Analyzing User Perspectives on Mobile App Privacy at Scale

Preksha Nema
Google
Bangalore, India
preksh@google.com

Pauline Anthonysamy
Google
Zurich, Switzerland
anthonysp@google.com

Nina Taft
Google
Mountain View, USA
ninataft@google.com

Sai Teja Peddinti
Google
Mountain View, USA
psaiteja@google.com

ABSTRACT

In this paper we present a methodology to analyze users' concerns and perspectives about privacy at scale. We leverage NLP techniques to process millions of mobile app reviews and extract privacy concerns. Our methodology is composed of a binary classifier that distinguishes between privacy and non-privacy related reviews. We use clustering to gather reviews that discuss similar privacy concerns, and employ summarization metrics to extract representative reviews to summarize each cluster. We apply our methods on 287M reviews for about 2M apps across the 29 categories in Google Play to identify top privacy pain points in mobile apps. We identified approximately 440K privacy related reviews. We find that privacy related reviews occur in all 29 categories, with some issues arising across numerous app categories and other issues only surfacing in a small set of app categories. We show empirical evidence that confirms dominant privacy themes – concerns about apps requesting unnecessary permissions, collection of personal information, frustration with privacy controls, tracking and the selling of personal data. As far as we know, this is the first large scale analysis to confirm these findings based on hundreds of thousands of user inputs. We also observe some unexpected findings such as users warning each other not to install an app due to privacy issues, users uninstalling apps due to privacy reasons, as well as positive reviews that reward developers for privacy friendly apps. Finally we discuss the implications of our method and findings for developers and app stores.

KEYWORDS

privacy, nlp, mobile apps, empirical

ACM Reference Format:
Preksha Nema, Pauline Anthonysamy, Nina Taft, and Sai Teja Peddinti. 2022. Analyzing User Perspectives on Mobile App Privacy at Scale. In *44th International Conference on Software Engineering (ICSE '22), May 21–29, 2022, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3510003.3510079

ICSE '22, May 21–29, 2022, Pittsburgh, PA, USA

ACM ISBN 978-1-4503-9221-1/22/05…$15.00
https://doi.org/10.1145/3510003.3510079

1 INTRODUCTION

In app stores, such as Google Play and Apple's App Store, users can write reviews to share their experience and opinions about the apps that they use. Reviews help other users to understand whether or not the app might be of interest to them. These reviews are also a feedback channel to developers who can learn how to improve their apps. App reviews can be a challenge to analyze as they are known to cover a broad range of topics, have widely varying quality (that is somewhat exacerbated by their unstructured form) [54]; thus it can be difficult for developers to parse out separate issues. Moreover, some issues, such as privacy related feedback, may have lower volume than other issues (e.g., battery performance), and thus may be less visible. However extracting privacy related feedback is of particular importance as by now developers are well aware that trust is heavily impacted by their privacy posture, and because privacy legislation and regulation are on the rise.

Unstructured app reviews provide a potentially rich source of content to learn about users' perspectives on privacy. User feedback can be mined at scale to extract product requirements [45, 68], however these methods do not focus on privacy. Traditional methods for gaining insights into user perspectives about privacy include conducting qualitative studies and surveys. Qualitative studies have the advantage of being in depth as they involve one-on-one interviews, however they are *limited in scale* to include typically 20-30 participants. Surveys, on the other hand, have the advantage being structured and repeatable, however are also limited in terms of scale typically to a few thousands. By leveraging automation and advanced Natural Language Processing (NLP) techniques, feedback by millions of users can be analyzed rapidly to extract privacy perspectives. This approach complements existing qualitative methods as it obtains privacy concerns on a new scale, yet cannot follow up with users for additional information. Our approach enables data driven decisions to be made - such as priority ranking across multiple issues. We therefore answer the following research questions:

- Can mobile app reviews be automatically analyzed at scale to identify privacy related ones?
- Can the identified privacy reviews be used to understand users' privacy concerns? How do they compare with concerns inferred from other qualitative methodologies?
- How can such a methodology be leveraged to inform the ecosystem, and more specifically mobile app developers?

This paper presents two main contributions. The first is a methodology to analyze user's privacy concerns with mobile apps via NLP

techniques **at scale**. The methodology includes a binary classifier to decide whether or not a particular review discusses a privacy topic, a mechanism to cluster reviews that discuss similar semantic topics, and a way to summarize the clusters by identifying reviews that are highly representative of all those belonging to a single cluster. Our methodology opens the door to a rich set of subsequent research on user privacy perspectives (e.g., relative ranking of privacy concerns, how these concerns evolve over time, why some privacy issues occur predominantly within specific app categories, user perspectives on different elements of personal data, etc). It also enables tools for developers to receive feedback that can help them improve the privacy of their products.

Our second contribution uses the methodology to provide **empirical evidence** (a first large-scale analysis) that confirms dominant privacy themes that have been identified in qualitative studies [3, 25, 36, 47, 79]. We apply our methods to 287M reviews, and report on the wide variation of privacy reviews across app categories, relative reviewing of specific app permissions and dominant privacy themes. We also discuss the implications of our findings for developers and how they can inform the design of privacy tools in app stores.

There are many challenges to this problem space. First, there is no labeled data. Second, the boundary of when a text is about privacy or not is a fuzzy one; for example, many security and privacy experts often do not agree on the amount of overlap between these two areas [13, 19]. Third, we need to be able to capture a broad range of privacy concerns, i.e., any topic that exists in established privacy taxonomies [4, 67]. Fourth, to learn a broad range of contexts, we need a classifier that can do more than memorization of keywords. Fifth, user reviews are well known to be variable in their writing style [54] which can lead to classification errors. We address all these challenges herein.

Our first contribution develops a multi-step method to address this problem. To generate our test, validation and training data we employ a method that combines Expert-Hand-Labeling and Heuristic Supervision [61] (Challenge #1). In this bootstrapping process we construct a set of regular expressions in consultation with linguistic experts, and based on our manual observations of how users express privacy concerns in reviews (Challenge #2, #3). We next develop a privacy classifier as an ensemble model that couples both USE [14] and BERT [20] deep learning models, which are trained on user texts and are known to generalize well (Challenge #3, #4, #5). Our classifier achieves 98% precision and 87% recall. We use K-means clustering to group similar privacy reviews, and propose a new metric to determine a good value of k that produces compact single-issue clusters (Challenge #3). We summarize the topic clusters, by extracting representative reviews for the cluster. Overall this yields a new method to understand user insights about privacy that complements traditional qualitative methods.

The development of automated privacy text classifiers (for user text) and the ability to separate privacy issues into distinct clusters in an unsupervised fashion, enables numerous types of large-scale analyses including (some of which are discussed in this paper) - ranking of issues, breadth of concerns across app types, unanticipated and emerging issues, sentiment with which users approach specific issues, user behavior (e.g. uninstalling an app), comparisons across app types (e.g. children's versus regular apps), issues per culture (as captured by language) and more. We discuss how such findings can inform developers about privacy shortcomings of their apps, and more generally, how it could inform privacy best practices across app stores.

Our second contribution is a large scale analysis using our methodology on 287M reviews, coming from approximately 2M apps from the Play Store. We identify approximately 440K reviews that discuss privacy. We found that the privacy-related reviews exist in all of the 29 app categories[1] in the Play store, and that these reviews capture a breadth of privacy concerns and perspectives. We find many reviews where users ask to know the purpose for a request to collect personal information or permission gated data (e.g., location, contacts, camera, etc). We observed four somewhat unexpected findings: (1) users warn each other not to install an app due to privacy reasons; (2) users uninstall apps due to privacy concerns; (3) users concerns about the selling of personal data are largely confined to a small set of app categories; and (4) some users reward developers whose apps are privacy friendly with privacy positive reviews.

We apply our clustering and summarization to ten categories of apps and identify the large compact clusters within each category. From these, we identity five dominant privacy concerns across multiple app categories: apps that appear to request unnecessary permissions, collection of personal information, tracking, privacy controls and apps that may be selling personal data. Next, we present an overview of these dominant themes using our automatically selected representative reviews that summarize each cluster. While the dominant privacy themes that emerged from our analysis are generally well-known, we demonstrate the ability to automate and scale this process. Our initial findings illustrate the potential of such automated analysis. We conclude the paper with a discussion on implications for developers and app stores, limitations, and a summary of remaining challenges needed to further mature this type of analysis.

2 RELATED WORK

The primary use of NLP in the privacy space in the last few years has been to analyze privacy policies [28, 31, 40, 49, 53, 56, 60, 63, 65, 66, 80]. Numerous efforts have focused on identifying inconsistencies between an application's source code and its privacy policy [40, 53, 65, 66, 80] or it's description [28, 56, 60]. Other uses include developing privacy chat bots [32] and automated question-answer systems based on deep-learning techniques [31, 63] to help users understand data practices. In [49], the authors explore the use of NLP to help users avoid inadvertently sharing personally identifying information in social media, emails, and text messages. In [57] the authors used NLP to analyze app description texts, within a larger mechanism to determine when two apps are similar; this was used to inform developers when they may be requesting an unnecessary permission. None of these privacy-oriented studies applied NLP to analyze the text in user reviews.

User reviews have been analyzed using NLP, but not in the context of privacy. Pagano and Maalej [54] conducted an empirical analysis of iOS app reviews and found that reviews are not easy to automatically analyze given their unstructured forms. Some efforts have built classifiers to identify informative app reviews [15] or

[1] Due to space limitations, the full list of apps has not been included in this paper.

reviews useful for app maintenance [55]; others identified inconsistencies between user reviews and ratings, as well as performed topic analysis of negative reviews to shed light on why users dislike a given app [26] and unsupervised topic discovery composed of semantically similar user comments based on bidirectional NLP algorithms [43]; work has also been done to automatically match bug reports with related app reviews [44]. In the Requirements Engineering space there have been large bodies of work that focused on feature/requirements extraction from user reviews and/or applied a sentiment analysis to find out how users like a certain feature [30, 45, 68]. [29] developed a review summarization framework to categorize app reviews into categories (e.g., bug reports and feature requests), and also extracts aspects and their sentiments (e.g., interface is good/poor). Tian et.al. have shown that including crowd sourced review information in app update notifications was more effective at alerting users of invasive or malicious app updates, especially for less trustworthy apps [72].

The prior research that is closest to ours is [8, 51, 52]. Besmer et al. [8] have trained a logistic regression model to detect privacy app reviews, and have shown that privacy reviews have lower star ratings and more negative sentiment, but have higher engagement (more upvotes/downvotes). In both [51, 52], the authors developed a SVM classifier to extract reviews from the Play Store on the two topics of security and privacy. The key purpose of [52] was to determine if app reviews discussing security and privacy lead to changes in the app, and [51] focused on how actual app behavior influences users' security and privacy concerns. Using static code analysis, the authors in [52] were able to demonstrate that the presence of security and privacy reviews are predictive of security and privacy app updates in 60% of the cases they looked at. This is very encouraging, as it means developers do respond to issues raised in these types of reviews. As part of that work, the authors had to develop a security and privacy review classifier to extract those reviews. The authors in [51] use dynamic analysis to show users security and privacy concerns are justified in that the apps do often exhibit troublesome behavior along the lines indicated in reviews.

Our work differs from these efforts in several ways. First, our focus is on privacy, not security and privacy combined. (We acknowledge that security and privacy are interwoven, at times, especially when security issues such as account hacking or password management have privacy consequences). Second, their classifiers rely on a basic SVM/Logistic Regression models and follow a bag-of-words approach. We use state of the art deep learning based NLP models (e.g., USE and BERT) that offer multiple advantages. These transformer models are trained on large corpuses of text (e.g., BERT is trained on Wikipedia) and thus these models have the potential to generalize beyond the labeled examples. As explained more fully in Section 3.2.1, a bag-of-words approach is likely to miss context, and that matters for identifying privacy-related reviews across a broad set of privacy topics. Third, we work with a larger dataset. We manually labeled 11K examples (compared to 2.4K in [8], 4K in [52] and 6K in [51]). We ran our inference analysis on 287M reviews. Fourth, our method incorporates clustering and summarization whereas these prior efforts only included classification. This is needed since our end goal - to provide a way to report users perspectives at scale - is different.

There is an enormous body of usable security research on user perceptions towards mobile app privacy. Most quantitative work have found that people are very protective of their personal information when using apps [36, 77], and actively engage with offered privacy controls to safeguard their information [10, 39, 73]. User frustration due to apps requesting unnecessary permissions has been well studied [23, 38, 39, 64, 73, 75]. In fact, users were often surprised by the abilities of applications to collect data in the background [35, 71], and were concerned with possible risks associated with permissions [24]. Special attention has been on studying the privacy concerns arising due to users' location being tracked [3, 17, 22].

These prior studies reveal similar privacy concerns that we observe in our research. However, they are all done by surveying few hundred participants or interviewing 30 or less participants. Our methodology scales up to large numbers of user reviews and tracks these issues efficiently across all apps, and our findings show the prevalence of specific privacy concerns across app categories.

3 METHODOLOGY

Figure 1 illustrates an overview of our approach. On the left are examples of reviews, some of which are about privacy and one of which is not (marked in yellow). Our first task is to extract those reviews that are related to privacy. This is achieved by constructing a binary classifier to distinguish between the privacy and non-privacy reviews. In figure 1, the review about taking pictures has been filtered out in the sample list of "privacy reviews". Our second task is to identify the common fine-grained privacy themes within the privacy-relevant app reviews. For this we leverage K-Means clustering; *'k'* is a parameter that needs to be chosen upfront, and we propose a custom metric to choose the best value for *'k'*. We use the TensorFlow [1] machine learning tool for training a binary classifier and clustering the privacy-related app reviews. In the following sections, we describe how we generate test, validation and training data; present the design of our privacy classifier; and describe our method for clustering and summarizing privacy-related reviews.

3.1 Dataset Curation

In order to develop test, validation, training and inference datasets, we collected approximately 580M Play reviews in English published on the Play store between April 2014 and Feb 2020. Our dataset was anonymized in terms of app names, as each review was only labeled by its app category; however we were given an aggregated app count of 2M. We started with zero labeled data. As per the many methods for generating labeled data, as outlined well in [61], we use Expert-Hand-Labeling (by subject matter experts) for our test and validation datasets to ensure these are of the highest quality since they are used to evaluate the performance of our machine learning models. Prior work [52] hints that privacy related reviews may constitute less than 1% of all reviews, hence we were facing an extremely imbalanced dataset. To bootstrap this procedure and generate candidate reviews that are likely to be about privacy we did the following. We relied on two well known privacy taxonomies [4, 67] to set the framing for our initial definition and scope of privacy issues. We curated an initial seed list of n-grams inspired by these taxonomies. Our manual labeling team consisted of the authors

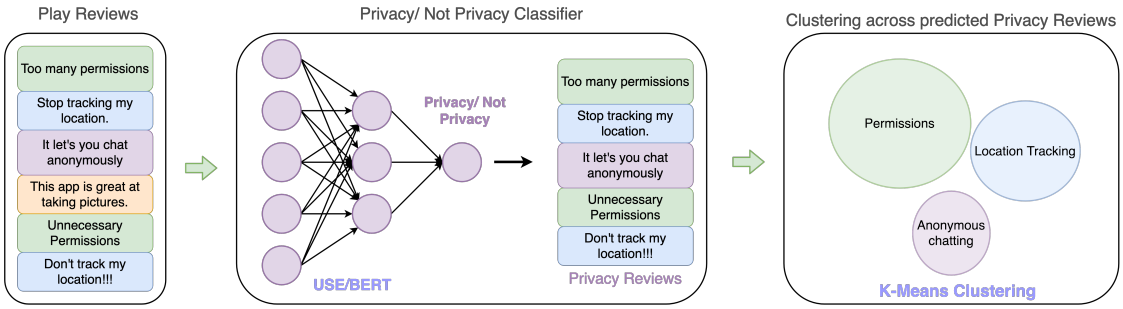

Figure 1: Method pipeline.

and three linguists. We filtered reviews based on this initial seed and conducted a preliminary manual evaluation. We looked to see the types of words and expressions that users employ to express privacy concerns.

There were two observations in this bootstrapping phase. First, filtering using 1-gram and 2-gram words can result in many false positives; for example, the single word "trust" can be used in "don't trust your teammates", which is not a statement about privacy. Thus, we decided to only use n-grams with $n >= 3$ for further filtering. Second, we converted our seed list of n-grams to regular expressions and expanded the set of expressions based on this manual exploration. Regex patterns allow us to succinctly capture grammatical variants of typical privacy statements. For every regex pattern, we checked to see if it occurred in at least 100 reviews that were privacy-related. Our regexes captured approximately 200 n-grams. We acknowledge that our list may not be complete and might have missed privacy issues or phrases used to discuss privacy.

We next selected ~11K app reviews for manual labeling. To ensure we ended up with enough labeled privacy examples, we selected 60% of the 11K reviews because they matched against our regex patterns, and the other 40% were ensured to not match any of the regexes. Each of the 11K reviews were manually examined and cross-labeled by three raters and a label (privacy or not privacy) was assigned. For the vast majority of reviews, all three labelers agreed. When this was not the case, discussion ensued until an agreement was reached. Among the 11,371 manually labeled reviews (ground truth), **6688** were labeled as **privacy** and **4683** were **not privacy**. The validation dataset used to evaluate the performance of the ML models after each training epoch was created by extracting **250 privacy** and **250 not privacy** reviews from this set. The remaining 10,871 reviews were treated as the test set for comparing the performance of different privacy classifiers.

Because each of our regexes (capturing 200+ n-grams) was verified – by checking for its appearance in at least 100 privacy-related reviews, and also based on our manual labeling exercise – we assume our regex list effectively constitutes a set of good quality heuristics. We use these to generate our training data, as per the method of *Heuristic-Supervision* as in [61]. We used roughly half of the reviews (~290M) to generate training samples. From this set we identified 250K app reviews that matched our privacy regex patterns and labeled them as privacy-related reviews; we then randomly sampled another 250K reviews that did not contain any of our privacy regexes and labeled them as not-privacy. Our privacy regex patterns could only identify 0.08% of reviews as being related to privacy. Finally, we used the second half of our collected reviews, namely 287M reviews, as our inference dataset. We performed classification, clustering and summarization on this set to understand users' top privacy concerns. The sizes of our training, validation and test datasets are shows in Table 1.

Table 1: Size of Data Sets

Datasets	Training	Validation	Test
Privacy	250 000	250	6438
Not-privacy	250 000	250	4433

Ethical Considerations: Our institution approved the use of this dataset because Play reviews are already public. In compliance with ethical training guidelines in our institution, we ensured that users' privacy were respected. We thus carried out the following. First, all researchers have been trained in ethical user research prior to this study. Second, the dataset was preprocessed to remove user identifiers and app names before the researchers were given access. The only accompanying metadata beyond the review text, was the app's category name and the publish timestamp. Third, access to this version of the dataset is limited to the authors of this paper.

3.2 Privacy Classifier

The simplest approach to extract privacy related reviews might be via a keyword list. However, it is non-trivial for multiple reasons to curate a comprehensive list of keywords. In addition to the false positive issue discussed in Section 3.1, the same word could mean different things depending on the context. For example, the occurrence of "invading" in a war game app review likely does not refer to a privacy concern, whereas it might in a review about a parental control app with a location tracking feature (e.g. "this app is invading my privacy"). Moreover, there could also be instances where a review does not contain privacy keywords but, based on its context, still be related to privacy. For instance, "don't want my friend accessing my email" does not contain any privacy specific keywords but is a privacy concern based on the context.

Traditional approaches to analyze text rely on the bag-of-words [33] representation or use word embeddings, such as Word2Vec [50] and GloVe [58], to encode text. Such systems do not encode information about the word sequence, and therefore cannot differentiate between reviews containing the same words in different order. For example, "delete cookies and website history" and "my article on history of cookies got deleted from website" use same words but

have different meanings. In addition to handling context, and word sequence, we also need to be able to correctly process reviews that are frequently unstructured, contain highly variable writing styles, grammatical errors and misspellings.

The above limitations establish the need to process reviews by incorporating robust NL Understanding components. In this work, we use state-of-the-art pre-trained natural language models, BERT [20] and Universal Sentence Encoder (USE) [14], to efficiently encode user reviews into an abstract representation. These models are better known to capture contexts. We provide here a brief background on the BERT and USE language models.

BERT: Bidirectional **E**ncoder **R**epresentations from **T**ransformers (BERT) [20] is a language representation model based on the Transformer architecture that has been trained on 3.3 billion word corpus. The model was trained on two tasks: "masked language modelling", where the aim is to predict the masked-out words of the input text using the information present in the surrounding words, and "next sentence prediction", where the model predicts the next sentence given the first sentence as the input. The large pre-trained BERT neural network model has had great success in NLU tasks, such as text summarization [41], question-answering [42], and has been shown to deliver impressive performance on downstream tasks even when working with small training datasets [69]. Hence, we fine-tune the pre-trained BERT model on our training set to create a binary privacy classifier.

USE: Universal **S**entence **E**ncoder is another deep neural network based model, that uses encoders (Transformer-based [74] or Deep-Averaging-Network based [34]) to learn meaningful sentence representations [14]. The model is trained on data from various sources, such as Wikipedia, discussion forums, web questions and answers, etc.; and has shown great performance in detecting fake news spreaders on Twitter [46] and in learning cross-lingual text representations [16]. Unlike BERT, we do not fine-tune the USE model, instead use the readily available pre-trained USE TF-Hub module [2] (which provides text embedding representations directly) and build a two-layer feed-forward neural network on top for creating our privacy classifier.

3.2.1 Models. We propose the following four model variants as candidates for our privacy classifier.

(Vanilla) BERT: We take the pre-trained BERT model, and add one additional network layer after the last layer for binary classification. We use a [CLS] token (start-of-sentence) to represent each review in its entirety. Thus our additional layer simply transforms the embeddings learnt for the [CLS] token to the two classes, i.e., privacy and not privacy. We fine-tune the BERT model on our training data for three epochs, and choose the best epoch model based on the performance on the validation dataset.

Sentiment-aware BERT (BERT-SST): From a preliminary analysis, we found that the privacy review texts generally have a negative sentiment. We use this information to further strengthen our BERT-based privacy classifier, by first fine-tuning BERT for sentiment classification task. Using similar model architecture mentioned above for (Vanilla) BERT, we first train the model on a text-sentiment dataset: Stanford Sentiment Treebank[3], that contains text and its binary (positive or negative) sentiment label. We then fine-tune this model on our training data to create the privacy classifier. We simply map the privacy class to the negative sentiment class, and not-privacy to the positive one. We fine-tune the sentiment-aware BERT model for three epochs and choose the best epoch model based on the performance on the validation set.

USE: We extract a new 512-dimensional embedding representation for each review by passing the reviews through the pre-trained USE model that is based on Deep Averaging Network encoder. We then pass the embedding through a feed-forward neural network with 2 layers and 512 hidden units each. The output of the 2^{nd} layer is then mapped to the two classes, i.e., privacy and not-privacy. The model is trained with the objective to maximize the probability of the correct class, and thus we use cross-entropy loss to optimize the network. We train the model for 20 epochs and chose the best model based on performance on the validation set.

Ensemble Model: In our experimentation, we noticed inconsistent label assignments by each of the three privacy classifiers above. (This is understandable as the underlying BERT and USE models are pre-trained on different datasets with distinct characteristics.) To better understand the scenarios where USE and BERT models were making different decisions (privacy or not-privacy), we manually examined about 1000+ reviews that had different labels from the two models. Below is an example of a statement that USE classified as *privacy* whereas BERT labled it *not-privacy*.

> *"you can record call automatically, record anonymous calls, record important calls ... you can choose the phone numbers in the phonebook or recording call automatic. the list of recorded files will be stored and streamlined for you in the phone call recorder."*

The following is an example of a review that BERT predicted as privacy whereas USE did not.

> *"not having tamil channels and sports channels which is the way of looting the trust from the customer service"*

The first example is ambiguous as making anonymous phone calls could be perceived as being related to privacy. However, this review mainly lists app features. The second example mentions trust however this isn't a privacy issue but instead perhaps one of feeling excluded due to a language not supported in the app. We wouldn't consider either of these to be about privacy.

To reduce such ambiguities and improve our confidence in the privacy review identification, our Ensemble model considers a review to be about privacy, if and only if, all three classifiers (USE, BERT and BERT-SST) labeled it as privacy. Note that this makes our model conservative (i.e., we will underestimate the number of privacy reviews) since we are choosing to focus on precision rather than recall. For our purposes of broadly understanding users' privacy concerns with mobile apps, we prefer to have less noise in our clusters. We acknowledge that a developer using such a method may opt for high recall to be sure not to miss any particular issue.

3.2.2 Performance of Models. We evaluate our four privacy classifier models (see Table 2) on the test dataset. High recall numbers indicate that the model is able to correctly identify most of the privacy-related reviews, and a high precision indicates that the model rarely labels a not-privacy review as being related to privacy. From the table, we see that the USE model has a high recall and a

[2] https://tfhub.dev/google/universal-sentence-encoder/1
[3] https://nlp.stanford.edu/sentiment/

Table 2: Performance of Models tested

Model	Accuracy	Precision	Recall	F1-score	AUC
USE	89.71	0.88	0.95	91.40	97.19
BERT	91.75	0.96	0.89	92.56	94.61
BERT-SST	93	0.93	0.90	91.70	90.6
Ensemble	91.85	0.98	0.87	92.49	98.18

relatively low precision, meaning that it has more false positives (not-privacy reviews labeled as privacy). On the other hand, both the BERT based models have a higher precision but low recall compared to USE. The BERT-SST model has higher accuracy than BERT, but looking at the F1 scores, vanilla BERT model performed better than BERT-SST. As expected the Ensemble model has the highest precision; it also obtains the highest AUC value, and a good F1 score. We use thus the Ensemble model in the remainder of this work for our classification task.

We did a qualitative analysis to see if our ensemble classifier was able to generalize its learning beyond the terms and expressions in our regex patterns. First we checked to see if our classifier learned any concepts not included in our regex patterns. For example, we did not include any terms related to "anonymity" or "anonymous" in our regexes, however we did find a number of reviews that mention "I like the anonymity ...". Our model likely learned that the words related to "anonymous" are often associated with privacy because of the following. The following real review - *"i don't want a personalized profile full of surveillance. anonymous access is a preferred"* - could have been flagged as privacy because of the word "surveillance" (that was in our regexes). Since this review also contains the word "anonymous", the classifier learns to associate this with the privacy label (given enough similar examples). Second, we compared the fraction of privacy reviews that our regexes alone match (0.08% in the 290M reviews used to generate our training data, Section 3.1), with those extracted by our classifier, namely (0.15% from our 287M reviews test data). This shows that our classifier does generalize beyond the terms and expressions in the regexes as it identifies roughly twice the amount of content as our regexes.

3.3 Clustering and Summarization

The next step in our pipeline is Clustering and Summarization of the privacy related reviews to tease out the different privacy concerns users describe. We know from prior work as well as our curated set of n-grams, there are a multitude of things users might write about, such as personal data collection, privacy controls, location tracking, a feeling of being spied on, third party data sharing, new privacy features, consents, etc. We refer to these as *privacy themes*. Since app reviews do not have fine-grained labels for such privacy themes, we use unsupervised learning (specifically clustering) to identify the common privacy issues. We use K-means clustering as our approach to clustering because it is simple and broadly used, and leave exploration of other clustering solutions as future work. We apply K-means to the set of privacy reviews per app category (Games, Parenting, Tools, etc). The motivation for studying categories independently is that the number of reviews across app categories was highly variable (ref Table 3). Analyzing all the reviews together would have not highlighted users' concerns in app

Table 3: Number / proportion of privacy reviews per category

App Category	Total # of Reviews	Proportion of Privacy Reviews
Dating	503656	0.81%
Parenting	133408	0.68%
House & Home	291982	0.59%
Communication	17742507	0.59%
Maps & Navigation	1737291	0.52%
Tools	25038585	0.45%
Health & Fitness	3569543	0.44%
Medical	881141	0.34%
Weather	1491430	0.31%
Auto & Vehicles	547314	0.30%
Social	16719868	0.27%
Events	102623	0.23%
Photography	9364745	0.22%
Libraries & Demo	226555	0.17%
Entertainment	17547266	0.13%
Beauty	199974	0.11%
Art & Design	598982	0.09%
Finance	10604716	0.08%
Personalization	7565285	0.06%
Music & Audio	10870967	0.06%
Shopping	10698465	0.05%
Games	112635058	0.05%
Sports	11292326	0.05%
Comics	553063	0.05%
Lifestyle	7150696	0.04%
Travel & Local	4595332	0.03%
Productivity	9037059	0.03%
Food & Drink	2322001	0.01%
News & Magazines	3281372	0.01%

categories with lower numbers of reviews. Analyzing the categories independently helped us identify issues which may be prominent in one category but not in others (e.g., tracking and selling data was not a key concern in the Games category).

Once these clusters are determined, we summarize the topic discussed in a cluster by selecting some highly representative reviews for each cluster. After independently reviewing the top representative reviews and labelling the clusters in each category, we performed a second round of annotation where we cross-checked the cluster labels across categories and mapped clusters discussing the same concerns to the broader topic.

K-Means clustering: We use the 512-dimensional USE embedding generated for each review to perform clustering. We use embeddings derived from our USE-based model instead of BERT-based models as it lead to a higher AUC score (refer Table 2). The clustering is performed within an app category, and *Cosine Distance* is used as the distance metric. Like any other clustering task, the challenge here is: *how to determine a good k value* for the number of target clusters, without knowing ahead of time how many distinct themes users may be writing about. There exist various metrics in the literature to choose k, such as the Silhouette Score[62], Dunn Index[21], CH-Score[12], etc. These metrics primarily reward well separated and compact clusters, which is also our goal. However, these metrics implicitly assume the following: 1) a sample belongs to only one cluster (or only belongs to one privacy theme); and 2) all samples belong to some cluster, i.e., no sample is considered to be an exception or outlier. For privacy related Play reviews, these assumptions do not always hold. Consider this example:

> *"1 star for forcing people to create an account. no it's not necessary for the user. it's apparently necessary for [APP_NAME_REDACTED]. permissions they demand aren't necessary and invasive. this app is more like spyware. clearly [APP_NAME_REDACTED] is making money off of people's personal and private information. i don't recommend this product. at minimum, if you're going to use this app, give them fake information. however they won't even let you use the app if you don't give them location tracking information. undoubtedly, they'll be another company in the news soon for exploiting customer privacy."*

This example review touches on multiple privacy themes, including unnecessary permissions, spyware, location tracking, and general privacy exploitation. There is potential that each of these fine-grained topics produce its own cluster when running K-Means; so this example review would be hard to place within one cluster as it might naturally be on the border of multiple clusters. Therefore, we aim to limit the influence of such reviews on the cluster formation by optimizing for the creation of well-formed clusters – clusters that predominantly discuss a single privacy issue. The next example illustrates the second assumption:

> *"i wish it worked better for pregnancy milestones, bump pics, etc. all the reminders assume your kid is born regardless of entering the due date. also, the pictures take forever to load. i uploaded multiple pics for one day and my sister thought it was just 1 pic because of the slow load time. COMPANY_NAME_REDACTED is faster and i can make a private group so my pics aren't super public."*

In this example, the user seems to be primarily expressing dissatisfaction about the reminder mechanisms and upload speeds. They also mention that they would like a new privacy feature, namely the ability to create private groups. The request for a private group option does make this a legitimate review about privacy. However, we did not find similar requests in other reviews in parenting apps, and thus this review is an outlier. We aim to limit the influence of such reviews in the clusters produced.

Given this, we have the following goals. Firstly, we aim for well separated clusters. Secondly, we aim to have a high number of clusters that have limited mixed-concern reviews and to minimize the number of outliers. To achieve this, we propose a *summarization metric* that will reward a value of k based on these criteria.

To address this first goal we use the following construct. Let S_i denote the center of cluster i. We define $dist_k$ as:

$$dist_k = \min\{\delta(S_i, S_j) \forall \{i, j\}\} \tag{1}$$

where $i \neq j$ and $i, j \in [0, k-1]$. $\delta(S_i, S_j)$ is the cosine distance between two cluster centers. $dist_k$ represents the minimum pairwise distance among all pairs of k cluster centers. Maximizing $dist_k$ ensures that the cluster centers be far apart in the vector space. We iterate through multiple values of k and select the one that maximizes this metric. This formula varies compared to existing metrics, where an average of inter-cluster distance is generally taken into account. Here we ensure that the minimum inter-cluster distance is the highest for the chosen k value.

To address our goals of limiting the influence of mixed-concern reviews and outliers, we want intuitively to "ignore" reviews that are *loosely associated* with a cluster, as well as the mixed-concern reviews as these are likely to be "borderline" between multiple

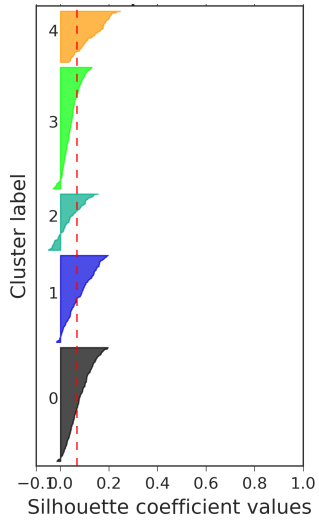

Figure 2: Silhouette Scores using K-Means for k=5

clusters. We aim to count the reviews within a cluster that are closely related, and refer to them as *upvotes* (defined below). We use the *silhouette score* to do this. Recall that the silhouette score $s(i)$ formulates how close each point is to its cluster center and how far it is from the nearest neighboring cluster, namely

$$s(i) = \frac{b(i) - a(i)}{\max(b(i), a(i))}$$

where $b(i)$ is the lowest average distance between i-th point and any cluster of which it is not a member, and $a(i)$ is the average distance between i and all the other points of the same cluster. As an example, we run K-Means with $k = 5$ on reviews identified as privacy in the *Parenting* category and compute the silhouette scores for each review, as shown in Figure 2. A negative silhouette score indicates that the review is assigned to a wrong cluster, as it is closer to the neighboring cluster. The dotted red line in Figure 2 represents the average silhouette score across all the reviews in the category. A low silhouette score (close to zero) indicates that the review is very close to the cluster boundary (data points/reviews which are on LHS of the red line, but still positive), and a high silhouette score indicates the review is closer to its own cluster center (data points/reviews on the RHS of the red line). We refer to reviews with silhouette score higher than the average silhouette score as *upvotes* for a given theme in a cluster. From Figure 2 we see that it could be useful to retain clusters 0, 1, and 4 that have a larger amount of upvotes and ignore clusters 2 and 3. Clusters 2 and 3 are likely to be poor quality since most of the silhouette scores are negative or below the average silhouette score. Hence in this illustrative example, we would aim to have 3 final clusters and simply not capture the ignored clusters which are unlikely to contribute to top issues.

We consider clusters to be compact when they have a high number of upvotes, and a low number of mixed-concern reviews in a given cluster. Therefore, we can rephrase our goal as aiming for a k that results in high number of *compact* clusters where a compact cluster is defined as one in which at least 30% of the samples are

upvotes. We refer to the number of compact clusters identified as M_k, for a given k.

We combine the above two principles, and define a new **Summarization Metric** as follows:

$$Summarization\ Metric\ = (dist_k * M_k) \quad (2)$$

We iterate through $k = 2, ..., 10$ and chose k for which the summarization score is highest, as we want to increase both $dist_k$ (distance between cluster centers) and M_k (number of compact clusters). We use M_c to denote the final number of compact clusters identified for the chosen k. We consider k between 2 and 10 because in our initial exploration using our methodology, we focus on dominant privacy concerns. However using a larger k would enable an analyst to look through the long-tail of privacy concerns.

We may end up with tens of thousands of reviews in a single cluster, thus it is important to summarize them in a way that captures the primary concern. We do this by selecting a few specific reviews that can be considered as representative of the reviews in the entire cluster. By summarizing this way, we capture the cluster topic *in the users' own words*. We rank reviews within a cluster according to their silhouette scores and use the top ten reviews with the highest scores as the representative reviews. We carefully analyzed these cluster representatives manually across all clusters and verified that these reviews indeed illustrate the main topic in each cluster. In the next sections, we select quotes/reviews that came from different categories to show the breadth of the issues across app categories.

3.4 Limitations

We used K-means as our first approach to clustering because it is simple and broadly used. However, our clusters are quite uneven in size, and thus more sophisticated approaches (such as Affinity Clustering [7]) could yield improved performance.

This kind of work is inherently hard because the definition of privacy is not exact. We relied on previously accepted privacy taxonomies, three professional linguists, and our own experience reading huge numbers of privacy reviews. A clearer sense of the distinction (or accepted overlaps) between security, censorship, harassment and privacy could take the form of agreed upon guidelines by a community of domain experts.

Although user perspectives within a cluster are automatically summarized by the representative sentences (ranked by silhouette scores), there is still a manual step in assigning thematic labels to each cluster (i.e privacy controls, selling data, etc.) - that we did by reading the top 20 reviews per cluster and finding a label based upon our interpretation of those reviews. NLP methods for topic labeling [48, 78] could automate this step - although their efficacy in the privacy domain needs to be evaluated.

4 PRIVACY THEMES

In this section, we present the top privacy themes that are associated with different app categories. To do this, we run our clustering analysis on the previously extracted privacy reviews. Due to space limitations, we present findings from 10 Play Store app categories instead of all 29 categories. We selected categories that were either large or ones where we expected privacy concerns might arise.

Recall that our clustering iterates through k=2...10 and picks the best k (number of clusters) according to our summarization metric. For each of our 10 app categories, we found that k varied from 6 to 9. We looked at the compact clusters across these 10 categories, and identified 5 themes that were dominant across multiple categories. Within each cluster considered, we ranked the reviews in those clusters by silhouette score. Recall that the top ranked reviews essentially represent the topic of the cluster as they are central to the cluster. For each of our selected clusters, we manually read the top 20 most representative reviews, and assigned a short thematic label to the cluster for ease of presentation and for summarization.

For example, we assigned the thematic label **too many permissions** to clusters whose representative reviews frequently mention that an app requests more permissions than what seems needed. These reviews are referring to the Android permissions such as location, contacts, microphone, camera, etc. This theme was present across all app categories, with location permission being the highest occurring sub-theme. In other large clusters, many reviews comment on the collection of **personal information**. Such reviews typically refer to information such as address, email, profile info, etc. Other themes we identified include **privacy controls, tracking** and **selling data**. The cluster of reviews assigned the **privacy-controls** theme captures reviews that either discuss the existing privacy controls of an app, or ask for a new privacy feature to be added to the app. Table 4 shows our themes and app categories; there is an asterisk in each table entry if that theme appears as the top-5 issue for the respective app category. A blank cell means that this topic did not surface as one of our M_c compact clusters for a given app category (there might still be reviews on this topic, but they are not large enough to generate a compact cluster).

While these privacy themes have been identified in qualitative studies before [3, 25, 36, 47, 79], **our analysis is the first large scale work to confirm these findings empirically**. Moreover, we are able to quantify the volume per theme thereby enabling us to rank these issues, and we are able to show which issues occur across multiple app categories (e.g. permissions) and which occur in only a few categories (e.g. selling data).

Given the scope of each of the themes, our purpose here is not to dive into details individually, but instead to summarize succinctly how users express these pain points. The included examples below come from our top ten representative reviews per cluster (topic), per app category. Because these examples are central to clusters of many thousands of similar reviews, they summarize the views of large groups of users. Our clusters ranged from a few thousand up to 20K in size. Among these representative reviews, we selected examples from different categories to illustrate that privacy concerns are rarely category specific. A more detailed look into each specific privacy theme is left as future work.

4.1 Concern 1: Too Many Permissions

The largest cluster, in each of our 10 categories, contains reviews that make comments about the app asking for too many permissions. This is expected as permissions are the gateway to accessing personal information, and prior work has pointed to excessive permissions being a major concern [23, 38, 39, 64, 73, 75]. Examples include:

Table 4: Privacy Topic themes occurring in App Categories

	Tools	Games	Communi-cations	Maps & Navigation	Enter-tainment	Parenting	Health & Fitness	Photo-graphy	Medical	Auto
Permissions	*	*	*	*	*	*	*	*	*	*
Personal Information		*		*	*	*	*	*	*	*
Privacy Controls	*	*	*			*		*		
Tracking			*		*	*		*	*	
Selling Data			*		*				*	

(1) *"i really enjoyed this app until i realized it had access to all of my photos , media and storage ! why the heck would a simple sudoku app need that ! ? i'm now more careful at checking the permissions before i install anything . i promptly deleted this app and installed a similar sudoku app that doesn't require such ridiculous permissions and it's just as good"* (Games)
(2) *"why does this app require access to my contacts? purpose? i take my privacy very seriously. no one should install this app if you value your privacy."* (Maps & Navigation)
(3) *"good app, but i hate it when apps request permissions it has no business for (in this case, access to my contacts), no thanks"* (Auto & Vehicles)

The first two reviews focus on whether the data collected is really needed for the functionality of the app. In example 1, the disconnect between the permissions asked for and the user's perception of the app functionality caused this user to uninstall the app, and switch to a more privacy-friendly app. In example 2, the user is asking for the purpose of collecting contacts (presumably because it is not clear from the context). The importance of sharing purpose with users has been established in the academic literature [6, 36, 73], and even in Android's best practice guidelines[4]; it is interesting to now see users effectively demanding that. It is important for developers to learn when users have such concerns, as they can be mitigated by providing explanations [40, 70] to address the "purpose" type questions. In example 3, users state they feel that data collection from some permission requests is "unacceptable" or a "risk".

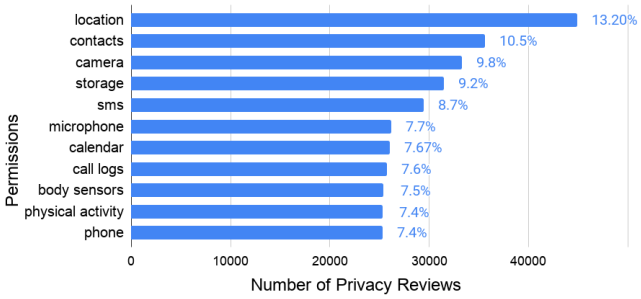

Figure 3: Privacy reviews that discuss permissions

We also examined which reviews mention a specific permission, such as microphone, location, etc. To do this we used simple keyword matching such as *phone, call log, contact, microphone, location, sms*, and checked if the word *permission* appeared somewhere in the review. Figure 3 shows the number of privacy reviews that mention each permission group as well as the percentage of reviews they represent. We see that the location permission is the most discussed permission, appearing in over 40,000 reviews (13% of the privacy reviews). Contacts is the second most discussed permission.

[4]https://developer.android.com/training/permissions/usage-notes

4.2 Concern 2: Too Much Personal Information

Reviews in which users complain that too much personally identifiable information (PII) is being collected is the second dominant privacy concern, and occurred in eight of our ten categories examined herein. While prior work has shown that users are concerned about too much personal information being collected, both in the context of mobile apps [36, 77] and generally [2, 11], unlike ours they do not rely on experiences users have using their own devices in the wild. Our analysis additionally shows that users warn others not to download an app explicitly because of PII collection, as shown in the following two examples:

(1) *"when it won't let you play the game unless you agree to let it use your information , it ain't worth playing . if you want privacy , don't download it"* (Games)
(2) *"dont do this!!! youre only giving the app your personal info! what you look like, your finger prints, everything! i downloaded this app just to warn everyone not to give this app permission to anything on your device! ... this is dangerous!"* (Entertainment)

Other users indicate their suspicion i.e., there is no good reason for the data collection. These suspicions, which are in essence requests for data collection justification, reflects the same issue we saw in the reviews about **too-many-permissions** – applied to different data items. In the first two examples below, the users are clearly quite annoyed. The user in the 3rd example implies that they might have uninstalled the app because of this reason.

(1) *"why the hell you need access to everything, you are service provider or intruder in privacy."* (Auto & Vehicles)
(2) *"horrific registration process; requiring personal information irrelevant to the application's purpose."* (Maps and Navigation)
(3) *"why does it need my device ids and the phone numbers of my callers. who i communicate with is none of their business. this app appears to be collecting more info than it requires to offer it's services ... they have no respect for my privacy, the app was useful."* (Health and Fitness)

Yet other users express themselves with terms relating to theft or harassment, as the following two examples convey. Clearly trust is impacted if users interpret an app's behavior this way.

(1) *"started stealing personal information. phone numbers and messages are being read by this application."* (Maps and Navigation)
(2) *"too much personal info: why do you need so much personal info you creeps? are you the offenders looking for prey? jeez"* (Parenting)

4.3 Other Dominant Privacy Concerns

Privacy Controls. Comments about privacy controls were a common issue across 5 app categories, namely in Communication, Parenting, Photography, Tools and Games. The fraction of reviews discussing this theme ranged from 24% for Communication apps to 6% for Games. The reviews about privacy controls show an interesting breadth, from explicitly requesting privacy controls to be easier (example 1), to frustration with privacy controls (example 2), to requesting new features (e.g. private chat in example 3). In example 2, the user is trying to hide some photos within app, yet does not appear to be able to do so successfully. To address the user frustration with using the offered privacy controls, personalized privacy assistants have been proposed [39, 73].

(1) *"don't share my pics on public domain without my permission, when people search for specific location to visit pics appear, so make privacy settings easy."* (Photography)
(2) *"lock the pics you dont wanna see then they just get copied right back to the gallery doesnt hide anything."* (Tools)
(3) *"i love this game ! it is so addictive and fun ! ... the only thing i would want to change add is private chat."* (Games)

Tracking. We saw tracking as a dominant concern in 5 app categories, namely Communications, Entertainment, Parenting, Photography and Medical apps. The fraction of privacy reviews that discuss tracking ranged from a minimum of 10% in Entertainment apps up to 38% in Parenting apps. Privacy concerns arising due to users' location being tracked have been well studied [3, 17, 22], but our analysis shows that location isn't the only attribute people are concerned about being tracked. Purchase history, contacts, and other personal information are also important. As the examples below show, users sentiment on this topic can range from annoyed to angry. Since our methods can extract reviews on this issue, it permits future work on sentiment analysis and perhaps a deeper exploration into which types of data are more sensitive. (Note, the third example mentions private information as well as spying and thus provides an example of reviews that fall on the boundary of two themes (one of the challenges in Section 3.3)).

(1) *"warning: this app is spyware and will track all your location info and purchases."* (Entertainment)
(2) *"spyware, data miner. will not connect until you grant access to your phone location and data. contacts and location are personal information. cameras don't need this to function. you lie to public!"* (Photography)
(3) *"i agree with protecting your child. but when your a teen like me you feel like you cant breath without constantly being watched. it's like being stalked by your own family, and as if you can't trust them. there's a boundary between protection and privacy. and this is stepping over the line."* (Parenting).

Selling Data. Users' have previously expressed concerns with service providers selling their data [18, 76]. We show here that upset due to the perception of personal data being "sold" to third parties also appears in mobile apps. Unexpectedly though, we found this to be dominant in only 3 app categories - that of Communications, Medical and Entertainment apps. In the Communications apps, we observed approximately 20% of the privacy reviews mentioning selling user data, making this a significant privacy issue for Communication apps. The examples below hint that this issue appears to make users feel undermined, as the comments are quite cynical. We also see that users imagine their data being sold off to a variety of recipients, including to companies, the NSA, and spammers.

(1) *"i would highly recommend staying away from this cash grab of an app and move to an app that actually cares about the data it collects about you. i feel like this company, and application are only used to track your data and information and selling it off to the highest bidder, sad thing is, they make you pay for it, so you're essentially paying to get your information stolen."* (Medical)
(2) *"they are selling patient's personal data to corporates. once u use [COMPANY-NAME-REDACTED], u will start getting mails and call from labs for medical tests. beware."* (Medical)
(3) *"what i do not understand is why [COMPANY-NAME-REDACTED] had to make another way that they could sell our private information to third parties like the NSA."* (Communications)

Privacy Positive Reviews. During the above exercise processing the top privacy pain points, we found - to our surprise - that while most privacy reviews are negative, we do see some privacy positive reviews. In these reviews, users mention that they like the privacy controls (examples 1 and 2) or that they are grateful the app is not collecting unnecessary information (example 3). The very presence of such reviews indicates that developers can be rewarded by privacy-friendly design. Examples include:

(1) *"this is one of the best photo sharing apps out there. no need to share your children's whole lives on social media and mess around with tons of privacy settings. you invite who you want to your album and can share privately with your partner or the whole family."* (Parenting)
(2) *"good way of keeping photos private."* (Photography)
(3) *"it was great fun playing this. i like that they don't want access to your private information unlike other apps."* (Games)

While these five themes were dominant concerns, there were smaller review clusters that touched upon other topics, such as "ads related to personal information" and "safety concerns due to personal information leakage". The diversity of privacy issues is broad thus making it challenging to provide examples of all topics, due to lack of space.

Overall we found many of these reviews to be fairly privacy savvy - many questioned the purpose of data collected and demanded justification, while others suggested specific privacy controls they would like to see added. Our quotes from representative reviews show that it is not uncommon for users to warn other others to stay away from an app for privacy related reasons. The dominant issue users are concerned about is the collection of too much personal data, be it from app permissions or PII. We found this to be true across nearly all app categories. We found reviews of users who uninstall apps due to privacy related reasons; this

is important to know as we suspect that developers are not completely aware when this happens. We also found privacy positive reviews and this indicates that developers can be publicly rewarded for privacy friendly behavior.

5 IMPLICATIONS

We illustrate how our method enables developers to improve the privacy of their products, and how app stores can leverage our findings to design better privacy tools and best practices.

Organizing and understanding user feedback for a particular app in a much more meaningful and actionable way. Today, the Play Developer Console [27] surfaces all privacy related user reviews under a theme called "Privacy". This makes it difficult to effectively understand user feedback and identify specific pain points. Instead, embedding our methodology into such developer feedback channels would allow app authors to address nuanced privacy concerns directly. For instance, concerns on:

- *Too many permissions*: if a developer sees many users complaining that they do not see the purpose for a permission request, then a guideline would suggest to provide a meaningful explanation or to remove the permission entirely.
- *Selling Data*: if a set of reviews shows much concern about selling personal data and the app does not engage in such behavior, then a developer would be advised to provide an educational intervention to clarify this misunderstanding.
- *Privacy controls*: if reviews identify missing features or capabilities then these can be translated into product requirements or bug fixes [43, 44].

Guiding the design of new tools for developers to provide transparency on their data-collection practices, such as the Apple Privacy Nutrition Labels [5], and its upcoming equivalent in the Play Store [59]. Understanding user concerns at scale and in depth via our methodology could inform the design of such transparency initiatives.

- Analyzing user concerns about *tracking, selling data, and too much personal information* could shed light on what type of labels are useful and how they should be designed. The two label systems above are a good step forward as they encourage developers to be upfront about their collection and sharing practices. Such labels could increase user trust and help them differentiate between the privacy stance of competing apps. Although proper use of such labels is turning out to be challenging [37].
- In the long term, our methodology will enable refining such labels and tailor them to the specific concerns of users. It will also allow monitoring of emerging privacy concerns.

Mechanism for providing insight into users perceptions of privacy. Beyond looking at top issues, one could evaluate how privacy issues vary by country, by language, by time, or compare privacy issues between children's apps versus regular apps. Running sentiment analysis on such privacy text data sets, using tools such as Stanford's NLTK [9], would enable large-scale comparison of sentiment across privacy issues and help developers and app stores prioritize what to fix first. These insights would be very valuable, especially for small scale apps that do not have resources to analyse their reviews another way.

Nudging developers towards better privacy practices. For instance, today the Play Developer Console nudges developers to remove permissions that are not used by apps in its peer groups [57]. By using our methodology, these nudges can be expanded to the other privacy concerns identified in this paper. For example, to mitigate tracking concerns, developers could be informed of malicious ad libraries as opposed to those that are much more privacy safe.

6 CONCLUSION

Understanding privacy trends is key to address systemic concerns across ecosystems and populations. To date, large-scale evidence of where the main privacy concerns lie for app users has been lacking. This paper is the first to provide a methodology that enables automated analysis and succinct summarization of privacy feedback, on a large scale. Our methodology can act as a mechanism for key ecosystem stakeholders to be responsive to evolving societal concerns regarding privacy. As new technologies come to the fore and as the risks of large-scale data harvesting become more apparent, this is a stepping stone towards systematic understanding of privacy concerns at scale.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/ Software available from tensorflow.org.

[2] Mark S Ackerman, Lorrie Faith Cranor, and Joseph Reagle. 1999. Privacy in e-commerce: examining user scenarios and privacy preferences. In *Proceedings of the 1st ACM Conference on Electronic Commerce*. Association for Computing Machinery, New York, NY, USA, 1–8.

[3] Hazim Almuhimedi, Florian Schaub, Norman Sadeh, Idris Adjerid, Alessandro Acquisti, Joshua Gluck, Lorrie Faith Cranor, and Yuvraj Agarwal. 2015. Your location has been shared 5,398 times! A field study on mobile app privacy nudging. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 787–796.

[4] A. Antón and J. Earp. 2003. A requirements taxonomy for reducing Web site privacy vulnerabilities. *Requirements Engineering* 9 (2003), 169–185.

[5] Apple's Privacy Nutrition Label 2022. App privacy details on the App Store. https://developer.apple.com/app-store/app-privacy-details/. Accessed: 2022-01-31.

[6] A. Barth, A. Datta, J. C. Mitchell, and H. Nissenbaum. 2006. Privacy and contextual integrity: framework and applications. In *2006 IEEE Symposium on Security and Privacy (S P'06)*. IEEE, Berkeley/Oakland, CA, USA, 15 pp.–198. https://doi.org/10.1109/SP.2006.32

[7] MohammadHossein Bateni, Soheil Behnezhad, Mahsa Derakhshan, MohammadTaghi Hajiaghayi, Raimondas Kiveris, Silvio Lattanzi, and Vahab Mirrokni. 2017. Affinity clustering: Hierarchical clustering at scale. *Advances in Neural Information Processing Systems* 30 (2017).

[8] Andrew R Besmer, Jason Watson, and M Shane Banks. 2020. Investigating user perceptions of mobile app privacy: An analysis of user-submitted app reviews. *International Journal of Information Security and Privacy (IJISP)* 14, 4 (2020), 74–91.

[9] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", US.

[10] Bram Bonné, Sai Teja Peddinti, Igor Bilogrevic, and Nina Taft. 2017. Exploring decision making with Android's runtime permission dialogs using in-context surveys. In *Symposium on Usable Privacy and Security ({SOUPS})*. USENIX Association, USA, 195–210.

[11] Tom Buchanan, Carina Paine, Adam N Joinson, and Ulf-Dietrich Reips. 2007. Development of measures of online privacy concern and protection for use on the Internet. *Journal of the American society for information science and technology* 58, 2 (2007), 157–165.

[12] Tadeusz Caliński and Jerzy Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods* 3, 1 (1974), 1–27.

[13] Ann Cavoukian. 2003. The Security-Privacy Paradox: Issues, misconceptions, and Strategies. (2003).

[14] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder. *CoRR* abs/1803.11175 (2018). arXiv:1803.11175 http://arxiv.org/abs/1803.11175

[15] Ning Chen, Jialiu Lin, Steven C. H. Hoi, Xiaokui Xiao, and Boshen Zhang. 2014. AR-Miner: Mining Informative Reviews for Developers from Mobile App Marketplace. In *Proceedings of the 36th International Conference on Software Engineering* (Hyderabad, India) *(ICSE 2014)*. Association for Computing Machinery, New York, NY, USA, 767–778. https://doi.org/10.1145/2568225.2568263

[16] Muthuraman Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Learning Cross-Lingual Sentence Representations via a Multi-task Dual-Encoder Model. *CoRR* abs/1810.12836 (2018). arXiv:1810.12836 http://arxiv.org/abs/1810.12836

[17] Saksham Chitkara, Nishad Gothoskar, Suhas Harish, Jason I Hong, and Yuvraj Agarwal. 2017. Does this app really need my location? Context-aware privacy management for smartphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3, Article 42 (2017), 22 pages.

[18] Margaret S Crocco, Avner Segall, Anne-Lise Halvorsen, Alexandra Stamm, and Rebecca Jacobsen. 2020. "It's not like they're selling your data to dangerous people": Internet privacy, teens, and (non-) controversial public issues. *The Journal of Social Studies Research* 44, 1 (2020), 21–33.

[19] Mina Deng, Kim Wuyts, Riccardo Scandariato, Bart Preneel, and Wouter Joosen. 2011. A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements. *Requir. Eng.* 16, 1 (2011), 3–32.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[21] J. C. Dunn. 1973. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics* 3, 3 (1973), 32–57. https://doi.org/10.1080/01969727308546046

[22] Kassem Fawaz and Kang G Shin. 2014. Location privacy protection for smartphone users. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. Association for Computing Machinery, New York, NY, USA, 239–250.

[23] Adrienne Porter Felt, Erika Chin, Steve Hanna, Dawn Song, and David Wagner. 2011. Android permissions demystified. In *Proceedings of the 18th ACM conference on Computer and communications security*. Association for Computing Machinery, New York, NY, USA, 627–638.

[24] Adrienne Porter Felt, Serge Egelman, and David Wagner. 2012. I've Got 99 Problems, but Vibration Ain't One: A Survey of Smartphone Users' Concerns. In *Proceedings of the Second ACM Workshop on Security and Privacy in Smartphones and Mobile Devices (SPSM '12)*. Association for Computing Machinery, New York, NY, USA, 33–44.

[25] Casey Fiesler and Blake Hallinan. 2018. "We Are the Product" Public Reactions to Online Data Sharing and Privacy Controversies in the Media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13.

[26] Bin Fu, Jialiu Lin, Lei Li, Christos Faloutsos, Jason Hong, and Norman Sadeh. 2013. Why People Hate Your App: Making Sense of User Feedback in a Mobile App Store. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Chicago, Illinois, USA) *(KDD '13)*. Association for Computing Machinery, New York, NY, USA, 1276–1284. https://doi.org/10.1145/2487575.2488202

[27] Google 2022. Play Developer Console. https://developer.android.com/distribute/console. Accessed: 2022-01-31.

[28] Alessandra Gorla, Ilaria Tavecchia, Florian Gross, and Andreas Zeller. 2014. Checking app behavior against app descriptions. In *Proceedings of the 36th International Conference on Software Engineering*. Association for Computing Machinery, New York, NY, USA, 1025–1035.

[29] X. Gu and S. Kim. 2015. "What Parts of Your Apps are Loved by Users?" (T). In *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE Press, USA, 760–770.

[30] Emitza Guzman and Walid Maalej. 2014. How Do Users Like This Feature? A Fine Grained Sentiment Analysis of App Reviews. In *2014 IEEE 22nd International Requirements Engineering Conference (RE)*. IEEE Press, 153–162. https://doi.org/10.1109/RE.2014.6912257

[31] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G Shin, and Karl Aberer. 2018. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27th {USENIX} Security Symposium ({USENIX} Security 18)*. USENIX Association, USA, 531–548.

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

[33] Zellig S. Harris. 1954. Distributional Structure. WORD 10, 2-3 (1954), 146–162. https://doi.org/10.1080/00437956.1954.11659520

[34] Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep Unordered Composition Rivals Syntactic Methods for Text Classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Beijing, China, 1681–1691. https://doi.org/10.3115/v1/P15-1162

[35] Jaeyeon Jung, Seungyeop Han, and David Wetherall. 2012. Short Paper: Enhancing Mobile Application Permissions with Runtime Feedback and Constraints. In *Proceedings of the Second ACM Workshop on Security and Privacy in Smartphones and Mobile Devices (SPSM '12)*. Association for Computing Machinery, USA.

[36] Patrick Gage Kelley, Lorrie Faith Cranor, and Norman Sadeh. 2013. Privacy as Part of the App Decision-Making Process. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 3393–3402. https://doi.org/10.1145/2470654.2466466

[37] Tianshi Li, Kayla Reiman, Yuvraj Agarwal, Lorrie Faith Cranor, and Jason I. Hong. 2022. Understanding Challenges for Developers to Create Accurate Privacy Nutrition Labels. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'22)*.

[38] Jialiu Lin, Shahriyar Amini, Jason I Hong, Norman Sadeh, Janne Lindqvist, and Joy Zhang. 2012. Expectation and purpose: understanding users' mental models of mobile app privacy through crowdsourcing. In *Proceedings of the 2012 ACM conference on ubiquitous computing*. Association for Computing Machinery, New York, NY, USA.

[39] Bin Liu, Mads Schaarup Andersen, Florian Schaub, Hazim Almuhimedi, Shikun Aerin Zhang, Norman Sadeh, Yuvraj Agarwal, and Alessandro Acquisti. 2016. Follow my recommendations: A personalized privacy assistant for mobile app permissions. In *Twelfth Symposium on Usable Privacy and Security ({SOUPS} 2016)*. USENIX Association, Denver, CO, 27–41.

[40] Xueqing Liu, Yue Leng, Wei Yang, Wenyu Wang, Chengxiang Zhai, and Tao Xie. 2018. A large-scale empirical study on android runtime-permission rationale messages. In *2018 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE Press, USA, 137–146.

[41] Yang Liu. 2019. Fine-tune BERT for Extractive Summarization. *CoRR* abs/1903.10318 (2019). arXiv:1903.10318 http://arxiv.org/abs/1903.10318

[42] Denis Lukovnikov, Asja Fischer, and Jens Lehmann. 2019. Pretrained transformers for simple question answering over knowledge graphs. In *International Semantic Web Conference*. Springer, Springer, USA, 470–486.

[43] Christoph Stanik; Tim Pietz; Walid Maalej. 2021. Unsupervised Topic Discovery in User Comments. In *29th IEEE International Requirements Engineering Conference (RE2021)* (2021-09-20). https://arxiv.org/abs/2108.08543

[44] Marlo Haering; Christoph Stanik; Walid Maalej. 2021. Automatically Matching Bug Reports With Related App Reviews. In *43rd International Conference on Software Engineering (ICSE21)* (2021-05-21). ACM, USA, 970–981. https://mast.informatik.uni-hamburg.de/wp-content/uploads/2021/02/ICSE2021_Haering_et_al_Matching_Bug_Reports_App_Reviews.pdfhttp://arxiv.org/abs/2102.07134

[45] Walid Maalej, Maleknaz Nayebi, and Guenther Ruhe. 2019. Data-Driven Requirements Engineering - An Update. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. IEEE Press, 289–290. https://doi.org/10.1109/ICSE-SEIP.2019.00041

[46] Soumayan Bandhu Majumder and Dipankar Das. 2020. Detecting Fake News Spreaders on Twitter Using Universal Sentence Encoder. In *CLEF*. Semantic Scholar, USA.

[47] Kirsten Martin and Katie Shilton. 2016. Putting mobile application privacy in context: An empirical study of user privacy expectations for mobile devices. *The Information Society* 32 (2016).

[48] Yashar Mehdad, Giuseppe Carenini, Raymond T. Ng, and Shafiq R. Joty. 2013. Towards Topic Labeling with Phrase Entailment and Aggregation. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*. ACL, USA.

[49] Nuhil Mehdy, Casey Kennington, and Hoda Mehrpouyan. 2019. Privacy Disclosures Detection in Natural-Language Text Through Linguistically-Motivated Artificial Neural Networks. In *International Conference on Security and Privacy in New Computing Environments*. Springer, Springer International Publishing, Cham, 152–177.

[50] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* 1, 1 (2013).

[51] Debjyoti Mukherjee, Alireza Ahmadi, Maryam Vahdat Pour, and Joel Reardon. 2020. An Empirical Study on User Reviews Targeting Mobile Apps' Security & Privacy. arXiv:2010.06371 [cs.CR]

[52] Duc Cuong Nguyen, Erik Derr, Michael Backes, and Sven Bugiel. 2019. Short Text, Large Effect: Measuring the Impact of User Reviews on Android App Security & Privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE Press, USA.

[53] Ehimare Okoyomon, Nikita Samarin, Primal Wijesekera, Amit Elazari Bar On, Narseo Vallina-Rodriguez, Irwin Reyes, Álvaro Feal, and Serge Egelman. 2019. On the ridiculousness of notice and consent: Contradictions in app privacy policies. *Proceedings of the Workshop on Technology and Consumer Protection (ConPro '19)* 1, 1 (2019).

[54] D. Pagano and W. Maalej. 2013. User feedback in the appstore: An empirical study. In *2013 21st IEEE International Requirements Engineering Conference (RE)*. IEEE Press, USA, 125–134.

[55] F. Palomba, P. Salza, A. Ciurumelea, S. Panichella, H. Gall, F. Ferrucci, and A. De Lucia. 2017. Recommending and Localizing Change Requests for Mobile Apps Based on User Reviews. In *2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE)*. IEEE Press, USA, 106–117.

[56] Rahul Pandita, Xusheng Xiao, Wei Yang, William Enck, and Tao Xie. 2013. {WHYPER}: Towards Automating Risk Assessment of Mobile Applications. In *Presented as part of the 22nd {USENIX} Security Symposium ({USENIX} Security 13)*. USENIX, USA, 527–542.

[57] Sai Teja Peddinti, Igor Bilogrevic, Nina Taft, Martin Pelikan, Úlfar Erlingsson, Pauline Anthonysamy, and Giles Hogben. 2019. Reducing Permission Requests in Mobile Apps. In *Proceedings of the Internet Measurement Conference* (Amsterdam, Netherlands) *(IMC '19)*. Association for Computing Machinery, New York, NY, USA, 259–266. https://doi.org/10.1145/3355369.3355584

[58] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. ACM, USA, 1532–1543. http://www.aclweb.org/anthology/D14-1162

[59] Play Safety Label 2021. New safety section in Google Play will give transparency into how apps use data. https://android-developers.googleblog.com/2021/05/new-safety-section-in-google-play-will.html. Accessed: 2021-08-27.

[60] Zhengyang Qu, Vaibhav Rastogi, Xinyi Zhang, Yan Chen, Tiantian Zhu, and Zhong Chen. 2014. AutoCog: Measuring the Description-to-Permission Fidelity in Android Applications. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* (Scottsdale, Arizona, USA) *(CCS '14)*. Association for Computing Machinery, New York, NY, USA, 1354–1365. https://doi.org/10.1145/2660267.2660287

[61] Alexander Ratner. 2019. Accelerating Machine Learning with Training Data Management. In *Stanford University PhD Thesis*. Stanford University, USA, 3. https://ajratner.github.io/assets/papers/thesis.pdf

[62] Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20 (1987), 53 – 65. https://doi.org/10.1016/0377-0427(87)90125-7

[63] Kanthashree Mysore Sathyendra, Shomir Wilson, Florian Schaub, Sebastian Zimmeck, and Norman Sadeh. 2017. Identifying the provision of choices in privacy policy text. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 2774–2779.

[64] Fuming Shih, Ilaria Liccardi, and Daniel Weitzner. 2015. Privacy tipping points in smartphones privacy preferences. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA.

[65] Rocky Slavin, Xiaoyin Wang, Mitra Bokaei Hosseini, James Hester, Ram Krishnan, Jaspreet Bhatia, Travis D Breaux, and Jianwei Niu. 2016. PVDetector: a detector of privacy-policy violations for Android apps. In *2016 IEEE/ACM International Conference on Mobile Software Engineering and Systems (MOBILESoft)*. IEEE, IEEE Press, US, 299–300.

[66] Rocky Slavin, Xiaoyin Wang, Mitra Bokaei Hosseini, James Hester, Ram Krishnan, Jaspreet Bhatia, Travis D Breaux, and Jianwei Niu. 2016. Toward a framework for detecting privacy policy violations in android application code. In *Proceedings of the 38th International Conference on Software Engineering*. ACM, US, 25–36.

[67] Daniel J. Solove. 2006. A Taxonomy of Privacy. *University of Pennsylvania Law Review* 154 (January 2006), 477–560.

[68] Christoph Stanik, Marlo Haering, Chakajkla Jesdabodi, and Walid Maalej. 2020. Which App Features Are Being Used? Learning App Feature Usages from Interaction Data. In *2020 IEEE 28th International Requirements Engineering Conference (RE)*. IEEE Press, 66–77. https://doi.org/10.1109/RE48521.2020.00019

[69] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification?. In *China National Conference on Chinese Computational Linguistics*. Springer, Springer, US, 194–206.

[70] Joshua Tan, Khanh Nguyen, Michael Theodorides, Heidi Negrón-Arroyo, Christopher Thompson, Serge Egelman, and David Wagner. 2014. The Effect of Developer-Specified Explanations for Permission Requests on Smartphone User Behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA.

[71] Christopher Thompson, Maritza Johnson, Serge Egelman, David Wagner, and Jennifer King. 2013. When It's Better to Ask Forgiveness than Get Permission: Attribution Mechanisms for Smartphone Resources. In *Proceedings of the Ninth Symposium on Usable Privacy and Security (SOUPS '13)*. Association for Computing Machinery, New York, NY, USA.

[72] Yuan Tian, Bin Liu, Weisi Dai, Blase Ur, Patrick Tague, and Lorrie Faith Cranor. 2015. Supporting privacy-conscious app update decisions with user reviews. In *Proceedings of the 5th Annual ACM CCS Workshop on Security and Privacy in Smartphones and Mobile Devices*. Association for Computing Machinery, New York, NY, USA, 51–61.

[73] Lynn Tsai, Primal Wijesekera, Joel Reardon, Irwin Reyes, Jung-Wei Chen, Nathan Good, Serge Egelman, and David Wagner. 2017. Turtleguard: helping android users apply contextual privacy preferences. *Proceedings of the Thirteenth USENIX Conference on Usable Privacy and Security* 1, 1 (2017), 145–162.

[74] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *CoRR* abs/1706.03762 (2017). arXiv:1706.03762 http://arxiv.org/abs/1706.03762

[75] Timothy Vidas, Nicolas Christin, and Lorrie Cranor. 2011. Curbing android permission creep. In *Proceedings of the Web*, Vol. 2. Springer, US, 91–96.

[76] Allison Woodruff, Vasyl Pihur, Sunny Consolvo, Laura Brandimarte, and Alessandro Acquisti. 2014. Would a privacy fundamentalist sell their {DNA} for $1000... if nothing bad happened as a result? the westin categories, behavioral intentions, and consequences. In *10th Symposium On Usable Privacy and Security ({SOUPS} 2014)*. USENIX, US, 1–18.

[77] Heng Xu, Sumeet Gupta, Mary Beth Rosson, and John Carroll. 2012. Measuring mobile users' concerns for information privacy. In *International Conference on Information Systems, ICIS 2012 (International Conference on Information Systems, ICIS 2012)*. Elsevier, US.

[78] Alex Yoo. 2018. Automatic Topic Labeling in 2018: History and Trends. https://medium.datadriveninvestor.com/automatic-topic-labeling-in-2018-history-and-trends-29c128cec17

[79] Serena Zheng, Noah Apthorpe, Marshini Chetty, and Nick Feamster. 2018. User Perceptions of Smart Home IoT Privacy. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (Nov 2018), 20 pages.

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