Unsupervised Summarization of Privacy Concerns in Mobile Application Reviews

Fahimeh Ebrahimi
febrah1@lsu.edu
Louisiana State University

Anas Mahmoud
amahmo4@lsu.edu
Louisiana State University

ABSTRACT

The proliferation of mobile applications (app) over the past decade has imposed unprecedented challenges on end-users privacy. Apps constantly demand access to sensitive user information in exchange for more personalized services. These—mostly unjustifiable—data collection tactics have raised major privacy concerns among mobile app users. Such concerns are commonly expressed in mobile app reviews, however, they are typically overshadowed by more generic categories of user feedback, such as app reliability and usability. This makes extracting user privacy concerns manually, or even using automated tools, a challenging and time-consuming task. To address these challenges, in this paper, we propose an effective unsupervised approach for summarizing user privacy concerns in mobile app reviews. Our analysis is conducted using a dataset of 2.6 million app reviews sampled from three different application domains. The results show that users in different application domains express their privacy concerns using domain-specific vocabulary. This domain knowledge can be leveraged to help unsupervised automated text summarization algorithms to generate concise and comprehensive summaries of privacy concerns in app review collections. Our analysis is intended to help app developers quickly and accurately identify the most critical privacy concerns in their domain of operation, and ultimately, alter their data collection practices to address these concerns.

CCS CONCEPTS

• **Security and privacy → Privacy protections**; • **Software and its engineering → Requirements analysis**.

KEYWORDS

Privacy, User Reviews, Mobile Apps

ACM Reference Format:
Fahimeh Ebrahimi and Anas Mahmoud. 2022. Unsupervised Summarization of Privacy Concerns in Mobile Application Reviews. In *37th IEEE/ACM International Conference on Automated Software Engineering (ASE '22), October 10–14, 2022, Rochester, MI, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3551349.3561155

1 INTRODUCTION

Mobile apps are designed with a set of user goals in mind. A user goal can be described as any abstract objective that the system under consideration should achieve [69]. For example, the goal of Sharing Economy apps (e.g., Uber and Airbnb) is to foster social capital and economic growth in resource-constrained communities [19, 45] while the goal of Health&Fitness apps is to promote healthy lifestyles among children and adults [16, 72]. However, driven by fierce market competition, app developers frequently deviate from their original goals. These deviations often come in the form of extreme privacy-invading tactics, such as constant location-tracking [39], unsolicited data collection [3, 71], or features that are intentionally engineered to lure users into sacrificing their privacy in exchange for more personalized experiences [1, 57].

Apps that do not adequately address their users' privacy concerns are often deemed untrustworthy or even abandoned by their users [22, 32, 55]. Therefore, in order for apps to survive market selection, app developers must constantly monitor their users' feedback and adjust their data collection strategies accordingly [36]. Users commonly communicate their feedback with app developers through textual app reviews [14, 20, 27, 47]. General-purpose review mining techniques, such as text classification and topic modeling, have been extensively used to classify such feedback into different types of actionable software maintenance requests [12, 25, 37, 43, 56, 63]. However, due to their sparsity and domain dependency, privacy concerns are frequently misclassified or under-recognized by these techniques [47]. For example, users of ridesharing apps (e.g. Uber and Lyft) might complain about the constant tracking of their location, while users of investing apps (e.g., Robinhood and Coinbase) might raise concerns about sharing their social security or bank information with the app. Such domain-specific feedback can be easily missed in the presence of more dominant categories of technical concerns (e.g., app crashing). Consequently, a *one-size-fits-all* approach may not be suitable for detecting privacy concerns across all application domains.

To address these challenges, in this paper, we propose a new unsupervised approach for summarizing privacy concerns in the mobile app market. Our approach is based on the assumption that privacy concerns are domain-specific. Therefore, leveraging the vocabulary that is commonly used by app users to express their privacy concerns in a specific domain can help generic text summarization algorithms generate more concise and representative summaries of these concerns. Our approach is evaluated using a large dataset of user reviews sampled from the domains of mental health, investing, and food delivery apps. Our long-term goal is to help app developers identify the critical privacy concerns in their domain of operation, alter their data collection practices to mitigate these concerns, and ultimately, survive user selection.

ASE '22, October 10–14, 2022, Rochester, MI, USA

ACM ISBN 978-1-4503-9475-8/22/10…$15.00
https://doi.org/10.1145/3551349.3561155

The remainder of this paper is organized as follows. In Section 2, we motivate our research. In Section 3, we describe our subject domains and experimental dataset. In Section 4, we present our procedure for extracting privacy keywords from app reviews. In Section 5, we propose a novel algorithm for summarizing privacy-related mobile app reviews. In Section 6, we discuss our key findings. In Section 7, we address the main limitations of the study. Finally, in Section 8, we conclude the paper and describe our future work.

2 BACKGROUND AND MOTIVATION

Privacy in the mobile app market has received significant attention from the research community over the past decade. However, recent systematic reviews have revealed that the majority of existing literature is focused on detecting privacy policy violations and preventing data leaks, while less attention has been paid to mining end-users' privacy concerns [20].

Earlier evidence on extracting privacy concerns in the mobile app market can be found in Khalid et al. [36]. The authors manually examined and classified thousands of one and two-star app reviews to get a better sense of end-users' complaints and their impact on app ratings. The analysis revealed that reviews including complaints about privacy-invading practices were often associated with the most negative impact on ratings. In another study, McIlroy et al.'s [47] qualitative analysis of 7,000 user reviews revealed that close to 17% of examined reviews raised privacy concerns. These concerns were expressed using more varied language than other types of technical issues.

Ciurumelea et al. [14] used iterative content analysis to develop a taxonomy of actionable issues in mobile app reviews, including compatibility, usage, resources, pricing, and privacy. In a more recent work, Hatamian et al. [27], proposed MARS, a tool for summarizing privacy-related mobile app reviews and classifying them into a set of predefined security threats, including spyware, phishing, and spam. Informative reviews were detected based on a keyword catalog seeded with the initial keywords: *privacy* and *security*. This catalog was iteratively expanded with more privacy-related keywords using word frequency analysis. Extracted keywords were then used to tune different text classifiers. MARS was able to classify 2,412 privacy-related reviews with a recall and precision of 91.30% and 94.84% respectively.

Besmer et al. [6] analyzed a massive dataset of mobile app reviews collected over the period of four years. The results showed that reviews that contained complaints about app privacy had lower star ratings and more negative sentiment than other reviews. The results also showed that users found privacy-related reviews to be more helpful than others. In another study, Mukherjee et al. [52] identified privacy-related app reviews using a generic set of privacy-related keywords. The authors found that only 0.5% of reviews were related to end-user security and privacy. Nguyen et al. [54] also used a set of 102 generic keywords to extract potential security and privacy concerns from 2,583 Google Play apps. A manual analysis of 4,000 reviews of these apps showed that 14% of them were either privacy or security-related. The authors also reported that preceding privacy reviews were a significant factor in predicting privacy-related app updates.

In summary, the majority of existing work employs generic text classification and NLP-based methods for detecting privacy concerns in mobile app reviews [46]. However, the results largely indicate that these generic solutions often struggle to capture domain-specific privacy concerns. For instance, supervised classification techniques rely on the presence of manually generated ground-truth datasets. Thus, these techniques are constrained to a single rubric of predefined categories [25, 37, 56]. Consequently, specific categories related to user privacy can be easily missed in the ground truth data. Unsupervised topic modeling techniques, such as Latent Dirichlet Allocation (LDA) [9], have also been applied to extract privacy information from app store reviews [30]. However, such techniques do not perform well with small, unstructured, and semantically-restricted text such as user reviews [8, 28, 70, 74].

Existing work has also revealed that users often express their privacy concerns using more varied language than other technical issues [47]. Consider for example the following three reviews selected from the domains of mental health, investing, and food delivery apps. The word *Facebook* clearly indicates a privacy concern in the mental health domain. However, in food delivery, the same word is used to express an issue related to customer support, while in the investing domain the word is used to express an issue related to user registration.

- **Mental Health**: "*Won't even let me sign up after collecting all of my Facebook data, just stole my identity.*"
- **Investing**: "*I got zero response back. I even blasted their Facebook but got nothing.*"
- **Food Delivery**: "*It doesn't recognize my facebook account so I can't even register for this.*"

Motivated by the limitations of existing work, in this paper, we propose an unsupervised approach for summarizing user privacy concerns in the mobile app market. We initially describe a systematic method for extracting privacy-related vocabulary from three different application domains. Extracted vocabulary is then leveraged to generate concise and comprehensive summaries of privacy concerns in app reviews.

3 DATA COLLECTION

Our underlying assumption in this paper is that app users express their privacy concerns using different terminology that is directly related to their apps' specific functionality. To verify this assumption, we collected a large-scale dataset of user reviews from three different application domains: mental health, food delivery, and investing. Investing apps have become increasingly popular in recent years due to the increasing interest in cryptocurrency trading. Zero-commission trading fees and continuous media coverage have brought in millions of new first-time traders. For example, one of the most popular investing apps, *Robinhood*, reported that close to 6 million new users joined the platform in 2021 [60]. Similarly, the domain of food delivery has experienced massive growth during the past two years. In particular, the demand for food delivery services has significantly increased due to the COVID-19 pandemic. For example, the four major food delivery apps - DoorDash, UberEats, GrubHub, and Postmates have all reported a significant increase in revenue generated during the lock-down order of 2020 [65]. This global health crisis has also led to a significant spike in the number of active users of mental health apps. People frequently resorted to these apps as a safer and inexpensive alternative to help them cope with the mental consequences of social isolation, unemployment, and economic hardships [15, 41].

To collect reviews from these three different domains, we identified the top-100 apps in the categories of Finance (investing), Food&Drink (food delivery), and Health&Fittness (mental health) on Google Play and the Apple App Store. Apps that met the following criteria were included in our dataset:

- For an app to be included in our analysis, we only considered apps with 10,000 reviews or more. This number of reviews was necessary to include only popular and well-established apps. We lowered this number to 5,000 for mental health apps as apps in this category do not get as many reviews.
- For the Finance category, banking *"all-in-one"* apps were excluded as the majority of these apps did not provide investing services. For Food&Drink, specific restaurant delivery apps, such as *Papa John's Pizza & Delivery* official app, were also excluded as they did not operate as independent delivery services. In the Health&Fitness category, physical health apps that did not explicitly support mental health were excluded.

After examining the top-100 apps in each category, eight investing, five mental health, and five food delivery apps were included. For each of these apps, we collected all textual reviews available on the Apple App Store and Google Play using Python web scrapers. Overall, 696,073, 1,708,831, and 204,374 reviews were collected for our set of investing, food delivery, and mental health apps respectively. The distribution of these reviews over apps is shown in Table 1.

4 EXTRACTING PRIVACY VOCABULARY

In this section, we empirically examine our assumption that privacy concerns in mobile app reviews are expressed using domain-specific vocabulary.

4.1 Privacy Term Extraction

Our analysis is conducted over low-rating (one and two stars) reviews in our dataset. These reviews are more likely to contain user complaints or useful feedback than high-rating reviews [29, 36, 47]. In total, 385,951, 511,032, and 43,647 reviews from the domains of investing, food delivery, and mental health are included in our analysis. Table 1 shows the total number of 1-2 star reviews for each app in our dataset.

Fig. 1 describes our indicator term (keyword) extraction process. The goal of our analysis is to generate a catalog of terms and phrases that signal privacy-related issues in different application domains. To generate such a catalog, we follow a systematic iterative process of word generation. We begin our analysis by seeding the catalog with the words *privacy*, *private*, and *security* [54]. These words are then used as search queries to locate potential privacy-related reviews in our dataset. The first iteration of the search returned 187, 753, and 629 reviews for the apps in the mental health, investing, and food delivery domains respectively. Each review is then manually examined by three judges to locate any other keywords (unigrams or bigrams) that are likely to be indicative of privacy issues. Each judge has to answer the questions: *does this review raise any form of privacy concern?* And if so, *what keyword(s) in the review are indicative of such concerns?* All judges hold professional degrees in software engineering as well as have an average of 6+ years of experience in app development. A pilot labeling session was held before running the actual analysis to explain the process and address any concerns. Terms generated after the first round are then used to retrieve the second set of reviews. Basically, any review that contained any of the identified keywords and did not appear in previous searches is included in the search results. This process is repeated until saturation, or no more new keywords are found.

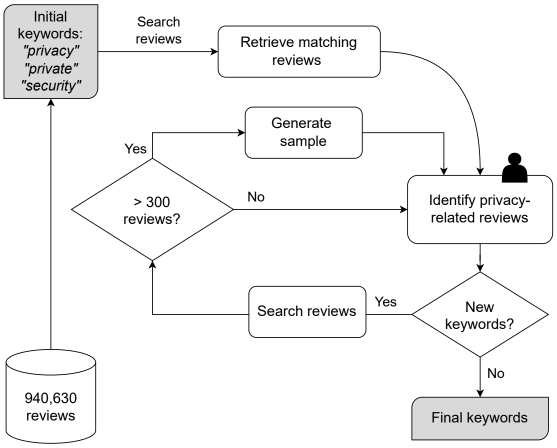

Figure 1: Our privacy keyword extraction procedure.

Notice that several keywords retrieved a large number of matches. For instance, the word *location* returned 10,829 reviews for the apps in the food delivery domain. Examining such a large number of reviews manually can be an exhaustive and error-prone task. Therefore, for such large sets, we only examine a statistically representative stratified sample of reviews. A sample size of 300 reviews is sufficient to maintain at least a 95% confidence level. Among the identified keywords, only nine of the keywords retrieved more than 300 reviews and needed the sampling phase. At the end of this process, 26 unique privacy-related keywords or phrases (e.g. *personal info*) were extracted from our dataset. These keywords are listed in Table 2. The table also shows the total number of reviews retrieved in each domain as well as the number of reviews that are privacy-related.

4.2 Keyword Analysis

Our keywords extraction process generated 26 unique privacy-related keywords. In Table 3, we show the percentage of privacy-related reviews that are retrieved by each of our keywords in each application domain. We use Chi-square (χ^2) to test for statistical significance in these results. Our null hypothesis (H_0) is that there is no difference in the percentage of privacy-related reviews retrieved by each privacy keyword between all domains. The alternative hypothesis (H_1) is in favor of the dependency between the domain

Table 1: The number of user reviews extracted (1-2 stars) for each app in our dataset.

Investing		Food Delivery		Mental Health	
App	**Reviews**	**App**	**Reviews**	**App**	**Reviews**
Robinhood	451,016 (325,534)	UberEats	748,584 (265,713)	Calm	106,181 (22,983)
Acron	76,761 (15,954)	DoorDash	598,513 (122,857)	Headspace	78,989 (16,376)
Stash	40,385 (10,683)	GrubHub	223,566 (63,776)	Sanvello	8,554 (698)
ETrade	15,807 (9,297)	Postmates	107,564 (53,579)	Talkspace	5,054 (2,928)
Fidelity	50,224 (9,034)	Seamless	30,604 (5,107)	Shine	5,596 (662)
TD Ameritrade	30,369 (8,973)				
Schwab	14,988 (4,596)				
Personal Capital	16,523 (1,880)				
Total	696,073 (385,951)	**Total**	1,708,831 (511,032)	**Total**	204,374 (43,647)

Table 2: The results of our indicator keyword extraction process, showing the number of reviews retrieved by each keyword at each round in each domain along with the number of privacy-related reviews (shown in parenthesis) identified.

Round	Keywords	Mental Health	Investing	Food Delivery
1st	*privacy, private, security*	187 (**140**)	753 (**291**)	629 (**183**)
2nd	*personal info, permission, user data, facebook, patient info, bank statement, bank login, credit card, SSN, social security*	869 (**200**)	1,805 (**586**)	1,402 (**289**)
3rd	*camera, microphone, GPS, location, job history, birth*	45 (**11**)	595 (**165**)	905 (**39**)
4th	*driver license, real name, last name, imei, identification info, email*	303 (**40**)	366 (**16**)	359 (**4**)
Total		1,404 (**391**)	3,519 (**1,058**)	3,295 (**515**)

and the keywords, in other words, the recall of different keywords when retrieving privacy-related reviews is significantly dependent on the domain. Since we have two variables (privacy-related and non-privacy-related) and three groups (domains), the degree of freedom of our test is set to $2 = (2 - 1) * (3 - 1)$. Given this degree of freedom and the confidence levels of 0.001, 0.01, and 0.05, χ^2 critical values are set to 13.816, 9.210, and 5.991, respectively. H_0 will be rejected if the χ^2 value is larger than the critical values. The last column of Table 3 shows the chi-square test results and the confidence level for each keyword. The results show that, for the majority of domain-specific keywords, we can reject H_0 with a confidence level of at least 0.05.

The results also show that the number of privacy-related reviews retrieved by the generic keywords (e.g., *privacy*, *security*, etc.) is not dependent on the application domain. Most of the reviews that contain the keyword *private* are not necessarily privacy-related. For instance, in the food delivery domain, only 23% of the reviews that contain this keyword raise privacy concerns, commonly appearing in reviews such as, *"I live on first floor first apt in a private building."* The same observation holds for other keywords that are frequently associated with privacy in the literature, such as *camera*, *security*, and *permission*. For example, users in the food delivery domain use the keyword *permission* to complain about their orders being canceled without their permission.

In general, the majority of our identified keywords are domain-specific. For example, the phrase *"last name"* appears as one of the main privacy-indicator terms in the mental health domain. All reviews (100%) that contain *"last name"* in this domain are privacy-related. However, in the food delivery and investing domains, less than 10% of reviews that contain *"last name"* are privacy-related. We also observe that some keywords are more domain-specific than others. For example, the term *"driver's license"* appears frequently in the privacy-related reviews of investing apps. Users are mainly complaining about apps asking for photocopies of their driver's license, such as, *"why do they need pictures of both sides of my driver's license, they already verified my bank account and who do they share this information with?"* However, none of the reviews that contain this keyword in the food delivery domain is privacy-related. In general, customers of food delivery apps use *"driver's license"* to ask questions about working for the app, such as, *"Can I work for this app if my driver's license was out of state?"* or complain about drivers, such as, *"I don't think my delivery kid has a driver's license."*

The results also show that some keywords are good signals of privacy issues only in two domains, but are mainly associated with false positives in the third domain. For example, the keyword *"GPS"* is almost always associated with privacy-related issues in the domains of investing and mental health. Users in these domains frequently use this keyword to express concerns about apps tracking their location. However, in the food delivery domain, users have no issue with delivery apps demanding access to their location to get their food delivered to their exact address. In general, despite their very common occurrence in the reviews of food delivery apps, keywords such as *GPS* and *location* are not indicative of privacy concerns in this domain, mainly appearing in reviews such as, *"only two restaurants for my location. horrible!"* or *"the driver needs to update their GPS to the new map."*

4.3 Keyword Co-occurrence Analysis

We observed while labeling the data that domain-specific keywords tend to co-occur in privacy-related reviews. For example, in the review *"it asks for your first and **last name**, **email** address and **facebook** account. I did not feel comfortable sharing **personal information**"*, a privacy concern of using a particular mental health app

Table 3: The percentage of privacy-related reviews in the reviews retrieved by each keyword in each application domain. The significance of word × domain dependency is measured using Chi-square (χ^2), $^{*}p \leq 0.05$, $^{**}p \leq 0.01$, $^{***}p \leq 0.001$

Keyword	Invest.	Mntl. Health	Food Deliv.	χ^2
microphone	93%	0%	40%	15.76^{***}
social security / SSN	52%	14%	15%	13.68^{**}
identification info	100%	0%	25%	6.36^{*}
bank login	11%	0%	0%	37.35^{***}
birth	26%	9%	0%	82.67^{*}
camera	23%	0%	11%	15.38^{*}
driver license	35%	0%	0%	1.5
credit card	4%	1%	3%	6.25^{*}
email	2%	13%	0%	59.11^{***}
facebook	15%	23%	2%	29.55^{***}
last name	10%	100%	4%	6.54^{*}
private	30%	55%	23%	15.8^{***}
job history	0%	100%	0%	-
imei	0%	100%	0%	-
patient info	0%	100%	0%	-
location	29%	44%	0%	102.73^{***}
GPS	100%	100%	1%	132.66^{***}
privacy	94%	98%	67%	4.93
personal info	71%	100%	94%	5.18
user data	100%	100%	50%	0.84
security	19%	19%	7%	1.92
permission	4%	27%	3%	3.18

is expressed using four indicator terms, *"last name", "email", "facebook",* and *"personal info".* To further examine this observation, we calculate the Normalized Pointwise Mutual Information (NPMI) between the set of privacy indicator terms extracted from our dataset. NPMI is an information-theoretic measure of information overlap, or statistical dependence, between two words [11]. Formally, NPMI between two words w_1 and w_2 can be measured as the probability of them occurring in the same text versus their probabilities of occurring separately. Assuming the collection contains N reviews, PMI can be calculated as:

$$NPMI(w_1, w_2) = \frac{\log_2\left(\frac{\frac{C(w_1,w_2)}{N}}{\frac{C(w_1)}{N}\frac{C(w_2)}{N}}\right)}{-\frac{C(w_1,w_2)}{N}} = \frac{\log_2\left(\frac{P(w_1,w_2)}{P(w_1)P(w_2)}\right)}{-P(w_1,w_2)} \quad (1)$$

where $C(w_1, w_2)$ is the number of reviews containing both w_1 and w_2, and $C(w_1)$ and $C(w_2)$ are the numbers of reviews containing w_1 and w_2 respectively. NPMI is normalized using the negative log-transformed count of the number of times w_1 and w_2 appear together. If w_1 and w_2 are frequently associated, the probability of observing them together will be much larger than the chance of observing them independently. This results in NPMI > 0. If there is no relation between w_1 and w_2, then the probability of observing w_1 and w_2 together will be much less than the probability of observing them independently (NPMI < 0).

The results of our NPMI co-occurrence analysis are shown in Fig. 2. The figure shows the semantic distance between our keywords projected on a 2D map. In the mental health domain, *"email", "last name", "GPS",* and *"Facebook"* commonly co-occur together in privacy reviews. These keywords retrieved the highest percentage of privacy-related reviews in the mental health domain (Table 3). Similarly, the keywords *"SSN",* and *"bank statement"* commonly co-occur in the privacy reviews of investing apps. We further observe that keywords from different application domains seem to be standing at the same semantic distance from the seed *"privacy".*

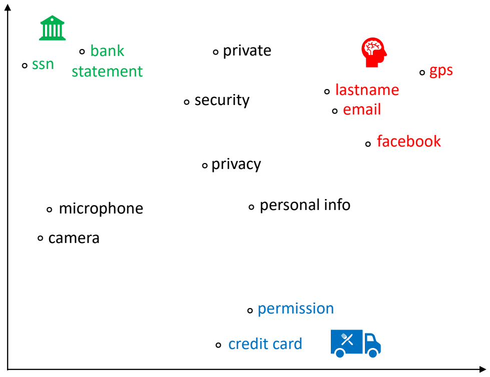

Figure 2: The NPMI distance between extracted privacy-related keywords from our different application domains.

In summary, our analysis in this section provides evidence that the keywords used by app users to express their privacy concerns are domain-specific. Such keywords are largely derived from the features of the app and its operational characteristics. While some keywords may provide a strong signal of privacy concerns in one domain, they may not be indicative of privacy issues in other domains. We also observe that domain-specific privacy-indicative keywords tend to frequently co-occur together in app reviews. In the next section, we show how such insights can be leveraged to generate concise summaries of privacy concerns in mobile app review collections.

5 REVIEW SUMMARIZATION

The first phase of our analysis has revealed that users' privacy concerns in mobile app reviews are commonly expressed using domain-specific vocabulary. Therefore, a *one-size-fits-all* approach for detecting privacy concerns across different application domains in the app store is destined to fail. This problem is further aggravated by the fact that privacy concerns are sparse, only appear in a small percentage of reviews, and are frequently overshadowed by more dominant categories of user feedback, such as concerns about the app's reliability (e.g., reporting bugs) or usability (e.g., requesting features) [6, 36, 52]. These limitations hinder the ability of supervised learning techniques (e.g., text classifiers) to detect privacy-related reviews as sparse categories of data can be easily missed in the training dataset. The problem also severely affects unsupervised topic modeling techniques, such as LDA, where generated topics are naturally representative of dominant themes in the data [67].

To work around these limitations, in this section, we propose and evaluate an unsupervised domain-specific approach for summarizing privacy concerns in the mobile app store. Our approach leverages domain knowledge to point our summarization approach toward the most prevalent privacy concerns in user reviews. Timely detection and handling of such concerns can be critical for app survival as recent evidence has shown that privacy-related reviews are commonly accompanied by negative sentiment and low ratings [6, 17, 36].

5.1 Text Summarization

The goal of summarization techniques is to capture the underlying dominant themes in a text corpus (source text) and represent them as cohesively and concisely as possible, in other words, to generate meaningful summaries [35, 40]. In the context of app reviews, automated summarization techniques are used to assimilate the perspectives of a large number of users to bring app developers' attention to any pressing issues that need to be addressed in future releases [33]. Summarization techniques can be either extractive or abstractive. Abstractive techniques aim to construct novel descriptions of the main ideas in a source text. Extractive techniques, however, group together specific key sentences and keywords from the source text to generate a concise summary of the text. Abstractive techniques are commonly known to be more sophisticated as lexical parsing and paraphrasing are needed to generate novel and meaningful summaries [26]. Therefore, they are known to perform better when applied to semantically rich text, such as scientific documents or news articles. However, user reviews are short and often expressed using informal and semantically restricted jargon [13, 64], making extractive techniques more effective in this context.

In general, extractive summarization techniques leverage the frequencies of individual words to estimate their importance to the source text [26]. The likelihood that a sentence from the source text will be considered as a representative summary is positively correlated to the average perceived importance of its words [35]. In semantically restricted text collections (e.g., user reviews), where individual text artifacts are short, the importance of words can be determined using a technique such as Hybrid TF.IDF [31]. TF.IDF consists of two components, Term Frequency (TF) and Inverse Document Frequency (IDF). TF is the raw frequency of a word in a document and IDF is an indicator of how much information this word provides. For a collection of short texts, Hybrid TF.IDF modifies the TF of words by dividing their frequency by the number of unique words in the entire text collection (N). This modification is necessary when dealing with user reviews as the probability of individual words occurring multiple times in a single review is relatively low. Formally, the Hybrid TF.IDF of a word w can be computed as:

$$HybridTF.IDF(w) = \frac{f(w,R)}{\sum f(w,r)} \times \log \frac{|R|}{df(w)} \tag{2}$$

where $f(w,R)$ is the term frequency of the word w in all reviews, $\sum f(w,r)$ is the total number of unique words in the review collection, R is the total number of reviews in the collection, and $df(w)$ is the number of reviews that contain w.

Given the above assumptions, Hybrid TF.IDF—as an extractive summarization technique—first calculates the importance of individual reviews as the average of their individual words' Hybrid TF.IDF values. The algorithm then selects the top K most important reviews as summaries. To control for redundancy, before a review r_i is added to the summary, the algorithm makes sure that r_i does not have a textual similarity above a certain threshold with the reviews already in the summary. The textual similarity between two reviews can be calculated as the cosine similarity between their TF.IDF vectors.

5.2 Privacy Review Summarization

Extractive summarization techniques work well for online text corpora [59]. However, due to their reliance on words' frequencies, sparse themes in the data tend to be either missed or underrepresented in the summary. To work around these limitations, we alter Hybrid TF.IDF in two ways. First, we adjust the weight of privacy indicator keywords in each domain to the maximum Hybrid TF.IDF value calculated for the domain. For example, after removing English stop-words and performing lemmatization, the maximum Hybrid TF.IDF calculated in the entire set of reviews for the mental health apps in our dataset is equal to 0.022. This weight is assigned to the list of privacy indicator keywords (Table 3) identified for this domain, including *"Facebook", "IMEI", "patient", "location", "email", "GPS", "last name"*, and *"real name"*. The same is applied to the privacy-indicator words identified for the investing domain and the food delivery domain. In addition, the generic catalog seeds (*"privacy", "private"*, and *"security"*) are also assigned to the same maximum weight in each domain.

The second adjustment is related to the way redundancy is controlled in the summary. To minimize the probability of retrieving reviews raising similar concerns, we enforce a similarity threshold calculated using the word embeddings of reviews. For a highly ranked review (based on its average Hybrid TF.IDF score) to be added to the summary, it has to stand at a specific minimum semantic distance from other reviews already in the summary. This distance is calculated in classical Hybrid TF.IDF using the cosine similarity between the TF.IDF vectors of reviews. However, relying on the textual similarity between reviews can lead to information loss. Word embeddings can overcome this problem by relying on the meaning of reviews rather than their lexical structure.

In our analysis, we use GloVe to calculate the word embeddings of individual reviews. GloVe [58] is a popular word embedding model that uses the similarities between words as an invariant to generate their vector representations. In general, word embeddings represent individual words in a corpus using multi-dimensional vectors of numeric values that are derived from the intrinsic statistical properties of the corpus. GloVe initially constructs a high dimensional matrix of words co-occurrence. Dimensionality reduction is then applied to the co-occurrence count matrix of the corpus. By applying a matrix factorization method on the count matrix, a lower dimension matrix is produced, where each row is the vector representation of a word. To conduct our analysis, we converted the list of pre-processed tokens in each user review into a vector of word embeddings using the pre-trained model of GloVe. We then used the generated word embeddings to represent the review. Word collection embeddings can be computed using operations on word vectors, such as their unweighted averaging/summation [50], Smooth Inverse Frequency (SIF) [2], and Doc2Vec [38, 62]. In our analysis, we used the simple unweighted averaging method to obtain an embedding for each review in our dataset [2, 61]. Averaging word vectors has been proven to be a strong baseline for paragraph representation, especially in cases when the order of words in the text is unimportant [34]. Algorithm 1 describes our privacy-concern summarization approach.

Algorithm 1 A description of our summarization algorithm.

```
for review $r_i$ in $R$ do                          ▷ R : reviews in the collection
    for word $w_j$ in $r_i$ do
        if $w_j \in$ privacy_keywords then          ▷ is $w_j$ an indicator keyword?
            $w_j$.weight = $max\_tf\_idf$            ▷ increase the weight
        else
            $w_j$.weight = Hybrid.TF.IDF($w_j$)     ▷ calculate Hybrid.TF.IDF
        end if
        $r_i$.total += $w_j$.weight                  ▷ sum up words weights
    end for
    $r_i$.weight = $r_i$.total / $|r_i|$             ▷ calculate review $r_i$ weight
end for

$R'$ = $R$.sort(DES)                ▷ sort reviews based on their Hybrid TF.IDF
$S = \{r_0\}$                       ▷ add the top ranked review to the summary
$count$ = 1                         ▷ review length

for $r_i$ in $R'$ do
    if GloVe.similarity($r_i$, $S$) < $\lambda$ then   ▷ $\lambda$: similarity threshold
        $S$.add($r_i$)                                  ▷ Add $r_i$ to summary $S$
        $count$++
        if $count$ == $k$ then                          ▷ $k$ is desired summary length
            break;
        end if
    end if
end for
```

5.3 Evaluation

To evaluate our proposed algorithm, we summarize a test dataset of reviews collected from a new set of apps sampled from our three application domains. Using a separate test dataset can help to validate our assumptions regarding the generalizability of our approach over each domain. Our test dataset includes 11,145 low star-rating user reviews collected from six apps following the inclusion/exclusion criteria described in Section 3. These apps are the mental health apps eMoods and Happify, the investing apps Wealthfront and Stockpile, and the food delivery app goPuff and Delivery.com. Table 4 describes the apps in our test dataset.

Before generating the summaries, for each of the reviews in our test dataset, English stop-words (e.g., *the, in, will*) are removed based on the list of stop-words provided in NLTK [42]. The remaining words are then lemmatized. We also exclude reviews that contain less than five words. This step is necessary in order to capture more informative summaries [37]. Table 5 shows the top summary reviews generated using Hybrid TF.IDF as well as our seeded summarization algorithm for each domain. For page-limit considerations, we only show the top five reviews. The table also shows the score calculated for each review and the concern category (quality requirement) raised in the review, if any [44].

The results show that the majority of summaries generated by Hybrid TF.IDF contains valid user concerns. However, none of these concerns are privacy-related in any of the domains. Instead, the top five spots in all three summaries are hijacked by the most dominant concerns in each domain, such as the app being inaccessible due to high fees or being unavailable to conduct a transaction. We further notice that there is a high level of redundancy in these summaries, even though a relatively low cosine similarity threshold of 0.2 was used. For instance, the summaries generated for the mental health, investing, and food delivery domains mainly expressed concerns regarding the apps being too expensive, untrustworthy, or having bad customer service respectively.

Table 4: The test dataset used in our analysis.

Domain	App	Avg. Rating	# of Reviews (1-2 star)
Investing	Wealthfront	4.8	4,437 (**320**)
	Stockpile	4.7	4,412 (**999**)
Mental Health	eMoods	4.8	1,774 (**75**)
	Happify	4.5	2,393 (**783**)
Food Delivery	goPuff	4.5	44,397 (**8,145**)
	Delivery.com	4.8	2,661 (**823**)
Total			60,074 (**11,145**)

A deeper look into the data reveals that the generated Hybrid TF.IDF summary reviews contain the most important words in the corpus (based on their Hybrid TF.IDF score). Table 6 shows the top 10 words with the highest Hybrid TF.IDF scores in each application domain. These words appear in the summary reviews in each domain. For example, the summary reviews for the food delivery domain contain the words *"delivery", "time"* and *"service"*, which are the most important in this domain according to their Hybrid TF.IDF. In general, none of the domain-specific privacy-related keywords (see Table 2) are among the top 100 words in our corpora of user reviews. In fact, according to Hybrid TF.IDF, the keyword *privacy* is ranked 863, 2045, and 2004 in the mental health, investing, and food delivery domains, respectively.

Table 5 also shows that our proposed seeded summarization approach managed to overcome Hybrid TF.IDF's limitations in all application domains. By adjusting the importance of the privacy-related keywords, we raised the probability of the privacy-related reviews being included in the summary. The table shows that precision of 80%-100% can be achieved in all domains at 5-review length summaries. We further notice that our generated summary reviews experience less redundancy than the summaries generated by Hybrid TF.IDF. For example, in the mental health domain, four out of the top five reviews raise privacy concerns about apps collecting personal information, demanding access to social media, and sharing user information with third-party entities. In the food delivery domain, the summary reviews raise concerns about apps collecting personal information (zip code), demanding access to phone contacts, and selling user information to third parties. In general, the low redundancy (higher coverage) in the summaries can be attributed to the fact that word embeddings are used to calculate the pairwise semantic similarity between reviews rather than relying on the textual similarity of their words. This helps to overcome the vocabulary mismatch problem affecting Hybrid TF.IDF. For example, the review *"I don't care how vetted this app is, no way are you getting my social and bank credentials"* is excluded from the summary of investing apps due to having high GloVe similarity with the summary review *"I'm worried because it has my SSN and bank login"*. Our analysis shows that a GloVe similarity score

Table 5: The top five reviews generated by Hybrid TF.IDF and our seeded summarization algorithm for all domains along with the average Hybrid.TF.IDF scores of each review and the quality concern expressed in the review.

Domain	Method	Top five reviews	Score	Concern
Mental Health	Hybrid TF.IDF	“The app crashes every time I try to open it. This is not making me happy.”	0.013	reliability
		“Absolutely not worth it unless you want to pay almost 300$ a year, don’t waste your time.”	0.012	accessibility
		“Make it a one time purchase, not a subscription.would get 5 stars if I could have paid 4.99 for the app once not every month.”	0.013	accessibility
		“The free activities are childishly simple and have nothing to do with actually creating a good mood, or anything other than taking up your time.”	0.010	accessibility
		“Hasn’t anyone ever heard the saying "money can’t buy happiness"?”	0.008	accessibility
	Hybrid TF.IDF + GloVe	“As soon as an app forces me to SIGN UP with Facebook (I gave up Facebook long ago and would never go back) or give my email... it’s over, I am not giving out any personal info.”	0.17	privacy
		“One more paid app, that doesn’t say clearly it is paid until collect your personal info”	0.17	privacy
		“Signups with email doesn’t recognize my valid email as correct.”	0.16	reliability
		“You shouldn’t need access to any info on my phone that affects my security.”	0.11	privacy
		“Patient, doctor confidentiality breeched. Your private chat is accessed by 3rd party without your permission.”	0.009	privacy
Investing	Hybrid TF.IDF	“If your looking to get your money taken this is a good app to start with.”	0.010	fraud
		“This app doesn’t let you sell until the end of the day so by that time price can go way down do not use this app”	0.010	usability
		“Would have wanted to rate it higher but i cant even use the app because i cant even use my email saying that there already is an account.”	0.009	reliability
		“Awful. High fees. Takes over 24 hours to trade every time”	0.008	usability
		“They locked my account. It says that it takes form 3 to 5 business days but it being 9”	0.008	usability
	Hybrid TF.IDF + GloVe	“app is glitchy & does not connect with bank properly. now I’m worried because it has my SSN and bank login info. would not recommend. Google doesn’t seem to vet any of these apps!”	0.33	privacy
		“I can receive money from my bank but I cant send money to my bank”	0.29	usability
		“Like the app but it was constantly using my camera. There’s no good reason for a stock trading app to use my camera especially when I’m not using it”	0.23	privacy
		“Let me link my bank account MANUALLY using routing/account numbers. I don’t want to give you MY BANK ID AND PASSWORD. ”	0.17	privacy
		“DONT USE THIS APP ITS A SCAM!! They SOLD my info and my social security number!!! ”	0.17	privacy
Food Delivery	Hybrid TF.IDF	“ Ordered then got a call that they don’t deliver to my area, after being told by the app that they did.”	0.016	accessibility
		“Cant get ahold of customer service, never got my delivery but was still charged. ”	0.016	usability
		“Used to be great when it only took 20 to 30 minutes, but now every order takes one to two hours. ”	0.014	usability
		“ Makes you sign up before you find out they’re not in your area. Waste of time”	0.012	accessibility
		“they made me wait one hour and i called and they told me i would ha e to wait another hour thats dumb hpnestly its happened twice already”	0.012	usability
	Hybrid TF.IDF + GloVe	“not in my area, wish you asked for the zip first before i gave you all my info”	0.29	privacy
		“Why do you need to download my contacts from email and my phone? Permissions are sketchy.”	0.25	privacy
		“ They sell screen recordings of customers to third partys”	0.17	privacy
		“This app wants too much personal information, unrelated to its purpose. Be aware. ”	0.14	privacy
		“Ever time I go 2 update my acct info it crashes!! Has never worked”	0.14	reliability

Table 6: The top 10 words with the highest Hybrid TF.IDF in each of our application domains.

Domain	Top 10 words
Mental Health	pay, free, get, use, try, subscription, money, want, time, trial
Investing	stock, market, money, buy, use, trade, get, make, fund, hedge
Food Delivery	service, get, delivery, time, food, use, customer, driver, bad, cancel

in the range [0.4 - 0.6] can achieve a balance between minimizing the redundancy of generated summaries and excluding important concerns.

To evaluate the performance of our seeded summarization approach at different length summaries, we generate summaries of lengths 5, 10, and 15 reviews for each set of reviews sampled from each of our domains. We assess the performance using two measures, precision and redundancy. Precision is calculated as the percentage of reviews in the summary that are privacy related, while redundancy is the percentage of reviews that raise privacy concerns already raised in other reviews in the summary. In an ideal scenario, the precision should always hover around 100% while redundancy should be kept to the minimum (depending on the number of privacy concerns in the reviews). However, in reality, with more reviews included in the summary, we should see a drop in the precision and an increase in the redundancy, but with an increase in the recall, or the number of privacy concerns recovered.

The results of our analysis are shown in Fig. 3. A summary of size 10 seems to achieve a balance between precision and redundancy. At 10 reviews, we are able to maintain a relatively high precision while keeping the redundancy under control. For example, increasing the summary length from 5 to 10 in the mental health domain generates 5 more privacy-related reviews among which three are redundant. However, the review *“Really difficult to delete an account, a complete violation of privacy”* reveals a new privacy concern (right to delete) that is not captured in the summary of size 5. Our analysis shows that setting the summary length to more than 10 can lead to a sizable drop in precision and a spike in redundancy.

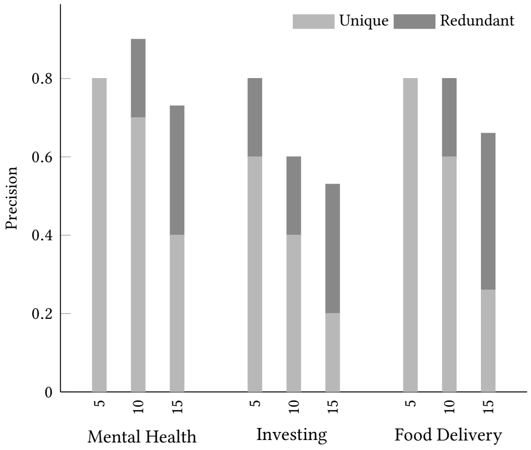

Figure 3: The performance of our summarization algorithm at different summary lengths as measured by precision and redundancy.

In general, our results indicate that most domains have between 4 - 6 unique privacy concerns. To confirm this observation, we generate the top 50 summary reviews for each application domain. Following a systematic coding process similar to the process described in Sec. 4.1, we categorize the different types of privacy concerns present in these reviews. The main question to answer is: *what types of privacy concerns are raised in this review?* The results are shown in Table 7. Among the generated review summaries, around 28% are not privacy-related. In the mental health domain, the mandatory request for Facebook credentials or email addresses is the most dominant concern. In the investing domain, the dominant concern is related to apps requesting access to users' cameras and microphones. In the food delivery domain, 22% of generated review summaries are raising concerns about the unnecessary data collection practices of apps. In general, 4-5 unique privacy concerns are detected in each domain.

6 DISCUSSION AND IMPACT

Static code analysis of large datasets of apps in the mobile app market has revealed that the majority of apps do not comply with the privacy claims in their privacy policies [7, 20, 71]. Apps use intrusive data collection strategies to harvest more information than they need. Such information is frequently exploited to build addictive or habit-forming apps or even shared with third-party ad agencies [5]. Recent research efforts have been focused on analyzing the privacy practices of mobile apps. The objective is to inform mobile app users about the potential misuse of their personal data [71]. However, less effort has been made to keep app developers up-to-date with their end-users privacy concerns. This can be mainly attributed to the fact that such concerns are typically buried within more common types of usability, reliability, or accessibility concerns.

In this paper, we propose a novel approach for extracting privacy concerns from mobile app user reviews. Our analysis is conducted using a large dataset of app reviews collected from multiple application domains. Our first observation is that, due to their sparsity, privacy concerns can be easily missed. Our results also show that relying on generic keywords, such as *privacy* or *security*, often leads to omission problems, where a large majority of domain-specific concerns are not correctly identified. Therefore, domain knowledge should be used to guide automated algorithms toward privacy-specific issues. Our work in this paper shows that privacy indicator keywords in an application domain can provide such knowledge. The idea of using keywords, or linguistic seeds, has long been used in text processing to address the label scarcity bottleneck [49, 67]. Seeds that are derived from domain knowledge can guide text classification and topic modeling algorithms towards critical, but under-represented themes in the data.

In terms of expected manual effort, an argument could be made about the first phase of our analysis, where privacy-indicator keywords have to be extracted manually. However, our results show that using only statistically representative samples of reviews is enough to extract the most important keywords, where most keywords should be detected after only a few rounds of manual inspection. Therefore, the expected manual effort can be less significant than manually labeling an entire dataset of reviews for supervised classification tasks.

Based on our findings, we can conclude with a large degree of confidence that leveraging domain knowledge in text summarization can help to produce more concise and more descriptive summaries of privacy concerns in mobile app reviews. Summarization techniques have the advantage of being unsupervised; no large datasets of ground-truth data need to be labeled based on pre-defined labels (e.g., SUR-Miner [23] and MARS [27]). Generated summaries can also be more easily interpretable than the results often generated by standard text classification methods, where classified reviews need to be individually vetted to extract the most common privacy complaints present in the reviews. Instead, summarization algorithms only produce a small number of representative reviews that encompass the main themes in the reviews.

Summarization techniques can also have an advantage over unsupervised topic modeling techniques, such as LDA. To support this claim, we examine the performance of LDA in capturing privacy concerns in our test dataset of reviews [51, 53, 68]. We first apply text processing to enhance the quality of generated topics. In particular, we remove non-ASCII characters and URLs and exclude English stop-words. Lemmatization is applied to the resulting list of words. LDA hyper-parameter α is set to be automatically learned from the corpus and β is set to 1/(number of topics). To ensure the stability of generated topics, the number of iterations (burnout) for the sampling process is set to 1000 [21]. To determine the number of topics, we rely on Gensim's coherence score. Topic coherence provides an objective measure to judge how good a given topic model is. Our analysis shows that, at around 5-8 topics, LDA generates the most cohesive topics for our dataset.

The most probable five topics generated by LDA for each domain are shown in Table 8; none of the generated topics seems to encompass any coherent theme related to privacy. For instance, the third topic generated for the investing apps includes privacy-related keywords such as *"card"* and *"info"*. However, the topic also includes other irrelevant terms such as *"stock"* and *"time"*, which commonly appear in reviews related to the availability of the service at a specific time frame. In general, most generated topics point toward

Table 7: The most common privacy concerns raised in the generated summaries (length of 50 reviews) for each of our subject application domain.

Dataset	Concern	Occurrence
Mental Health	Requesting email/Facebook data	30%
	Asking for sensitive information before the free trial	20%
	Collecting patient information (health info, job history, lastname)	16%
	Requesting unnecessary permissions (location, device data)	12%
	Sharing and selling users' data	8%
	Other	12%
	Not privacy-related	24%
Investing	Requesting access to camera and microphone	26%
	Collecting identification information (SSN, birthdate, driver's license)	24%
	Collecting Financial information (bank statements, income, tax info, login info)	22%
	Other	10%
	Not privacy-related	30%
Food Delivery	Asking for personal information before trial use	22%
	Requesting unnecessary permissions (microphone, camera, contacts)	20%
	Asking for users' credit card information	10%
	Selling and sharing personal information to third parties	10%
	Other	8%
	Not privacy-related	32%

issues that have been detected by classical Hybrid TF.IDF, including issues related to apps being expensive, such as the topic {*get, time, work, free, try, make, pay, money, version, one*} generated for the set of investing apps, or bad customer service, such as the topic {*service, order, customer, available, never, area, deliver, time, restaurant, item*} generated for the set of food delivery apps. These results can be largely attributed to the sparsity of privacy reviews (topics) in the data and their limited length [8, 28, 73]. Prior evidence has shown that LDA does not perform well when the input documents are short in length [8, 28, 73]. This leads LDA to downgrade topics related to privacy in favor of more prevalent topics, such as usability or reliability.

In terms of impact, our work in this paper bridges an important gap in mining mobile app users' privacy concerns. Our expectation is that such information can help app developers to adjust their release engineering strategies to mitigate their end users' privacy concerns and sustain their trustworthiness [17, 22, 32, 55]. This can be very critical for domains such as public health, where apps typically demand access to more personal information than the average app. For instance, health departments across the world have been using virus-tracking mobile apps to track down Covid-19 outbreaks. However, recent surveys have shown that a large percentage of the world population abstained from installing these apps due to privacy and mistrust concerns [4, 10]. Addressing these concerns can enhance these apps' adoption rates, thus contributing to the world's ongoing effort to fight the Covid-19 pandemic [66].

7 THREATS TO VALIDITY

The study presented in this paper has several limitations that could potentially limit the validity of the results. The main threat to the external validity of our study stems from the fact that only the top apps from three application domains were considered in our analysis. This could limit the generalizability of our results over other apps, domains, or even less popular apps from these domains. However, given their large user-bases, popular apps often receive significantly more feedback than less-popular apps [48]. Therefore, privacy issues are more likely to manifest over these apps than smaller ones. Furthermore, we evaluated our approach over an unseen-before test dataset of apps. This has helped to enhance the confidence in the external validity of our approach.

Table 8: The top five topics generated by LDA for the app reviews in our test dataset.

Dataset	Topics	Most probable words
Mental Health	Topic 1	get, time, work, free, try, make, pay, money, version, one
	Topic 2	time, use, pay, free, even, like, mood, need, activity, month
	Topic 3	pay, game, money, day, log, happiness, option, email, account, version
	Topic 4	like, time, better, work, get, make, way, pay, subscription, feel
	Topic 5	work, people, like, money, time, free, premium, help, great, happy
Investing	Topic 1	account, money, stock, even, fee, get, take, time, day, bank
	Topic 2	stock, money, day, account, sell, bank, customer, market, trade
	Topic 3	account, stock, time, money, buy, bank, card, log, back, sell
	Topic 4	stock, trade, time, price, money, buy, fee, make, use, email
	Topic 5	account, bank, get, service, email, money, customer, stock, use, still
Food Delivery	Topic 1	use, order, work, time, even, get, like, service, address, screen
	Topic 2	area, deliver, order, get, time, say, number, still, phone, email
	Topic 3	service, order, customer, available, never, area, deliver, time
	Topic 4	order, time, hour, deliver, never, minute, get, driver, service, food
	Topic 5	deliver, order, price, item, get, like, time, driver, customer, food

A potential threat to the internal validity of our study might originate from the fact that, in the first phase of our analysis, domain experts were used to manually label privacy-related reviews and privacy indicative keywords. To enhance the validity of this process, a discussion session was held before running the labeling sessions to make sure that all experts were clear on their assignments and that all of their questions and concerns were addressed. These sessions included labeling samples of reviews to test-run our procedure. Furthermore, each expert only had to examine a statistically representative sample of reviews if the number of retrieved reviews was more than 300. No time constraint was enforced to minimize fatigue. Overall, these measures helped to preserve the integrity of the manually-labeled data; a small conflict rate of $\sim 5\%$ was detected between domain expert classifications.

Other internal validity threats might arise from the app review sampling process. In particular, we only included low-rating reviews in our analysis. This might have led to the exclusion of some informative reviews from the data. However, as recent evidence has shown, reviews expressing user concerns, and especially privacy concerns, are often associated with low star-ratings [24, 36, 67]. Therefore, excluding high rating-reviews is highly unlikely to lead to concern omission.

8 CONCLUSION

In this paper, we proposed a novel unsupervised summarization approach for detecting privacy concerns in mobile app reviews. In the first phase of our analysis, we used an iterative word generation process to extract keywords indicative of privacy issues in three different mobile app domains. Our analysis showed that users in different application domains use different vocabulary to express their privacy concerns. This vocabulary is mainly related to the features of the app and its operational characteristics. In the second phase of our analysis, extracted keywords were used as seeds to help Hybrid TF.IDF, a generic text summarization technique, extract privacy-related reviews. Our evaluation showed that seeding Hybrid TF.IDF with domain-specific keywords helped to generate privacy-focused summaries. The results also showed that using word embeddings to calculate the semantic similarity between extracted summary reviews reduced the redundancy of generated summaries for each application domain. Our proposed approach is intended to help mobile app developers working in agile teams to quickly and accurately identify the most pressing privacy issues in their domain of operation, and ultimately, propose design solutions to mitigate these issues and enhance their chances of survival. A replication package is submitted to enable independent replications of our study[1].

The work presented in this paper will be extended along three main directions. First, we will continue to evaluate the proposed approach against other existing approaches and over other application domains. Our objective is to generate catalogs of comprehensive taxonomies and even NLP patterns that are indicative of privacy issues in the mobile app market [18]. Second, the generated taxonomies will be used to systematically tune different text summarization, modeling, and classification techniques and identify near optimal configurations (e.g., summary length, similarity thresholds, etc.) to calibrate these techniques. Third, working prototypes will be developed to assess, through longitudinal studies, the usability and long-term practical significance of our approach. Ultimately, our objective is to help app developers understand how their end-users perceive their app's privacy practices and how these practices can impact their ratings and retention rates [17].

ACKNOWLEDGMENTS

This work was supported by the U.S. National Science Foundation (Award CNS 1951411) and in part by the LSU Provost's Fund for Innovation in Research - Faculty Travel Grant Program.

[1] https://seel.cse.lsu.edu/data/ase22.zip

REFERENCES

[1] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, Yang Wang, and Shomir Wilson. 2017. Nudges for privacy and security: Understanding and assisting users' choices online. *Comput. Surveys* 50, 3 (2017), 44.

[2] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2016. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations.*

[3] Abdulbaki Aydin, David Piorkowski, Omer Tripp, Pietro Ferrara, and Marco Pistoia. 2017. Visual configuration of mobile privacy policies. In *Inter. Conf. on Fundamental Approaches to Software Engineering.* 338–355.

[4] Muhammad Ajmal Azad, Junaid Arshad, Syed Muhammad Ali Akmal, Farhan Riaz, Sidrah Abdullah, Muhammad Imran, and Farhan Ahmad. 2021. A First Look at Privacy Analysis of COVID-19 Contact-Tracing Mobile Applications. *IEEE Internet of Things Journal* 8, 21 (2021), 15796–15806.

[5] Yizhaq Benbenisty, Irit Hadar, Gil Luria, and Paola Spoletini. 2021. Privacy as first-class requirements in software development: A socio-technical approach. In *IEEE/ACM International Conference on Automated Software Engineering.* 1363–1367.

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

[7] Jaspreet Bhatia and Travis Breaux. 2018. Semantic incompleteness in privacy policy goals. In *International Requirements Engineering Conference.* 159–169.

[8] Lidong Bing, Wai Lam, and Tak-Lam Wong. 2011. Using query log and social tagging to refine queries based on latent topics. In *International Conference on Information and Knowledge Management.* 583–592.

[9] David Blei, Andrew Ng, and Michael Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022.

[10] Surekha Borra. 2020. *COVID-19 Apps: Privacy and Security Concerns.* Springer Singapore.

[11] Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *German Society for Computational Linguistics* 30 (2009), 31–40.

[12] Ning Chen, Jialiu Lin, Steven Hoi, Xiaokui Xiao, and Boshen Zhang. 2014. AR-miner: mining informative reviews for developers from mobile app marketplace. In *International Conference on Software Engineering.* 767–778.

[13] Jackie Cheung. 2008. Comparing abstractive and extractive summarization of evaluative text: controversiality and content selection. *Thesis in the Department of Computer Science of the Faculty of Science, University of British Columbia* 47 (2008).

[14] Adelina Ciurumelea, Andreas Schaufelbühl, Sebastiano Panichella, and Harald Gall. 2017. Analyzing reviews and code of mobile apps for better release planning. In *Inter. Conf. on Software Analysis, Evolution and Reengineering.* 91–102.

[15] Lisa Cosgrove, Justin Karter, and Zenobia Morrill. 2020. Psychology and Surveillance Capitalism: The Risk of Pushing Mental Health Apps During the COVID-19 Pandemic. *Journal of Humanistic Psychology* 60, 5 (2020), 611–625.

[16] Laura Dennison, Leanne Morrison, Gemma Conway, and Lucy Yardley. 2013. Opportunities and Challenges for Smartphone Applications in Supporting Health Behavior Change: Qualitative Study. *Journal of Medical Internet Research* 15, 4 (2013), e86.

[17] Andrea Di Sorbo, Giovanni Grano, Corrado Aaron Visaggio, and Sebastiano Panichella. 2021. Investigating the criticality of user-reported issues through their relations with app rating. *Journal of Software: Evolution and Process* 33, 3 (2021), e2316.

[18] Andrea Di Sorbo, Sebastiano Panichella, Corrado A. Visaggio, Massimiliano Di Penta, Gerardo Canfora, and Harald C. Gall. 2021. Exploiting Natural Language Structures in Software Informal Documentation. *IEEE Transactions on Software Engineering* 47, 8 (2021), 1587–1604.

[19] Tawanna Dillahunt and Amelia Malone. 2015. The Promise of the Sharing Economy Among Disadvantaged Communities. In *Annual ACM Conference on Human Factors in Computing Systems.* 2285–2294.

[20] Fahimeh Ebrahimi, Miroslav Tushev, and Anas Mahmoud. 2020. Mobile App Privacy in Software Engineering Research: A Systematic Mapping Study. *Information and Software Technology* 133 (2020).

[21] Thomas Griffiths and Mark Steyvers. 2004. Finding scientific topics. *National Academy of Sciences* 101, 1 (2004), 5228–5235.

[22] Jie Gu, Yunjie (Calvin) Xu, Heng Xu, Cheng Zhang, and Hong Ling. 2017. Privacy concerns for mobile app download: An elaboration likelihood model perspective. *Decision Support Systems* 94 (2017), 19–28.

[23] Xiaodong Gu and Sunghun Kim. 2015. What Parts of Your Apps Are Loved by Users?. In *International Conference on Automated Software Engineering.* 760–770.

[24] Hui Guo and Munindar Singh. 2020. Caspar: extracting and synthesizing user stories of problems from app reviews. In *International Conference on Software Engineering.* 628–640.

[25] Emitza Guzman, Muhammad El-Haliby, and Bernd Bruegge. 2015. Ensemble methods for app review classification: An approach for software evolution (n).

In *International Conference on Automated Software Engineering*. 771–776.

[26] Udo Hahn and Inderjeet Mani. 2000. The challenges of automatic summarization. *Computer* 33, 11 (2000), 29–36.

[27] Majid Hatamian, Jetzabel Serna, and Kai Rannenberg. 2019. Revealing the unrevealed: Mining smartphone users privacy perception on app markets. *Computers & Security* 83 (2019), 332–353.

[28] Liangjie Hong and Brian Davison. 2010. Empirical Study of Topic Modeling in Twitter. In *Workshop on Social Media Analytics*. 80–88.

[29] Leonard Hoon, Rajesh Vasa, Jean-Guy Schneider, and Kon Mouzakis. 2012. A preliminary analysis of vocabulary in mobile app user reviews. In *Australian Computer-Human Interaction Conference*. 245–248.

[30] Hanyang Hu, Shaowei Wang, Cor-Paul Bezemer, and Ahmed E. Hassan. 2019. Studying the Consistency of Star Ratings and Reviews of Popular Free Hybrid Android and IOS Apps. *Empirical Softw. Engg.* 24, 1 (2019), 7–32.

[31] David Inouye and Jugal Kalita. 2011. Comparing twitter summarization algorithms for multiple post summaries. In *International Conference on Privacy, Security, Risk and Trust and International Conference on Social Computing*. 298–306.

[32] Pew Internet. 2012. Apps and privacy: More than half of app users have uninstalled or decided to not install an app due to concerns about their personal information. Retrieved May 2022 from https://www.pewresearch.org/internet/2012/09/05/privacy-and-data-management-on-mobile-devices-2/

[33] Nishant Jha and Anas Mahmoud. 2018. Using frame semantics for classifying and summarizing application store reviews. *Empirical Software Engineering* 23, 6 (2018), 3734–3767.

[34] Tom Kenter, Alexey Borisov, and Maarten Rijke. 2016. Siamese cbow: Optimizing word embeddings for sentence representations. *arXiv preprint arXiv:1606.04640* (2016).

[35] Elham Khabiri, James Caverlee, and Chiao-Fang Hsu. 2011. Summarizing user-contributed comments. In *International Advancement of Artificial Intelligence Conference on Weblogs and Social Media*.

[36] Hammad Khalid, Emad Shihab, Meiyappan Nagappan, and Ahmed Hassan. 2015. What do mobile app users complain about? *IEEE Software* 32, 3 (2015), 70–77.

[37] Zijad Kurtanović and Walid Maalej. 2017. Mining user rationale from software reviews. In *International Requirements Engineering Conference*. 61–70.

[38] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*. 1188–1196.

[39] Li Li, Tegawendé Bissyandé, Mike Papadakis, Siegfried Rasthofer, Alexandre Bartel, Damien Octeau, Jacques Klein, and Le Traon. 2017. Static analysis of Android apps: A systematic literature review. *Information and Software Technology* 88 (2017), 67–95.

[40] Clare Llewellyn, Claire Grover, and Jon Oberlander. 2014. Summarizing newspaper comments. In *International Advancement of Artificial Intelligence Conference on Weblogs and Social Media*. 599—-602.

[41] Robert Longyear and Kostadin Kushlev. 2021. Can mental health apps be effective for depression, anxiety, and stress during a pandemic? *Practice Innovations* 6, 2 (2021), 131–137.

[42] Edward Loper and Steven Bird. 2002. NLTK: The natural language toolkit. In *COLING/ACL on Interactive Presentation Sessions*. 69–72.

[43] Walid Maalej and Hadeer Nabil. 2015. Bug report, feature request, or simply praise? On automatically classifying app reviews. In *International Requirements Engineering Conference*. 116–125.

[44] Anas Mahmoud and Grant Williams. 2016. Detecting, classifying, and tracing non-functional software requirements. *Requirements Engineering* 21, 3 (2016), 357–381.

[45] Chris Martin. 2016. The sharing economy: A pathway to sustainability or a nightmarish form of neoliberal capitalism? *Ecological Economics* 121 (2016), 149–159.

[46] William Martin, Federica Sarro, Yue Jia, Yuanyuan Zhang, and Mark Harman. 2017. A Survey of App Store Analysis for Software Engineering. *IEEE Transactions on Software Engineering* 43, 9 (2017), 817–847.

[47] Stuart McIlroy, Nasir Ali, Hammad Khalid, and Ahmed Hassan. 2016. Analyzing and automatically labelling the types of user issues that are raised in mobile app reviews. *Empirical Software Engineering* 21, 3 (2016), 1067–1106.

[48] Stuart Mcilroy, Weiyi Shang, Nasir Ali, and Ahmed Hassan. 2017. User Reviews of Top Mobile Apps in Apple and Google App Stores. *Commun. ACM* 60, 11 (2017), 62–67.

[49] Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. 2018. Weakly-Supervised Neural Text Classification. In *ACM International Conference on Information and Knowledge Management*. 983–992.

[50] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*. 3111–3119.

[51] Shahab Mokarizadeh, Mohammad Rahman, and Mihhail Matskin. 2013. Mining and Analysis of Apps in Google Play. In *International Conference on Web Information Systems and Technologies*. 527–535.

[52] Debjyoti Mukherjee, Alireza Ahmadi, Maryam VahdatPour, and Joel Reardon. 2020. An Empirical Study on User Reviews Targeting Mobile Apps' Security & Privacy. *arXiv preprint arXiv:2010.06371* (2020).

[53] Maleknaz Nayebi, Homayoon Farrahi, Ada Lee, Henry Cho, and Guenther Ruhe. 2016. More insight from being more focused: Analysis of clustered market apps. In *International Workshop on App Market Analytics*. 30–36.

[54] Duc 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 *Symposium on Security and Privacy*. 555–569.

[55] John Grundy Mohamed Abdelrazek Omar Haggag, Sherif Haggag. 2021. COVID-19 Vs Social Media apps: Does privacy really matter?. In *International Conference on Software Engineering - Software Engineering in Society*.

[56] Sebastiano Panichella, Andrea Di Sorbo, Emitza Guzman, Corrado Visaggio, Gerardo Canfora, and Harald Gall. 2015. How can I improve my app? Classifying user reviews for software maintenance and evolution. In *International Conference on Software Maintenance and Evolution*. 281–290.

[57] Elias Papadopoulos, Michalis Diamantaris, Panagiotis Papadopoulos, Thanasis Petsas, Sotiris Ioannidis, and Evangelos Markatos. 2017. The Long-Standing Privacy Debate: Mobile Websites vs Mobile Apps. In *Inter. Conf. on World Wide Web*. 153–162.

[58] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing*. 1532–1543.

[59] Elizabeth Poché, Nishant Jha, Grant Williams, Jazmine Staten, Miles Vesper, and Anas Mahmoud. 2017. Analyzing user comments on YouTube coding tutorial videos. In *International Conference on Program Comprehension*. 196–206.

[60] PYMNTS. 2021. High-Speed Traders Pay Robinhood $331 Million In Q1 To Execute Trades. Retrieved October 15, 2021 from https://www.pymnts.com/earnings/2021/high-speed-traders-pay-robinhood-331-million-dollars-q1-execute-trades/

[61] Amir Sadeghian and Alireza Sharafat. 2015. Bag of words meets bags of popcorn. (2015).

[62] Hani Safadi, Weifeng Li, Pouya Rahmati, Saber Soleymani, Krzysztof Kochut, and Amit Sheth. 2020. Curtailing Fake News Propagation with Psychographics. *SSRN Electronic Journal* (2020).

[63] Andrea Sorbo, Sebastiano Panichella, Carol Alexandru, Junji Shimagaki, Corrado Visaggio, Gerardo Canfora, and Harald Gall. 2016. What would users change in my app? Summarizing app reviews for recommending software changes. In *International Symposium on Foundations of Software Engineering*. 499–510.

[64] Lauren Squires. Language in society. Enregistering internet language. *2010* 39, 4 (Language in society), 457–492.

[65] Levi Sumagaysay. 2020. The pandemic has more than doubled food-delivery apps business. Now what? Retrieved October 15, 2021 from https://www.marketwatch.com.

[66] Ramona Trestian, Guodong Xie, Pintu Lohar, Edoardo Celeste, Malika Bendechache, Rob Brennan, Evgeniia Jayasekera, Regina Connolly, and Irina Tal. 2021. Privacy in a Time of COVID-19: How Concerned Are You? *IEEE Security and Privacy* 19, 5 (2021), 26–35.

[67] Miroslav Tushev, Fahimeh Ebrahimi, and Anas Mahmoud. 2022. Domain-Specific Analysis of Mobile App Reviews Using Keyword-Assisted Topic Models. In *International Conference on Software Engineering*.

[68] Svitlana Vakulenko, Oliver Müller, and Jan Brocke. 2014. Enriching iTunes App Store categories via topic modeling. In *International Conference on Information Systems*. 1–11.

[69] Axel van Lamsweerde. 2009. *Requirements Engineering: From System Goals to UML Models to Software Specifications*. Wiley.

[70] Rajesh Vasa, Leonard Hoon, Kon Mouzakis, and Akihiro Noguchi. 2012. A preliminary analysis of mobile app user reviews. In *Computer-Human Interaction Conference*. 241–244.

[71] Xiaoyin Wang, Xue Qin, Mitra Bokaei, Rocky Slavin, Travis Breaux, and Jianwei Niu. 2018. Guileak: Tracing privacy policy claims on user input data for Android applications. In *Inter. Conf. on Software Engineering*. 37–47.

[72] Joshua West, P. Cougar Hall, Carl Hanson, Michael Barnes, Christophe Giraud-Carrier, and James Barrett. 2012. There's an App for That: Content Analysis of Paid Health and Fitness Apps. *Journal of Medical Internet Research* 14, 3 (2012), e72.

[73] Grant Williams, Miroslav Tushev, Fahimeh Ebrahimi, and Anas Mahmoud. 2020. Modeling user concerns in Sharing Economy: the case of food delivery apps. *Automated Software Engineering* 27 (2020), 229–263.

[74] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *International Conference on World Wide Web*. 1445–1456.