Learning a Privacy Incidents Database

Pradeep K. Murukannaiah
Rochester Institute of Technology
134 Lomb Memorial Drive
Rochester, NY 14623-5608
pkmvse@rit.edu

Chinmaya Dabral
North Carolina State University
890 Oval Drive
Raleigh, NC 27695-8206
csdabral@ncsu.edu

Karthik Sheshadri
North Carolina State University
890 Oval Drive
Raleigh, NC 27695-8206
kshesha@ncsu.edu

Esha Sharma
North Carolina State University
890 Oval Drive
Raleigh, NC 27695-8206
esharma2@ncsu.edu

Jessica Staddon
North Carolina State University
890 Oval Drive
Raleigh, NC 27695-8206
jessica.staddon@gmail.com

ABSTRACT

A repository of privacy incidents is essential for understanding the attributes of products and policies that lead to privacy incidents. We describe our vision for a novel privacy incidents database and our progress toward building a prototype. Key challenges in gathering such a database include bootstrapping and sustainability. We propose a semi-automated framework that can recognize privacy incidents and related information from various online sources such as news, blogs, and social media. The crux of our framework is an *incident classifier* that identifies whether a piece of text in natural language is related to a privacy incident or not. We curate a dataset consisting of 1324 news articles of which 543 articles are about one or more privacy incidents. We train the incident classifier on this dataset, considering a variety of feature engineering, feature selection, and classification techniques. We find that our incident classifier yields an F_1 measure of 93.1%, which is about 12% higher than the keyword search-based baselines we adopt.

CCS CONCEPTS

• **Security and privacy** → **Human and societal aspects of security and privacy**; • **Information systems** → *Data analytics*;

KEYWORDS

Privacy, incidents, database, analytics

1 INTRODUCTION

Many databases exist for *security* incidents. Indeed, the patterns and characteristics of security incidents, as captured by these databases, are a significant driver of security technology innovation. Patterns are detected by analyzing repositories of malware/viruses/worms (e.g. [15, 28, 44]), incidents affecting control/SCADA systems [40], general security alerts and updates [46] and data breaches (e.g., [34]). For example, the malware database VX Heavens [10] is referenced in almost 8,000 research papers according to Google Scholar, many of which tested algorithms on data supplied by VX Heavens before it was shut down in 2012.

Although there is some overlap between privacy and security incidents, most types of privacy incidents are not represented in these security incident repositories. In particular, incidents of cyberbullying/stalking, revenge porn, social media oversharing, data reidentification and surveillance, generally do not involve a security incident and so are not included in the current repositories. Table 1 demonstrates the diversity of privacy incident types.

Even for areas in which security incident databases include privacy incidents (e.g. data breaches), analysis is difficult. Current databases are not synchronized, making it difficult to compare across them and calculate accurate statistics. For example, for the year 2014, the Privacy Rights Clearinghouse data [34] finds less than 400 data breach incidents [22], Romanosky finds approximately 1200 [37], and, using proprietary data from 70 companies and organizations, Verizon finds over 2000 breaches [47]. The difficulty of tracking the frequency and consequences of data breaches in the absence of a comprehensive database was noted as early as 2007 in a U.S. Government accountability report [45].

While a complete analysis requires a comprehensive database, even an ad hoc incident analysis indicates that there are common elements of incidents that, when identified, improve system privacy. For example, a number of Internet companies have introduced new privacy policies and user consent approaches that have received negative reactions from end users or regulators. Privacy policy changes at Spotify [2] and Google [7], and user consent mechanisms at Facebook [16], have all been criticized as having deficiencies in user comprehension, visual presentation and user-utility. Analysis of these and other consent-related privacy incidents may suggest design patterns that diminish the chances of similar incidents going forward. For example, the fact that many of the criticisms of these "dark design patterns" (cf. [21]) are similar in nature has led the FTC to specify user interface requirements for disclosures, including that they be made in "print that contrasts highly with the background on which they appear" [16], and the Article 29 Working Party recommends that privacy policies be "immediately visible

HotSoS, Hanover, MD, USA

978-1-4503-5274-1/17/04…$15.00
DOI: http://dx.doi.org/10.1145/3055305.3055309

and accessible, for instance visible without scrolling and accessible via one click, from each service landing page" [7].

Incident patterns may also suggest engineering improvements. Recent years have seen repeated incidents of accidental sharing of personally identifiable information (PII), occurring at government agencies [3], healthcare providers [11], online retailers [14] and pharmaceutical companies [17]. Analyzing such incidents may suggest better technological approaches to detecting sensitive data before they are shared.

We are building the first comprehensive database of privacy incidents. While the potential to identify the trends and patterns described above using an incident database is a huge data mining opportunity, building such a database is a substantial data mining challenge, in itself, in terms of both efficiently gathering past privacy incidents and sustaining incident coverage going forward. To address this challenge we have manually evaluated over 1300 articles to build positive and negative training sets from which to learn a privacy incidents classifier. Since our high-dimensional data is vulnerable to classifier over-fitting, we use the information-theoretic measure of mutual information to reduce the feature set. Our best, SVM-based, classifier, achieves precision and recall that are both significantly better than keyword-based classifiers that flag articles as privacy incident-related if they contain "privacy" or related keywords. This classifier thus greatly reduces the human-review time needed when news articles are used to maintain and augment a privacy incidents database over time.

Related Work

As mentioned earlier, privacy incidents involving the disclosure of sensitive personal, financial and health information (e.g. social security numbers) have been collected by several organizations including the Privacy Rights Clearinghouse [34], Verizon [47], the Identity Theft Resource Center [23], and Advisen (e.g., [6]). Existing analyses of data breaches include the consideration of organization response to breaches and the cost of such breaches (e.g., [5, 37, 42]), visualizations of breach data [1], and, more recently, the consumer perspective on breach notifications [4].

Data breaches as collected and analyzed in these works, involve the breach of a company database storing such data for multiple users, often due to a security attack. However, breaches that do not involve a company or organization database (e.g. wifi payload collection [12]) and non-breach privacy incidents are not documented by such repositories.

The privacy research community is increasingly focused on gathering and analyzing privacy incidents (e.g., [18, 37]). Our work supports those efforts by providing an efficient means for gathering a repository of privacy incidents.

Finally, we note that this paper overlaps with and extends an unpublished working paper [31].

2 VISION AND PROTOTYPE

We adopt a deliberately broad definition of "privacy incident" [32], in particular, as *an event involving accidental or unauthorized collection, use or exposure of sensitive information about an individual,* ***or*** *an event that creates the perception that unauthorized collection, use or exposure of sensitive information about an individual may happen or is happening,* ***and*** *the event involves data in digital form.*

With this definition we choose to include both realized privacy "harms" (e.g., as discussed in [9]) and perceived or expected harms. The latter are important to include as they indicate policies, product features or practices that are disliked or misunderstood by users.

As an example of the importance of including perceived privacy risk, consider the revamped privacy policy introduced by the streaming music service Spotify in August 2015. The new policy included vague language about data usage and sparked a lot of criticism, leading to a reversal in September of 2015 [2]. There is no evidence any concrete privacy harm happened while this transitory privacy policy was in place, but under our definition it is a privacy incident because it was perceived as harmful and is therefore a useful data point toward understanding users' privacy preferences and perceptions, improving policies and based on that understanding.

At a high level, a privacy incident involves end user(s), one or more software systems and, potentially, adversarial user(s). These entities may cause an incident in several ways including: a system error (e.g. a bug or misconfiguration), a user error (e.g. accidental disclosure), system design (e.g. poorly chosen default settings), a system misuse or unintended use (e.g. unwanted personalization features), and a system abuse or attack (e.g. a hacking incident or a man-in-the-middle attack). These causes may overlap. For example, a design flaw may lead to user errors, and an attacker may exploit system misconfiguration. One goal of the incident database is to better understand the frequencies of and correlations between these causes. Table 1 demonstrates the diversity of privacy incidents.

One goal of the database is to support trend identification, in particular, correlation of privacy incidents with external antecedents and consequences. For example, Acquisti et al. [5] show that data breaches have a negative impact on a company's market value on the announcement day of the breach. Romanosky et al. [38] show that the adoption of data breach disclosure laws have significantly reduced identity theft caused by data breaches. The privacy incidents database will markedly reduce the data collection effort required to conduct such investigations.

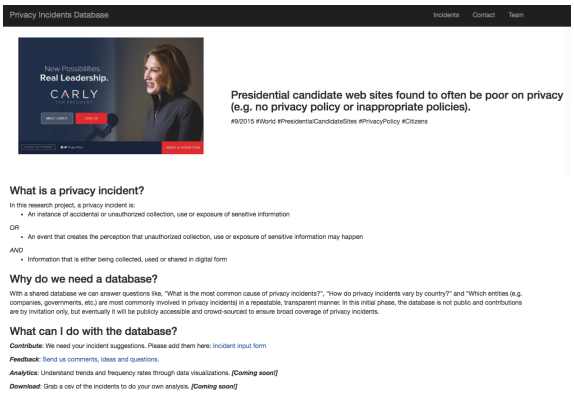
Because many of these incidents details only emerge over time, and may be spread out over many articles, we seek to automate the identification of articles about any privacy incident, rather than just incidents that are new to the database.

Our prototype (go.ncsu.edu/privacyincidents) contains 276 incidents dating from 1990 to the present, at the time of writing. For each incident there is a short description, attribute tags associated with the incident (e.g. names of entities involved), a link to public information about the incident, and, in some cases, a case study that answers the above questions in more depth. It is a closed database only editable by the authors and their students. Figure 1 shows the current landing page and some of the incidents.

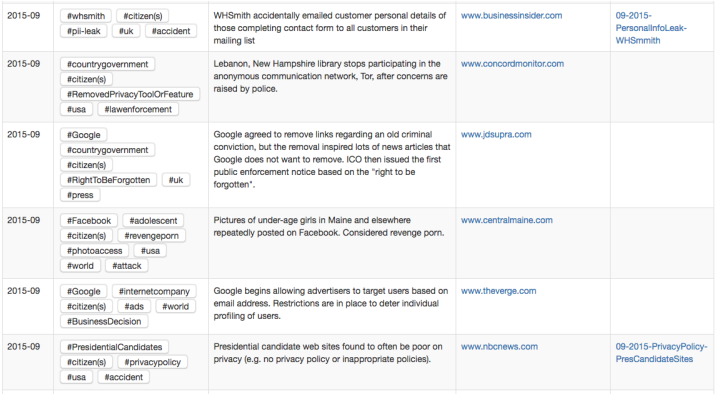
The complexity of incident root causes, as well as variation in the domains of incidents, the entities involved, and the populations impacted, argues for a collaboratively maintained database. We envision a Wikipedia-style space in which contributors can suggest incidents, gather incident evidence and reach consensus on incident attributes. The geographical range of privacy incidents suggests that the database will benefit from diverse contributors

Table 1: Examples of privacy incidents demonstrating the variety of root causes and other characteristics, bolded for readability. Incidents are grouped by predominant cause.

Root Cause	Entities	Date First Known	Description
System Error	Facebook LiveJournal, Xanga, Digg MySpace, Hi5	5/2010	Code used by the services allows **referrer** links containing **user IDs** to be sent to **advertisers** when users click on ads, thus identifying them. [13]
User Error	Kazaa	6/2002	Many Kazaa users found to be **sharing personal files** on the Kazaa network. [19]
System Design	Fitbit	7/2011	**Default privacy setting** allows FitBit profiles to be surfaced by search engines, **revealing sexual activities** of FitBit users. [25]
Unintended Use/ System Misuse	Netflix	12/2009	Researchers **de-anonymized** Netflix prize dataset, causing a **closeted gay mom** to be outed based on her Netflix viewing pattern. [35]
System Misuse/ Attack	FBI	1/2010	FBI found to have improperly gained access to **calling records of citizens**. In some cases, reporters were targeted as part of leak investigations. [39]



(a) Lading page



(b) Partial listing of the incidents.

Figure 1: Screenshots of the current prototype (http://go.ncsu.edu/privacyincidents). Visitors can click on the fourth column link on the incidents list page to visit a public article about the incident and case studies authored by the team are linked to in the fifth column. Clicking on a tag, highlights the same tag associated with different incidents.

and the interdisciplinary aspect of privacy means a wide variety of communities must be engaged in maintaining the database.

3 SEMI-AUTOMATED INCIDENT IDENTIFICATION FROM NEWS

Information about privacy incidents is available from a variety of sources such as news articles, blogs, and social media. However, finding privacy incidents manually, e.g., via ad hoc keyword search,

is nontrivial due to the vast amount of information available on the Web and high potential for false positives when relying on keyword search to identify incidents. For example, a search for the keyword "privacy" would flag articles such as, "Johnny Manziel asks for privacy as he enters rehab" [26], and "Park visitors disturb privacy of animals during mating season" [49], neither of which meet our definition of privacy incident, as the former is primarily concerned with news about a celebrity entering a hospital and the latter does not concern the privacy of humans.

We posit that information about privacy incidents has distinguishing features such as keywords, entities, and sentiment. Our objective is to develop automated techniques that exploit such features to enable efficient and large-scale identification of articles related to privacy incidents. Figure 2 provides an overview of our vision of a semi-automated framework for identifying privacy incidents. The crux of the framework is the *incident classifier*, a machine-learned model, that automatically classifies a given piece of information as related to a privacy incident or not. Further, the framework is semi-automated in the sense that an expert (human user) reviews the incidents identified by the classifier before they are added to our database.

We identify the key challenges in realizing an incident classifier as: (1) curating a *training set*, consisting of both articles that are and are not related to privacy incidents; (2) engineering *features* to represent training instances in a vector space; and (3) training a *classifier* to distinguish privacy incidents from nonincidents.

3.1 Training Set

As mentioned earlier, we focus on news articles as the source of information for identifying information related to privacy incidents. Many leading news agencies have public APIs (application programming interface), e.g., [20, 30], providing programmatic search and access to the news they publish, both current and historical.

Given a news article, our objective is to classify the article as either primarily concerned with one or more privacy incidents or not concerned with any privacy incidents. To do so, we develop an incident classifier via supervised machine learning . A supervised model requires a curated training set: a set of news articles in which each article is labeled as either primarily concerned with a privacy incident(s) (positive example, included in the set, P) or not (negative example, included in the set N). To make this determination, we rely on the definition of a privacy incident from Section 2.

To form the initial pool of articles from which P and N were extracted, we employed the following three data gathering methods. First, we randomly selected articles using the New York Times [30] and Guardian [20] APIs, expecting these articles, largely, to be part of the set N (since most news is unrelated to privacy). Second, we used the keyword "privacy" and keywords found to be closely associated with privacy news [41] to retrieve articles via the APIs. Third, we gathered articles that had been manual tagged as privacy articles by the New York Times. We also manually reviewed articles to identify examples that either narrowly failed to meet our definition for subtle reasons (e.g., articles regarding physical, non-digital, privacy incidents and incidents regarding security but not privacy) and articles that represented particular privacy issues falling in the Solove categories [43].

From the initial pool of 1104 articles gathered by these methods, we extracted 4 sets of 100 and 1 set of 198 articles. For each set, 2 coders independently reviewed the articles and coded them as either closely related to a privacy incident or unrelated to any privacy incident. Overall, the inter-coder agreement was high, with an average Cohen's Kappa [48] of 0.9608. Given this high agreement, we relied on a single coder for the remaining 506 articles.

For those articles reviewed by two coders, each article found as related to a privacy incident by both coders, was put in set P; similarly each article that both coders agreed as not related to privacy incidents was added to N. Any article on which there was a disagreement was not used in training set. Similarly, articles reviewed by a single coder were assigned to P or N based on the code. Finally, we also included 249 articles that had already been gathered by the Privacy Incidents Database project [29, 31], as part of P. We did not code these 249 articles since they were already reviewed by the Privacy Incidents Database project and determined to be positive examples.

As a result of this process and some de-duplication, P consists of 543 articles and N consists of 781 articles (total of 1324 out of the original pool of 1353). The most common publishers in our training data are the New York Times and the Guardian. However, the final training set ($P \cup N$) also includes a variety of other news publishers and tech blogs.

3.2 Feature Engineering

Our training set consists of the textual contents of a set of news articles. In order to train the incident classifier, we need to represent each news article as a set of *features*. Figure 3 summarizes the steps we followed to do so.

3.2.1 Text Preprocessing. We employ techniques from natural language processing to preprocess our dataset. First, we extract all sentences in a news article and perform *parts-of-speech* (PoS) tagging. We retain only content words (nouns, verbs, adjectives, and adverbs) and further remove stopwords [36], based on the intuition that only the remaining words help classification. Next, we perform *lemmatization* to reduce inflectional forms of words to their base or dictionary forms (known as *lemma*), e.g., the words "collecting" and "collected" are reduced to their root form "collect." Lemmatization helps reduce the number of words in an article, and consequently, the number of features used for classification.

3.2.2 Unigrams and Bigrams. After preprocessing, each news article consists of the lemmas of content words in that article. Next, we extract two types of tokens. A *unigram* is a unique lemma across all articles in the dataset. A *bigram* is a unique pair of consecutive lemmas across all articles in the dataset. Table 2 shows the number of tokens in our dataset before and after preprocessing steps.

Table 2: Number of tokens in our dataset.

Unique words (before preprocessing)	59,864
Unigrams (after preprocessing)	34,537
Bigrams (after preprocessing)	449,988

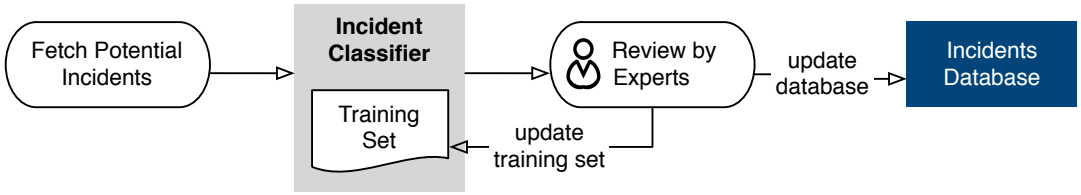


Figure 2: Our vision of a semi-automated technique for populating the privacy incidents database.

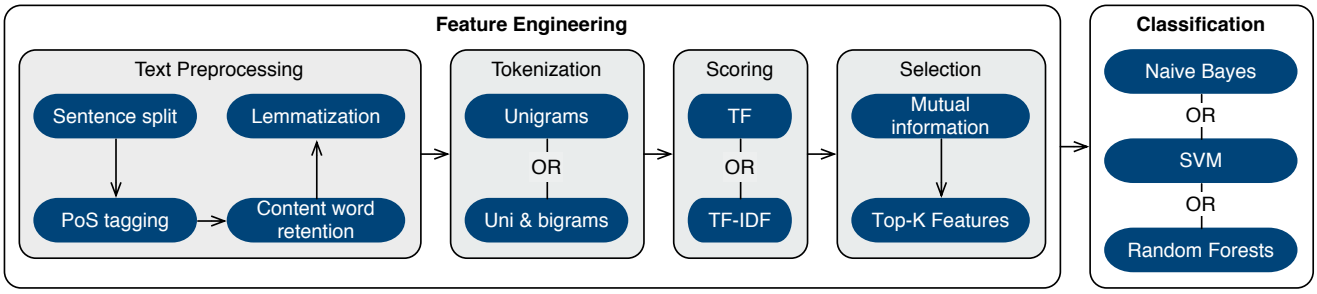


Figure 3: An overview of the steps we follow to engineer features from news articles for performing classification. The steps marked as OR represent some of the variations we consider.

3.2.3 TF and TF-IDF Scores. Our next task is to represent each news article as a feature vector. To do so, we treat each token (unigram and/or bigram) in the entire dataset as a feature. To compute feature values, we employ the TF scores or TF-IDF scores [27, Chapter 6] described below. Note that although there are multiple variants of these scores, we employ one set of commonly used variants to demonstrate our overall approach.

The *term frequency* (TF) is defined as:

$$\mathrm{TF}(t,a) = \begin{cases} 1 + \log f_{t,a} & \text{if } f_{t,a} > 0 \\ 0 & \text{Otherwise,} \end{cases} \quad (1)$$

where $f_{t,a}$ is the frequency of the token t in article a. The logarithm of raw frequency is used for sublinear scaling with the intuition that a term occurring ten times in an article may not be ten times as important as a term occurring once.

Next, the *inverse document frequency* (IDF) is defined as:

$$\mathrm{IDF}(t,A) = \log \frac{N}{|a \in A : t \in a|}, \quad (2)$$

where t is a token in article a, A is the set of all articles in the dataset, N is the size of A, and $|a \in A : t \in a|$ is the number of articles in which t appears.

Finally, the TF-IDF score is the product of term frequency (Equation 1) and inverse document frequency (Equation 2):

$$\text{TF-IDF}(t,a,A) = \mathrm{TF}(t,a) \times \mathrm{IDF}(t,A). \quad (3)$$

In a nutshell, the TF-IDF score indicates the importance of a token in an article, considering all articles in the dataset. That is, for a given token, the TF-IDF score increases as the token appears more frequently in the article, but decreases as the token appears in more articles.

3.2.4 Feature Selection. Our dataset consists of 1324 instances (positive and negative examples combined), but a substantially larger number of features—considering only unigrams yields 34,537 features; considering bigrams, in addition, increases the number of features more than ten fold. A dataset in such a high-dimensional space poses two key challenges. First, training a classifier can be inefficient. Second, and more importantly, a classifier trained in a high-dimensional space can be prone to overfitting. That is, given a large number of features, it is likely that some of them are "rare" and exist only in the training data. Such features may yield a classifier highly accurate on the training data, but error-prone on new data.

To address these challenges, we select a subset of features before performing classification. Specifically, we employ the expected *mutual information* (MI) of a token t and a class c [27, Chapter 13] to determine whether to select a feature or not. MI is defined as:

$$\mathrm{MI}(U;C) = \sum_{e_t \in \{1,0\}} \sum_{e_c \in \{1,0\}} \mathrm{P}(e_t, e_c) \times \log_2 \frac{\mathrm{P}(e_t, e_c)}{\mathrm{P}(e_t) \times \mathrm{P}(e_c)}, \quad (4)$$

where U is a random variable, taking the values of $e_t = 1$ for articles containing token t and $e_t = 0$ for articles not containing t; and C is a random variable, taking values of $e_c = 1$ for articles in class c (articles related to privacy incidents) and $e_c = 0$ for articles not in c (articles not related to privacy incidents).

In a nutshell, the expected mutual information of a token t and class c measures the extent to which the presence of t contributes to classifying an article as belonging to class c. Thus, we rank all features by their mutual information values and select top K features. Table 3 shows top 20 features in our dataset ordered by mutual information.

Table 3: The top 20 features in our dataset ordered by their MI values (features with an underscore are bigrams).

Token	MI	Token	MI
privacy	0.3531	datum	0.2280
information	0.2225	company	0.1278
collect	0.0949	personal_information	0.0880
personal	0.0847	phone	0.0846
breach	0.0834	use	0.0824
access	0.0801	user	0.0753
customer	0.0736	privacy_policy	0.0712
address	0.0710	surveillance	0.0709
request	0.0621	law	0.0610
investigation	0.0575	print_april	0.0560

3.3 Classification

In Section 3.2, we transformed news articles in natural language to data instances in vector space. These data instances consist of two non-overlapping sets: news articles related to privacy incidents and those not related to privacy incidents. We refer to these two sets of articles as *privacy* class and *non-privacy* class, respectively.

3.3.1 Classifiers. We incorporate *classifiers* to identify whether a news article belongs to the privacy or the non-privacy class. We learn the parameters of the classifier considering articles from the sets P and N in Section 3.1 as privacy and non-privacy classes, respectively, using the feature engineering of Section 3.2. We exploit one of the following well-known classifiers.

Naive Bayes (NB) [33, Chapter 5] is a probabilistic classifier that exploits Bayes' theorem to compute the probability of an instance belonging to a class. To estimate class-conditional probabilities, NB makes a strong (naive) assumption that features are conditionally independent of each other given the class.

Support Vector Machines (SVM) [33, Chapter 5] constructs a hyperplane to separate positive and negative data instances. Training an SVM model involves learning the parameters of a hyperplane that maximizes the *margin* between classes (informally, the margin is the extent of separation between the classes given the hyperplane). We employ SVM with a linear kernel.

Random Forests (RF) [8] operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes of the individual trees. Each tree is trained in a greedy fashion on a randomly chosen subset of the training data, selecting the TF-IDF feature that maximises information utility. We employ a standard RF algorithm trained on 400 trees each of depth 20 (these parameters were refined experimentally).

3.3.2 Baselines. We develop keyword-based classifiers to serve as baselines for evaluating the other classifiers above.

A keyword-based classifier identifies a news article as privacy or non-privacy depending on the presence or absence of certain keywords, respectively. Identify a set of keywords is the crux of this approach. Then, we predict a news article as belonging to the privacy class if it contains one of the keywords identified, and as non-privacy class, otherwise. To identify the set of keywords we employ one of the following techniques.

Privacy keyword technique employs the term "privacy" and its inflectional forms as the set of keywords.

Privacy and Solove keywords technique employs the term "privacy" and in addition, terms based on the names of Solove categories and subcategories [43] as the keyword set. Table 4 shows the Solove keywords we add and their counts in the privacy class (we also include the inflectional forms of these words, but do not show them in the table, for brevity).

Top-K keywords technique employs K most frequent tokens in the privacy class as the set of keywords. We perform the text preprocessing steps described in Section 3.2.1 on the articles in the privacy class before identifying the keywords. Here, we only consider nouns based on our observation that other frequent content words may not to be useful for classification (e.g., the verb "say" was the most frequent content word in the privacy class, but will likely not influence classification). Table 5 shows Top 20 keywords in the privacy class of our dataset.

Table 4: Counts of keywords representing Solove's categories and subcategories for articles in the privacy class.

Token	Count	Token	Count
		Info. dissemination	24
Info. collection	935	Breach confidentiality	694
Surveillance	545	Disclosure	289
Interrogation	3	Exposure	162
Info. processing	161	Increased accessibility	38
Aggregation	34	Blackmail	21
Identification	354	Appropriation	72
Insecurity	15	Distortion	3
Secondary use	2	Invasion	66
Exclusion	21	Intrusion	118
		Decisional interference	27

Table 5: The 20 most frequent tokens (nouns only) in our dataset, considering articles in the privacy class.

Token	Count	Token	Count
datum	2349	information	2023
privacy	1897	company	1838
security	1440	user	1330
government	1261	people	1089
google	1077	facebook	1005
apple	992	law	876
app	858	court	806
case	805	service	778
year	772	phone	735
site	689	time	667

4 EVALUATION

As Section 3.2 describes, our feature engineering pipeline consists of several variation points. First, we perform a comprehensive evaluation to identify an optimal set of steps. Then, we compare the

performance of our incident classifier with those of the baselines. We evaluate the performance of a classifier via precision, recall, and F_1 scores defined below.

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \ \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$
$$F_1\text{-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}, \quad (5)$$

where TP, TN, FP, and FN are true and false positives and negatives.

Our dataset consists slightly more negative instances (781) than positive instances (543). We randomly excluded 238 (= 781 − 543) negative instances to have a balanced training set of 543 positive and negative instances, each. All results we report in this section are based on ten-fold cross-validation on the balanced training set.

4.1 Feature Engineering

After text preprocessing, we have options of (1) unigram vs. unigrams and bigrams; (2) TF vs. TF-IDF scores; and (3) different K values for selecting top-K features. Figure 4 and Table 6 compare the performances of classifiers trained with features engineered via different combinations of these three options. For the values K, considering the large number of possibilities, we experiment in powers of two (1, 2, 4, 8, ... up to maximum number of features). We make three major observations from these results.

First, the choice of TF-IDF scores against TF scores makes little difference to performance. Indeed, using TF scores yields slightly better results than using TF-IDF scores.

Second, employing both unigrams and bigrams yields better results than employing only unigrams. In particular, the difference in precision is quite noticeable. This suggests that using bigrams yields certain features whose values are unique to a class, whereas the values of their constituent unigrams are not unique to a class. For example, the bigram "third_party" has higher TF scores, on average, for the privacy class, whereas the TF scores for the unigrams "third" and "party" are about the same for the two classes.

Third, the performance of the classifier increases as the number of features increases, initially. However, after a certain threshold, further increasing the number of features reduces both precision and recall. We conjecture that models trained with more features than this threshold overfit the training data. Thus, as Table 6 demonstrates, choosing an optimal number of features yields a classifier with better performance than the one choosing all features yields.

4.2 Classification

We identified that choosing unigrams and bigrams as features, TF scores as their values, and selecting a subset of features yields best classifier results for our dataset. Considering this combination, Table 7 compares performances of the three machine-learned classifiers (Naive Bayes, SVM, and Random Forests) and the three baselines we evaluate. We observe the following from these results.

First, we find that machine-learned classifiers based on our feature engineering perform better than keyword-based baselines. Specifically, the SVM-based classifier performs best, overall, with the mean F_1 score of 93.1%, which is about 12% higher than the best F_1 score among the three baselines. This suggests that systematically building an incident classifier (data curation, feature engineering, and classification) is worth the effort.

Second, we note that even with an optimal number of features, the datasets we train the machine-learned classifiers on consist of considerably more features than the number of data instances. SVM is known to work-well in high-dimensional spaces. We conjecture this to be a reason why SVM performs better than Naive Bayes and Random Forests in our setting.

Third, we observe that employing "privacy" as the keyword yields a classifier with a high precision (91.8%) on the privacy class. This indicates most articles mentioning "privacy" tend to be related to a privacy incident. However, we note the recall of this classifier for the privacy class is only 70%. Thus, many articles that are about a privacy incident do not explicitly mention the word "privacy."

Fourth, we observe the keyword-based classifiers based on Privacy and Solove keywords, and Top-3 keywords both yield a higher recall on the privacy class compared to the Privacy keyword-based classifier. That is, whereas searching only for "privacy" recalls only 70% of articles about privacy incidents, adding more keywords recalls more than 90% articles about privacy incidents (notice from Table 5 that privacy is one of the top-3 keywords). However, it is important to note that, in these cases, as the recall improves precision reduces (i.e., false positives increase), consistently. Figure 5 demonstrates this trend of increased recall and reduced precision, as K increases in the Top-K keyword-based classifier.

5 DISCUSSION AND CONCLUSION

We envision a collaboratively maintained Privacy Incidents Database that supports research, practice, and policy making. We developed an incident classifier that recognizes information about privacy incidents from news articles. Since finding privacy-related information online is like searching for needles in a haystack, the classifier can be of great value in bootstrapping the incidents database and sustaining it (since the classifier automatically identifies the information to be reviewed by experts, it reduces the overall human effort required to add an incident to the database).

Our evaluation suggests that machine-learned incident classifiers outperform keyword search based approaches in recognizing news articles about privacy incidents. The main challenge with keyword based approaches is that identifying a good set of keywords is nontrivial. Although we sought to systematically develop keyword sets, the sets have limitations. For example, we added terms from the Solove categories and subcategories [43], literally, as keywords. However, an article about "secondary use" (a Solove subcategory) may not explicitly contain the phrase "secondary use." Similarly, considering Top-K keywords may yield keywords that are in both privacy and non-privacy articles (e.g., "Google," "Facebook," and "Apple" in Table 5, and are thus, not effective in classifying news articles. Nonetheless, the keyword based approaches achieve a high recall, even better than the machine-learned classifiers. As Figure 5 shows, with as few as Top-8 keywords, the corresponding classifier's recall on the privacy class is about 99%. Since an expert reviews articles tagged by the classifier before adding them to the database, high recall is desirable. However, since the precision of these classifiers is quite low (e.g., the precision of the Top-8

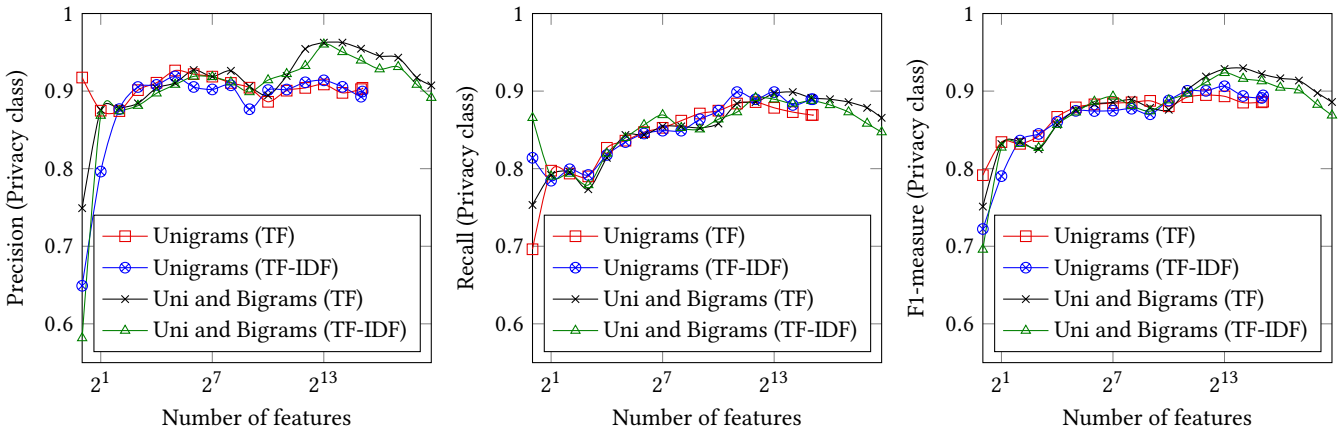


Figure 4: Comparing SVM classifiers trained on different top-K features selected based on information gain.

Table 6: Performance of an SVM-based classifier, considering different feature engineering techniques.

Feature		Privacy Class			Non-privacy Class			Mean		
Type	Count	Prec.	Recall	F_1	Prec.	Recall	F_1	Prec.	Recall	F_1
Unigrams (TF)	34536 (all)	0.904	0.869	0.886	0.874	0.908	0.891	0.889	0.889	0.889
Unigrams (TF)	4096 (optimal)	0.904	0.886	0.895	0.888	0.906	0.897	0.896	0.896	0.896
Unigrams (TF-IDF)	34536 (all)	0.899	0.890	0.895	0.891	0.901	0.896	0.895	0.895	0.895
Unigrams (TF-IDF)	8192 (optimal)	0.914	**0.899**	0.906	0.900	0.915	0.908	0.907	0.907	0.907
Uni & Bigrams (TF)	449987 (all)	0.907	0.866	0.886	0.871	0.912	0.891	0.889	0.889	0.889
Uni & Bigrams (TF)	8192 (optimal)	**0.962**	0.897	**0.929**	**0.903**	**0.965**	**0.933**	**0.933**	**0.931**	**0.931**
Uni & Bigrams (TF-IDF)	449987 (all)	0.891	0.847	0.869	0.854	0.897	0.875	0.873	0.872	0.872
Uni & Bigrams (TF-IDF)	8192 (optimal)	0.960	0.890	0.924	0.897	0.963	0.929	0.929	0.926	0.926

Table 7: Performances for different classification techniques (having 8192 selected TF unigrams as features) and baselines.

	Privacy Class			Non-privacy Class			Mean		
Classifier	Precision	Recall	F_1	Precision	Recall	F_1	Precision	Recall	F_1
Naive Bayes	0.863	0.924	0.892	0.919	0.853	0.884	0.891	0.889	0.888
SVM	**0.962**	0.897	**0.929**	**0.903**	**0.965**	**0.933**	**0.933**	**0.931**	**0.931**
Random Forests	0.931	0.820	0.872	0.839	0.939	0.886	0.885	0.879	0.879
Privacy keyword	0.918	0.700	0.794	0.757	0.937	0.838	0.838	0.819	0.816
Privacy & Solove keywords	0.664	0.930	0.775	0.783	0.530	0.663	0.774	0.730	0.719
Top-3 keywords	0.721	**0.932**	0.813	0.904	0.639	0.749	0.812	0.785	0.781

keyword based classifier in the privacy class is only 53.5%). The tradeoff between increasing recall and reducing human effort (lower precision means more false positives and more reviewing work for experts) remains to be studied.

We have deliberately trained our classifiers on a large set of positive (i.e. privacy incident-related) articles, rather than on training data that is representatively balanced with positive and negative. We did this both because a representative set would have few if any positive examples, given the low support of privacy incidents, and because, as the complexity of privacy taxonomies [24, 43] demonstrates, privacy incidents are varied in nature and so argue for more training data. That said, using a large positive training set has the potential of making the classifier's job harder as well, particularly, because we included many "grey" examples (i.e., incidents that concern security but not privacy). Indeed, doing so may make it appear more likely that an article both has a feature and is related to a privacy incident than is actually the case. In future work, we plan to experiment with different proportions of training data and test the classifiers on an ongoing basis as in Figure 2.

Finally, we identify three avenues for future work. First, in addition to news, social media (e.g., Twitter) is an increasingly popular

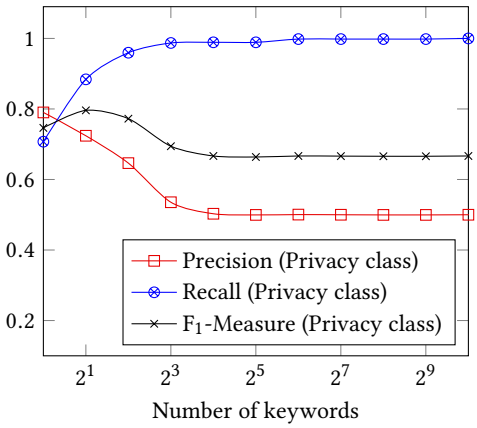


Figure 5: Performances of keyword-based classifiers trained on different top-K keywords.

source of privacy incidents related information. An interesting direction is to extend our incident classifier to tag content on social media as relating privacy incidents or not. Second, we performed a binary classification of privacy or non-privacy. However, it can also be valuable to perform finer classification, e.g., recognizing the Solove category to which a piece of information belongs. A key challenge in performing such classification task is building a sufficiently large expert-annotated dataset. Third, our eventual objective is to support a variety of analyses on the data in the Privacy Incidents Database. An interesting direction is to develop such analysis tools considering the data in our database. Example analyses tasks include, identifying the organizations or software systems associated with an incident (e.g., via named entity recognition); identifying similar incidents (e.g., via clustering); and identifying the sequence of events associated a privacy incident (e.g., via temporal analysis).

ACKNOWLEDGEMENT

Thanks to the US Department of Defense (Science of Security Lablet grant) for partial support and to Kimberly Milner for helping assemble the training data for the classifiers.

REFERENCES

[1] World's biggest data breaches, selected losses greater than 30,000 records (updated 11th august 2015). http://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks.
[2] A. Mamiit. Spotify revamps new privacy policy after user backlash. http://developer.nytimes.com/, September 5, 2015.
[3] A. Southall. U.S. workers are on alert after breach of data. http://www.nytimes.com/2010/11/07/us/07breach.html?_r=0, November 6, 2010.
[4] L. Ablon, P. Heaton, S. Romanosky, and D. C. Lavery. *Consumer attitudes toward data breach notifications and loss of personal information.* Rand Corporation, 2016.
[5] A. Acquisti, A. Friedman, and R. Telang. Is there a cost to privacy breaches? An event study. *ICIS 2006 Proceedings*, page 94, 2006.
[6] Advisen. Cyber loss: Outsourcing data to the cloud doesn't diminish risk. http://www.advisenltd.com/2015/09/10/cyber-loss-outsourcing-data-to-the-cloud-doesnt-diminish-risk, September, 2015.
[7] Article 29 Data Protection Working Party. Google privacy policy: WP29 proposes a compliance package. http://www.cnil.fr/english/news-and-events/news/article/google-privacy-policy-wp29-proposes-a-compliance-package/, September 2014.
[8] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct. 2001.
[9] R. Calo. The boundaries of privacy harm. *Indiana Law Journal*, 86(3):1131–1162, 2011.
[10] G. Cluley. VX Heavens, old-school virus-writing website, raided by police, March 28, 2012.
[11] D. Palmer. ICO investigating 56 Dean Street clinic for disclosing details of 780 HIV patients in data breach. http://www.computing.co.uk/ctg/news/2424278/ico-investigating-health-clinic-for-disclosing-details-of-780-hiv-patients-in-mass-email, September 2, 2015.
[12] E. Bailey Jr. Google in the hot seat over WiFi privacy breach. http://www.huffingtonpost.com/2010/06/08/google-privacy-slammed-ov_n_604084.html, June 8, 2010.
[13] E. Steel and J. Vascellaro. Facebook, myspace confront privacy loophole. the wall street journal. may 21, 2010. http://www.wsj.com/news/articles/SB10001424052748704513104752567012154655 96.
[14] L. England. People are freaking out because WHSmith is accidentally emailing customer contact details to other customers, September 2, 2015.
[15] Enigma Software Group. Enigma software's threat database. http://www.enigmasoftware.com/threat-database/.
[16] FTC. Agreement containing consent order in the matter of facebook, inc, a corporation. https://www.ftc.gov/sites/default/files/documents/cases/2011/11/111129/facebookagree.pdf. File no. 092 3184. United States Of America Federal Trade Commission.
[17] FTC. Eli Lilly settles FTC charges concerning security breach. https://www.ftc.gov/news-events/press-releases/2002/01/eli-lilly-settles-ftc-charges-concerning-security-breach, January 18, 2002.
[18] S. Garfinkel and M. F. Theofanos. A collection of non-breach privacy events, February, 2016.
[19] N. S. Good and A. Krekelberg. Usability and privacy: a study of kazaa p2p file-sharing. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 137–144. ACM, 2003.
[20] T. Guardian. Guardian open platform. http://open-platform.theguardian.com/. Accessed: 2016-3-3.
[21] H. Brignull, M. Miquel, J. Rosenberg, J. Offer. Dark patterns: Fighting user deception worldwide. http://darkpatterns.org/.
[22] N. Huq. Follow the data: Dissecting data breaches and debunking myths. *Trend-Micro Research Paper*, September, 2015.
[23] ITRC. Identity theft resource center. http://www.idtheftcenter.org/.
[24] B.-J. Koops, B. C. Newell, T. Timan, I. Škorvánek, T. Chokrevski, and M. Galič. A typology of privacy. *University of Pennsylvania Journal of International Law, Forthcoming*, 2016.
[25] L. Rao. Sexual activity tracked by Fitbit shows up in Google search results. http://techcrunch.com/2011/07/03/sexual-activity-tracked-by-fitbit-shows-up-in-google-search-results/, July 3, 2011.
[26] T. Lutz. Johnny Manziel asks for privacy as he enters rehab. *The Guardian*, February 2, 2015.
[27] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval.* Cambridge University Press, New York, 2008.
[28] Microsoft. Microsoft's malware protection center. https://www.microsoft.com/security/portal/mmpc/.
[29] NCSU, UNCC, Clemson University, and Rochester Institute of Technology. The privacy incidents database research project. https://research.csc.ncsu.edu/privacyincidents/index.php.
[30] NYTimes. The New York Times API. http://developer.nytimes.com/. Accessed: May 2016.
[31] P. K. Murukannaiah, J. Staddon, H. Lipford and B. Knijnenburg. PrIncipedia: A privacy incidents encyclopedia. Working paper at the 2016 Privacy Law Scholars Conference.
[32] P. K. Murukannaiah, J. Staddon, H. Lipford and B. Knijnenburg. (Work in Progress) Is this a privacy incident? Using news exemplars to study end user perceptions of privacy incidents. In *Proceedings of the Workshop on Usable Security*, pages 1-7 (To appear). 2017.
[33] T. Pang-Ning, M. Steinbach, V. Kumar, et al. *Introduction to data mining.* Addison-Wesley Longman, Boston, 2006.
[34] PRC. Privacy Rights Clearinghouse. http://www.privacyrights.org/.
[35] R. Singel. Netflix spilled your brokeback mountain secret, lawsuit claims. wired, december 2009. http://www.wired.com/2009/12/netflix-privacy-lawsuit/.
[36] Ranks NL. Default English stopwords list. http://www.ranks.nl/stopwords. Accessed: 2016-12-19.
[37] S. Romanosky. Examining the costs and causes of cyber incidents. In *Twelfth Annual Forum on Financial Information Systems and Cybersecurity: A Public Policy Perspective*, January, 2016.
[38] S. Romanosky, R. Telang, and A. Acquisti. Do data breach disclosure laws reduce identity theft? *Journal of Policy Analysis and Management*, 30(2):256–286, 2011.
[39] C. Savage. F.b.i. violated rules in obtaining phone records, report says. http://www.nytimes.com/2010/01/21/us/21fbi.html?_r=0, January 20, 2010. Accessed: 2016-05-16.

[40] Security Incidents Organization. RISI online incident database. http://www.risidata.com/Database.
[41] K. Sheshadri, N. Ajmeri, and J. Staddon. No (privacy) news is good news: An analysis of privacy news in the u. s. and u. k. from 2010-2016. *Under review*, 2016.
[42] Shey, Heidi. Market overview: Customer data breach notification and response services,âĂİ forrester, august 5, 2015.
[43] D. J. Solove. A taxonomy of privacy. *University of Pennsylvania Law Review, Vol. 154, No. 3*, January 2006.
[44] Symantec. Symantec security response threat writeups. https://www.symantec.com/security_response/landing/azlisting.jsp.
[45] U. S. Government Accountability Office. Data breaches are frequent, but evidence of resulting identity theft is limited; however, the full extent is unknown, 2007. Report to Congressional Requesters. Washington, D.C., GAO-07-737.
[46] US-CERT. United States Computer Emergency Readiness Team. https://www.us-cert.gov/.
[47] Verizon. 2015 data breach investigations report, 2015.
[48] A. J. Viera and J. M. Garrett. Understanding interobserver agreement: The kappa statistic. *Fam Med 37, 5 (2005)*, pages 360–363, 2005.
[49] H. Warwick. Photographers – don't pap our wild animals, they need some privacy too. *The Guardian*, November 10, 2015.