.

Categories and Subject Descriptors: H.2.8 [**Database Applications**]: Data mining; K.4.1 [**Public Policy Issues**]: Privacy

General Terms: Legal Aspects

Additional Key Words and Phrases: Privacy policy, data mining, classification

ACM Reference Format:
Razieh Nokhbeh Zaeem, Rachel L. German, and K. Suzanne Barber, 2016. PrivacyCheck: Automatic Summarization of Privacy Policies Using Data Mining. *ACM Trans. Inter. Tech.* 9, 4, Article 39 (March 2010), 18 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

When consumers give personally identifiable information (PII) on the Internet, they often have no idea what companies will do with it. Federal and state laws require most businesses to publicly post a privacy policy stating how they use users' PII. The Federal Trade Commission (FTC) has successfully prosecuted companies for deceptive and misleading practices when using personal data in ways contrary to their stated privacy policies. As a result, most companies have privacy policies that are easily ac-

This work is supported by the State of Texas IDWise Project. 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.
Author's addresses: R. Nokhbeh Zaeem, R. L. German, and K. S. Barber, Center for Identity, University of Texas at Austin, {razieh, rachel, sbarber}@identity.utexas.edu.

DOI: http://dx.doi.org/10.1145/0000000.0000000

cessible online. The problem is that many people never read these lengthy and often technical documents.

Research shows that while most users know about privacy policies, less than half of them have *ever* read a privacy policy [Meinert et al. 2006]. Studies that used self-reported data from users found that only 4.5% claim to always read them [Milne and Culnan 2004]. However, the more reliable server side observation of websites reveals even more astonishing statistics that only 1% or less of users click on a website's privacy policy [Kohavi 2001]. More recent studies using advanced eye tracking techniques show that the same still holds true today: users barely take effort to read privacy policies thoroughly [Steinfeld 2016].

The fact that most users do not read privacy policies might be attributed to the privacy policies' poor readability [Ermakova et al. 2014]. A study of the readability of privacy policies showed that the average privacy policy required two years of college level education to comprehend [Graber et al. 2002; Milne et al. 2006]. An analysis of 80 privacy policies for top health websites found that none of the websites had a privacy policy that was comprehensible by most English-speaking individuals in the United States [Graber et al. 2002]. In addition, a study of online privacy policies found that privacy policies are getting longer and harder to read, with the readability score of the privacy policies decreasing over time [Milne et al. 2006]. In fact, reading privacy policies is so time consuming that if users were to read each new privacy policy they encounter in a year, it would take them over 200 hours [McDonald and Cranor 2008].

In order to improve users' ability to comprehend privacy policies, we developed PrivacyCheck [UT CID 2015], a valuable Privacy Enhancing Technology (PET) that gives users a quick and easily understood overview of the essential content of a company's online privacy policy. PrivacyCheck is a browser add-on that is intended to provide a graphical, at-a-glance summary of privacy policies. When the user provides the URL of the company's privacy policy page in the browser, PrivacyCheck utilizes a data mining algorithm to return icons that indicate the level of risk for several factors that impact the security and privacy of a user's identity.

PrivacyCheck summarizes a privacy policy with respect to a list of ten privacy factors (Section 2.1). For each of these factors, e.g., email address, PrivacyCheck answers a basic question, e.g., "How does the site handle your email address?". The answers to these questions are mapped to three levels of risk: red (high risk), yellow (medium risk), and green (low risk). For example, if a website does ask for users' email addresses, but states in the privacy policy that it uses them only for the intended service, it is ranked at the yellow risk level with respect to this PII factor. We interviewed privacy experts and utilized previous research to compile the list of factors.

PrivacyCheck automatically predicts risk values for each privacy factor using a classification data mining (supervised machine learning) model. Our key insight in developing PrivacyCheck was to train a data mining model against privacy policies for each factor, and then use it to predict risk values for the factor when checking new privacy policies. To train the models, a seven-person team of researchers, graduate and undergraduate students read 400 privacy policies randomly selected from the NYSE, Nasdaq and AMEX company listings, and manually assigned risk levels with respect to each of the ten factors.

In this paper, we present the following contributions:

— We use data mining to summarize online privacy policies.
— We present PrivacyCheck, a browser extension that readily extracts risk levels for privacy factors from privacy policies and shows them graphically.
— We carefully investigate 400 privacy policies (from 10% of all companies listed on the NYSE, Nasdaq, and AMEX stock markets) and use them to train data mining

models. The biggest corpus of this kind we could find in the literature includes just 115 privacy policies [Wilson et al. 2016b; Usable Privacy 2016].
— We evaluate PrivacyCheck using 50 other policies and show how it exceeds similar tools and certifications.

This paper is organized as follows. Section 2 explains the user interface and technical foundation of PrivacyCheck. Section 3 puts PrivacyCheck in context by surveying other available PET tools and services. Section 4 evaluates PrivacyCheck and compares it to the other tools discussed in Section 3. Finally, Section 5 outlines some of the use cases of PrivacyCheck and Section 6 concludes the paper.

2. PRIVACYCHECK BROWSER EXTENSION

PrivacyCheck seeks to use data mining to automatically summarize important factors discussed in privacy policies. To that end, it receives the privacy policy's URL from the user, pre-processes the policy's text, and sends the processed text to data mining servers. Then PrivacyCheck receives a privacy policy summary from the data mining server and displays it as colorful icons accompanied by short text snippets.

PrivacyCheck is currently implemented as a browser extension for Google Chrome and is publicly available [UT CID 2015]. Figure 1 shows a snapshot of the PrivacyCheck browser extension. The user first navigates to the URL of a privacy policy and then opens the browser extension and clicks its start button. PrivacyCheck extracts the text of the privacy policy, pre-processes it, and sends it to the data mining model to determine the level of risk for each of the ten privacy factors. PrivacyCheck then displays the risk levels as red (high risk), yellow (medium risk), and green (low risk), which are more elaborately explained once the user hovers over each item (as seen in Figure 1).

2.1. PrivacyCheck Questions

The ten questions that PrivacyCheck answers are:

(1) How does the site handle your email address?
(2) How does the site handle your credit card number and home address?
(3) How does the site handle your Social Security number?
(4) Does the site use or share your personally identifiable information for marketing purposes?
(5) Does the site track or share your location?
(6) Does the site collect personally identifiable information from children under 13?
(7) Does the site share your information with law enforcement?
(8) Does the site notify you or allow you to opt out when their privacy policy changes?
(9) Does the site allow you to edit or delete your information from its records?
(10) Does the site collect or share aggregated data related to your identity or behavior?

Table I shows the risk levels for each of the privacy factors. The red risk level is also assigned in case information about a given factor is not disclosed.

In order to choose the questions that PrivacyCheck seeks to answer, we evaluated related work and performed a survey.

2.1.1. Previous Work on Privacy Factors. The Organization for Economic Co-operation and Development (OECD) is one of the first to provide Guidelines on the Protection of Privacy, including eight privacy principles [Regard 1980]: Collection Limitation Principle, Data Quality Principle, Purpose Specification Principle, Use Limitation Principle, Security Safeguards Principle, Openness Principle, Individual Participation Principle, and Accountability Principle.

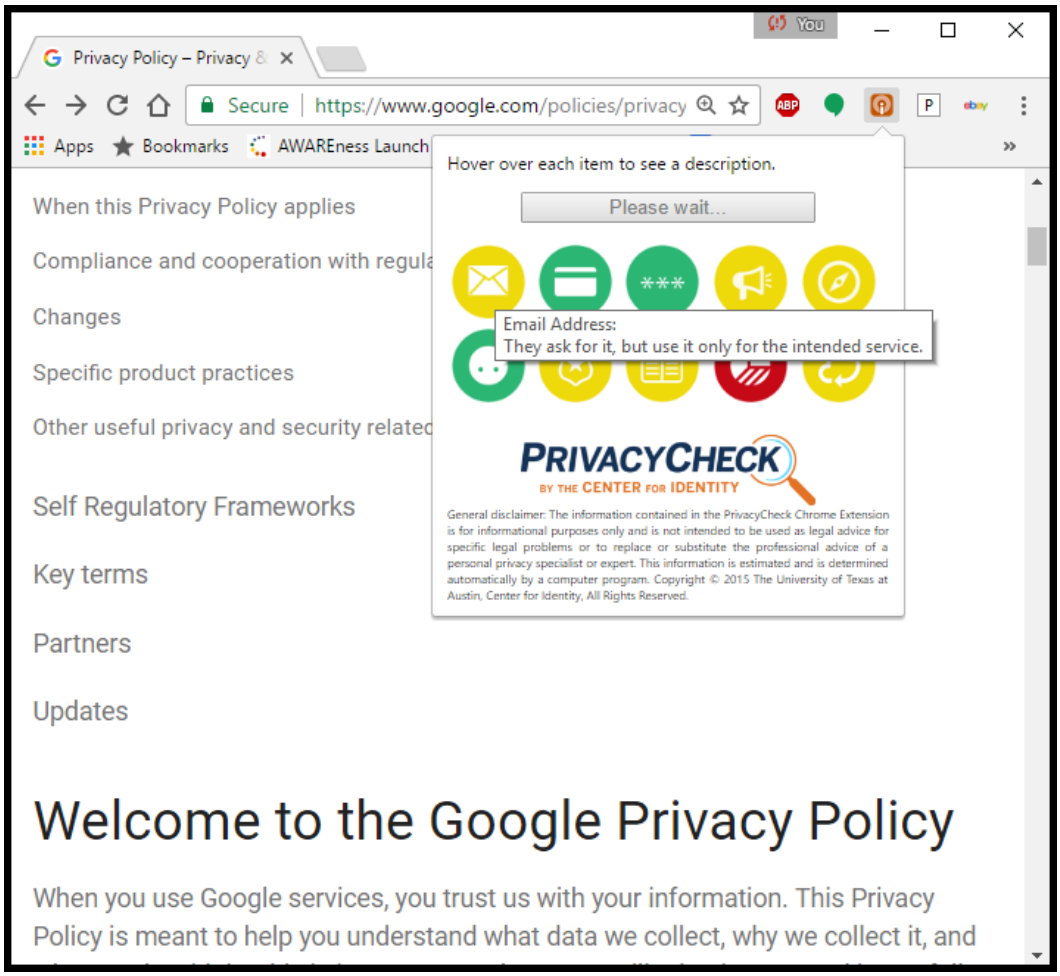

Fig. 1. A snapshot of the PrivacyCheck Chrome extension.

Table I. Risk level interpretations for privacy factors.

Factor	Green Risk Level	Yellow Risk Level	Red Risk Level
(1) Email Address	Not asked for	Used for the intended service	Shared w/ third parties
(2) Credit Card Number	Not asked for	Used for the intended service	Shared w/ third parties
(3) Social Security Number	Not asked for	Used for the intended service	Shared w/ third parties
(4) Ads and Marketing	PII not used for marketing	PII used for marketing	PII shared for marketing
(5) Location	Not tracked	Used for the intended service	Shared w/ third parties
(6) Collecting PII of Children	Not collected	Not mentioned	Collected
(7) Sharing w/ Law Enforcement	PII not recorded	Legal docs required	Legal docs not required
(8) Policy Change	Posted w/ opt out option	Posted w/o opt out option	Not posted
(9) Control of Data	Edit/delete	Edit only	No edit/delete
(10) Data Aggregation	Not aggregated	Aggregated w/o PII	Aggregated w/ PII

The FTC recommends that privacy policies follow Fair Information Practice Principles (FIPPs) [FTC 2000]: Notice, Choice, Access, Security, and Enforcement. We also reviewed public submissions and staff reports from several workshops and round-tables that the FTC held in 2010 and 2012 [FTC 2010; 2012] that suggested these privacy factors: Aggregation, Encryption, Third Party Sharing, Sharing with Law Enforcement, Security, Access, Control, Usage, Ads, Retention, and Location.

More recent work (Usable Privacy) [Wilson et al. 2016a] defines these categories for annotating privacy policies: First Party Collection/Use, Third Party Sharing/Collection, User Choice/Control, User Access/Edit/Deletion, Data Retention, Data Security, Policy Change, Do Not Track, and finally International/Specific Audiences.

We also looked up several online services like Disconnect Me Privacy Icons [Disconnect Me 2014] which includes: Expected Use, Expected Collection, Precise Location, Data Retention, Do Not Track, Children Privacy, SSL Support, Heartbleed, and TRUSTe Certification.

2.1.2. Factors Survey. Since it was not practical to include all the privacy factors we gathered from the literature, we interviewed privacy experts to identify factors that are most important when summarizing a privacy policy. We interviewed 16 full time employees and graduate students of the Center for Identity at UT Austin, who actively work in the field of privacy and security. The participants were asked to score each of the potential factors from 1 to 4. The full questionnaire is shown in Appendix A. Using the results of the survey, we enlisted the factors that the interviewees cared most about and designed PrivacyCheck to answer the questions about those factors. During the training phase, we carefully reviewed each of the selected privacy policies and manually assigned it answers to the privacy factor questions and the answers' corresponding risk levels (Section 2.3).

2.2. Architecture

Figure 2 shows a high level architecture of the PrivacyCheck extension. The browser client (written in HTML and Java-script) initially determines whether the given URL indeed points to a privacy policy. In order to do so, it follows the algorithm explained in Section 2.4 to get the related text. It then sends the related text to the data mining server. The data mining server checks the text against a trained classification model to find out if this is a privacy policy. The result of the classification model is sent back to the browser extension. The browser extension checks the result and, if not a privacy policy, alerts the user that the URL does not point to a privacy policy. If the URL is determined to be for a privacy policy, however, the browser extension follows the same algorithm (Section 2.4) to extract related paragraphs for each privacy factor. The related text snippets for privacy factors are asynchronously sent to the data mining server. The data mining server has a classification model trained for each factor (Section 2.5), which it uses to classify the text snippet according to the risk levels (high, medium, and low). Once the browser extension client receives the result for a factor, it shows it as red, yellow, or green for high risk, medium risk, or low risk, respectively.

2.3. Corpus of 400 Privacy Policies

We compiled a set of 400 privacy policies that we use for several purposes: collecting keywords (Section 2.4), designing answers and risk levels (Section 2.1), and training the classification models (Section 2.5).

A scientific methodology for selecting companies is central to collecting a comprehensive and generalizable corpus of their privacy policies. We aimed for a selection of companies that:

(1) Were reputable, i.e., listed by well known industrial entities.

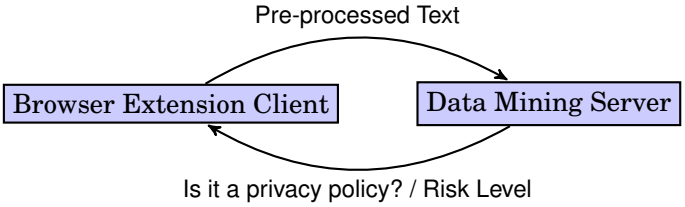

Fig. 2. High level architecture of the PrivacyCheck extension.

(2) Were categorized based on a standard and commonly used industrial classification.
(3) Covered a wide range of categories and industries across that classification.

To achieve a selection that meets these goals, we focused on the companies listed by the NYSE, Nasdaq, and AMEX stock markets (Section 2.3.1) using the Industry Classification Benchmark (Section 2.3.2).

2.3.1. NYSE, Nasdaq, and AMEX. The New York Stock Exchange (NYSE), Nasdaq, and the American Stock Exchange (AMEX) are American stock exchange markets, and are respectively the first, second, and third largest stock exchanges by market capitalization in the US. The Nasdaq Company List includes companies listed on Nasdaq, as well as NYSE and AMEX. As of this writing, the companies listed by these three stock markets add up to 6,349 worldwide, most of them (5,626 companies) in North America (the United States, Canada, and Mexico) [Nasdaq 2015].

We used the Nasdaq company list for training PrivacyCheck. The list uses the Industry Classification Benchmark and includes the company's name, the ICB industry to which it belongs, and a link to its website.

2.3.2. Industry Classification Benchmark. The Industry Classification Benchmark (ICB) is an industry classification taxonomy. It segregates markets into 10 industries which are in turn partitioned into 114 sub-sectors. Each company is allocated to a sub-sector that most closely resembles its majority source of revenue [ICB 2006]. Over 70,000 companies and 75,000 securities worldwide are categorized by ICB. ICB is used globally, including by NYSE, Nasdaq, and AMEX [ICB 2006].

2.3.3. Company Selection. The NYSE, Nasdaq, and AMEX collectively list a total of 6,349 companies as of the date of this paper. We randomly selected 10%, 635 companies, evenly across industries. Once we had the list of companies to study, we first found the URLs of their privacy policies. We reached the company's website using the link posted on the NYSE, Nasdaq or AMEX company list. If that link was broken, we performed a Google search with the company name, manually locating the company's website. To get to the privacy policy, we searched for the word "Privacy" on the English version of the company's homepage. If we could not find the privacy policy in this way we performed a Google search for "Privacy" only on the company's website (using the "site" advanced option of the search query). We then manually located the correct URL to the company's privacy policy. Out of these 635 companies, 12 companies did not have a website and another 175 companies did not have a privacy policy posted on their website. Finally, 48 companies shared the privacy policy of other companies that we had selected, because of sharing a parent company or a law firm. Therefore, we were able to locate 400 unique privacy policies. All 400 links were manually checked to make sure they point to the latest version of the company's privacy policy.

Table II. Keywords for privacy factors.

Factor	Keywords
(0) Being a Privacy Policy	privacy, policy
(1) Email Address	email, mail, third, party, share, sell, promote, affiliate
(2) Credit Card Number	credit, card, bill, debit, pay, third, party, share, sell, promote, affiliate
(3) Social Security Number	social, security, number, ssn, third, party, share, sell, promote, affiliate
(4) Ads and Marketing	ad, market, third, party, share, sell, promote, affiliate
(5) Location	locate, geo, mobile, gps, third, party, share, sell, promote, affiliate
(6) Collecting PII of Children	age, child
(7) Sharing w/ Law Enforcement	law, regulate, legal, government, warrant, subpoena, court, judge
(8) Policy Change	notice, change, update, post
(9) Control of Data	choice, edit, delete, limit, setting, account, access, update
(10) Data Aggregation	aggregate, non-identifiable

2.4. PrivacyCheck Client Side: Text Pre-processing

We utilized a text pre-processing algorithm to extract parts of privacy policies that are related to each of the privacy factors. The text pre-processing algorithm (shown in Appendix B) breaks the text into paragraphs, removes punctuation, converts uppercase to lowercase, removes stop words, performs stemming, and keeps only the paragraphs that have at least one keyword related to a particular factor. For instance, for the privacy factor Email Address, only those paragraphs that have at least one keyword related to Email Address are kept. Table II lists related keywords for each factor.

In Table II, the first row indicates keywords used to detect a privacy policy, and the subsequent rows indicate the keywords associated with the various privacy factors. These keywords are selected by considering, not only the privacy factor itself (e.g., Email Address), but also the questions we seek to answer about each factor to assign risk levels (e.g., whether email addresses are shared with third parties). Hence keywords like *third*, *party*, *share*, and *sell* are included for Email Address and other privacy factors that answer similar questions.

We investigated the 400 privacy policies in order to determine the list of keywords related to each of the privacy factors. The method used for selecting keywords was largely manual, and consisted of identifying the most frequently-used words in the paragraphs that seemed related to a given privacy factor. Investigating other methodologies and assessing their selection of such keywords is a future avenue of work.

2.5. PrivacyCheck Server Side: Data Mining Models

Once the text pre-processing is complete, PrivacyCheck sends each text snippet to the data mining server. We trained 11 data mining models, one for detecting if the corresponding page is a privacy policy and one for each of the ten privacy factors.

To train the models, we leveraged the 400 privacy policies. A team of seven researchers, graduate and undergraduate students read all of these policies, which totaled to close to 700K words, and scored each policy according to Table I, using the red/yellow/green risk levels. We performed quality control by assigning 15% of the privacy policies (60 policies, randomly selected) to two team members in order to train them to be consistent in assigning risk levels. It is important to note that the ground truth of how a company deals with users' PII is assumed to be what its privacy policy states. Matching the practice of the company with its privacy policy is beyond the scope of this paper.

Next, we applied the pre-processing algorithm to process the 400 policies and put together a training file for each factor. The training file includes the corresponding text snippet (containing only those paragraphs that have a related keyword) and the manually-determined risk level for each of the 400 policies. In order to train the model

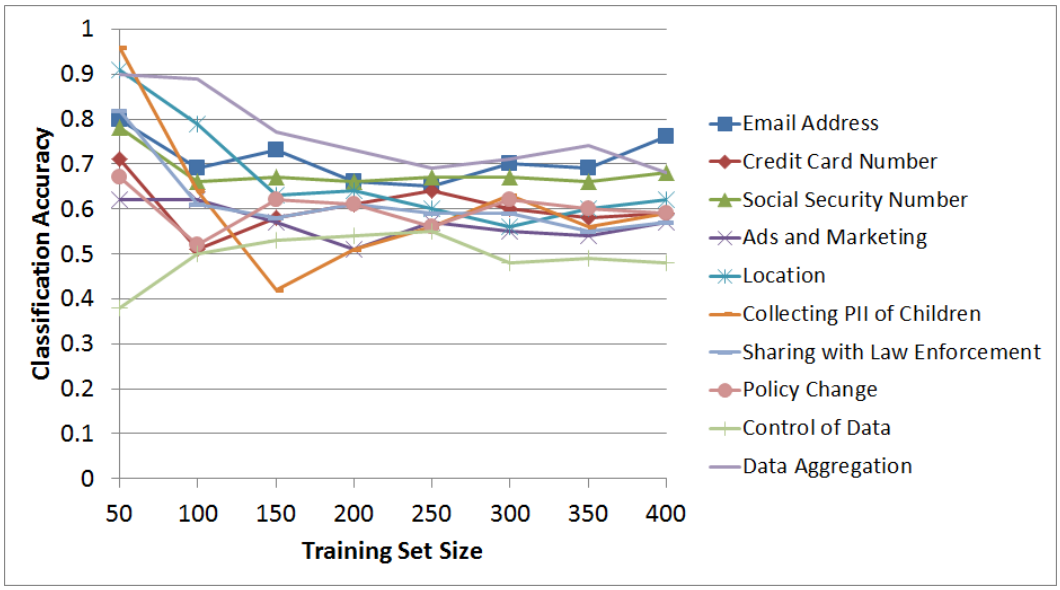

Fig. 3. Classification accuracy of models when trained against different numbers of privacy policies.

to determine whether or not the web page is indeed a privacy policy, we utilized the 400 privacy policies with 400 randomly selected web pages that were not privacy policies. We randomly generated these non-privacy-policy pages using a random web page generator and made sure that they indeed are not privacy policies. All the web pages underwent the same text pre-processing using the keywords for Being a Privacy Policy in Table II to compile the training file that includes text snippets and the class (i.e., is a policy/is not a policy) for each of the 800 web pages.

We uploaded these training files to Google Prediction API [Google 2014a] to train classification models. Google Prediction deliberately keeps the classification algorithm it uses unspecified, as it trains multiple black box models and chooses the best performer. We trained a classification model against each file, the independent variable being the text snippet and the dependent variable being the class (i.e., the risk level). Each privacy policy had one feature for every model: the text snipped generated by the text pre-processing algorithm for the corresponding privacy factor. We consider including more features to be a promising future avenue of work.

Figure 3 shows the Classification Accuracy (F1 score) of the models when trained against different numbers of privacy policies selected from the corpus. The model Classification Accuracy (CA) is a number between 0 and 1 reported when training a model on Google Prediction, where 1 is 100% accuracy. This is an estimate, based on the amount and quality of the training data, of the prediction accuracy of each model. As this figure shows, the CA of most of the models converge as the number of privacy policies used for training reaches 350, justifying our decision of training models against a corpus of 400 policies.

In order to estimate the prediction ability of the models, we performed 5-fold cross validation using the corpus of 400 privacy policies. In each iteration of cross validation, the ten models were trained using 320 privacy policies and then tested against the remaining 80 policies. The ground truth, as always, was the result of manually investigating the policy. Table III shows the cross validation results. The models matched the ground truth 40% to 73% of the time.

Table III. PrivacyCheck 5-fold cross validation results, the percentage of PrivacyCheck correctly predicting ground truth.

	Email Address	Credit Card Number	Social Security Number	Ads and Marketing	Location	Collecting PII of Children	Sharing w/ Law Enforcement	Policy Change	Control of Data	Data Aggregation
Iteration 1	60%	43%	61%	48%	65%	36%	43%	49%	33%	88%
Iteration 2	63%	45%	50%	55%	43%	49%	45%	41%	40%	56%
Iteration 3	90%	54%	58%	64%	60%	56%	53%	66%	44%	69%
Iteration 4	66%	44%	48%	45%	48%	59%	50%	61%	46%	60%
Iteration 5	85%	70%	65%	54%	74%	54%	45%	56%	36%	64%
Average	73%	51%	56%	53%	58%	51%	47%	55%	40%	67%

3. RELATED WORK

In this section, we review tools, services, and privacy enhancing technologies that help users protect their privacy, without having to read privacy policies in detail, in the following categories:

(1) Privacy seals require web page operators to enroll in order to evaluate their privacy policies.
(2) New formats encourage web page operators to adopt machine-readable notation to be automatically interpreted.
(3) Crowd sourced services have an online community that reads and rates privacy policies.
(4) Data mining tools leverage machine learning and natural language processing to semi-automatically annotate privacy policies.
(5) Tracking monitors observe web pages in action, instead of investigating their privacy policies.

3.1. Privacy Seals

Privacy seals are logos of organizations or agencies that evaluate and rate privacy policies. For example, TRUSTe [TRUSTe 2015] is a data privacy management company that examines privacy policies and helps businesses align their privacy policies with legal requirements. While the information provided by TRUSTe is manually extracted, TRUSTe and other similar services suffer from two major drawbacks: (1) they significantly lack comprehensive web coverage; even though TRUSTe owns 67.94% of the market share among all similar services it covers only roughly 55,000 in one million web pages [Datanyze 2015], and (2) web page operators have to sign up and potentially pay for such services which hinders their universal adoption.

Another privacy seal is provided by Better Business Bureau (BBB) [BBBOnline 2015], a non-profit organization that provides free business reviews. However, BBB accredited businesses pay a fee for accreditation review.

Overall, researchers have expressed concerns with privacy seals in general: insufficient scrutiny of privacy seal organizations, negative self-selection of websites that participate in a seal, and users' ignorance regarding privacy seals [Kobsa 2007].

3.2. New Formats

The Platform for Privacy Preferences Project (P3P) [Cranor et al. 2006b] is a standard for websites to express their privacy policies in a both human and machine readable format. Following such standards enables automatic interpretation of privacy policies. P3P and similar standards require web page operators to adopt new formats. As a result, P3P has always suffered from lack of industry participation. Consequently, the P3P working group was closed in 2006 and P3P 1.1 was never finalized [Cranor 2012].

Currently, P3P is used in only 70,000 in one million web pages [BuiltWith 2015]. The failure of P3P in attracting industry participation, however, is not limited to the number of websites that do not support it. A large fraction of the websites that support P3P chose to include a minimal and mis-representative version of their privacy policy just to prevent Internet Explorer, the major web browser supporting P3P, from blocking their cookies. In fact, thousands of websites were found to use an identical erroneous policy recommended by a Microsoft support website to avoid cookie blocking by Internet Explorer [Cranor 2012].

Researchers have attempted visualizing privacy policies represented in P3P (e.g., Privacy Bird extension for Internet Explorer [AT&T 2002; Cranor et al. 2006a]), some with only little success in improving the comprehension [Reeder et al. 2008]. Internet Explorer 6.0 or later includes a feature that textually presents privacy policies formatted as P3P.

Nutrition Label [Kelley et al. 2010; Kelley et al. 2009; Cranor 2012] introduced another new format that asks website operators to consolidate their privacy policies in a one page standardized format inspired by the nutrition facts panel found on food and drug packages. Complying with this new format also places an additional, unwanted burden on website operators.

3.3. Crowd Sourced Services

Terms of Service; Didn't Read (ToS;DR) [ToS;DR 2012] is a free software project that started in 2012 to address the problem that very few users actually read the terms of service for websites they use. In this project, an online community of volunteers read, discuss, and rate privacy policies. The ratings and discussions are available online and as free software in the form of browser extensions for Mozilla Firefox, Google Chrome, Apple Safari, and Opera. Even though privacy policies addressed in this project are read and rated by humans and discussed thoroughly, the in-depth and occasionally long discussions pose a new challenge to the usefulness of the ratings: one might as well read, selectively, the original privacy policy itself. Furthermore, the coverage of ToS;DR is even more limited than privacy seals and new formats; only 66 privacy policies are rated so far [ToS;DR 2012].

3.4. Data Mining Tools

Building on ToS;DR, Privee [Zimmeck and Bellovin 2014] combines crowd sourcing with rule and machine learning classifiers to classify privacy policies that are not already rated in the crowd sourcing repository. However, the performance of Privee was found to be limited by the ambiguity of natural language. Privee most closely resembles our work, although (1) it uses a smaller corpus for training (only 66 policies from ToS;DR), (2) its privacy factors include only collection, sharing, ads, retention and encryption, which is less than half of what PrivacyCheck covers, and (3) its training does not enjoy the consistency of a small team working closely together.

The Usable Privacy Project [Sadeh et al. 2013] takes advantage of natural language processing, machine learning, privacy preference modeling, crowd sourcing, and formal methods to semi-automatically annotate privacy policies. This project annotates [Wil-

son et al. 2016c; Wilson et al. 2016b] a corpus of policies with attributes and data practices as the first step [Usable Privacy 2016]. The Usable Privacy Project has also used [Ammar et al. 2012] a statistical classifier trained using ToS;DR data to answer two basic questions about privacy policies automatically. While the project plans to release tools that *automatically* digest information from privacy policies to show in an easy-to-use format, no such tool, other than the corpus [Usable Privacy 2016], is available yet.

Similarly, others [Clarke et al. 2012] have proposed semi-automated extraction of privacy policy features using crowd sourcing, natural language processing, and privacy preference modeling.

3.5. Tracking Monitors

Ghostery [Ghostery 2015] is a software available as free browser extensions for Mozilla Firefox, Google Chrome, Internet Explorer, Opera, and Apple Safari. Ghostery tracks cookies, tags, web bugs, pixels, and beacons and then notifies the user of their presence as well as the companies that operate them, giving the user the choice to make informed decisions about blocking them. Similarly, Adblock Plus [AdblockPlus 2015] is a free extension that blocks adds and disables tracking. Adblock Plus is available for Android, Google Chrome, Mozilla Firefox, Internet Explorer, Opera, and Apple Safari among other browsers. Ghostery, Adblock Plus, and other similar services are fundamentally different from PrivacyCheck in that they focus on the actions a website takes, instead of the legal privacy policy it posts. Hopefully, such actions are aligned with the privacy policy, but that is not guaranteed. In addition, tracking monitors do not indicate the usage of the information gathered from users.

4. EVALUATION

In this section, we use three different methods to evaluate how accurately PrivacyCheck works in practice. First, we manually investigate what it shows for 50 new privacy policies (Section 4.1). Second, we compare it to some of the alternative tools discussed in Section 3 for those 50 new privacy policies (Section 4.2). Third, we consider the feedback of the user base community currently employing PrivacyCheck as a browser extension (Section 4.3).

4.1. Testing PrivacyCheck

In order to evaluate how well PrivacyCheck summarizes privacy policies, we tested it against 50 new privacy policies not seen in the training phase. To choose privacy policies for testing, we performed a Google search with terms “privacy policy” [Google 2014b] (on November 13, 2014 from the U.S. using Google Chrome on Windows) and selected the first 50 non-sponsored search results that we had not used in the training phase. The set of privacy policies included well-known websites (e.g., Google, Facebook, Twitter, CNN, Wikipedia) and less-known websites (e.g., Ello, OwnPhones, and Automattic). We thoroughly read each of these new privacy policies and manually determined the risk level for each of the factors. Then we ran PrivacyCheck and recorded the risk levels it indicated.

PrivacyCheck correctly identified all of the 50 test websites as privacy policies. Table IV compares the Ground Truth (according to what the policy states, and found by reading the policy) with what PrivacyCheck finds automatically. Depending on the privacy factor considered, PrivacyCheck matched the ground truth 42% to 76% of the time.

Table IV. PrivacyCheck (PC) testing results, GT stands for ground truth found by reading privacy policies.

	Email Address	Credit Card Number	Social Security Number	Ads and Marketing	Location	Collecting PII of Children	Sharing w/ Law Enforcement	Policy Change	Control of Data	Data Aggregation
GT = Red, PC = Red	0	0	0	0	2	0	13	3	9	0
GT = Red, PC = Yellow	13	0	0	17	4	0	11	0	0	11
GT = Red, PC = Green	0	1	0	0	7	2	1	1	0	0
GT = Yellow, PC = Red	0	2	0	0	0	0	7	14	10	0
GT = Yellow, PC = Yellow	36	24	2	21	10	0	16	25	12	38
GT = Yellow, PC = Green	0	10	4	1	10	11	2	4	0	0
GT = Green, PC = Red	0	3	13	0	0	0	0	1	12	0
GT = Green, PC = Yellow	1	5	1	11	2	0	0	1	7	1
GT = Green, PC = Green	0	5	30	0	15	37	0	1	0	0
PC Matches GT	72%	58%	64%	42%	54%	74%	58%	58%	42%	76%

4.2. PrivacyCheck vs. Other Tools

We consider representative tools from the categories discussed in Section 3. We ignore Privacy Seals and Tracking Monitors, because the former only indicate participation in a seal and the latter do not provide static information about a website, but rather dynamically track it in action. Therefore, we evaluate one New Format (P3P), one Crowd Sourced Tool (ToS;DR), and two Data Mining Tools (Privee and Usable Privacy) using the 50 test policies as shown in Table V. For PrivacyCheck[1], Table V shows the percentage of the factors that PrivacyCheck judged correctly for each privacy policy.

4.2.1. Comparing PrivacyCheck with P3P. We used Internet Explorer 11 to investigate P3P formatted privacy policies. Only one website (Microsoft) provided its policy in the P3P format, which was summarized as a 70 line description by Internet Explorer. The summary included what kind of information is collected and why, who has access to the information, how long it is retained, whether users have access to the information, and how disputes are handled. This summary is, by design, 100% accurate, as it is provided by the privacy policy itself using the machine readable format. Our evaluation with the 50 test policies confirmed the lack of participation in the P3P format, explained in Section 3.2.

4.2.2. Comparing PrivacyCheck with ToS;DR. Even though only 66 privacy policies are rated in the ToS;DR project, 14 of them were used in the testing phase of our research. The reason is that we used top ranking companies through a Google search to choose the 50 test websites, and crowd sourcing efforts like ToS;DR mostly concentrate on well known and popular companies as well. Out of these 14 policies, a classification was available for only seven. For the other seven websites, even though a classification was not available, thumbs up and thumbs down (positive and negative) points were given. Examples of thumbs up points are: (1) users can request access and deletion of personal

[1]PrivacyCheck was trained using the entire training set (400 policies). None of the 50 test policies were present in the training set.

Table V. PrivacyCheck vs. other tools. The numbers indicate the percentages of matching between a given tool and the ground truth (– means results unavailable). The companies are sorted alphabetically.

Company	PrivacyCheck	P3P	ToS;DR	Privee	Usable Privacy
500Pics	60	–	100	ToS;DR	–
AddThis	70	–	–	67	–
Adobe	60	–	–	67	–
AdRoll	60	–	–	83	–
AOL	50	–	–	67	100
Apple	70	–	100	ToS;DR	–
AT&T	70	–	–	100	–
Automattic	50	–	–	50	–
BlackBerry	60	–	–	67	–
CNN	50	–	–	83	–
Delicious	60	–	100	Discrepancy	–
Dell	80	–	–	83	–
Dep of Justice	90	–	–	100	–
Disney	70	–	–	83	100
EA	60	–	–	100	–
Ebay	100	–	–	67	–
Ello	60	–	–	67	–
Facebook	40	–	100	ToS;DR	–
GitHub	70	–	100	ToS;DR	–
Google	50	–	100	ToS;DR	100
Hilton	60	–	–	67	–
IBM	70	–	–	67	–
Lego	60	–	–	83	–
LinkedIn	40	–	–	83	–
Microsoft	50	100	100	ToS;DR	–
Monster	80	–	–	50	–
MyKolab	40	–	100	ToS;DR	–
National Weather	80	–	–	33	–
NBCUniversal	30	–	–	50	100
OwnPhones	40	–	–	100	–
Pandora	60	–	–	50	–
Pinterest	50	–	–	67	–
PNC	50	–	–	83	–
RocketFuel	60	–	–	67	–
RoyalAirForce	50	–	–	100	–
Slack	50	–	–	67	–
Snapchat	50	–	–	50	–
Sony Music	20	–	–	83	–
SoundCloud	50	–	100	ToS;DR	–
Staples	70	–	–	83	–
TwitPic	50	–	100	Discrepancy	–
Twitter	60	–	100	ToS;DR	–
UT Austin	70	–	–	100	–
Verison	80	–	–	67	–
Walmart	60	–	–	83	100
Wikipedia	70	–	100	Discrepancy	–
Wordpress	60	–	100	Discrepancy	–
Yahoo	70	–	100	ToS;DR	100
Ziff Davis	50	–	–	83	–
Zynga	80	–	–	67	–

information, (2) terms and privacy policy pages are organized and formatted well, and (3) the service does not track users at all. Examples of thumbs down points include: (1) terms may be changed any time without notice to the user, (2) the service tracks users on other websites, and (3) the service may sell user data as part of a business transfer. The classifications and points were 100% accurate, as they were crowd sourced. It is important to note, however, that for a general selection of privacy policies, the lack of coverage is very severe for crowd sourced tools like ToS;DR, as previously discussed in Section 3.3. Moreover, the summaries provided by ToS;DR were up to multiple pages long.

4.2.3. Comparing PrivacyCheck with Privee. The Privee extension for Chrome [Zimmeck 2014] is supposed to show the exact information as ToS;DR when it is available. However, we found 4 discrepancies where it did not (Table V). If ToS;DR does not have a record for a website, Privee uses its machine learning classifiers to learn from crowd sourced information provided by ToS;DR and build on it. The Privee extension labels each policy with an overall letter grade, from A to C, by considering the following six factors: Collection, Profiling, Ad Tracking, Ad Disclosure, Retention, and Encryption. We manually checked the privacy policy to see for what percentage of these factors Privee matches the ground truth.

4.2.4. Comparing PrivacyCheck with Usable Privacy. We used the recently released tool website Explore Usable Privacy [Usable Privacy 2016], which annotates the privacy policy of 115 websites under nine filters: (1) First Party Collection/Use, (2) Third Party Sharing/Collection, (3) User Choice/Control, (4) User Access, Edit and Deletion, (5) Data Retention, (6) Data Security, (7) Policy Change, (8) Do Not Track, and (9) International and Specific Audiences. The results presented on this site were obtained using annotations crowd sourced from ten law students. As Tables V shows, the Usable Privacy website includes privacy practice statements of six of the test policies. The website displays only the number of statements under each of the above filters, but one can access the actual annotated statements of the policy by downloading the data set. Similarly to ToS;DR, the crowd sourced annotations of Usable Privacy are inclined towards well-known companies and are, presumably, 100% accurate when available. Also similar to ToS;DR, Usable Privacy suffers from lack of coverage.

4.2.5. Discussion of Privacy Factors. Among the tools considered, PrivacyCheck, Privee, and Usable Privacy have a fixed set of privacy factors/filters. PrivacyCheck and Usable Privacy have many overlapping factors: PrivacyCheck covers all Usable Privacy filters but Data Retention, Data Security, Do Not Track, and International/Specific Audiences. PrivacyCheck also completely covers every privacy factor of Privee but Retention and Encryption. Different design methods resulted in different sets of privacy factors for each of the tools. Each of these factors is important and a combination of tools provides a better big picture than each tool separately.

4.2.6. Discussion of Web Coverage. P3P, ToS;DR and Usable Privacy results are presumably 100% accurate when available because they are crowd sourced/manually investigated. However, they significantly lack coverage: out of the 50 test websites, P3P results are available for only one, ToS;DR for only 14, and Usable Privacy for only six.

4.2.7. Discussion of Accuracy. Privee was on average 74% accurate, whereas PrivacyCheck was 60% accurate when applied on this set of 50 test policies. Investigating and improving the accuracy of PrivacyCheck is the most important future work direction for this work.

4.2.8. Threats to Validity. A threat to the validity of this evaluation is that we employed a different method to gather the set of 50 test privacy policies than the method we

Table VI. PrivacyCheck user reviews.

User	Date	Rating	Review
Mike W	Apr 18, 2016	5/5	Fills a real need. Actually reading privacy policies isn't realistic, but just blindly trusting them isn't good idea either, so it's nice to have software figure it out for you automatically.
A Pantheon	May 12, 2015	5/5	Works as intended. I don't know how others can't get this to work. It's even self explanatory!!!
Alison Pruntel	May 7, 2015	1/5	Can't get this to work. Downloaded the add-on to Chrome, navigate to privacy policy on my website, click start and get message from PrivacyCheck that "this is not a privacy policy." However, it is a page clearly labeled "Privacy Policy." Maybe it only works for shopping sites?

used to gather the set of 400 training privacy policies. The selection of 50 test policies through a Google search provided the most popular companies and enabled us to more closely evaluate how PrivacyCheck's counterparts work in practice, as many of these counterparts (P3P, ToS;DR, and Usable Privacy) work on only a tiny fraction of the most popular companies. Finally, the cross validation results and test results were very close: in cross validation PrivacyCheck matched the ground truth with an accuracy of 40% to 73% and using the testing policies it matched the ground truth 42% to 76% of the time, suggesting that PrivacyCheck performs similarly with both datasets.

4.3. PrivacyCheck User Base

PrivacyCheck is currently installed on 406 Chrome browsers and has a rating of 4.56 (rated by 9 users) on a scale of 1 to 5. Table VI shows all of its three textual reviews written by independent users [UT CID 2015]. According to the reviews, users found it "working as intended" and "self explanatory". One problem that users had with PrivacyCheck was that it does not work on privacy policies that do not discuss the ten factors. We feel that this shortcoming can be overcome by investigating more privacy factors.

5. APPLICABILITY: PRIVACYCHECK AND THE ECONOMICS OF PRIVACY

In a very recent article Acquisti et al. [Acquisti et al. 2016] reviewed the economics of privacy and reported that, in digital economies, consumers often receive imperfect, incorrect, or asymmetric information regarding what data is collected about them and how that data will be used. Hence, consumers' ability to make informed decisions about their privacy is severely hindered. Our work on PrivacyCheck directly addresses this need and seeks to improve users' understanding of what they agree to in a privacy policy. By using PrivacyCheck, consumers are able to quickly gather accurate knowledge of how companies are using their personal information. Thus, PrivacyCheck enables consumers to achieve clarity in a formerly opaque corner of the world of privacy.

In addition, many businesses, particularly small businesses, pick a privacy policy from a default list of options available on the Internet. While this practice is convenient, a company's being unaware of how they have pledged to manage, store, and use consumer data can have harmful consequences. For instance, a small business facing a data breach might not know how they should inform their customers in a manner that is compatible with the default privacy policy they picked without understanding the technical details. Furthermore, small businesses may lack the technical expertise to understand how the cryptic language in their privacy policy relates to what they *intend* or *want* to do with consumer data. By using PrivacyCheck on their own privacy policy, small businesses can better understand their promises to the consumers regarding

the handling of personal information. PrivacyCheck can also enable small businesses to better understand their enacted privacy policy, and therefore, better communicate this vision of handling personal information to their consumers. In these ways, PrivacyCheck offers a novel and technically simple way to understand and communicate the details of a privacy policy for both businesses and consumers.

6. CONCLUSIONS

In this paper, we presented a novel data mining-based technique, accompanied with its free implementation as a browser extension, to automatically sum up online privacy policies and show them as graphical icons with short descriptions. We identified, through a literature review and a survey of privacy experts, ten essential questions users should ask about how businesses use their PII. Our browser extension, PrivacyCheck, automatically answers these ten questions for any given privacy policy using data mining classification models that are trained on 400 policies and operate on a server. PrivacyCheck assigns a risk level (green, yellow, or red) to the privacy policy for each of the ten factors in question. Unlike the other somewhat similar counterparts we discussed in this paper, PrivacyCheck is readily and universally applicable on privacy policies. We evaluated PrivacyCheck and found that its results were accurate 40% to 73% of the time in cross validation. Finally, PrivacyCheck proved to be useful to an independent body of users, being installed hundreds of times on the Google Chrome web browser. As the most promising future work avenue, we envision applying a considerable variety of Machine Learning algorithms on our training set in order to improve the accuracy of PrivacyCheck.

ACKNOWLEDGMENTS

The authors would like to thank the Texas Legislature for its visions and investments in research advances and technology to support consumer empowerment in understanding and managing their online privacy and identity assets.

The authors also would like to thank Raphael De Los Santos, Usman Mahmood, Ali Ziyaan Momin, Blake Muir, and Haoran Niu for reading privacy policies to train PrivacyCheck, and James E. Zaiss for proofreading.

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

Alessandro Acquisti, Curtis Taylor, and Liad Wagman. 2016. The economics of privacy. *Journal of Economic Literature* 54, 2 (2016), 442–492.

AdblockPlus. 2015. Adblock Plus Surf the web without annoying ads! (2015). Retrieved June 3, 2015 from https://adblockplus.org/

Waleed Ammar, Shomir Wilson, Norman Sadeh, and Noah A Smith. 2012. Automatic categorization of privacy policies: A pilot study. *Research Showcase @ CMU* (2012).

AT&T. 2002. Privacy Bird. (2002). Retrieved June 15, 2015 from http://www.privacybird.org

BBBOnline. 2015. Better Business Bureau. (2015). Retrieved June 15, 2015 from http://www.bbb.org/central-texas/bbb-education-foundation

BuiltWith. 2015. P3P Policy Usage Statistics. (2015). Retrieved June 3, 2015 from http://trends.builtwith.com/docinfo/P3P-Policy

Nathan Clarke, Steven Furnell, Julio Angulo, Simone Fischer-Hübner, Erik Wästlund, and Tobias Pulls. 2012. Towards usable privacy policy display and management. *Information Management & Computer Security* 20, 1 (2012), 4–17.

Lorrie Cranor, Marc Langheinrich, Massimo Marchiori, Martin Presler-Marshall, and Joseph Reagle. 2006b. The Platform for Privacy Preferences 1.1 (P3P1.1) Specification. (2006).

Lorrie Faith Cranor. 2012. Necessary but not sufficient: Standardized mechanisms for privacy notice and choice. *J. on Telecomm. & High Tech. L.* 10 (2012), 273.

Lorrie Faith Cranor, Praveen Guduru, and Manjula Arjula. 2006a. User interfaces for privacy agents. *ACM Transactions on Computer-Human Interaction (TOCHI)* 13, 2 (2006), 135–178.

Datanyze. 2015. Truste market share in the Alexa top 1M. (2015). Retrieved June 3, 2015 from https://www.datanyze.com/market-share/security/truste-market-share

Tatiana Ermakova, Annika Baumann, Benjamin Fabian, and Hanna Krasnova. 2014. Privacy Policies and Users' Trust: Does Readability Matter?. In *20th Americas Conference on Information Systems (AMCIS)*.

FTC. 2000. Privacy Online: Fair Information Practices in the Electronic Marketplace: A Federal Trade Commission Report to Congress. (2000). Retrieved October 21, 2015 from https://www.ftc.gov/reports/privacy-online-fair-information-practices-electronic-marketplace-federal-trade-commission

FTC. 2010. Exploring privacy: an FTC roundtable discussion. (2010). Retrieved May 21, 2015 from https://www.ftc.gov/sites/default/files/documents/public_events/exploring-privacy-roundtable-series/privacyroundtable_march2010_transcript.pdf

FTC. 2012. Protecting Consumer Privacy in an Era of Rapid Change: Recommendations For Businesses and Policymakers. (2012). Retrieved May 21, 2015 from https://www.ftc.gov/reports/protecting-consumer-privacy-era-rapid-change-recommendations-businesses-policymakers

Ghostery. 2015. Join over 40 million Ghostery users and download the web's most popular privacy tool. (2015). Retrieved June 3, 2015 from https://www.ghostery.com/en/home

Google. 2014a. Google Prediction API v 1.6. (2014). Retrieved June 3, 2015 from https://cloud.google.com/prediction/docs

Google. 2014b. Google Search Engine. (2014). Retrieved November 13, 2014 from https://www.google.com/?gws_rd=ssl#q=privacy+policy

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.

ICB. 2006. Industry Classification Benchmark (ICB): a single standard defining the market. (2006). Retrieved October 7, 2015 from http://www.icbenchmark.com

Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W Reeder. 2009. A nutrition label for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security*. ACM, 4.

Patrick Gage Kelley, Lucian Cesca, Joanna Bresee, and Lorrie Faith Cranor. 2010. Standardizing privacy notices: an online study of the nutrition label approach. In *Proceedings of the SIGCHI Conference on Human factors in Computing Systems*. ACM, 1573–1582.

Alfred Kobsa. 2007. Privacy-enhanced web personalization. In *The adaptive web*. Springer, 628–670.

Ron Kohavi. 2001. Mining e-commerce data: the good, the bad, and the ugly. In *International conference on Knowledge discovery and data mining*. ACM, 8–13.

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.

David B Meinert, Dane K Peterson, John R Criswell, and Martin D Crossland. 2006. Privacy policy statements and consumer willingness to provide personal information. *Journal of Electronic Commerce in Organizations* 4, 1 (2006), 1.

George R Milne and Mary J Culnan. 2004. Strategies for reducing online privacy risks: Why consumers read (or don't read) online privacy notices. *Journal of Interactive Marketing* 18, 3 (2004), 15–29.

George R Milne, Mary J Culnan, and Henry Greene. 2006. A longitudinal assessment of online privacy notice readability. *Journal of Public Policy & Marketing* 25, 2 (2006), 238–249.

Nasdaq. 2015. Nasdaq. (2015). Retrieved September 3, 2015 from http://www.nasdaq.com

Robert W Reeder, Patrick Gage Kelley, Aleecia M McDonald, and Lorrie Faith Cranor. 2008. A user study of the expandable grid applied to P3P privacy policy visualization. In *Proceedings of the 7th ACM workshop on Privacy in the electronic society*. ACM, 45–54.

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

Disconnect Me. 2014. disconnect Me Privacy Icons. (2014). Retrieved March 15, 2016 from https://disconnect.me/icons

Usable Privacy. 2016. Usable Privacy Project Website. (2016). Retrieved September 28, 2016 from https://usableprivacy.org/

UT CID. 2015. PrivacyCheck. (2015). Retrieved May 16, 2016 from https://chrome.google.com/webstore/detail/privacycheck/poobeppenopkcbjejfjenbiepifcbclg

Norman Sadeh, Alessandro Acquisti, Travis D Breaux, Lorrie Faith Cranor, Aleecia M McDonalda, Joel R Reidenbergb, Noah A Smith, Fei Liu, N Cameron Russellb, Florian Schaub, and others. 2013. *The usable*

privacy policy project. Technical Report. Technical Report, CMU-ISR-13-119, Carnegie Mellon University.

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

ToS;DR. 2012. Terms of Service; Didn't Read. (2012). Retrieved March 4, 2015 from https://tosdr.org

TRUSTe. 2015. TRUSTe. (2015). Retrieved March 4, 2015 from http://www.truste.com

Shomir Wilson, Florian Schaub, Aswarth Dara, Sushain K. Cherivirala, Sebastian Zimmeck, Mads Schaarup Andersen, Pedro Giovanni Leon, Eduard Hovy, and Norman Sadeh. 2016a. Demystifying Privacy Policies Using Language Technologies: Progress and Challenges. *TA-COS 16: LREC Workshop on Text Analytics for Cybersecurity and Online Safety* (2016).

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, and others. 2016b. The creation and analysis of a website privacy policy corpus. In *Annual Meeting of the Association for Computational Linguistics*. 1330–13340.

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

Sebastian Zimmeck. 2014. Privee Chrome Extension. (2014). Retrieved July 13, 2015 from https://chrome.google.com/webstore/detail/privee/lmhnkfilbojonenmnagllnoiganihmnl

Sebastian Zimmeck and Steven M. Bellovin. 2014. Privee: An Architecture for Automatically Analyzing Web Privacy Policies. In *23rd USENIX Security Symposium (USENIX Security 14)*. USENIX Association, San Diego, CA, 1–16. https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/zimmeck

Received February 2007; revised March 2009; accepted June 2009

Online Appendix to: PrivacyCheck: Automatic Summarization of Privacy Policies Using Data Mining

RAZIEH NOKHBEH ZAEEM, University of Texas at Austin
RACHEL L. GERMAN, University of Texas at Austin
K. SUZANNE BARBER, University of Texas at Austin

DOI: http://dx.doi.org/10.1145/0000000.0000000

A. ONLINE SURVEY OF IMPORTANT PRIVACY FACTORS

We are developing a browser extension that takes in the privacy policy of a website, and automatically analyzes its (usually long and boring) text. The extension then summarizes the privacy policy for you and shows it using visual icons and colors. For example, if the privacy policy allows the website to collect your email address and sell it to third parties, the extension displays an email icon in red or with a danger sign. In order to make a good and useful extension, we need to know what parts of a privacy policy users care most about. This form is designed to collect your feedback on what you care about. For the purpose of trimming down the extension, please try to discriminate between the items that are most important and those that are somewhat important. Limit responses of "care a great deal" to those items that you feel it is most important to keep private.

Answer the questions on a scale of 1 to 4, with 1 being "do not care" and 4 being "care a great deal".

A.1. The information that you enter when interacting with a website

How much do you care about the way that a website deals with your...

— Name
— Email address
— Phone number
— Billing information (credit card number)
— Social security number
— Driver's license number
— Personal health information, employer or health care plan information
— Education and work history
— Personally identifiable information, if you are under 13 years old

A.2. The information that a website collects automatically

How much do you care if a website gathers and uses information about your...

— Device and software data, for example device type, operating system, browser type and version, browser plug-in types and versions, IP address, MAC address, time zone setting, and screen resolution
— Cookies, for example cookie number, and Flash cookies (also known as Flash Local Shared Objects)
— Viewed or searched products
— Purchase history and credit history information from credit bureaus
— Browsing pattern, for example URL click stream to/through/from their website, page response times, download errors, length of visits to pages, page interaction information (such as scrolling, clicks, and mouse-overs)
— Social networking accounts
— Login and password for other websites

A.3. The information that a website can collect when you are on a mobile device

How much do you care about the way that a website deals with your...

— Exact location

A.4. Usage

Do you want the extension to inform you if the website uses any of the information mentioned above for...

— Processing orders for products or services, and responding to questions
— Improving customer services
— Delivering personalized content within the site, providing search results and links (including paid listings and links)
— Ads, marketing, communication regarding updates, offers, and promotions
— Monitoring and ensuring site integrity and security, protecting the rights or safety of other users
— Aggregating non-identifiable information for business analysis
— Complying with the law and governmental requests
— Credit risk reduction, and collecting debt
— Transferring of assets if the company is acquired
— Determining your geographic location, providing location-based services
— Measuring the effectiveness of ads and user interactions with them

A.5. Others

Do you want the extension to inform you about the website's policy for...

— Updating their privacy policy
— Allowing you to update or delete your information
— Enforcing the privacy policy
— Retaining data

B. TEXT PRE-PROCESSING ALGORITHM

The input to this algorithm is the text from the web page T, a set of privacy factors F that we would like to consider, and the set of keywords K_f for each factor f (Table II elaborates on these keywords). The algorithm's output is a text snippet S_f for each of these factors. Line 1 breaks the web page text into paragraphs. Then, for each paragraph, the algorithm performs the following. Line 3 replaces all non alphanumeric characters with spaces, effectively removing all the punctuation marks to create the **p**unctuation-**l**ess paragraph pl. Line 4 converts this punctuation-less paragraph to **l**ower**c**ase lc. The next line removes all the stop words (**s**top word-**l**ess sl) and replaces any sequence of spaces (generated through previous text manipulating lines or originally present in the text) with one single space. Finally Line 6 performs stemming to keep only the word roots in the **f**inal **p**aragraph fp. Line 7 puts those word roots in W. Line 8 iterates through the factors and for each factor does the following steps. If any word of this paragraph contains any keyword for the factor then the entire paragraph fp is kept in S_f for that factor and the algorithm moves on to the next factor. Finally, after iterating over each factor and over each paragraph, the algorithm returns S which contains all S_f's.

ALGORITHM 1: Text pre-processing

input : Web Page Text T, Set of Privacy Factors F, Set of Keywords for Each Factor K_f
output: Text Snippet for Each Privacy Factor S_f

```
1  P ← T.split('\n')
2  foreach paragraph p in P do
3  |  pl ← p.replace(/[^A-Za-z0-9-]/g," ") // punctuation-less removes any non alphanumeric
   |     character
4  |  lc ← pl.toLowerCase() // converts to lowercase
5  |  sl ← lc.replace(/\b(i|me|my|...|should|now)\b/g,' ').replace(/\ s{2,}/g," ")
   |     // stopword-less removes stop words and replaces any double or more spaces
   |     with a single space
6  |  fp ← sl.stem() // final paragraph includes only the word stems
7  |  W ← fp.split(' ')
8  |  foreach factor f in F do
9  |  |  nextFactor:
10 |  |  foreach word w in W do
11 |  |  |  foreach keyword k in K_f do
12 |  |  |  |  if w.contains(k) then
13 |  |  |  |  |  S_f ← S_f + fp+' '
14 |  |  |  |  |  break nextFactor
15 |  |  |  |  end
16 |  |  |  end
17 |  |  end
18 |  end
19 end
20 return S
```

WWW.IDENTITY.UTEXAS.EDU