Information Modelling and Knowledge Bases XXXIV
M. Tropmann-Frick et al. (Eds.)

doi:10.3233/FAIA220507

Permissions vs. Privacy Policies of Apps in Google Play Store and Apple App Store

Boštjan BRUMEN, Aljaž ZAJC, Leon BOŠNJAK
University of Maribor (www.um.si), Faculty of Electrical Engineering and Computer science, Smetanova 17, Si-2000 Maribor, Slovenia
bostjan.brumen@uni-mb.si

Abstract. The "free" business model prevails in mobile apps available through the major channels, hinting at the possibility that users "pay" for the use of the mobile apps by sharing their private data with the developers and platform providers. Several types of personal data and permissions of mobile applications were analyzed. We examined 636 apps in several categories, such as medical, health & fitness, business, finance, and entertainment. The types of personal data being requested by the apps were collected from their privacy policies and the list of permissions was scraped from the platform's store. We implemented a privacy policy word processing algorithm, the purpose of which was to gain a better insight into the types of data collected. Using the algorithm results, we also performed statistical analyses, based on which we found, expectedly, that free mobile applications collect more data than paid ones. However, there are discrepancies between the permissions we obtained from the privacy policy texts and those stated on the Google Play and Apple App Store websites. More permission requirements emerged from the privacy policy texts than were shown on corresponding app stores, which is a worrying result.

Keywords. privacy, mHealth, mobile applications, Google Play, Apple App Store, permissions

1. Introduction

The field of personal data is an increasingly important topic in today's information age. Almost every person today has a smartphone with several applications installed. Applications typically require the users to allow different permissions; apps access, store and process personal data based on those permissions. However, the users are generally unaware of what types of data and permissions they provide to mobile app developers and platform owners. Application developers, especially those using the "free" business model, finance their development by selling the acquired data [1-3]. This can violate human rights (the right to privacy) if users are unaware of it or it is being done without their explicit consent. A survey found that most privacy policy texts state that user data can be accessed by a third party, namely 71% on the iOS platform and 46% on the Android platform [4].

Furthermore, about 50% of the analyzed applications did not have a text on data privacy that could inform users about the use of their data [2]. Therefore, application developers and owners should be held accountable and minimize the damage resulting from inappropriate data processing. We should advocate for the participation of

developers, platform owners, privacy policy writers, and other stakeholders so that they work together and adhere to standards.

In this work, we examined which app permissions users must give to each application for its use; these permissions are stated on the platform (Google Play or Apple App Store). Next, we analyzed the privacy policies, extracted the keywords related to permissions, and compared the expressed permissions on the platform with those extracted from the privacy policies. We were particularly interested in the categories of health-related applications (mHealth), as such data are particularly sensitive. Android (and iOS) platforms have the mHealth app category broken down into two separate categories, "Medical" and "Health&Fitness." Additionally, we checked apps' privacy policies and permissions in categories "Entertainment," "Finance," and "Business."

In this paper, the research question is, how accurate are the stated lists of required permissions compared to those listed in privacy policies, and if mHealth apps are any different from the three other categories.

The rest of the paper is organized as follows. Section 2 presents the literature review dealing with medical apps, data collection, and privacy. In Section 3, we describe our research method and deliver the results. In Section 4, we conclude the paper with final remarks.

2. Literature review

Mobile applications related to health and healthy living (shorter "mHealth" for mobile health) are becoming increasingly popular. They provide the user with convenient and fast access to health and healthy living-related data. Although these applications bring many benefits to the user, they can also threaten the security and privacy of personal data. A 2015 Kerbs and Duncan survey found that 58% of U.S. citizens have health-related mobile apps installed on their phones [5]. They use mobile health applications for various purposes, such as lifestyle, diet, fitness, medication guidelines, diagnoses, various treatments, nutrition, etc. These applications improve peoples' daily lives and reduce costs, but at the same time, they can collect large amounts of personal data [5]. Research in [20] showed that in a sample of 61 mHealth applications, all required access (permissions) to various services on the phone. Many of these apps also encouraged users to share their information online. About half of the applications reviewed did not have information on the privacy policy texts. In these texts, companies write how they will obtain, use, disclose and manage the user's personal data. The privacy policies (should) also state to which third parties this information will be provided [2].

The results of the survey showed that 60% of users who download an app from the Google Play Store or iOS App Store decided not to install the app because it requires a lot of personal information from them; 43% of users chose to delete it application for the same reason [4].

Another study was done on a small number of selected applications for depression. In 116 applications, 4% of those that received a transparency rating according to their criteria were acceptable, and 28% received a questionable rating. However, other applications, which accounted for the majority, i.e., 68%, were assessed as unacceptable. Less than half of these applications (49%) had privacy policy text. Based on the research, it was found that apps on the App Store are more likely to contain privacy policy text than those uploaded to the Google Play platform. Applications requiring users to provide

identification-related information are said to have the text of the privacy policy in more cases than those that did not [6].

In an article [7], an online survey among students was conducted. The purpose of the study was to find out how many students agree with the terms of using the mobile app and the privacy policy texts without having to read them. The study also examined whether students grant permission to an application without rejecting it for security reasons. The survey was conducted on 170 students. The results showed that as many as 62% approve of the terms of use and the privacy policy without reading its text. The reason for this is mainly that the texts are too long. However, the results are pretty different in terms of the permissions required by each application. As many as 92% of respondents said they refuse an individual permit if they believe that the application does not require access to that permission [7].

A study [4] found that more than two-thirds of the privacy policy texts analyzed in mHealth applications stated that user data could be shared with a third party. The collection of personal data was mentioned in most of the texts analyzed. However, 29% of the total applications studied did not mention what types of data they would collect in the privacy policy texts [4].

Another study involving 15838 mHealth apps [8] has shown that data collection operations and transmissions in apps traffic involved external service providers (third parties), and the top 50 third parties were responsible for most of the data collection operations. Worryingly, 23.0% of user data transmissions occurred on insecure communication protocols and 28.1% of apps provided no privacy policies.

In another work [9] the authors studied 23 mHealth apps in depth; all collected personal health-related data and allowed behavioral tracking, and 61% of the apps allowed location tracking. Only 70% had a privacy policy displayed. In addition, 3 apps apps collected data before obtaining consent.

If mobile health applications were to be more widely accepted among users and in the health system, security and privacy issues would need to be addressed more consistently and thoroughly [10]. As [11] noted, the privacy assessment of mHealth apps is a complex task, as the criteria used by different authors are very heterogeneous. In-app information and privacy policies are primarily utilized by the scientific community to extract privacy information from mHealth apps. However, the privacy policies are not being read by a vast majority of users. It is suspected that users rely on the permissions being asked by the apps and they perceive is as a way to control their data.

In the following sections, we present a method of analyzing and comparing the permissions as stated in the Google Play Store and Apple App Store versus those extracted from the privacy policies texts.

3. Method

This section presents how we collected and analyzed the permissions and privacy policies of the studied applications found in the Google Play Store and Apple App Store.

3.1. Data collection and preparation

We used the software libraries google-play-scraper 8.0.4 [12] and app-store-scraper 0.17.0 [13] to obtain the data from Google Play and Apple App Store, respectively. With the help of these two libraries, we created a program in the Python programming

language that scraped data about applications from the Google Play Store and Apple App Store websites. The program downloaded the data for the most popular applications from each selected category. Because we are interested in mHealth applications, we have chosen the "Medical" and "Health & Fitness" categories. Additionally, to be able to compare if apps from these two categories are any different from others, we additionally selected apps from the "Business," "Entertainment, " and "Financial" categories.

We had several difficulties retrieving the website data due to certain limitations that the Google Play Store and Apple App Store had. Namely, the sites did not allow the bulk collection of data from our side. After a certain amount of transferred data, they started to block access. We were forced to transfer applications' data gradually. We did this by downloading each category separately. Within the category, we further divided the downloading of data into free and paid applications. This was a very time-consuming process, as we were forced to set up a VPN to change IP addresses to allow for a seamless transmission. Program code #1 shows part of the entire program for scraping data about Android applications. This procedure is the same for each category.

```
gplay.list({
 category: gplay.category.HEALTH_AND_FITNESS,
 collection: gplay.collection.TOP_FREE,
 fullDetail: true,
 num: 160
}).then((apps) => {
 apps.forEach(function (element) {
  gplay.permissions({ appId: element["appId"] }).then((permissions) => {
   permissions.forEach(function (permission) {
    element[permission["type"]] = "yes"
   });
   android_health_and_fitness_free.push(element);
   //shranimo v .csv
   (async () => {
    const csv = new ObjectsToCsv(android_health_and_fitness_free);
    await csv.toDisk('./android/android_health_and_fitness_free.csv', {
allColumns: true });
   })();
  });
 });
});
```

Program code 1: Scraping of data of Android applications in the Health & Fitness category

Once we obtained the data of free and paid applications for each category, we then combined them into one .csv file. We discarded columns that were irrelevant to our research. We stored the list of all possible permissions and those required by each individual application.

However, only the URL link was present for the privacy policy text information, not the full privacy policy text. The URL link should lead to the page where this text is supposed to be located. We have created a new program that uses the Scrapy 2.3 software library [14] to download privacy policy texts. It does this by reading the URLs stored in the shared .csv file. A new column named PrivacyPolicyText is created in which the text is saved. The program cleans the content of each privacy policy text by removing the unnecessary HTML tags. The process is presented in Program code #2.

```
app_data = []

 def start_requests(self):
  logging.getLogger('scrapy').setLevel(logging.WARNING)
  urls = []

  self.df = pd.read_csv("./data/vsi_ios.split-19.csv", sep=",",
index_col=0) # Preberemo CSV
  self.df = self.df.head(100)
  data_dict = self.df.to_dict()
  all_apps = data_dict['url']

  for name, url in all_apps.items():
   if not pd.isnull(url): # samo tisti url-ji ki obstajajo
     #samo tiste domene ki obstajajo
     try:
     request = requests.get(url, timeout=6)
     except:
     continue
     if request.status_code == 200:
     urls.append(url)
     self.app_data.append({"name": name, "url": url})

  for url in urls:
   yield scrapy.Request(url=url, callback=self.parse_seznam_aplikacij)

 def parse_seznam_aplikacij(self, response):
  #print(response.request.url)
  privacyPolicyUrl = response.css('div.small-hide ul.inline-list
li.inline-list__item a::attr(href)').extract()[-1]
  if privacyPolicyUrl == None:
   privacyPolicyUrl = None

  try:
   name = response.xpath('//h1[@class="product-header__title app-
header__title"]/text()').extract_first().strip()
  except:
   name = None

  app = {}
  app['trackName'] = name
  app['privacyPolicy'] = privacyPolicyUrl
  yield app
```

Program code 2: Acquisition of URL links to privacy policies

In total, we identified 3073 applications, as follows:

- "Medical": 383 for iOS and 232 for Android,
- "Business": 374 for iOS and 233 for Android,
- "Entertainment": 371 for iOS and 255 for Android,
- "Health/Fitness": 370 for iOS and 255 for Android,

- "Finance": 360 for iOS and 239 for Android.

We immediately excluded two applications that did not have a proper name (value for the name field was missing).

The exclusion criteria were that the app should have a privacy policy in the first place and be written in English. So, once we had all the app data collected in one file, we decided to review some of the privacy policy texts manually. We randomly selected twenty applications and read their privacy policy texts. In addition to the reading itself, we also observed how well the text is refined and whether it contains any unreadable characters or not.

After just a few texts, we discovered that the texts were not in the form we expected. Either they included illegible characters, or only parts of the text were downloaded, or there were no texts at all. We then began to investigate the cause of this problem. We visited the URL link where the privacy policy text was supposed to be located manually. Many URL links did not work because the pages did not exist, or the links were to the main web page of a company and not to a subpage containing the text of the privacy policy.

In some cases, the URL link did not lead to the app's page but to the PDF document that contained that text. The Scrapy program did not copy the contents of a PDF document because it works only with the HTML structures. Another problem was that access to specific sites has been disabled for visitors outside the USA. This could be partially solved by using a VPN connection, but there were no software solutions for other previously mentioned problems. We were forced to find a solution.

The solution was to download all the privacy policy texts manually. We went through the list of all applications and visited their URL links, which should contain the privacy policy texts. If the URL link had privacy policy text, we copied and pasted it into the appropriate location in our application file. If the URL contained a link to a PDF, we saved its textual contents to the .csv file. If the URL link led to the address page of the company that developed the application, we tried to find a link to a subpage on the page that would contain the text of the privacy policy and then copy and save it to our file. This manual process was very time-consuming. It took about a month. With this manual process, we thus obtained many more privacy policy texts than we had in the automated process, and their content was readable and useful for further studies.

From the total list of all apps, we then selected only those that were available on both platforms. Program code #3 shows the process.

```
ios_apps = df_ios["title"].tolist()
ios_developer = df_ios["developer"].tolist()
ios_podatki = merge(ios_apps, ios_developer)

android_apps = df_android["title"].tolist()
android_developer = df_android["developer"].tolist()
android_podatki = merge(android_apps, android_developer)


enake_aplikacije = []

enake_apps_df_proizvajalec = pd.DataFrame(columns=df_ios.columns)#prazen
df z imeni stolpcev

for ios in ios_podatki:
 for android in android_podatki:
  if len(str(ios[1]).lower()) < len(str(android[1]).lower()):
   if (str(android[1]).lower().find(str(ios[1]).lower().split(' ')[0]) !=
-1):
    if len(str(ios[0]).lower()) < len(str(android[0]).lower()):
     if (str(android[0]).lower().find(str(ios[0]).lower()) != -1):
      enake_aplikacije.append((ios[0],android[0]))
      enake_apps_df_proizvajalec = pd.concat([df_ios.loc[df_ios['title']
== ios[0]], enake_apps_df_proizvajalec])
      enake_apps_df_proizvajalec =
pd.concat([df_android.loc[df_android['title'] == android[0]],
enake_apps_df_proizvajalec])
    else:
     if (str(ios[0]).lower().find(str(android[0]).lower()) != -1):
      enake_aplikacije.append((ios[0],android[0]))
      enake_apps_df_proizvajalec = pd.concat([df_ios.loc[df_ios['title']
== ios[0]], enake_apps_df_proizvajalec])
      enake_apps_df_proizvajalec =
pd.concat([df_android.loc[df_android['title'] == android[0]],
enake_apps_df_proizvajalec])
```

Program code 3: "Find same apps on both platforms" procedure

We ended up with a population of 636 apps with privacy policies in English from the aforementioned five categories (please note in the iOS group there are 2 apps more than in Android due to Pro and Lite versions of two apps that were available for the iOS but not for the Android). **Figure 1** depicts the process.

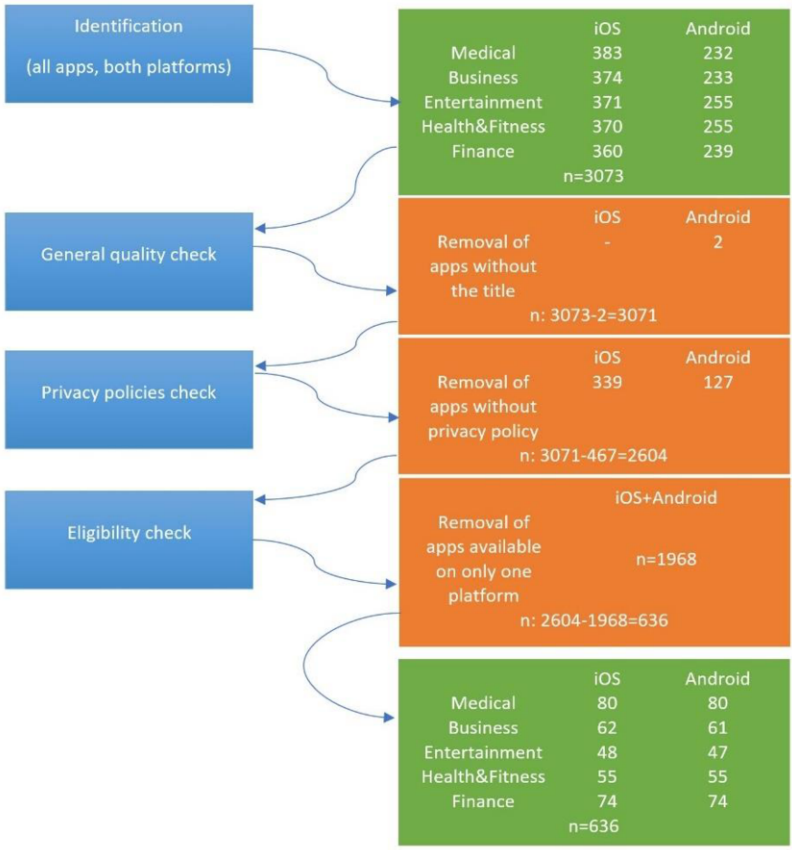

Figure 1: App selection process with inclusion and exclusion criteria

3.2. Data processing

We implemented our Privacy Policy processing algorithm for each of the 636 applications. In the article [15], the authors came up with the idea that the texts of a privacy policy could be processed using the `nltk` (Natural language toolkit) module [15]. With the help of this machine learning module, the authors of [7] analyzed the comments of the applications published on Google Play. Namely, they wanted to find out which comments are related to the security vulnerabilities of the applications. This would make it easier for developers to find security vulnerabilities. They succeeded with the implementation of MARS (Mobile App Reviews Summarization). The latter uses the `nltk` module for its operation. Their algorithm works by first filtering and preprocessing comments, then breaking each sentence into words (tokenization), finding and removing irrelevant words (stop words), and the stem of a word (stemming) [15]. We used a similar process to implement our algorithm, as was also the case in other research, e.g., [3, 6].

First, we prepared a list of synonyms for each permission from the obtained list of all permissions. These synonyms were determined for each permission separately. We did this by reading some of the privacy policy texts and looking for the information that

appears in them and is tied to it. E.g., for "Personal Name," we searched if a name is mentioned somewhere in the text. This could have come in various forms, such as Personal name, First name, Given name, etc. All of these words represent synonyms for our "Personal Name" permission, representing the fact that an app is collecting personal names. We determined synonyms for all permissions. However, because we wanted to have as many synonyms as possible, we wrote a program that scraped the appropriate synonyms given the word. So, in the JavaScript programming language, we created a web scraper using the powerthesaurus-API 2.0.3 [16] library, which visited www.thesaurus.com and downloaded all its synonyms based on the input word. The synonyms returned from the web page were saved in a .txt file. These synonyms had to be manually reviewed because they were not all relevant. Some had nothing to do with the original word, so we excluded them.

For example, for the permission "Medical Diagnoses" we identified the following synonyms:

- health information
- symptoms
- health plan
- health condition
- accident information
- health conditions
- diseases
- syndrome
- symptom
- abnormality
- health situation
- health status
- medical state
- physical conditions
- biological information
- illness

Next, we did a preprocessing of the text of the privacy policies. We redesigned the base text and a list of synonyms to be suitable for processing with the `nltk` module. This step is significant as such texts can be very long and demanding to process. The preprocessing process includes the following steps:

1. tokenization: each sentence in the text is broken down into several tokens or words, which will later be used to remove irrelevant words (stop words) and convert them to the origin of the word (stemming). Synonyms are also broken down into several tokens.
2. stop words removal: many unimportant words in the text do not help us analyze and search for personal data and permits. Examples of such words would be: 'i', 'me', 'my', 'myself', 'we', 'our', 'ours',' ourselves', 'you', 'you're', '. ',' * '… We remove these irrelevant words from our text. All words are capitalized in lowercase.
3. stemming: we converted the word into the original form (root). An example would be the words "argue," "argued," and "arguing" - all of which change into the basic form of "argu" in this process [15].

For example, the aforementioned list of synonyms for "Medical Diagnoses" looks as follows after the processing:

- health inform
- symptom
- health plan
- health condit
- accid inform
- health condit
- diseas
- syndrom
- symptom
- abnorm
- health situat
- health statu
- medic state
- physic condit
- biolog inform
- ill

Once we had all the data ready for processing, we were able to run the word search algorithm in the privacy policy texts.

The algorithm works by using a function that accepts as a parameter a preprocessed synonym for a particular column representing permission or personal data. It also takes preprocessed privacy policy text for each application.

The algorithm generally divides one initial field of synonyms into four fields containing strings with the same number of words. A similar thing happens with the preprocessed text of the privacy policy. The whole text is broken into strings of one word, two words, three words, and four words. These differently long strings of words are each stored in their own box according to their length. The algorithm then first compares the contents of the synonyms field and the privacy policy text fields containing a one-word string. Each set of synonyms is compared to each set of the privacy policy text. If it finds the appearance of a synonym in the text for an individual column, the function returns True, else False. If True is returned, the search is aborted, and the same process is performed for the second column. The same function is called again, which gets the same privacy policy text as the input parameter, but different synonyms belonging to the second column. In this way, the program is executed for all columns. However, if no synonym is found in the privacy policy text, the program returns False.

The search is repeated so that the fields containing two words in the synonym string and two words in the string containing the privacy policy text information are now compared. Next, the search is repeated for a three-word string and a four-word string. Based on our list of prepared synonyms, there was no occurrence of a synonym longer than four words.

Table 1 shows the results of running our algorithm on two test apps, with listed personal data and permissions, as detected through the privacy policies.

Table 1: A list of permissions and personal data for two test apps

Personal data and permissions	Application name	
	ZOOM Cloud Meetings	Microsoft Teams
Calendar	True	True
Reminder	False	True
Contact	True	True
Photo	True	True
Bluetooth	True	False
Microphone	True	True
Camera	True	True
Location	True	True
HealthKit	False	False
Media Library	False	True
Motion	False	True
Documents	True	True
Language	False	True
Notifications	True	True
Personal Name	True	True
Age	True	True
Phone number	True	True
ID Card	True	True
E-mail	True	True
Physical address	True	True
Medical diagnoses	False	False
Medical ID numbers	False	False
Medical history	False	False
Insurance	False	False
User activity	True	True
Device info.	True	True
Payment	True	True
Social media	False	True
Ads	True	True
IAP	False	False

3.3. Results

In the following sub-sections, we present the results of the processing.

3.3.1. Presence of privacy policies

Firstly, we show the results of the analysis of the presence of privacy policies. Table 2 presents the number and the percentage of apps without privacy policy, and total number of apps, in iOS and Android segment, for each of the selected categories.

From Table 2, it can be seen that 467 or 15 percent of all apps (n=3071) did not have a privacy policy, or it was hidden in such a way that we could not have found it even after an extensive search at the developer's website. This is a surprisingly high percentage.

We were interested whether the percentages of missing privacy policies differ by categories by chance or not. Chi-Square test of independence [17] was performed for both Android and iOS platform.

Table 2: Apps without privacy policies

Category	Android without privacy policy	iOS without privacy policy	Total
Business	11% (26/233)	17% (63/374)	15% (89/607)
Entertainment	15% (37/255)	23% (84/371)	19% (122/627)
Finance	8% (19/239)	19% (70/360)	15% (89/598)
Health&Fitness	6% (15/255)	12% (44/370)	9% (59/625)
Medical	13% (30/232)	20% (78/383)	18% (108/614)
Total	10% (127/1212)	18% (339/1858)	15% (467/3071)
Chi-Square (value)	13,411	16,813	
Sig.	0,009	0,002	

We can conclude that there is a significant association between the apps' category and whether or not they have a privacy policy available ($\chi(4) = 13{,}411$, $p = 0{,}009$ for Android and $\chi(4) = 16{,}813$, $p = 0{,}002$).

However, there is no significant association between apps' category and its missing percentage of privacy policies across the two platforms ($\chi(4) = 0{,}976$, $p = 0{,}913$).

3.3.2. Personal data collected

Browsing through the privacy policies with our algorithm we extracted the following personal data are being collected.

Table 3: Personal info collected by apps, Android platform

Personal data Android	Health& Fitness	Medical	Business	Enter-tainment	Finance	Chi-S	Sig.
Personal Name	89%	77%	91%	82%	89%	1,626	0,804
Age	66%	69%	61%	60%	68%	1,031	0,905
Phone number	97%	94%	96%	96%	98%	0,091	0,999
ID Card	42%	27%	47%	24%	28%	12,417	**0,015**
E-mail	89%	83%	91%	98%	89%	1,129	0,863
Physical address	53%	52%	49%	38%	38%	4,826	0,306
Medical diagnoses	10%	8%	4%	36%	55%	86,159	**<0,001**
Medical IDs	10%	6%	16%	47%	53%	72,924	**<0,001**
Medical history	28%	13%	16%	53%	69%	66,223	**<0,001**
Insurance	21%	4%	24%	22%	41%	30,768	**<0,001**
User activity	95%	96%	92%	89%	88%	0,543	0,969
Device info.	97%	96%	99%	96%	95%	0,095	0,999
Payment	98%	79%	96%	78%	80%	4,557	0,336
Social media	65%	60%	69%	56%	45%	5,797	0,215
Ads	79%	90%	80%	76%	78%	1,479	0,830
IAP	10%	1%	13%	3%	27%	14,333	**0,002**

In the Table 3 (above) and Table 4 (below), we can see that applications on both platforms collect a lot of personal data. The percentages show how many applications in each category collect certain data.

Mostly, more than 90%, apps are collecting phone numbers, e-mail addresses, logging of user activity, phone data and payment methods. Apps also do a lot of advertisements, which is shown in the "Ads" line. We can especially highlight the lines Phone number and Device info, because here the percentages for all categories are very high (higher than 93%).

We can also see which personal data they collect the least. In the table, the line IAP (in-app purchases) stands out. Developers therefore have the lowest demand for these data - the highest for the Medical category is barely 27% in Android and 15 % for the Entertainment in iOS.

Table 4: Personal info collected by apps, iOS platform

Personal data iOS	Health& Fitness	Medical	Business	Enter-tainment	Finance	Chi-S	Sig.
Personal Name	87%	81%	85%	80%	91%	0,953	0,917
Age	62%	66%	61%	62%	63%	0,236	0,994
Phone number	97%	94%	93%	95%	99%	0,243	0,993
ID Card	41%	30%	42%	27%	28%	6,345	0,175
E-mail	90%	89%	84%	96%	91%	0,882	0,935
Physical address	53%	47%	47%	38%	36%	4,498	0,343
Medical diagnoses	7%	11%	1%	38%	54%	93,099	**<0,001**
Medical IDs	7%	9%	12%	42%	54%	75,758	**<0,001**
Medical history	28%	15%	12%	55%	69%	70,693	**<0,001**
Insurance	25%	6%	26%	24%	38%	22,050	**<0,001**
User activity	95%	94%	87%	88%	88%	0,633	0,959
Device info.	97%	96%	95%	91%	100%	0,447	0,978
Payment	95%	79%	93%	76%	78%	3,881	0,422
Social media	62%	55%	55%	55%	40%	4,891	0,299
Ads	82%	87%	74%	73%	75%	1,877	0,758
IAP	2%	15%	1%	13%	5%	22,889	**<0,001**

Data related to mHealth can be seen in the above table in the lines Medical diagnoses, Medical IDs and Medical history. We can observe that also the insurance related data are most often collected in the Medical category. It relates mostly to health insurance, health credit, health coverage, etc. Expectedly, the health data is most requested and collected by apps in the Health & Fitness and Medical categories.

Data related to various forms of payment and transactions using credit cards, bank transfers, payment with cryptocurrencies, etc. are presented in the Payment line. Most of the applications that require data from users in connection with payment belong to, expectedly, the Business and Finance categories. More than 80% of applications require this type of data.

We performed a Chi-square goodness-of-fit test statistical analysis on our data in Table 4 and Table 5. It showed that there is a significant difference between the categories of both platforms for five categories of personal data. These are Medical diagnoses, Medical IDs, Medical history, Insurance and IAP. Additionally, on the Android platform another category is outstanding, i.e., the ID Card (mostly collected in the Health&Fitness category). The results of this analysis can be seen in the Sig column; all results are less than the 0,05. For all other personal data, there are no significant differences between the categories.

Table 5: Percentage of apps requiring a given permission, Android

Permission Android	Health& Fitness	Medical	Business	Entertainment	Finance	Chi-S	Sig.
Calendar	58%	69%	79%	69%	64%	3,522	0,475
Reminder	67%	80%	89%	73%	84%	3,883	0,422
Contact	96%	95%	95%	94%	91%	0,157	0,997
Photo	47%	39%	61%	69%	53%	10,201	**0,037**
Bluetooth	9%	4%	2%	2%	5%	7,545	0,110
Microphone	46%	61%	71%	63%	68%	6,065	0,194
Camera	67%	71%	82%	69%	87%	4,106	0,392
Location	89%	90%	97%	92%	89%	0,495	0,974
HealthKit	15%	6%	0%	0%	0%	3,857	**0,050**
Media Library	82%	64%	82%	73%	69%	3,432	0,488
Motion	6%	4%	11%	8%	4%	5,333	0,255
Documents	69%	71%	76%	77%	82%	1,413	0,842
Language	33%	31%	47%	65%	23%	27,457	**< 0,001**
Notifications	36%	39%	42%	38%	38%	0,497	0,974

3.3.3. Requested permissions

Next, we checked which permissions apps generally request. In Table 6 and Table 7, we can see that, in general, the apps on both platforms require quite a few permissions. Some of these permissions are required by almost all applications; for example, on Android platform, for the Contact and the Location, more than 95% of applications need access to the user's contacts. More than 89% of applications require a user to enable to track her location.

In addition to a general review of individual requirements through categories, we also performed a statistical analysis with the Chi-square goodness-of-fit test. With this analysis, we determined whether there are significant differences between the categories and the permits. The analysis showed that there are significant differences between the categories for the following permissions on the Android platform: Photo, HealthKit, and Language ($p<0,05$, $df=4$), and for HealthKit, and Language on the iOS platform ($p<0,05$, $df=4$).

Table 6: Percentage of apps requiring a given permission, iOS

Permission iOS	Health& Fitness	Medical	Business	Entertainment	Finance	Chi-S	Sig.
Calendar	80%	66%	54%	60%	69%	5,848	0,211
Reminder	87%	82%	85%	66%	79%	3,444	0,487
Contact	97%	92%	88%	93%	96%	0,545	0,969
Photo	62%	62%	47%	47%	47%	7,158	0,128
Bluetooth	2%	2%	4%	7%	4%	4,421	0,352
Microphone	72%	60%	60%	46%	59%	5,710	0,222
Camera	85%	66%	76%	67%	68%	3,608	0,462
Location	98%	92%	81%	86%	90%	1,826	0,768
HealthKit	0%	0%	0%	15%	6%	3,857	0,050
Media Library	72%	64%	65%	82%	65%	3,351	0,501
Motion	12%	4%	4%	4%	5%	8,414	0,078
Documents	79%	75%	80%	69%	73%	1,074	0,898
Language	44%	55%	16%	33%	31%	23,989	<0,001
Notifications	49%	36%	35%	35%	45%	4,300	0,367

3.3.4. Paid vs. free apps

Additionally, we checked whether there are any differences between paid and "free" apps.

Table 7: Percentages of required permissions, free vs. paid apps, Android

Permission Android	Free apps	Paid apps	Chi-Square (value)	Sig.
Calendar	75%	34%	15,422	**0,000**
Reminder	86%	51%	8,942	**0,003**
Contact	94%	93%	0,005	0,942
Photo	58%	29%	9,667	**0,002**
Bluetooth	5%	3%	0,500	0,480
Microphone	71%	22%	25,817	**0,000**
Camera	84%	42%	14,000	**0,000**
Location	94%	78%	1,488	0,222
HealthKit	5%	2%	1,286	0,257
Media Library	75%	64%	0,871	0,351
Motion	7%	2%	2,778	0,096
Documents	79%	61%	2,314	0,128
Language	40%	27%	2,522	0,112
Notifications	45%	12%	19,105	**0,000**

Table 8: Percentages of required permissions, free vs. paid apps, iOS

Permission iOS	Free apps	Paid apps	Chi-Square (value)	Sig.
Calendar	73%	32%	16,010	**0,000**
Reminder	86%	48%	10,776	**0,001**
Contact	94%	90%	0,087	0,768
Photo	56%	29%	8,576	**0,003**
Bluetooth	4%	3%	0,143	0,705
Microphone	67%	24%	20,319	**0,000**
Camera	79%	46%	8,712	**0,003**
Location	92%	76%	1,524	0,217
HealthKit	5%	2%	1,286	0,257
Media Library	71%	61%	0,758	0,384
Motion	7%	2%	2,778	0,096
Documents	79%	59%	2,899	0,089
Language	35%	31%	0,242	0,622
Notifications	46%	15%	15,754	**0,000**

We can see in Table 7 and Table 8 that the percentages for requested permissions are higher for free applications. After conducting a Chi-square goodness-of-fit test analysis, we found significant differences between paid and free applications for some permits: Calendar, Reminder, Photo, Microphone, Camera, and Notifications. Interestingly, the same results apply for both platforms. For all these permits, the p-value was less than 0.05. Free apps require more specific permissions than paid ones.

3.3.5. Android vs iOS

Next, we checked whether there are any differences between the platforms regarding the permissions. That is, if the same applications are requiring different permissions based on the platform.

Table 9: Comparison of permissions, Android vs. iOS

Permission	Percentage		Chi-Square (value)	Sig.
	Android	**iOS**		
Calendar	68%	12%	39,200	0,000
Calendar	67,7%	65,6%	0,314	0,575
Reminder	79,3%	78,5%	0,055	0,814
Contact	94,0%	93,1%	0,255	0,613
Photo	52,4%	50,8%	0,155	0,693
Bluetooth	4,4%	3,8%	0,148	0,701
Microphone	62,1%	59,3%	0,509	0,476
Camera	75,9%	72,6%	0,909	0,340
Location	91,2%	89,0%	0,913	0,339
HealtKit	4,1%	4,1%	0,000	0,987
Media Library	73,0%	69,1%	1,210	0,271
Motion	6,3%	5,7%	0,099	0,753
iCloud	0,3%	0,3%	0,000	0,996
Documents	75,2%	75,1%	0,002	0,964
Language	37,6%	34,1%	0,870	0,351
Notifications	38,6%	40,4%	0,221	0,639
Face ID, Touch ID	0,6%	0,6%	0,000	0,995

It can be observed that the value of Sig. in Table 9 is not less than or equal to 0.05 for any permit. This means that there are no significant differences between the Android and iOS operating systems in relation to permissions.

3.3.6. Stated requirements vs. data collected

The central part of our analysis was to check whether the stated requirements listed on the Google App Store website matched those found in privacy policies.

Table 10: Comparison of permissions, privacy policy vs. Google Play website

Permission	Percentage		Chi-Square (value)	Sig.
	Privacy Policy	Google Play		
Calendar	68%	12%	39,200	**0,000**
Camera	76%	58%	2,418	0,120
Contact	94%	35%	26,984	**0,000**
Device Info.	97%	41%	22,725	**0,000**
Location	73%	58%	1,718	0,190
Microphone	62%	40%	4,745	**0,029**
Phone	96%	46%	17,606	**0,000**
Media Library	73%	84%	0,771	0,380
Ads	80%	14%	46,340	**0,000**

From Table 10 above, we can see that the permissions stated on the Google Play website and those extracted from the privacy policy texts are quite different. The percentage shows how much each permission is required in the privacy policy texts and on the Google Play website. As we can see, the percentages are higher in the privacy policy texts of all permits. We performed a Chi-square goodness-of-fit test analysis and found significant differences between the following permissions: Calendar, Contact, Device Info, Microphone, Phone, and Ads. For all these permits, the p-value is less than 0.05.

Unfortunately, there is no information about permissions in the App Store for applications on the iOS platform. However, since the results from the previous sub-section indicate that there are no statistically significant differences, we can generalize the findings for both platforms.

4. Discussion and conclusion

In this paper, we analyzed 636 apps from the Google App store with respect to the permissions these apps require. We compared the list of permissions the apps require, as stated on the store's website, to those discovered by analysis of the privacy policies.

The paper's novelty is applying a (semi) automated linguistic analysis on privacy policies coupled with permission requirements analysis to unearth (deliberately!?) hidden information on which permissions apps require. Best to our knowledge, no previous research on privacy policies was done using the specific linguistic technique on permissions. We focused our study to medical and health & fitness-based apps because, in principle, they handle the most sensitive data, and compared them to other categories, such as business, entertainment, and finance.

Generally, health-related apps require all the "expected" permissions, such as calendars, reminders, and documents. Surprisingly, though, not so many apps required

the HealthKit permission, but on the other hand, they overwhelmingly required the location and contact data permissions. Surprisingly high is the requirement for camera permission.

The comparison of the list of permissions as stated on the Google Play Store and those identified to be needed through the privacy policy texts revealed a worrying fact: most apps require more permissions than stated, and the differences are pretty high. The differences between the stated and the needed permissions for most permissions are statistically significant.

Again, the lenient regulations are not protecting users in the app markets. As with previous studies, prominent market players are not protecting users. Big players in the market have made protecting personal data deliberately tricky, if not impossible, and are protecting app developers. It seems no one is checking the statements of permissions on the app store, and no one is checking whether a privacy policy is available. Sadly, almost 10 % of the analyzed apps did not have a privacy policy available or accessible.

More stringent regulations should be effected, requiring the big players and developers alike to disclose their policies with respect to private data fully, permissions they require, and to be fully accountable for the misleading information.

Acknowledgment

The author acknowledges the financial support from the Slovenian Research Agency (research core funding No. P2-0057) and the University of Maribor (www.um.si, core funding).

References

[1] Ghazinour, K., et al., *A Privacy-aware Platform for Sharing Personal Information on Wearable Devices.* Procedia Computer Science, 2016. **98**: p. 205-210.
[2] Parker, L., et al., *How private is your mental health app data? An empirical study of mental health app privacy policies and practices.* International journal of law and psychiatry, 2019. **64**: p. 198-204.
[3] Brumen, B., *Content Analysis of Medical and Health Apps' Privacy Policies*, in *Information Modelling and Knowledge Bases XXXIII*. 2022, IOS Press. p. 188-203.
[4] Robillard, J.M., et al., *Availability, readability, and content of privacy policies and terms of agreements of mental health apps.* Internet Interventions, 2019. **17**: p. 100243.
[5] Krebs, P. and D.T. Duncan, *Health app use among US mobile phone owners: a national survey.* JMIR mHealth and uHealth, 2015. **3**(4): p. e4924.
[6] O'Loughlin, K., et al., *Reviewing the data security and privacy policies of mobile apps for depression.* Internet interventions, 2019. **15**: p. 110-115.
[7] Moallem, A. *Do you really trust "Privacy Policy" or "Terms of Use" agreements without reading them?* in *International Conference on Applied Human Factors and Ergonomics*. 2017. Springer.
[8] Tangari, G., et al., *Mobile health and privacy: cross sectional study.* BMJ, 2021. **373**: p. n1248.
[9] Alfawzan, N., et al., *Privacy, Data Sharing, and Data Security Policies of Women's mHealth Apps: Scoping Review and Content Analysis.* JMIR Mhealth Uhealth, 2022. **10**(5): p. e33735.
[10] Hussain, M., et al., *A security framework for mHealth apps on Android platform.* Computers & Security, 2018. **75**: p. 191-217.
[11] Benjumea, J., et al., *Privacy Assessment in Mobile Health Apps: Scoping Review.* JMIR Mhealth Uhealth, 2020. **8**(7): p. e18868.
[12] Olano, F. *google-play-scraper 8.0.4*. 2022; Available from: https://www.npmjs.com/package/google-play-scraper
[13] Olano, F. *App Store Scraper*. 2021; Available from: https://github.com/facundoolano/app-store-scraper. Last visited 2022-07-10.
[14] Scrapy. *Scrapy 2.3*. 2022; Available from: https://pypi.org/project/Scrapy/

[15] Hatamian, M., J. Serna, and K. Rannenberg, *Revealing the unrevealed: Mining smartphone users privacy perception on app markets.* Computers & Security, 2019. **83**: p. 332-353.
[16] Wormer, T. and Z. Sikelianos. *powerthesaurus-api 2.0.3*. 2022; Available from: https://www.npmjs.com/package/powerthesaurus-api.
[17] Argyrous, G., *Statistics for research: With a guide to SPSS, 3rd ed.* 2011, Thousand Oaks, CA, USA: SAGE Publications Ltd.