Characterizing Human Aspects in Reviews of COVID-19 Apps

Mattia Fazzini*, Hourieh Khalajzadeh†, Omar Haggag†, Zhaoqing Li*,
Humphrey Obie†, Chetan Arora‡, Waqar Hussain†, John Grundy†
*University of Minnesota, Minneapolis, MN, USA; {mfazzini|li000114}@umn.edu
†Monash University, Melbourne, Australia; {hourieh.khalajzadeh|omar.haggag|humphrey.obie|waqar.hussain|john.grundy}@monash.edu
‡Deakin University, Melbourne, Australia; {chetan.arora}@deakin.edu.au

ABSTRACT

To successfully satisfy user needs, software developers need to suitably capture and implement user requirements. A critical and often overlooked characteristic of user requirements are "human aspects", which are personal circumstances affecting the use of software (e.g., age, gender, language, etc.). To better understand how human aspects can impact the use of software, this work presents an empirical study focusing on app reviews of COVID-19 contact tracing apps. We manually analyzed a dataset of 2,611 app reviews sampled from the reviews associated with 57 COVID-19 apps. To analyze the reviews, we performed qualitative and quantitative analyses. The analyses characterize the human aspects contained in the reviews and investigate whether the apps suitably address the human aspects. We identified 716 reviews related to human aspects and grouped these into nine categories. Of these 716 reviews, 8% report bugs, 14% describe future/improvement requests, and 22% detail the user experience. Our analysis of the results reveal that human aspects are important to users and we need better support to account for them as software is developed.

ACM Reference Format:
Mattia Fazzini*, Hourieh Khalajzadeh†, Omar Haggag†, Zhaoqing Li*,, Humphrey Obie†, Chetan Arora‡, Waqar Hussain†, John Grundy†. 2022. Characterizing Human Aspects in Reviews of COVID-19 Apps. In *IEEE/ACM 9th International Conference on Mobile Software Engineering and Systems (MOBILESoft '22), May 17–24, 2022, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3524613.3527814

1 INTRODUCTION

End-users (or users in short) can leverage software solutions to perform a wide variety of daily activities, such as reading the news, shopping online, streaming content, and communicating with family and friends. To ensure that software solutions successfully address diverse user needs, software developers need to suitably elicit and account for user requirements during the software development process. Although developers generally account for different types of user requirements [53], they tend to overlook and account for requirements related to *human aspects* [24]. End-user human aspects are a crucial characteristic for the successful development and adoption of fit-for-purpose software solutions. For example, according to the internet usage report from the Pew Research center in 2019 [9], 75% of United States adults over the age of 65 are now using the internet. This highlights that there is an opportunity and need for better tailoring software solutions to elderly users. Furthermore, a recent study [39] shows that teenagers using software often find that the visual design of certain websites is not suited to their taste. Such issues may cause certain user groups to abandon a certain solution because it does not account for their preferences, needs, or personal circumstances. Overlooking human aspects can also lead to tragic consequences. Instagram [32] was partly blamed for the death of a teenager [25] because the software platform did not consider the emotional impact of software usage on end-users by not suitably handling the explicit imagery of self-harm.

Although recent research has worked on suitably integrating some human aspects into some parts of the software development process [15, 23, 35, 51], developers currently still have little understanding of what different human aspects are most important for app take-up and usage, and whether these aspects are sufficiently considered in the development stages [24]. As a first step toward bridging this gap, we conducted an empirical study to characterize human aspects based on the information provided in the reviews of 57 COVID-19 contact-tracing mobile applications (or apps in short). We decided to focus on reviews from COVID-19 contact-tracing apps, as millions of users with very diverse backgrounds and circumstances (e.g., age, gender, and language) used these apps [33, 38, 61].

We collected reviews from the 57 apps and manually analyzed a sample of 2,611 reviews. To identify the human aspects appearing in the reviews and to understand whether these aspects are currently considered in the apps, we performed two qualitative analyses based on deductive, inductive, and axial coding. We first identified reviews wherein the users discussed the relation of the human aspects to the use of the app and characterized these aspects. In our second analysis, we classified the relevant reviews to investigate whether human aspects are currently sufficiently considered as the apps are developed. Finally, we also analyzed the ratings of human-aspect-related reviews to understand whether human aspects are associated with positive or negative user assessments.

Through our analyses, we identified 716 out of 2,611 reviews that are related to human aspects and categorized them into nine categories: *age*, *disability*, *emotion*, *gender*, *language*, *location*, *privacy*, *socioeconomic status*, and *miscellaneous.* Our analyses also established that reviews related to human aspects are discussed both positively and negatively by the users. The reviews further

MOBILESoft '22, May 17–24, 2022, Pittsburgh, PA, USA

ACM ISBN 978-1-4503-9301-0/22/05…$15.00
https://doi.org/10.1145/3524613.3527814

substantiate the claim that human aspects are not always considered in software development. In fact, 14% of the reviews related to human aspects are submitted as feature requests. Among other review types, we found that 22% of the reviews are related to user experience, and 8% report bugs.

We also provide a discussion of our findings to inform researchers and practitioners interested in providing better support for integrating human aspects in software development. Specifically, we discuss how our results can inform research on automatically extracting human aspects from app reviews, mapping human aspects to software features, and devising techniques to better elicit and address human aspects during software development. Although our results and findings are based and apply to COVID-19 apps, we hope that the findings can also be helpful for researchers and practitioners working on human aspects in other domains.

The main contributions of our work include:

- We investigate human aspects appearing in app reviews of COVID-19 apps and provide a categorization of these aspects. To the best of our knowledge, this is the first in-depth categorization of human aspects based on app reviews;
- We analyze whether human aspects have been adequately addressed in the apps, and provide insights on better integrating these aspects in the development processes;
- We make the data and tools from our study publicly available [18] to foster further research on the topic.

2 MOTIVATION

We define a *human aspect* as any personal circumstance affecting the use of an app. Our study focuses on human aspects emerging in *app reviews* (or reviews in short). A review captures user feedback on an app and generally contains a title, a textual description, and a star rating. The star rating ranges between one and five, and it is a quick way for the user to assess an app with respect to their review's content. In the star rating, five stars represent the highest positive assessment. If an app review discusses a human aspect, we say that the review is *human-aspect-related*. Figure 1 shows two examples of what we consider as human-aspect-related reviews. The two reviews are from Corona-Warn-App [14].

Figure 1a provides a human-aspect-related review that discusses how the socioeconomic status of a user might prevent the user from using the app ("*[...] Not everyone can and wants to buy the latest mobile phone [...]*"). Specifically, because the app was using contact tracing services [2] that worked only on newer smartphone models, users with older smartphone models, or the ones unable to buy the latest models (according to the review) were automatically precluded from using the app. Furthermore, this review presents an example of how human aspect (i.e., socioeconomic status) might be negatively impacted by a compatibility issue [29, 42, 62, 63]. We posit that the user submitted the review with the intent to ask for an app improvement. The review mentions that the app should also work on older phones to allow more users to use the app ("*[...] it should also work with older smartphones [...]*"). The negative textual feedback from the user is also reflected in the one star rating of the review (or 1★ in short).

Figure 1b presents another human-aspect-related review that praises the app because the app took into consideration the needs of

Title: *Requires a new iPhone*
Text: *If as many people as possible are to use this app it should also work with older smartphones. Not everyone can and wants to buy the latest mobile phone to use the app. You shouldn't be surprised if you don't reach many citizens.*
Rating ★ ☆ ☆ ☆ ☆

(a) First review for Corona-Warn-App on the Apple store.

Title: *Barrier-free*
Text: *Find the app clearly even people with disabilities have been thought of. The texts are easy to understand. It is also fully usable for blind people.*
Rating ★ ★ ★ ★ ★

(b) Second review for Corona-Warn-App on the Apple store.

Figure 1: App reviews discussing human aspects.

people with disabilities. Furthermore, the review provides feedback on the user experience, as it reports how the user interacted with the app. Finally, the positive feedback expressed in the text of the review is also reflected in the 5★ star rating of the app.

Such app reviews highlight how different types of human aspects (i.e., socioeconomic status and disability) considered (or not considered) in an app might impact the use of the app. In this work, we investigate the characteristics of human aspects impacting the use of COVID-19 apps because (i) these apps have been used worldwide by millions of diverse users; (ii) they have had to be developed and deployed quickly to address the pandemic crisis; (iii) they use a variety of development methods, platforms, libraries etc; (iv) they have generated a large number of app reviews; and (v) they are a timely case study of app development that needs to carefully take into account diverse end user human aspects.

3 METHODOLOGY

To characterized the human aspects impacting the use of COVID-19 apps, we investigate the following key research questions (RQs):

- **RQ1: What are the most prevalent human aspects discussed in COViD-19 app reviews?** Different types of human aspects might impact the use of an app. In this RQ, we analyze app reviews to categorize the types of human aspects that affected the use of an app and investigate the frequency with which different aspects are mentioned in the reviews.
- **RQ2: What is the rating associated with reviews related to human aspects?** With a review, a user can express a positive or a negative assessment for an app. In this RQ, we analyze the star rating of reviews to characterize user assessments of reviews related to human aspects. Additionally, we also investigate user assessments in relation to human aspects of different categories.
- **RQ3: What are the types of reviews containing human aspects?** Users write different types of reviews. For example, some reviews report bugs, while others request new features. In this RQ, we categorize the types of the reviews related to human aspects to better understand whether human aspects were sufficiently well considered as apps were developed.
- **RQ4: Are different human aspects associated with different review types?** We investigate if different types of human aspects were accounted differently as apps were developed.

Table 1: Characteristics of the COVID-19 apps and reviews considered in the study.

ID_A	App Name	Country	Users	*Reviews*		*Filtered Reviews*		*Samples*	
				Apple Store	*Google Play Store*	*Apple Store*	*Google Play Store*	*Apple Store*	*Google Play Store*
A01	Aarogya Setu	India	100,000,000	3,901	58,666	1,599	7,182	66	68
A02	Beat Covid	Gibraltar	9,000	4	32	3	5	3	5
A03	BeAware	Bahrain	400,000	72	1,195	45	316	28	57
A04	CareFiji	Fiji	27,000	7	112	2	30	2	22
A05	COCOA	Japan	4,000,000	2,039	30	923	17	64	14
A06	Corona-Warn-App	Germany	14,000,000	9,379	26,392	3,866	11,281	67	68
A07	COVID Alert	Canada	-	777	-	382	-	58	-
A08	CovidRadar	Mexico	-	26	58	14	22	12	17
A09	COVIDSafe	Australia	6,440,000	3,277	7,308	1,094	2,277	65	67
A10	Ehteraz	Qatar	2,531,620	632	9,567	260	2,152	55	67
A11	eRouska	Czech Republic	277,000	79	1,084	42	597	27	62
A12	HaMagen	Israel	2,000,000	263	652	127	282	45	55
A13	Hayat Eve Sığar	Turkey	14,186,000	5,451	40,777	3,750	18,762	67	68
A14	Immuni	Italy	2,200,000	3,107	11,504	1,474	4,430	66	68
A15	MorChana	Thailand	355,000	35	297	18	83	15	38
A16	MyTrace	Malaysia	100,000	30	385	18	111	15	43
A17	NHS COVID-19 App	England	-	603	1,251	274	405	55	59
A18	NZ COVID Tracer	New Zealand	588,800	374	1,840	179	628	50	62
A19	ProteGO	Poland	41,665	110	1,120	85	589	39	62
A20	Rakning C-19	Iceland	140,000	42	73	20	29	16	21
A21	Shlonik	Kuwait	-	88	450	37	144	25	47
A22	Smittestopp	Denmark	-	90	333	44	146	27	47
A23	StaySafe	Philippines	1,200,000	113	769	41	194	26	51
A24	StopCovid	France	1,900,000	805	4,663	455	1,922	60	66
A25	Stopp Corona	Austria	600,000	320	1,446	130	474	45	60
A26	SwissCovid	Switzerland	500,000	473	1,142	242	656	54	62
A27	Tawakkalna	Saudi Arabia	-	-	15,011	-	3,312	-	67
A28	Trace Together	Singapore	2,100,000	883	2,466	344	693	57	63
A29	ViruSafe	Bulgaria	55,000	49	456	32	229	22	53
A30	VirusRadar	Hungary	10,000	-	242	-	98	-	41
				33,029	189,321	15,500	57,066	1,131	1,480

3.1 Dataset

Our study required a dataset of COVID-19 app reviews. To the best of our knowledge, no suitable and readily available dataset existed when we started the study (August 2020). To create our dataset, we first analyzed a curated list of COVID-19 contact-tracing apps [38] and then selected 57 apps acknowledged as official, i.e., endorsed by the national government. The 57 apps include apps from 30 countries, with 27 countries having apps on both the Apple Store (AS) and Google Play Store (GS), one country (Canada) having the app only on the AS, and two countries (Saudi Arabia and Hungary) having their apps only on the GS. For the 27 countries with two apps, we identified that the countries used the same app name on both stores. In our study, we use a total count of 57 apps instead of 30 apps (27+1+2) because the apps target different platforms and their development processes might lead to differ study results. Table 1 reports the characteristics of the apps considered in our study. For each country, the table provides an identifier for the name of the app (ID_A), the app name (*App Name*), the name of the country (*Country*), and the number of users that installed the AS or the GS version of the app from the country (*Users*).

To build our dataset of app reviews, we used a two-step approach. We first collected the reviews and then selected relevant ones using a keyword-based filtering approach. To collect the reviews, we used an available tool [28] that, for each of the 57 apps, (i) downloads the reviews from the store of the app, (ii) detect the language of the reviews, and (iii) translates the content of non-English reviews to English. For each review, we collected its title, review text, and the users' rating, which ranges between 1–5. To translate the text and title of the reviews, we leveraged the Google Translate Ajax API [22]. After downloading all reviews, we had a set of 222,350 reviews (33,029 from apps on the Apple Store and 189,321 from apps on the Google Play Store). Table 1 lists the number of reviews per app under the *Reviews* header.

Given the dataset size and the manual tasks characterizing the qualitative analyses of our study, we decided to filter our dataset to include only the reviews that were likely to be related to human aspects. To this end, we leveraged another available tool [51] that uses a keyword-based approach to identify relevant reviews. We did not use a machine learning-based approach to identify relevant reviews as no labeled dataset is available for the task. Although a keyword-based approach might lead to the inclusion of unrelated reviews, our objective was to have a set of reviews from which it was feasible to identify relevant ones for the study. We consider investigating approaches to better filter related reviews as an interesting but orthogonal research direction and plan to work on those approaches in future work. Specifically, we plan to leverage the dataset created in this work to define machine learning-based approaches for the task. The output of that research could lead to an even more comprehensive dataset of human-aspect-related reviews.

Given a review and a set of keywords, the keyword-based tool finds whether a review is relevant by using a two-step approach. First, the tool preprocesses each review to (i) correct misspelled words, (ii) performs stopword removal, and (iii) carries out stemming Second, the tool marks the review as relevant if its text contains one of the keywords of interest. We identified relevant keywords by manually analyzing a statistically significant sample of the app reviews in our dataset and also included synonyms of relevant keywords. We created the sample using stratified random sampling and the sample contained 384 reviews (not represented in Table 1),

Table 2: Codes to categorize human aspects in app reviews.

Code	*Summary Description*
Age	The review discusses how the age of an app user relates to the use of the app.
Disability	The review reports how physical impairments or mental conditions relate to the use of the app.
Emotion	The review includes a user reaction to the use of the app.
Gender	The review mentions how the gender of a user affected the use of the app.
Language	The review discusses how a user's language relates to the app language and how this situation affected the use of the app.
Location†	The review reports how the location of a user impacts the use of the app.
Privacy†	The review discusses how the user personal information relates to the use of the app.
Socioeconomic Status	The review provides details on how the socioeconomic status of a user relates to the use of the app.
Miscellaneous	The review discusses other user characteristics or abilities affecting the use of the app.

† We categorized reviews mentioning users not being able to access the app from a specific location in the Location category and reviews describing access to the user location in the Privacy category.

which is statistically significant at 95% confidence level (CL) and 5% margin of error (ME). We collected keywords from the reviews by first checking whether a review discussed a human aspect per the definition provided in Section 2 and then extracted those keywords that highlighted the human aspect discussed in the review. After collecting the keyword terms from the reviews, we extended the set of terms with synonyms related to the original terms [64]. To perform this task, we used an online thesaurus [58], the thesaurus of job titles [37], the Merriam-Webster dictionary [16], and WordNet [48]. This process gave us 2,672 keywords. The fairly high number of keywords is due to the fact that our keyword selection included 2,177 keywords for job professions. After running the tool, our dataset contained 72,566 reviews (15,500 from apps on the AS and 57,066 from apps on the GS) and the number of reviews for each app is reported in Table 1 under the *Filtered Reviews* header.

3.2 Human Aspects Categorization

This first part of our study aims to identify and characterize human aspects in app reviews. We performed a qualitative analysis based on inductive and axial coding [13, 47]. Inductive coding is a systematic approach for manually coding (i.e., labeling) textual content and the set of codes is identified during the analysis. Axial coding is a technique that helps relate codes to one another and find higher-level codes representing abstractions of the original codes. In our analysis, a code categorizes a human aspect mentioned in a review and we assign codes to only those reviews that discuss how a human aspect affected the use of the app.

Our analysis was divided into two parts and performed by three of the paper's authors (called raters hereafter). In the first part of the analysis, the three raters created the codebook for the analysis — a document detailing the rules for assigning a specific code to a review. In the second part of the analysis, the three raters coded a sample of the reviews in the filtered dataset. To define the codebook, the three raters analyzed a sample of 383 reviews (not represented in Table 1). The sample size is statistically significant at 95% CL and 5% ME, and the sample was created using stratified random sampling (rounding the strata samples when needed). The three raters created the codebook iteratively by having weekly meetings in which they discussed and revised the human aspect categories

Table 3: Codes used to categorize the types of reviews containing human aspects.

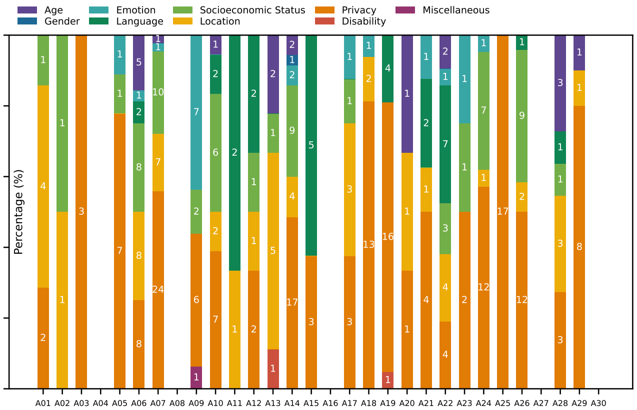
Code	*Summary Description*
Bug Report	The review describes problems with the app that should be fixed, such as a crash or a performance issue.
Feature/Improvement Request	In the review, the user asks for a feature to be added or improved.
User Experience	The review describes how the user used a functionality in the app.
Other	The review discusses other aspects related to the use of the app.

emerged from the analysis. Specifically, for each review processed during this step of the analysis, the raters (i) carefully read the review, (ii) checked whether it fit the general definition of human aspect (reported in Section 2), and (iii) created (if did not exist already) a category for classifying the human aspect discussed in the review. While creating a new category, the raters (i) defined a label for the category, (ii) created short and detailed descriptions for the category, and (iii) added the review as a typical examplar for the category if they thought it would help them classify other reviews. The analysis produced a codebook with nine codes and we provide the codes and their summary descriptions in Table 2. (We publish the complete codebook in our online appendix [18].) The creation of the codebook took two person-month to complete.

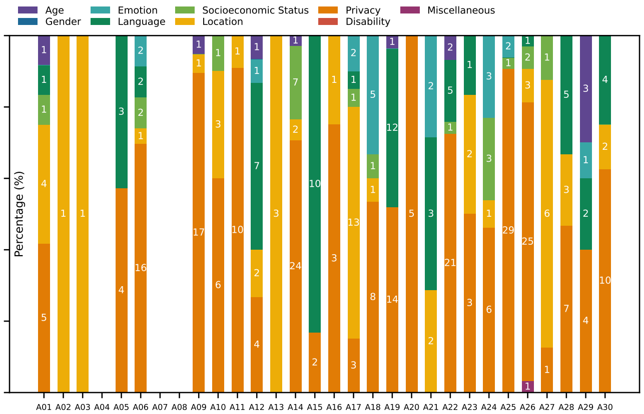
After creating the codebook, the three raters analyzed and coded a sample of reviews for each app and app store for a total of 2,611 reviews. We analyzed review samples as the effort required to undertake the coding process was quite high. Table 1 provides, under the *Samples* header, the size of the review samples we analyzed for each app and app store. The samples are statistically significant at 95% CL and 10% ME and do not contain any of the reviews analyzed to create the codebook. In this case, we did not use stratified random sampling but only random sampling as we created samples for each app and app store. During the coding process, the raters could assign multiple labels to the same review as a review might be related to different human aspects. Furthermore, we employed negotiated agreement as a mean to address the reliability of coding [8]. Using this technique, the three raters collaboratively agree on the code of a review. Because we used negotiated agreement, measures such as inter-rater agreement are not applicable. We used negotiated agreement as it is advantageous in research like ours where generating new insights is the primary concern [49]. In the analysis, the raters could reach an agreement for all the reviews they analyzed. Through this analysis, we identified that 716 of the 2,611 reviews are related to human aspects and contained information that could help the developers understand the issue. Analysis of the 2,611 reviews took four person-months to complete.

3.3 Review Types Categorization

The second part of our study aims at characterizing whether and how human aspects are considered as the apps were developed. To this end, we investigate how human aspects relate to the purpose why reviews were reported (e.g., to report a bug). We denominate the purpose as the *"review type"*. To characterize the review types, we conducted an additional qualitative analysis. The analysis was performed by the same raters that completed the analysis described in Section 3.2 and is based on deductive, inductive, and axial coding [13, 47]. In the analysis, a code represents a review type and we



(a) Distribution of human aspects for reviews on the Apple store.



(b) Distribution of human aspects for reviews on the Google Play store.

Figure 2: Distribution of human aspects across different apps.

started with deductive coding as related work [45] already offered a set of initial codes categorizing review types. Specifically, we used four codes from related work [45] as our initial set of codes for the codebook. To finalize the codebook, the three raters processed the reviews that they already analyzed to create the codebook for the analysis of human aspects. We used this set as the raters were already familiar with the reviews. After performing inductive coding, which followed a similar methodology as the one presented in Section 3.2, the raters created six codes and through axial coding the raters reduced the number of codes to four. Table 3 provides a summary description of the four codes in our codebook. Although the number of codes is the same as one in related work [45], one of the categories in our codebook differs. Specifically, our codebook does not contain the *Rating* category (which identifies reviews containing text reflections of the reviews' star ratings) but includes a more general *Other* category, which aims at including reviews that do not fall into the other three categories. The other three categories are: *bug report*, *feature/improvement request*, and *user experience*. With these categories, we aim to investigate whether human aspects relate to software bugs, whether human aspects are not considered in the development of the apps, or how human aspects relate to the features implemented in the apps. (The complete codebook is in our online appendix [18].)

After creating the codebook, the three raters coded the 716 reviews related to human aspects and used negotiated agreement to ensure reliability of coding.

4 FINDINGS

4.1 RQ1: What are the most prevalent human aspects discussed in app reviews?

In our manual analysis of 2,611 app reviews, we identified 716 reviews related to human aspects. Through the analysis, we classified the aspects into nine categories: *age*, *disability*, *emotion*, *gender*, *language*, *location*, *privacy*, *socioeconomic status*, and *miscellaneous*. Overall, **privacy aspects are the ones discussed the most**, appearing in 54% of the human-aspect-related reviews. Next, **location aspects are discussed** in 14% of reviews, followed by **socioeconomic status and language aspects**, with 11% each. Emotion aspects are discussed in 5% of the reviews, age aspects in 4% of the reviews, and all the remaining aspects (disability, gender, and miscellaneous) in less than 1% of the reviews (4 reviews total).

Figure 2 provides the distribution of the human-aspect categories for each of the apps considered. The figure groups the distributions by app store. For each app and app store, the figure reports the percentage of human aspects in a certain category as compared to the total number of human-aspects identified in the reviews. The numbers on the bars show the number human-aspects-related reviews in a category. In agreement with our overall results, Figure 2 shows that **aspects related to privacy are most frequently discussed** for most of the apps. However, there are also apps for which this is not the case. For example, for Hayat Eve Sığar (A13) on the AS and Tawakkalna (A27) on the GS, **location aspects are discussed the most**. Furthermore, for HaMagen (A12) on the GS and MorChana (A15 on the GS, **language aspects are the ones** that appear most frequently. For all apps, aspects related to **socioeconomic status are moderately discussed**. There are **only three apps** (A13, A14, A19 on the AS) with reviews related to **age, gender, and disability**.

Figure 2 also shows that **the number of human-aspect-related reviews greatly varies across apps and app stores**. For example, ViruSafe (A29) has nine out of 53 reviews that are related to human aspects for its GS app, while it has ten out of 22 reviews that are related to human aspects for its AS app. Beat COVID (A02) and BeAware (A03) are other examples having more reviews for their GS apps, but more human-aspect-related reviews for their apps on the AS. However, there are also apps for which the opposite is true. For example, COCOA (A05) has seven out of 14 reviews that are related to human aspects for its app on the GS, while only nine out of 64 reviews related to human aspects for its app on the AS. The results show that **app usage is impacted by human aspects**. The variability associated with the results motivates future research on providing techniques to help standardize how human aspects are considered in apps.

4.2 RQ2: What is the rating associated with reviews related to human aspects?

We considered the human-aspect-related reviews with one or two stars as negatively rated, with four or five stars as positively rated,

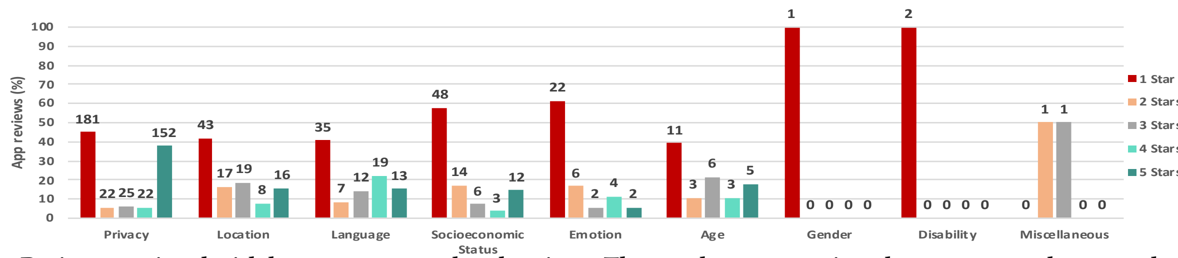


Figure 3: Rating associated with human-aspect-related reviews. The graph groups reviews by category and reports the percentage of reviews for each rating. The number of app reviews per rating is shown at the top of each bar.

and with three stars as neutrally rated. Figure 3 shows the star rating per human-aspect category. **Of the 716 human-aspect-related reviews, 55% were negatively rated, 34% were positively rated, and the remaining 9% were neutrally rated**. Gender, disability, and miscellaneous aspects are only raised in five app reviews, four negatively, and one neutrally. Among other categories, **emotion aspects are the ones associated with the highest percentage of negatively rated reviews (77.77%)**, followed by **socioeconomic, location, privacy and language** with 74.69%, 58.25%, 50.49%, and 48.83% respectively. Reviews containing **age aspects were negatively rated by half of the users** who provided reviews related to this category. Looking at positively rated reviews, **privacy is the most positively rated aspect**, with 43.28% of the reviews being positively rated, followed by language (37.02%), age (28.57%), location (23.3%), socioeconomic (18.07%), and emotion (16.66%). In this section, we discuss the rating associated with each of the human aspects, from the most frequently occurring aspect (i.e., privacy) to the ones appearing less frequently.

4.2.1 Privacy. Users that submitted negatively rated reviews discussing privacy aspects, had concerns about the privacy of these apps. Some users believe COVID-19 apps are violating their privacy by accessing their data and location, even in the background when the app is not being used. Users concern that COVID-19 apps are tracking their location, either by directly accessing the actual location of the users using GPS or by using Bluetooth to trace the location and the proximity user distance. An example:

> *[...] The problem is that the highly sensitive issue of* ***data protection and privacy is not given enough consideration*** *[...] I cannot recommend anyone to use this app until all of these points have been clarified [...]* - Stopp Corona, AS, 1★

An interesting finding is that, as shown in Figure 3, despite privacy being the most prevalent aspect and being negatively rated in 50.49% of the cases, this aspect category was the most positively rated type among all the categories that emerged from our categorization. Specifically, almost 43% of the reviews that mentioned privacy aspects were positively rated. Reasons behind such user satisfaction could be a result of the attention that researchers and practitioners are giving to the privacy of mobile apps [7, 11, 43, 46, 57] and the focus that COVID-19 apps had on privacy [56, 59]. Moreover, the users who were satisfied with the privacy of COVID-19 apps compared these apps to social media apps. They pointed out that although social media apps have documented and well-demonstrated privacy issues, people using these apps tend not to worry about the privacy of the apps. Finally, some users defended the idea and the importance of using contact tracing apps to fight the pandemic. They emphasized the importance of users' engagement, even though, people might have different privacy concerns. An example of a positively rated review:

> *The source code is available and has been dissected. No data transmitted because* ***everything remains locally stored in the phone****. Possibility to erase everything. [...]* ***Personal data intact the application does not ask for any personal information****. [...]* - StopCovid, AS, 5★

These results show that app privacy can be expected and required differently by various users depending on their preferences, goals, concerns, and personal circumstances. The results motivate further research on automated techniques to document and explain how software relates to certain human aspects.

4.2.2 Location. Among the reviews mentioning location aspects, 58.25% were negatively and 23.3% were positively rated. By investigating the reviews, we found that these issues are mainly related to the inability of using the COVID-19 apps due to geographical circumstances associated with the users. For example, COVIDSafe initially supported only Australian phone numbers during the registration process. This situation prevented users with no Australian mobile numbers from registering and using the app. These issues were resolved in later updates, after being reported by many visitors and students. Location-related issues were also raised by users in European apps, especially for the cases where users live close to the country borders. These users needed to use two apps, as they crossed the border daily, and faced challenges. This situation was repeatedly reported for Corona-Warn-App, AS, as some of its users needed to pass the borders to Denmark daily. An example of a negatively rated review mentioning the location aspect is:

> *I cant use this app.* ***I am on business visa. I dont have Qatar ID. How can use this without id?*** - Ehteraz, GS, 2★

These results show that developers might not be aware of the challenges users face concerning certain human aspects. This indicates a need for better support to elicit such aspects.

4.2.3 Language. 48.83% of the reviews discussing language-related aspects were negatively rated, while only 37.2% were positively rated. We found that users were dissatisfied with the language support provided by different COVID-19 apps. In the analysis, we found that the majority of the apps only supported one language. Consequently, some users, unable to communicate in that one language, felt excluded. This issue was reported seven times for the Smittestopp app on the AS. Despite being reported multiple times, the app developers did not address the issue at the time we performed our study, highlighting a need for better techniques to report and account for such issues. Looking at positively rated reviews, we could identify that some of the apps offered support for multiple languages, and the users appreciated this feature. Furthermore, from positively rated reviews, it also emerged that some apps offered continued support for this human aspect. For example, the COVIDSafe apps on the AS and the GS were periodically updated to include new languages. Examples of negatively and positively rated reviews discussing language aspects are:

> *Why is there no way to select an interface language in the app? **I am a foreigner. I speak french and english. My interface is in German which I don't know yet. How to use the app?*** - SwissCovid, AS, 2★

> ***Giving this 5 stars because it is in English.** Thank you very much for that* - COCOA, GS, 5★

These results include that user language considerations need greater attention in apps intended for wide community use, such as COVID-19 apps.

4.2.4 *Socioeconomic Status.*

Among the reviews discussing aspects from the socioeconomic status category, 74.69% were negatively rated and 18.07% were positively rated. The large percentage of negative reviews was mostly due to the fact that the majority of the apps we considered used contact-tracing libraries [2] that required the apps to run on recent operating system versions. This characteristic made the apps compatible only with recent device models as these devices were the ones running suitable operating systems. This design choice prevented some users from using the apps as not everyone had access or was able to buy the latest smartphones. From a manual analysis of the reviews, we identified that elderly users were often affected by this issue [50, 52]. An example of negatively rated review is:

> *This app should have been designed to work on **older phones.** [...]* - COVID Alert, AS, 1★

Looking at positively rated reviews, some users praised the use of the contact-tracing libraries as they were designed with privacy in mind. An example of positively rated review is:

> *To all the people complaining about **older devices**, well that has more to do with **Apple & Google services** that this app uses. The only way to **not invade privacy** and still do what this app promises needs newer devices. [...]* - COVID Alert, AS, 5★

These reviews highlight that human aspects might lead to conflicting requirements. Developers hence need the tools to precisely track such requirements as software is developed.

4.2.5 *Emotions.*

More than 77% of the reviews describing aspects from the emotion category were rated negatively. These users were mostly frustrated and dissatisfied due to the software issues such as bugs and instability issues that lead the app to be inaccessible or unusable. For example, a user was very frustrated not being able to register for the app due to not receiving the one-time password (OTP) code. Other users were very concerned and frustrated because of their privacy and how COVID-19 apps handle their data and location. Only 16.66% of the users left positively rated reviews. Other positively rated reviews praised the simplicity of the app's interfaces. A negative example:

> *The same overlap **notification keeps popping up every couple of hours** [...] even tough I keep marking it as not relevant making the app tiresome and annoying.* - HaMagen, GS, 1★

Such reviews show the importance of COVID-19 apps fostering positive emotions to ensure take up and usage.

4.2.6 *Age.*

Users from different age groups needed to be able to use COVID-19 apps. We found some COVID-19 apps have problems in being used by people from different age groups, and this was reflected in the ratings associated with the apps. 50% of the reviews related to age aspects were negatively rated, while only 28.57% were rated positively. One user of Virusafe on the GS reported being unable to register accounts for children since the app does not allow the registration of children under 14 years. Another example, a review from Smittestopp on the GS mentioned that elderly users were not able to use the app since it required them to authenticate using an identification method called NemID [31], but not all elderly users have that ID. This second example highlights a problem in eliciting the requirements from one of the main stakeholders. The review associated with this example is:

> *You **exclude everyone without Nem ID** which is usually **older and particularly vulnerable** which I am sorry for as it excludes me from using it among other things.* - Smittestopp, GS, 1★

Apps designed for wide community use such as COVID-19 apps need careful consideration of varying aged users.

4.2.7 *Other categories.*

Gender, disability, and miscellaneous aspects were only discussed in a small number of reviews (five reviews). Gender and disability were negatively rated with one star and miscellaneous aspects were rated with two and three stars. We also found a positively rated review discussing a disability aspect. We identified the review as we created the analysis codebook and presented this review in Figure 1b in Section 2. The two negatively rated reviews containing disability aspects discussed how the apps were not suitably designed for visually impaired users:

> ***I cannot use it because I am visually impaired**. It needs to be more accessible.* - Hayat Eve Sığar, AS, 1★

> *[...] the **blind cannot select and accept the regulations** and will not proceed* - ProteGO Safe, AS, 1★

4.3 RQ3: What are the types of reviews containing human aspects?

Based on the qualitative analysis described in Section 3.3, we could group the review types of human-aspect-related reviews into four categories. **44% of the reviews were connected to app features**. Breaking down this percentage, 22% discussed the user experience, 14% submitted a feature/improvement request, and 8% reported a bug. The other reviews (56%) provided general feedback on the app (*other* category). Reviews providing general feedback **did not explicitly discuss aspects of the other three categories**, and often provided opinions/ratings for the apps. These reviews would benefit from an interactive feedback system where developers can further understand how human aspects relate to concrete software

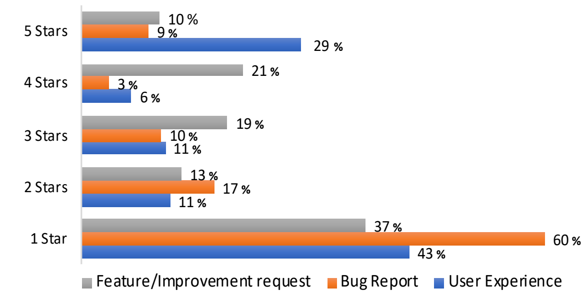


Figure 4: Ratings associated with the reasons for which users report human aspects.

engineering tasks. The star ratings associated with the review types are shown in Figure 4. We did not include the *other* category as our discussion focuses on the remaining three review types. For each review type, the figure reports the percentage of reviews having a certain star rating with respect to all the reviews of a specific type. To highlight the characteristics of the reviews from different review types, this section discusses the ratings of the reviews and presents relevant examples. As we did for RQ2, we considered reviews with four or five starts as positively rated, and reviews with one or two stars as negatively rated.

4.3.1 User Experience. In most of cases (54.14%) **users submitted negatively rated reviews about user experience**, and only 35% of the reviews related to user experience were positively rated. Negatively rated reviews from this category can help app developers better understand and improve specific usage scenarios. For example, a negatively rated review (including an emotion aspect) described that the user found it cumbersome and problematic to use the app, as the Bluetooth technology used by the app prevented the user's smartphone from pairing with the hands free system of the user's car. Specifically, the review was:

> *[...]* ***the Bluetooth keeps preventing my phone from pairing with hands free in car and other Bluetooth devices.*** *I generally have to either shutdown the app or go into my phone settings every time I get in the car to pair with hands free.* ***Not only is this annoying but could be a safety issue for people driving.*** - COVIDSafe, AS, 2★

4.3.2 Feature/Improvement Request. **Half of the reviews discussing feature/improvement requests were negatively rated** with only 31% positively rated. In our manual analysis, we found that both positively and negatively rated reviews could provide valuable information to app developers to better account for human aspects. As an example, a user of the Aarogya Setu app (on the AS) submitted a positively rated review that praised how the app accounted for privacy and also suggested to use Apple's new exposure-tracking features since they worked even when the app is not open. Some users provided negatively rated reviews as the apps lacked features deemed essential. For example, a review related to a socioeconomic aspect asked app developers to support adding accounts for children as it was not possible for the user to have additional phones:

> *Can you please* ***make an option where we can add accounts*** *[...] This would be really helpful for families that have 3 children and* ***don't have the finances to buy a phone for each of them."*** - Ehteraz, AS, 2★

Table 4: Human aspects and the reasons they are submitted.

Human Aspects	*User Experience*	*Bug Report*	*Feature/Improvement request*	*Other*
Age	6	1	7	15
Gender	0	0	0	1
Emotion	12	15	3	10
Language	7	8	36	34
Socioeconomic Status	15	7	10	54
Location	16	18	18	60
Privacy	101	9	33	254
Disability	1	0	0	1
Miscellaneous	2	0	0	0
Total	160	58	107	429

Another user, negatively discussed a privacy aspect and asked for a feature in order to be able to trust the app:

> *Again* ***still doesn't work with Apple's password manager****. [...] please update the app to allow the native password manager on iOS randomly generate a password [...]* - NZ COVID Tracer, AS, 2★

We classified this last review as a feature/improvement request as it discussed allowing a certain interaction with a feature (the password manager) of the OS.

4.3.3 Bug Reports. Unsurprisingly, **most of the reviews discussing a bug report are negatively rated** (77.58%), while only 12% are positively rated. Reviews reporting bugs reveal how certain human aspects might affect the adoption of the apps and how app developers can improve their apps. A frustrated user writes:

> *[...]* ***It does not support older phones with small screens****. My partner has a nexus 5x which does not allow her to verify her account* ***as the "next button" step is not visible.*** - NZ COVID Tracer, GS, 1★

This example also reflects on a user experience where the lack of focus on supporting smaller phone sizes leads to the exclusion of a whole class of disadvantaged users. This highlights the need for a design approach that handles some of the trade-offs that software developers need to consider to accomplish overall application goals. These results highlight that **focusing on human aspects as software is developed might lead to an improved user experience**. Furthermore, an ecosystem of techniques focusing on human aspects might avoid releasing software with critical bugs.

4.4 RQ4: Are different human aspects associated with different review types?

The distribution of review types varies across different human aspects and Table 4 reports the distribution. Most of the users reporting reviews related to **age aspects share their user experience** (six out of 29 human-aspect-related reviews) or ask for a new feature or improvement (seven out of 29). Reviews reporting **emotion aspects are mostly due to bugs** (15 out of 40) or related to user experience (12 out of 40). Reviews discussing **language aspects are mostly asking for features or improvements** (36 out of 85). Reviews related to **socioeconomic aspects are reported to discuss the user experience** (15 out of 86), ask for improvements in the apps (ten out of 86), and describe bugs (seven out of 86). Reviews containing **privacy aspects are mostly associated with user experience** (101 out of 397), followed by feature/improvement request (33 out of 397). Overall, **most of the human-aspect-related reviews discussing human aspects are associated with user**

experience (160 out of 754), followed by feature/improvement requests (107 out of 754), and bug reports (58 out of 754).

Five categories of human aspects (age, emotion, language, socioeconomic status, location and privacy) have at least one review of each review type (user experience, bug report, and feature/improvement request). Looking at human aspect categories individually, we observed that, overall, **different human aspects have been considered differently as apps were developed**. For example, we observed that the focus of COVID-19 apps on user privacy [56, 59] lead to a high percentage of user experience reviews as compared to bug reports. For reviews related to language aspects, we, instead, observed that a large portion of the reviews focused on feature/improvement requests to support different languages. We believe that language aspects might have been overlooked as the concept of travel did not align well with the design of COVID-19 apps [60].

These results show that software engineering approaches tailored to human aspects need to be developed on multiple fronts. These techniques might help taking into account different human aspects and ensure software is inclusive of all users.

5 DISCUSSION

In this section, we provide a discussion of our results to inform researchers and practitioners interested in providing better support for integrating human aspects in software development.

Reviews of COVID-19 apps discuss different human aspects: Our analysis identified 716 human-aspect-related reviews out of the 2,611 reviews that we analyzed. If this number would extend to the whole set of reviews available to us (222,350 reviews), the human-aspect-related reviews would account for roughly 8.9% of the reviews. This number reveals that human aspect reviews are relevant for app users. Different human aspects are discussed in different COVID-19 apps. Some of the human aspects are discussed more commonly, such as privacy, location, language, and socioeconomic status. Although there is a limited discussion for some other aspects (i.e., gender, and disability), the discussion appearing in the reviews related to these aspects reveal significant concerns with the apps, which should be addressed as apps are developed.

Privacy, the most prevalent human aspect identified in our study, is the most positively reported aspect. This could be due to the focus that researchers and practitioners have on the privacy of mobile applications [7, 11, 43, 46, 57]. However, the number of users negatively reporting privacy aspects is significant and emphasizes the need for further research on this topic. Our study also highlighted that users have varying preferences in relation to privacy and developers need to be able to take the users' diverse backgrounds, abilities, and concerns into account [3]. Among our results, there were also cases in which the resolution of the problem discussed in a human-aspect-related review was not fully in the hands of the app developers (e.g., apps using contact-tracing libraries that required the apps to run on recent operating system versions). In this case, developers could work on alternative solutions or, when not possible, work early with other developers to help shape the technology so that human aspects are suitably considered. Overall, the aspects and related-issues that emerged reinforce the need for better software engineering techniques to address human aspects in emergency/public health/public service apps.

Prevalence of human-aspect-related reviews varies among apps and stores: Although location and socioeconomic status were two aspects directly being discussed in this study, these can also indirectly influence other human aspects. This can be a consequence of the users' location and culture, the technologies they have access to, and software engineering practices that lead to the design of the app in a specific country. Also, some human-aspect-related issues lead to other issues, for example, not being able to access an app from a certain location forces the user to use another app that does not support the languages they understand. This highlights that human aspects can be coupled with each other and having techniques to account for such relations might help developers to identify and solve multiple issues at once.

Various reasons lead to human aspect related reviews: User experience is the most prevalent reason, both negatively and positively rated, while the bug report is the most negatively rated reason. This shows that developers need to not only take the user experience into account, they should continuously improve the app based on bug reports and requested features to address human-aspect-related issues. Such issues might arise due to the lack of understating of the end-users needs. Software engineers typically have differing characteristics from most of their end-users, in terms of their age, gender, socioeconomic status, and their physical/mental impairments [23, 24]. Alshayban et al. indicated the lack of accessibility awareness in app developers [1]. This situation influences the degree to which the app developers understand and incorporate end-users needs. This reflects the need for a methodology to continuously monitor how our apps impact the users in non traditional ways. Moreover, users need better ways to report their human aspect related issues to developers. For example, there could a dedicated field in reporting systems, which may also provide an automated categorization of the issue based on our results. The option to have a structured section for human aspects would also raise awareness of these among users and app developers.

Better approaches to identify human aspects from natural language are needed: Human-centered design aims to create solutions centered around the people who use the product [17]. Online app reviews provide a rich resource for users who were not originally considered in the design process to effectively communicate their needs and express their concerns and opinions regarding the system. However, analysing the large number of app reviews is beyond the developers' capacity. In our study, we extended an existing keyword-based approach [51] to identify reviews likely related to human aspects. However, the tool gave us a significant number of false positives, i.e., only 716 of the 2,611 reviews we manually analyzed were actually related to human aspects. This indicates that there are currently limited techniques to identify and account for human aspects. Better tools and techniques are required for users to report human aspect related issues, and for developers to gather such information without a manual review process.

6 THREATS TO VALIDITY

Internal Validity. The main internal validity concerns in our study are related to non-English reviews, and the notion of relevance between review rating and human aspects discussed in the review. We translated non-English reviews to English, which could have potentially led to miss-categorization or missing out on reviews. To mitigate this threat, we leveraged Google Translate – known to have accurate performance on a large number of languages [21].

Our analysis relates reviews' ratings to human aspects. This might be imprecise for reviews discussing multiple human aspects, or the ones that discuss a human aspect and other topics. This potential threat did not largely affect our results as 695 out of the 716 human-aspects-related reviews focused only on one aspect. Furthermore, the raters verified that the review's description was in line with the review rating, and no mismatch was found. Our comparison of AS and GS apps did not account for the user perception of the app platform and that could lead to different results.

Construct Validity. The main construct validity concerns in our study is related to the manual analysis of reviews. Our manual tasks might have introduced errors. We mitigated this by having multiple authors involved in the manual analyses we performed.

Conclusion Validity. The main conclusion validity concern in our study is related to the keyword-based filtration process. We selected the keywords for filtering relevant reviews over a statistically significant sample, however the reviews we analyze might not have included some keywords to identify relevant human aspects. We attempted to mitigate this threat by expanding our set of keywords with relevant synonyms from different sources so as to identify as complete as possible set of reviews likely related to human aspects.

External Validity. Our results might not generalize to other COVID-19 apps or apps in general. We attempted to mitigate these threats by analyzing statistically significant samples from 57 nationally endorsed apps from both the Apple App Store and Google Play Store. Additionally, our results also depend on the development processes used to create the considered apps and those processes might be different from the ones used to create other apps.

7 RELATED WORK

App reviews, in recent years, have been used for analysing security and privacy issues [65], extracting feature requests, bug reports and requirements-related information [27, 34, 36, 44, 45], and studying user satisfaction and sentiments [20, 26, 41]. Our work most closely relates to the analysis of app reviews for COVID-19 contact-tracing apps. Below, we position our work against these research strands.

Rekanar et al. [54] manually analyzed 1,287 app reviews from the Google/Apple app stores and performed sentiment analysis and identified users' focus in those reviews for the Irish contact-tracing app. The authors reported that the overall perception of the users was mostly positive towards the app, and users' reviews helped highlight data protection and transparency issues. Haggag et al. [27, 28] analyzed 2 million app reviews to understand users' privacy concerns and compared them with users perceptions of security and privacy on social media platforms. Haggag et al. [28] reported that inaccessibility and instability of the contact-tracing apps decreased their popularity and user uptake. Bano et al. [3, 4] analyzed user reviews of 16 contact-tracing apps to determine the success or failure criteria for such apps. The authors report that a mix of technical (such as Bluetooth and battery) and non-technical (such as lack of consideration for the socio-cultural landscape of countries) issues contributed to the success or failure of these apps. Garousi et al. [19] performed exploratory analysis of nine European countries using a commercial tool based on ≈40,000 app reviews. Similar to Bano et al. [4], they highlighted the technical and non-technical issues in the apps, as reported by the users.

Some research strands focus on the privacy of contact tracing apps [5, 6, 10, 12, 30, 40, 55]. Our work is different in that we do not focus on privacy or other aspects purely from a technical point of view. We consider such aspects from the point of view of various diverse end-users of the apps with differing requirements and issues. Moreover, we consider a comprehensive set of human-centric aspects rather than just focusing on one aspect.

Obie et al. [51] analysed ≈22,000 app reviews from Google Play store using natural language processing techniques to understand user reported issues. Using Schwartz theory of basic values from social sciences, they detected violations of user values caused by the feature offered in their selected mobile apps. The reported values violations included *curiosity*; a general lack of the desired information to satisfy users questions or queries e.g. lack of prompt notifications, updated information and statistics about the app in use; *honesty* and *transparency* e.g. charging fee right after the free trial version without any notification and a general lack of app's *helpfulness* or usefulness. The authors did not study or report violation of values relating to factors such as age, gender, or physical and mental abilities.

This is the first comprehensive study of app reviews to identify and discuss human aspects. While others [4, 28] did note some human aspects in their analysis, none of the existing works analyzed the app reviews for human aspects systematically at this scale, i.e., 57 official contact-tracing apps.

8 CONCLUSION

We presented an empirical study that characterized human aspects in reviews from COVID-19 contact tracing apps. We manually analyzed 2,611 reviews from 57 apps and identified 716 human-aspect-related reviews. We categorized human aspects into nine different categories, identified that human-aspect-related reviews are discussed both positively and negatively, and confirmed that human aspects are not always suitably considered as apps are developed. In key future work, we plan to devise a technique to automatically identify human-aspect-related reviews leveraging our dataset. We will perform an empirical study to extend our work by analyzing human aspects from other sources (e.g., GitHub issues) and in different software domains. We will also investigate an approach to relate human-aspect-related reviews to app features, combining natural language processing with static analysis techniques to identify feature descriptions in reviews and connect them to code in the apps. We will define an approach to identify and extract code examples from apps that accounted for human aspects and provide such examples as suggestions to developers of other relevant apps. Future work could look at human aspects in other apps by following a similar methodology. Finally, we want to work on techniques to help better incorporate human aspects during app development.

ACKNOWLEDGEMENT

This work is partially supported by Australian Research Council Laureate Fellowship FL190100035 and Discovery Project DP200100020.

REFERENCES

[1] Abdulaziz Alshayban, Iftekhar Ahmed, and Sam Malek. 2020. Accessibility issues in Android apps: state of affairs, sentiments, and ways forward. In *2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE)*. IEEE, 1323–1334.

[2] Apple and Google. 2020. *Privacy-Preserving Contact Tracing*. Retrieved May 15, 2021 from https://covid19.apple.com/contacttracing

[3] Muneera Bano, Chetan Arora, Didar Zowghi, and Alessio Ferrari. 2021. The Rise and Fall of COVID-19 Contact Tracing Apps: when NFRs Collide with Pandemic.. In *29th IEEE International Requirements Engineering Conference (RE)*. IEEE.

[4] Muneera Bano, Didar Zowghi, and Chetan Arora. 2020. Requirements, Politics, or Individualism: What Drives the Success of COVID-19 Contact-Tracing Apps? *IEEE Software* 38, 1 (2020), 7–12.

[5] Lars Baumgärtner, Alexandra Dmitrienko, Bernd Freisleben, Alexander Gruler, Jonas Höchst, Joshua Kühlberg, Mira Mezini, Richard Mitev, Markus Miettinen, Anel Muhamedagic, et al. 2020. Mind the GAP: Security & privacy risks of contact tracing apps. In *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*. IEEE, 458–467.

[6] Yoshua Bengio, Daphne Ippolito, Richard Janda, Max Jarvie, Benjamin Prud'homme, Jean-François Rousseau, Abhinav Sharma, and Yun William Yu. 2021. Inherent privacy limitations of decentralized contact tracing apps. *Journal of the American Medical Informatics Association* 28, 1 (2021), 193–195.

[7] Petter Bae Brandtzaeg, Antoine Pultier, and Gro Mette Moen. 2019. Losing control to data-hungry apps: A mixed-methods approach to mobile app privacy. *Social Science Computer Review* 37, 4 (2019), 466–488.

[8] John L Campbell, Charles Quincy, Jordan Osserman, and Ove K Pedersen. 2013. Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research* 42, 3 (2013), 294–320.

[9] Pew Research Center. 2019. *Internet/Broadband Fact Sheet*. Retrieved May 15, 2021 from https://www.pewresearch.org/internet/fact-sheet/internet-broadband

[10] Eugene Y Chan and Najam U Saqib. 2021. Privacy concerns can explain unwillingness to download and use contact tracing apps when COVID-19 concerns are high. *Computers in Human Behavior* 119 (2021), 106718.

[11] Yi Chen, Mingming Zha, Nan Zhang, Dandan Xu, Qianqian Zhao, Xuan Feng, Kan Yuan, Fnu Suya, Yuan Tian, Kai Chen, et al. 2019. Demystifying hidden privacy settings in mobile apps. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 570–586.

[12] Hyunghoon Cho, Daphne Ippolito, and Yun William Yu. 2020. Contact tracing mobile apps for COVID-19: Privacy considerations and related trade-offs. *arXiv preprint arXiv:2003.11511* (2020).

[13] Juliet Corbin and Anselm Strauss. 2014. *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage publications.

[14] Corona-Warn-App. 2020. *Corona-Warn-App Open Source Project*. Retrieved May 15, 2021 from https://www.coronawarn.app/en

[15] Maheswaree Kissoon Curumsing, Niroshinie Fernando, Mohamed Abdelrazek, Rajesh Vasa, Kon Mouzakis, and John Grundy. 2019. Emotion-oriented requirements engineering: A case study in developing a smart home system for the elderly. *Journal of Systems and Software* 147 (2019), 215 – 229.

[16] Merriam-Webster Dictionary. 2021. *Merriam-Webster Dictionary*. Retrieved May 15, 2021 from https://www.merriam-webster.com

[17] Tayba Farooqui, Tauseef Rana, and Fakeeha Jafari. 2019. Impact of human-centered design process (HCDP) on software development process. In *2019 2nd International Conference on Communication, Computing and Digital systems (C-CODE)*. IEEE, 110–114.

[18] Mattia Fazzini, Hourieh Khalajzadeh, Omar Haggag, Zhaoqing Li, Humphrey Obie, Chetan Arora, Waqar Hussain, and John Grundy. 2022. *Artifacts for Characterizing Human Aspects in Reviews of COVID-19 Apps*. Retrieved Mar 28, 2022 from https://doi.org/10.5281/zenodo.6392198

[19] Vahid Garousi, David Cutting, and Michael Felderer. 2020. Mining user reviews of COVID contact-tracing apps: An exploratory analysis of nine European apps. *arXiv preprint arXiv:2012.13589* (2020).

[20] Necmiye Genc-Nayebi and Alain Abran. 2017. A systematic literature review: Opinion mining studies from mobile app store user reviews. *Journal of Systems and Software* 125 (2017), 207–219.

[21] Google. 2020. *Recent Advances in Google Translate*. Retrieved May 15, 2021 from https://ai.googleblog.com/2020/06/recent-advances-in-google-translate.html

[22] Googletrans. 2021. *googletrans 3.0.0*. Retrieved May 15, 2021 from https://pypi.org/project/googletrans

[23] John Grundy, Hourieh Khalajzadeh, and Jennifer Mcintosh. 2020. Towards Human-centric Model-driven Software Engineering.. In *ENASE*. 229–238.

[24] John Grundy, Hourieh Khalajzadeh, Jennifer McIntosh, Tanjila Kanij, and Ingo Mueller. 2020. HumaniSE: Approaches to Achieve More Human-Centric Software Engineering. In *International Conference on Evaluation of Novel Approaches to Software Engineering*. Springer, 444–468.

[25] The Guardian. 2019. *Instagram bans 'graphic' self-harm images after Molly Russell's death*. Retrieved May 15, 2021 from https://www.theguardian.com/technology/2019/feb/07/instagram-bans-graphic-self-harm-images-after-molly-russells-death

[26] Emitza Guzman and Walid Maalej. 2014. How do users like this feature? a fine grained sentiment analysis of app reviews. In *2014 IEEE 22nd international requirements engineering conference (RE)*. IEEE, 153–162.

[27] Omar Haggag, John Grundy, Mohamed Abdelrazek, and Sherif Haggag. 2022. Better Addressing Diverse Accessibility Issues in Emerging Apps: A Case Study using COVID-19 Apps. In *9th IEEE/ACM International Conference on Mobile Software Engineering and Systems 2022 (MobileSoft 2022)*.

[28] O. Haggag, S. Haggag, J. Grundy, and M. Abdelrazek. 2021. COVID-19 vs Social Media Apps: Does Privacy Really Matter?. In *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*. IEEE, Los Alamitos, CA, USA, 48–57.

[29] Dongjie He, Lian Li, Lei Wang, Hengjie Zheng, Guangwei Li, and Jingling Xue. 2018. Understanding and detecting evolution-induced compatibility issues in Android apps. In *2018 33rd IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 167–177.

[30] Katie Hogan, Briana Macedo, Venkata Macha, Arko Barman, Xiaoqian Jiang, et al. 2021. Contact Tracing Apps: Lessons Learned on Privacy, Autonomy, and the Need for Detailed and Thoughtful Implementation. *JMIR Medical Informatics* 9, 7 (2021), e27449.

[31] Nem ID. 2021. *Nem ID*. Retrieved May 15, 2021 from https://www.nemid.nu/dk-en

[32] Instagram. 2021. *Instagram*. Retrieved May 15, 2021 from www.instagram.com

[33] Steve Jacob and Justin Lawarée. 2020. The adoption of contact tracing applications of COVID-19 by European governments. *Policy Design and Practice* (2020), 1–15.

[34] Nishant Jha and Anas Mahmoud. 2019. Mining non-functional requirements from app store reviews. *Empirical Software Engineering* 24, 6 (2019), 3659–3695.

[35] Aria YukFan Jim, Hyun Shim, Jue Wang, Lionel Richie Wijaya, Rongbin Xu, Hourieh Khalajzadeh, John Grundy, and Tanjila Kanij. 2021. Improving the Modelling of Human-Centric Aspects of Software Systems– A Case Study of Modelling End User Age in Wirefame Designs. In *Evaluation of Novel Approaches to Software Engineering (ENASE)*.

[36] Timo Johann, Christoph Stanik, Walid Maalej, et al. 2017. Safe: A simple approach for feature extraction from app descriptions and app reviews. In *2017 IEEE 25th International Requirements Engineering Conference (RE)*. IEEE, 21–30.

[37] johnpcarty. 2021. *Thesaurus of Job Titles*. Retrieved May 15, 2021 from https://github.com/johnpcarty/Thesaurus-of-Job-Titles

[38] Bobbie Johnson. 2021. *The Covid Tracing Tracker: What's happening in coronavirus apps around the world*. Retrieved May 15, 2021 from https://www.technologyreview.com/2020/12/16/1014878/covid-tracing-tracker

[39] Alita Joyce and Jakob Nielsen. 2019. *Teenager's UX: Designing for Teens*. Retrieved May 15, 2021 from https://www.nngroup.com/articles/usability-of-websites-for-teenagers

[40] Douglas J Leith and Stephen Farrell. 2021. Contact tracing app privacy: What data is shared by Europe's GAEN contact tracing apps. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 1–10.

[41] Huiying Li, Li Zhang, Lin Zhang, and Jufang Shen. 2010. A user satisfaction analysis approach for software evolution. In *2010 IEEE International Conference on Progress in Informatics and Computing*, Vol. 2. IEEE, 1093–1097.

[42] Li Li, Tegawendé F Bissyandé, Haoyu Wang, and Jacques Klein. 2018. Cid: Automating the detection of api-related compatibility issues in android apps. In *ISSTA*.

[43] Bin Liu, Jialiu Lin, and Norman Sadeh. 2014. Reconciling Mobile App Privacy and Usability on Smartphones: Could User Privacy Profiles Help?. In *Proceedings of the 23rd International Conference on World Wide Web* (Seoul, Korea) *(WWW '14)*. ACM, New York, NY, USA, 201–212. https://doi.org/10.1145/2566486.2568035

[44] Mengmeng Lu and Peng Liang. 2017. Automatic classification of non-functional requirements from augmented app user reviews. In *Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering*. 344–353.

[45] Walid Maalej, Zijad Kurtanović, Hadeer Nabil, and Christoph Stanik. 2016. On the automatic classification of app reviews. *Requirements Engineering* 21, 3 (2016), 311–331.

[46] Kirsten Martin and Katie Shilton. 2016. Putting mobile application privacy in context: An empirical study of user privacy expectations for mobile devices. *The Information Society* 32, 3 (2016), 200–216. https://doi.org/10.1080/01972243.2016.1153012 arXiv:https://doi.org/10.1080/01972243.2016.1153012

[47] Matthew B Miles, A Michael Huberman, and Johnny Saldaña. 2018. *Qualitative data analysis: A methods sourcebook*. Sage publications.

[48] George A. Miller. 1995. WordNet: A Lexical Database for English. *COMMUNICATIONS OF THE ACM* 38 (1995), 39–41.

[49] Elizabeth R Morrissey. 1974. Sources of error in the coding of questionnaire data. *Sociological Methods & Research* 3, 2 (1974), 209–232.

[50] Amber L Mueller, Maeve S McNamara, and David A Sinclair. 2020. Why does COVID-19 disproportionately affect older people? *Aging* 12, 10 (2020), 9959.

[51] Humphrey O. Obie, Waqar Hussain, Xin. Xia, John Grundy, Li Li, Burak Turhan, Jon Whittle, and Mojtaba Shahin. 2021. A First Look at Human Values-Violation in App Reviews. In *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*. 29–38.

[52] Donatella Rita Petretto and Roberto Pili. 2020. Ageing and COVID-19: What is the role for elderly people?

[53] Klaus Pohl. 2010. *Requirements Engineering - Fundamentals, Principles, and Techniques*. Springer.

[54] Kaavya Rekanar, Ian R O'Keeffe, Sarah Buckley, Manzar Abbas, Sarah Beecham, Muslim Chochlov, Brian Fitzgerald, Liam Glynn, Kevin Johnson, John Laffey, et al. 2021. Sentiment analysis of user feedback on the HSE's Covid-19 contact tracing app. *Irish Journal of Medical Science (1971-)* (2021), 1–10.

[55] Frantz Rowe. 2020. Contact tracing apps and values dilemmas: A privacy paradox in a neo-liberal world. *International Journal of Information Management* 55 (2020), 102178.

[56] Marcel Salathé, Christian L Althaus, Nanina Anderegg, Daniele Antonioli, Tala Ballouz, Edouard Bugnion, Srjan Capkun, Dennis Jackson, Sang-Il Kim, James Larus, et al. 2020. Early evidence of effectiveness of digital contact tracing for SARS-CoV-2 in Switzerland. *medRxiv* (2020).

[57] Victor Sucasas, Georgios Mantas, Ayman Radwan, and Jonathan Rodriguez. 2016. An OAuth2-based protocol with strong user privacy preservation for smart city mobile e-Health apps. In *2016 IEEE International Conference on Communications (ICC)*. IEEE, 1–6.

[58] Thesaurus. 2021. *Thesaurus*. Retrieved May 15, 2021 from www.thesaurus.com

[59] Carmela Troncoso, Mathias Payer, Jean-Pierre Hubaux, Marcel Salathé, James Larus, Edouard Bugnion, Wouter Lueks, Theresa Stadler, Apostolos Pyrgelis, Daniele Antonioli, Ludovic Barman, Sylvain Chatel, Kenneth Paterson, Srdjan Čapkun, David Basin, Jan Beutel, Dennis Jackson, Marc Roeschlin, Patrick Leu, Bart Preneel, Nigel Smart, Aysajan Abidin, Seda Gürses, Michael Veale, Cas Cremers, Michael Backes, Nils Ole Tippenhauer, Reuben Binns, Ciro Cattuto, Alain Barrat, Dario Fiore, Manuel Barbosa, Rui Oliveira, and José Pereira. 2020. Decentralized Privacy-Preserving Proximity Tracing. arXiv:2005.12273 [cs.CR]

[60] UNWTO. 2020. *100% of Global Destinations Now Have COVID-19 Travel Restrictions, UNWTO Reports*. Retrieved May 15, 2021 from https://www.unwto.org/news/covid-19-travel-restrictions

[61] Michel Walrave, Cato Waeterloos, and Koen Ponnet. 2020. Adoption of a contact tracing app for containing COVID-19: a health belief model approach. *JMIR public health and surveillance* 6, 3 (2020).

[62] Lili Wei, Yepang Liu, and Shing-Chi Cheung. 2016. Taming Android Fragmentation: Characterizing and Detecting Compatibility Issues for Android Apps. In *Proceedings of the 31st IEEE/ACM International Conference on Automated Software Engineering* (Singapore, Singapore). Association for Computing Machinery, New York, NY, USA, 226–237.

[63] Lili Wei, Yepang Liu, and Shing-Chi Cheung. 2019. Pivot: learning api-device correlations to facilitate android compatibility issue detection. In *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*. 878–888.

[64] Zhibiao Wu and Martha Palmer. 1994. Verbs Semantics and Lexical Selection. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics* (Las Cruces, New Mexico). Association for Computational Linguistics, USA, 133–138.

[65] Sebastian Zimmeck, Ziqi Wang, Lieyong Zou, Roger Iyengar, Bin Liu, Florian Schaub, Shomir Wilson, Norman M Sadeh, Steven M Bellovin, and Joel R Reidenberg. 2017. Automated Analysis of Privacy Requirements for Mobile Apps.. In *NDSS*.