The Journal of Systems & Software 184 (2022) 111136

Contents lists available at ScienceDirect

The Journal of Systems & Software

journal homepage: www.elsevier.com/locate/jss

In practice

Mining user reviews of COVID contact-tracing apps: An exploratory analysis of nine European apps☆

Vahid Garousi [a,b,*], David Cutting [a], Michael Felderer [c,d]

[a] Queen's University Belfast, UK
[b] Bahar Software Engineering Consulting Corporation, UK
[c] University of Innsbruck, Austria
[d] Blekinge Institute of Technology, Sweden

ARTICLE INFO

Article history:
Received 11 December 2020
Received in revised form 6 June 2021
Accepted 25 October 2021
Available online 4 November 2021

Keywords:
Mobile apps
COVID
Contact-tracing
User reviews
Software engineering
Software in society
Data mining

ABSTRACT

Context: More than 78 countries have developed COVID contact-tracing apps to limit the spread of coronavirus. However, many experts and scientists cast doubt on the effectiveness of those apps. For each app, a large number of reviews have been entered by end-users in app stores.
Objective: Our goal is to gain insights into the user reviews of those apps, and to find out the main problems that users have reported. Our focus is to assess the "software in society" aspects of the apps, based on user reviews.
Method: We selected nine European national apps for our analysis and used a commercial app-review analytics tool to extract and mine the user reviews. For all the apps combined, our dataset includes 39,425 user reviews.
Results: Results show that users are generally dissatisfied with the nine apps under study, except the Scottish ("Protect Scotland") app. Some of the major issues that users have complained about are high battery drainage and doubts on whether apps are really working.
Conclusion: Our results show that more work is needed by the stakeholders behind the apps (e.g., app developers, decision-makers, public health experts) to improve the public adoption, software quality and public perception of these apps.

1. Introduction

As of October 2020, more than 78 countries and regions have developed so far (or are developing) COVID contact-tracing apps to limit the spread of coronavirus.[1] The list is quickly growing, and as of this writing, 19 of those apps are open source.

Contact-tracing apps generally use Bluetooth signals to log when smartphones, and hence their owners, are close to each other, so if someone develops COVID symptoms or tests positive, an alert can be sent to other users they may have infected. An app can be developed using two different approaches: centralized or decentralized. Under the centralized model, the data gathered is uploaded to a remote server where matches are made with other contacts should a person start to develop COVID symptoms. This is the method that countries such as the UK were initially pursuing.

By contrast, the decentralized model gives users more control over their information by keeping it on the phone. It is there that matches are made with people who may have contracted the virus. This is the model promoted by Google, Apple and an international consortium, advised in part by the MIT-led Private Automated Contact Tracing (PACT) project (pact.mit.edu) (Scudellari, 2020). Both types have their pros and cons. Since early summer 2020, a split emerged between the two approaches. However, privacy and platform support issues have pushed countries to use the decentralized model.

The apps have been promoted as a promising tool to help bring the COVID outbreak under control. However, there are many discussions in the media, the academic (peer-reviewed) literature (Martuscelli and Heikkilä, 2020), and also the grey literature about the 'efficacy of contact-tracing apps' (try a Google search for the term inside quotes). A systematic review (Braithwaite et al., 2020) of 15 studies, which had studied the efficacy of contact-tracing apps, found that "*there is relatively limited evidence for the*

☆ Editor: Neil Ernst.

* Corresponding author at: Queen's University Belfast, UK.
E-mail addresses: v.garousi@qub.ac.uk (V. Garousi), d.cutting@qub.ac.uk (D. Cutting), michael.felderer@uibk.ac.at (M. Felderer).

[1] www.xda-developers.com/google-apple-covid-19-contact-tracing-exposure-notifications-api-app-list-countries/ and en.wikipedia.org/wiki/COVID-19_apps

https://doi.org/10.1016/j.jss.2021.111136
0164-1212/

impact of contact-tracing apps". A French news article[2] reported that, as of mid-August 2020, "*StopCovid [the French app] had over 2.3 million downloads [of a population of 67 million people] and only 72 notifications were sent [by the app]*".

One cannot help but wonder the reasons behind the low efficacy and low adoption of the apps by the general public in many countries. The issue is a multi-faceted, complex, and interdisciplinary issue, as it relates to fields such as public health, behavioral science (Anon, 2020), epidemiology, and software engineering.

The software engineering aspect of contact-tracing apps is quite diverse in itself, e.g., whether different apps developed by different countries will cooperate/integrate (when people travel across counties/borders), and whether the app software would work as intended (e.g., will it record the nearby phone IDs properly, and will it send the alerts to all the recorded persons?). The decentralized nature of the system makes such a verification a challenging task. Some other related developments include a news article reporting that a large number of developers worldwide have found a large number of defects in one of the contract-tracing apps (England's open-source app).[3]

Another software engineering angle of the issue is the availability of a high number of user reviews in the two major app stores: the Google Play Store for Android apps and the Apple App Store for the iOS apps. A user review often contains information about the user's experience with the app and opinion of it, feature requests, or bug reports (Genc-Nayebi and Abran, 2017). Many insights can be mined by analyzing the user reviews of these apps to figure out what end-users think of COVID contact-tracing apps, and that is what we analyze and present in this paper. Studies have shown that reviews written by the users represent a rich source of information for the app vendors and the developers, as they include information about bugs and ideas for new features (Iacob and Harrison, 2013). Mining of app store data and app reviews has become an active area of research in software engineering (Genc-Nayebi and Abran, 2017) to extract valuable insights. User ratings and reviews are user-driven feedback that may help improve software quality and address missing application features.

Among the insights that we aim at deriving in this study are the ratios of users which as per their reviews, have been happy or unhappy with the contact-tracing apps and the main issues (problems) that most users have reported about the apps. The nature of our analysis is "exploratory" (Runeson and Höst, 2009) in nature, as we want to explore the app reviews and extract insights from them which could be useful for the different stakeholders, e.g., app developers, decision-makers, researchers, and the public, to benefit from or act upon.

Also, the focus of our paper is software engineering "in society" (Kazman and Pasquale, 2019), since it is clear that contact-tracing apps are widely discussed in the public media and are used by millions of people worldwide, and also have the potential to have major influences on people's lives in the challenges that the COVID pandemics has brought upon all the people of the world. Furthermore, many resources have argued that "*these apps are safety-critical*",[4] since "*a faulty proximity tracing app could lead to false positives, false negatives, or maybe both*." It is thus very important that these apps and their user reviews be carefully studied to ensure that upcoming updates of existing apps or new similar apps have the highest software quality.

Another motivating factor for this study is ongoing research and consulting engagement of the first author in relation to the Northern Irish contact-tracing app (called "StopCOVID NI"[5]). Since May 2020, he has been a member of an Expert Advisory Committee for the StopCOVID NI app. Some of his activities so far have included peer review and inspection of various software engineering artifacts of the app, e.g., UML design diagrams, test plans, and test suites (see page 13 of an online report by the local Health Authority[6]). In that Expert Advisory committee, the members have felt the need to review and mine insights from user reviews in app stores to be able to provide a feedback loop to the committee and the software engineering team of the app. Thus, the current study will provide benefits in that direction (to the committee), and also, by analyzing other apps from other countries, we will provide insight for other stakeholders (researchers and practitioners) elsewhere too.

The methodology applied in this paper is an "exploratory" case study focussing on user feedback (ratings and comments) for nine widely used European apps as a representative subset of the 50+ available worldwide. A series of nine research questions (Section 3.2) were created and addressed using combinations of automated sentiment analysis, numerical ratings, download figures, and manual sampling of textual reviews.

Key results include a general dissatisfaction users' have for the various contact tracing apps (with the notable exception of the NHS Scotland app) and that users are often confused with the interface and operation, i.e., the apps are overcomplex. A consistent concern commonly raised across the apps was that of power consumption causing battery drain, something that had been widely reported in the media for early releases. Where geographical boundaries were close, and users would be expected to cross, such as in the United Kingdom, a lot of negative comments related to the lack of interoperability (if the user is over the border, their app won't work to record contacts).

The remainder of this paper is structured as follows. In Section 2, as background information, we provide a review of contact-tracing apps and then a review of related works. We discuss the research approach, research design, and research questions of our study in Section 3. Section 4 presents the results of our study. In Section 5, we discuss a summary of our results and their implications for various stakeholders (app developers, decision-makers, researchers, the public, etc.). Finally, Section 6 concludes the paper and discusses our ongoing and future works.

2. Background and related work

As the background and related work, we review the following topics in the next several sub-sections:

- Usage of computing and software technologies in the COVID pandemic (Section 2.1)
- A review of contact-tracing apps and how they work (Section 2.2)
- Related work on mining of app reviews (Section 2.3)
- Closely related work: Mining of COVID app reviews (Section 2.4)

After discussing those related works, we will position this work with regard to the related work in Section 2.5. For the interested reader, we also provide a further review of related work in the appendix in the following groups:

- Grey literature on software engineering of contact-tracing apps

2 www.lefigaro.fr/secteur/high-tech/stopcovid-2-3-millions-de-telechargements-et-seulement-72-notifications-envoyees-20200819

3 eandt.theiet.org/content/articles/2020/05/developers-find-new-flaws-in-source-code-of-nhs-contract-tracing-app/

4 www.eff.org/deeplinks/2020/04/challenge-proximity-apps-covid-19-contact-tracing

5 covid-19.hscni.net/stop-covid-ni-mobile-app/

6 covid-19.hscni.net/wp-content/uploads/2020/07/Expleo-StopCOVIDNI-Closure-Report-V1.0.pdf

Table 1
Digital technologies in the public-health response to COVID-19 pandemic (from Budd et al., 2020)

Public-health need	Digital tool or technology	Example of use
Digital epidemiological surveillance	Machine learning	Web-based epidemic intelligence tools and online syndromic surveillance
	Survey apps and websites	Symptom reporting
	Data extraction and visualization	Data dashboard
Rapid case identification	Connected diagnostic device	Point-of-care diagnosis
	Sensors including wearables	Febrile symptoms checking
	Machine learning	Medical image analysis
Interruption of community transmission	Smartphone app, low-power Bluetooth technology	Digital contact tracing
	Mobile-phone-location data	Mobility-pattern analysis
Public communication	Social-media platforms	Targeted communication
	Online search engine	Prioritized information
	Chat-bot	Personalized information
Clinical care	Tele-conferencing	Telemedicine, referral

- Formal and grey literature on overall quality issues of contact-tracing apps
- Behavioral science, social aspects, and epidemiologic aspects of the apps

2.1. Usage of computing and software technologies in the COVID pandemic

A number of digital, computing, and software technologies have been developed and are in use in the public health response to COVID-19 pandemic (Budd et al., 2020). A survey paper in the Nature Medicine magazine (Budd et al., 2020) reviewed the breadth of digital innovations (computing and software systems) for the public-health response to COVID-19 worldwide, their limitations, and barriers to their implementation, including legal, ethical, and privacy barriers, as well as organizational and workforce barriers. The paper argued that the future of public health is likely to become increasingly digital. We adopt a summary table from that paper in Table 1 (Budd et al., 2020).

As the table shows, there are various public health needs and various digital tools/technologies to address those needs. Contact-tracing mobile apps are just one of the digital tools/technologies to address one of those needs, i.e., interruption of community transmission.

Other than contact-tracing mobile apps, other types of software systems have also been developed and used, related to the COVID pandemic, e.g., a system named Dot2Dot,[7] which "*is a software tool to help health authorities trace and isolate people carrying an infectious disease*" and a mobile app named *COVIDCare NI*,[8] developed in Northern Ireland, by the regional healthcare authority. The app provides various features to users, e.g., accessing personalized advice based on user's answers to a number of symptom-check questions, deciding if the user needs clinical advice and how to access it, and easily find links to trusted information resources on COVID-19 advice and mental health resources.

2.2. A review of contact-tracing apps and how they work

As discussed in Section 1, more than 78 countries and regions have developed so far (or are developing) COVID contact-tracing apps to limit the spread of coronavirus.[9] The list is quickly growing, and as of this writing, 19 of those apps are open source.

Almost all proximity-detecting contact-tracing apps use Bluetooth signals emitting from nearby devices to record contact events (Budd et al., 2020). However, in August 2020, news[10] came out that a WiFi-Based contact-tracing app has been developed in the University of Massachusetts Amherst. In addition to Bluetooth and WiFi technologies, other technologies such as GPS (Wang et al., 2020), IP addresses (Wen et al., 2020), and ultrasound (Zarandy et al., 2020) have also been suggested to be used in COVID contact-tracing apps.

A contact-tracing app can be developed using either of two different approaches: centralized or decentralized. Centralized contact-tracing apps share information about contacts and contact events with a central server (often set up by the healthcare authority of a region or country). A centralized app uploads information when a user reports testing positive for COVID. Decentralized apps upload only an anonymous identifier of the user who reports testing positive for COVID. This identifier is then broadcast to all users of the app, which compares the identifier with on-phone contact-event records. If there is a match on the mobile app of a given user, that app gives a notification to the user. Taken from a paper in this area (Budd et al., 2020), Fig. 1 depicts the process of how these apps work. Another paper (Ahmed et al., 2020) has modeled the tracing process of a decentralized app as a UML sequence diagram (also shown in Fig. 1).

The most widely used framework for developing decentralized contact-tracing apps is the "Exposure Notification System (ENS)"[11,12] framework/API, originally known as the "Privacy-Preserving Contact Tracing Project", which is a framework and protocol specification developed by Google Inc. and Apple Inc. and to facilitate digital contact-tracing during the COVID-19 pandemic. The framework/API is a decentralized reporting-based protocol built on a combination of Bluetooth Low Energy (BLE) technology (Gomez et al., 2012) and privacy-preserving cryptography. But let us also note that there have been various critiques of Google/Apple's ENS framework (Hoepman, 2020).

In addition to the Google/Apple's ENS framework, other[13] frameworks and protocols have also been proposed and used for developing contact-tracing apps, e.g.,. Pan-European Privacy-Preserving Proximity Tracing (PEPP-PT) project (github.com/pepp-pt/), and BlueTrace/ OpenTrace (bluetrace.io). A comprehensive survey of contact-tracing frameworks and mobile apps is presented in Martin et al. (2020).

As of May 2020, at least 22 countries had received access to the protocol. Switzerland and Austria were among the first to back

7 www.dot2dot.app
8 play.google.com/store/apps/details?id=net.hscni.covid19ni
9 www.xda-developers.com/google-apple-covid-19-contact-tracing-exposure-notifications-api-app-list-countries/ and en.wikipedia.org/wiki/COVID-19_apps
10 www.eurekalert.org/pub_releases/2020-08/uoma-cso080720.php
11 www.google.com/covid19/exposurenotifications/
12 developer.apple.com/exposure-notification/
13 en.wikipedia.org/wiki/COVID-19_apps#List_of_frameworks.

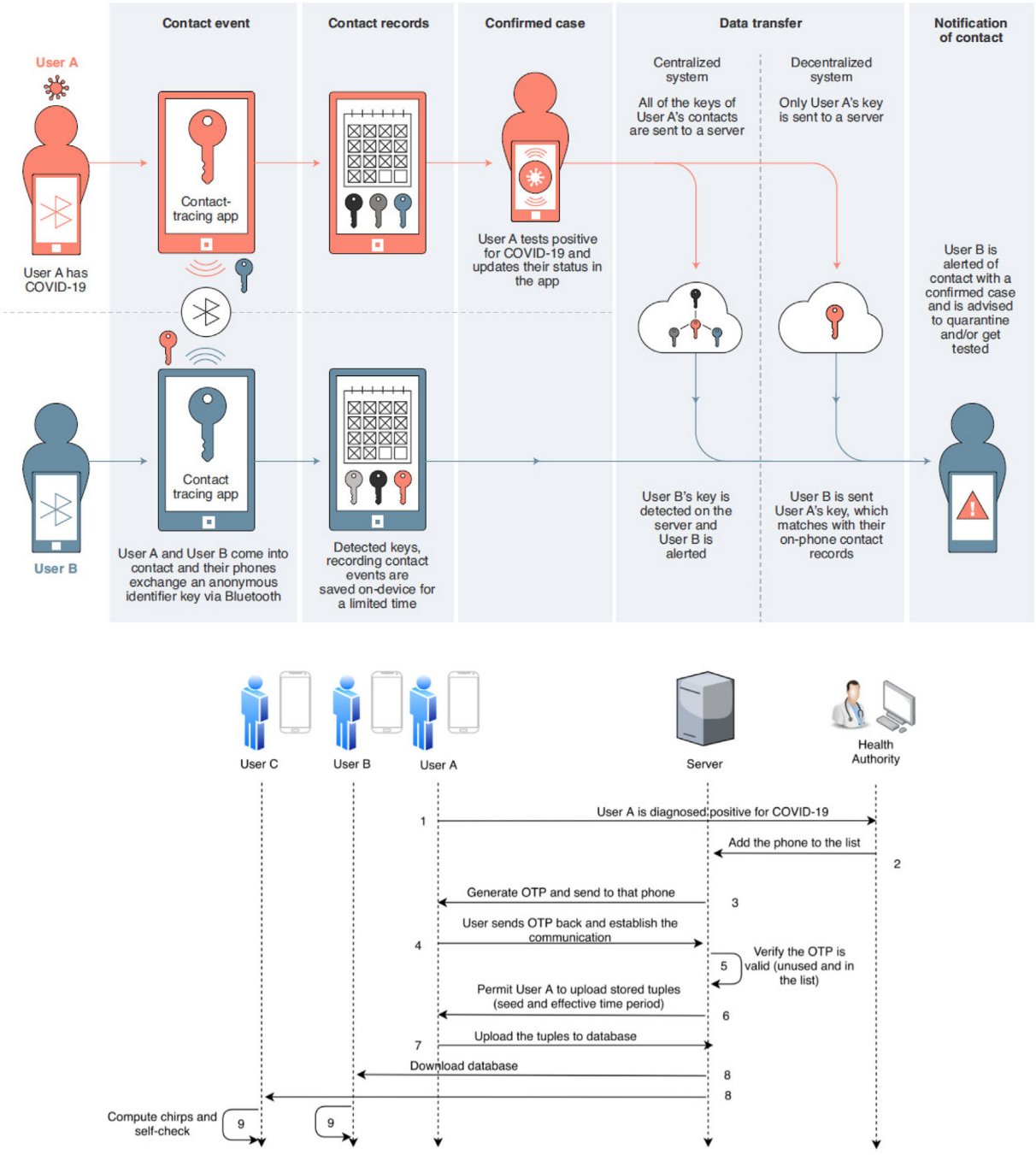


Fig. 1. How the COVID-19 contact-tracing apps work on Bluetooth-enabled smartphones (Budd et al., 2020). Tracing process of a decentralized app. *Source:* The sequence diagram is taken from Ahmed et al. (2020).

the protocol.[14] Shortly after, Germany announced it would back Exposure Notification, followed by Ireland and Italy.

More concretely, to know what features these apps provide, we show, as examples, several screenshots from the user interface of the *Protect Scotland* app and the *COVID Tracker Ireland* app in Fig. 2. We have taken these screenshots from the apps' pages[15] in the Google Play Store. For the case of the *Protect Scotland* app, we can see that it only provides the "basic"/core contact-tracing features (use cases), i.e., tracing, adding test results, and sending notifications to recorded (traced) contacts. However, the *COVID Tracker Ireland* app provides some extra features in addition to the "core" features, e.g., showing the number of registrations since the app's launch, COVID cases by county, etc.

Since early 2020, COVID has severely impacted the work and lives of almost everyone on the planet, and contact-tracing apps have been widely discussed in online media, social media, and news outlets. As of this writing (mid-December 2020), a Google search for "contact-tracing app"[16] returned 2,110,000 hits on the web, many of which are news about these apps in the media.

14 www.reuters.com/article/health-coronavirus-europe-tech/switzerland-austria-align-with-gapple-on-corona-contact-tracing-idUSL3N2CA36L

15 play.google.com/store/apps/details?id=gov.scot.covidtracker and play.google.com/store/apps/details?id=com.covidtracker.hse

16 www.google.com/search?q=%22contact-tracing+apps%22

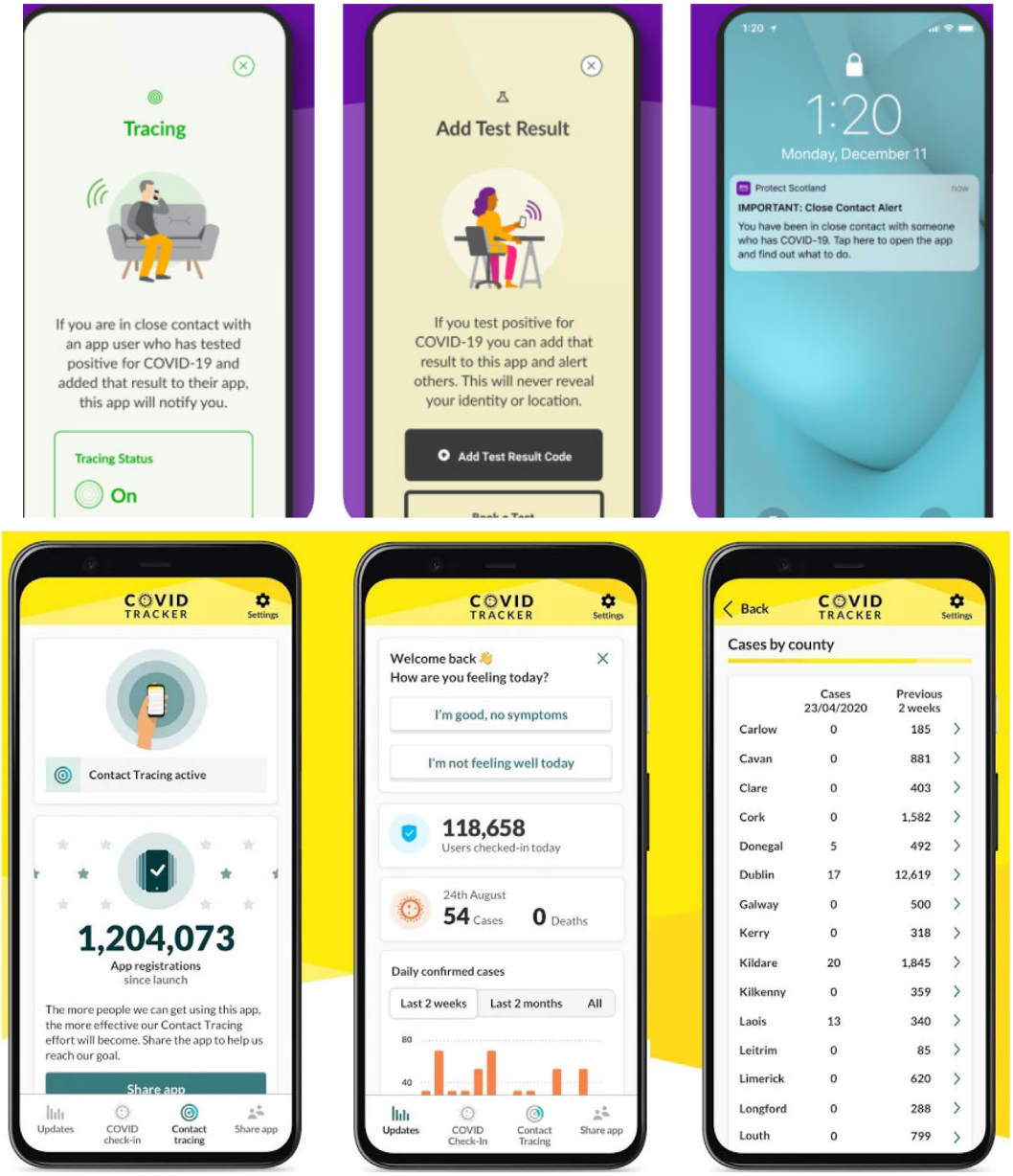


Fig. 2. Screenshots from the graphical user interface (GUI) of the *Protect Scotland* and *COVID Tracker Ireland* apps.

Also, in the relatively short timeframe since early 2020, many research papers have been published about these apps. As of this writing (mid-December 2020), a search in Google Scholar for "contact-tracing app"[17] returned 1010 papers, which have been published in different research areas, e.g., public health, behavioral science (Anon, 2020), epidemiology, and software engineering. We show a short list of a few interesting papers in the following, from that large set of papers:

- Contact tracing mobile apps for COVID-19: Privacy considerations and related trade-offs (Cho et al., 2020)
- One app to trace them all? Examining app specifications for mass acceptance of contact-tracing apps (Trang et al., 2020)
- A survey of covid-19 contact tracing apps (Ahmed et al., 2020)
- COVID-19 contact tracing apps: the 'elderly paradox' (Rizzo, 2020)
- On the accuracy of measured proximity of Bluetooth-based contact-tracing apps (Zhao et al., 2020)
- Vetting Security and Privacy of Global COVID-19 Contact Tracing Applications (Sun et al., 2020)
- COVID-19 Contact-tracing Apps: A Survey on the Global Deployment and Challenges (Li and Guo, 2020)

Also, various reports and news articles have discussed the high costs involved in the engineering (development and testing) of contact-tracing apps. For example, for the Australian app, the cost was estimated to be 70 million Australian dollars ($49 m USD).[18] For the UK NHS contact-tracing app, the cost was reported to be more than £35 million pounds.[19] The development cost of

17 scholar.google.com/scholar?%22contact-tracing+app%22

18 www.bbc.co.uk/news/technology-53485569

19 www.digitalhealth.net/2020/09/total-cost-of-nhs-contact-tracing-app-set-to-top-35-million/

the Irish app (*COVID Tracker Ireland*) was reported[20] to be about £773 K pounds only.

2.3. Related work on mining of app reviews

User feedback has long been an important component of understanding the successes or failures of software systems, traditionally in the form of direct feedback or focus groups and more recently through social media or the distribution channels themselves, i.e., feedback in app or software stores (Morales-Ramirez et al., 2015). A systematic literature review (SLR) (Genc-Nayebi and Abran, 2017) of the approaches used to mine user opinion from app store reviews identified a number of approaches used to analyze such reviews and some interesting findings such as correlation between app rating and downloads (apps rated as high-quality gain more users), but there are significant issues identifying the overall sentiment of many reviews through automated processing. Many of their reviewed studies identified the key difference in the ease with which ratings can be used numerically compared with the difficulties in "understanding" unstructured textual commentaries, especially in different societal and linguistic settings.

With ratings being seen as key to the success of apps (Genc-Nayebi and Abran, 2017), it is important to understand the concerns and issues that lead users to most commonly complain or leave poor reviews, work which is undertaken in Khalid et al. (2014). With respect to what users complain about in mobile apps, Khalid et al. (2014) qualitatively studied 6390 low-rated user reviews for 20 free-to-download iOS apps. They uncovered 12 types of user complaints. The most frequent complaints were functional errors, feature requests, and app crashes. Of particular note in the context of this paper and COVID apps is that privacy and ethics concerns are also a common type of complaint, with an example review given of an unnamed app that it is *"yet another app that thinks your contacts are fair game"* (Khalid et al., 2014). Beyond the iOS focus of Khalid et al. (2014), most successful apps co-exist in at least two ecosystems (Apple and Google) and share the same brand even if they may not share the same codebase.

Hu et al. (2019) seek to analyze reviews of the "same" app from both Android and iOS and compare the cross-platform results, finding that nearly half (32 out of 68) of hybrid apps (where the codebase is largely shared between platforms) "receive a significantly different distribution of star ratings across both studied platforms". The authors state that this shows a great deal of variability in how users perceive the apps even with the same fundamental features and interface depending on the users' platform.

When mining app store data also country-specific differences in mobile app user behavior were identified (Lim et al., 2014). The authors collected data from more than 15 countries, including the USA, China, Japan, Germany, France, Brazil, United Kingdom, Italy, Russia, India, Canada, Spain, Australia, Mexico, and South Korea. Analysis of data provided by 4824 participants showed significant differences between app user behaviors across countries. For example, users from the USA are more likely to download medical apps, users from the United Kingdom and Canada are more likely to be influenced by price, users from Japan and Australia are less likely to rate apps. Also, in this paper, we analyze app reviews from several countries and, therefore, should be aware of country-specific differences when analyzing the data.

Being able to take advantage of user feedback to learn lessons and improve current or future apps have also been studied. Several papers have taken advantage of this data, especially where the volume of reviews and ratings may make manual analysis impractical. Scherr et al. (2019) presented a lightweight framework built on the use of emojis as representative of emotive feeling and expression of an app, building from their initial findings that large numbers of textual reviews also included emojis. Beyond general opinions, Guzman and Maalej (2014) presented an approach to look at user sentiment with relation to specific features and use techniques such as Natural Language Processing (NLP) to gain this insight.

Within the health domain, Stoyanov et al. (2015) defined a mobile app rating scale called "MARS" with a specific focus on descriptors aligned to health apps, including mental health ranging from UX to quality perceptions and technical considerations. This approach has the potential to be widely applied to health-related apps and used as a base of comparison between them.

Mining of app-store data, including app reviews, has become an active area of research in software engineering (Genc-Nayebi and Abran, 2017). Papers on this topic are typically published in the Mining Software Repositories (MSR) community. Authors in Genc-Nayebi and Abran (2017) provide a systematic literature review (SLR) on opinion mining studies from mobile app store user reviews. The SLR shows that mobile app ecosystems and user reviews contain a wealth of information about user experience and expectations. Furthermore, it is highlighted that developers and app store regulators can leverage the information to better understand their audience. This also holds for COVID contact-tracing apps as applied in this paper. However, the SLR also highlights that opinion spam or fake review detection is one of the largest problems in the domain. Further studies on app ratings cover topics on quality improvement through lightweight feedback analyses (Scherr et al., 2019), sentiment analysis of app reviews (Guzman and Maalej, 2014), and consistency of star ratings and reviews of popular free hybrid Android and iOS apps (Hu et al., 2019).

2.4. Closely related work: Mining of COVID app reviews

In terms of related work, three insightful blog posts under a series entitled "*What went wrong with Covid-19 Contact Tracing Apps*" have recently appeared in the IEEE Software blog.[21] The articles reported analyses of user reviews of three such apps: Australia's CovidSafe App, Germany's Corona-Warn App, and the Italian app. They presented thematic findings on what went wrong with the apps, e.g., lack of citizen involvement, lack of understanding of the technological context of Australian people, ambitious technical assumptions without cultural considerations, privacy and effectiveness concerns.

2.5. Positioning this work with related work

Based on the review of each above category of related work in the above sections and also in the appendix, we can position this work with regard to related work as follows: This paper is the closest to the studies which have mined COVID app reviews, and then the large body of knowledge on mining of app reviews, in general. Since an important aspect of our work is quality assurance of COVID apps, based on user reviews, our work also related to the formal literature and grey literature on overall quality issues of contact-tracing apps, as reviewed in the appendix. In more general terms, this work is also positioned within the area of mining reviews to infer software quality and adoption.

A recent paper (Rekanar et al., 2020) presented a sentiment analysis of user review on the Irish (*COVID Tracker Ireland*) app.

[20] www.theguardian.com/world/2020/jul/20/cheap-popular-and-it-works-irelands-contact-tracing-app-success

[21] blog.ieeesoftware.org/2020/09/what-went-wrong-with-covid-19-contact.html

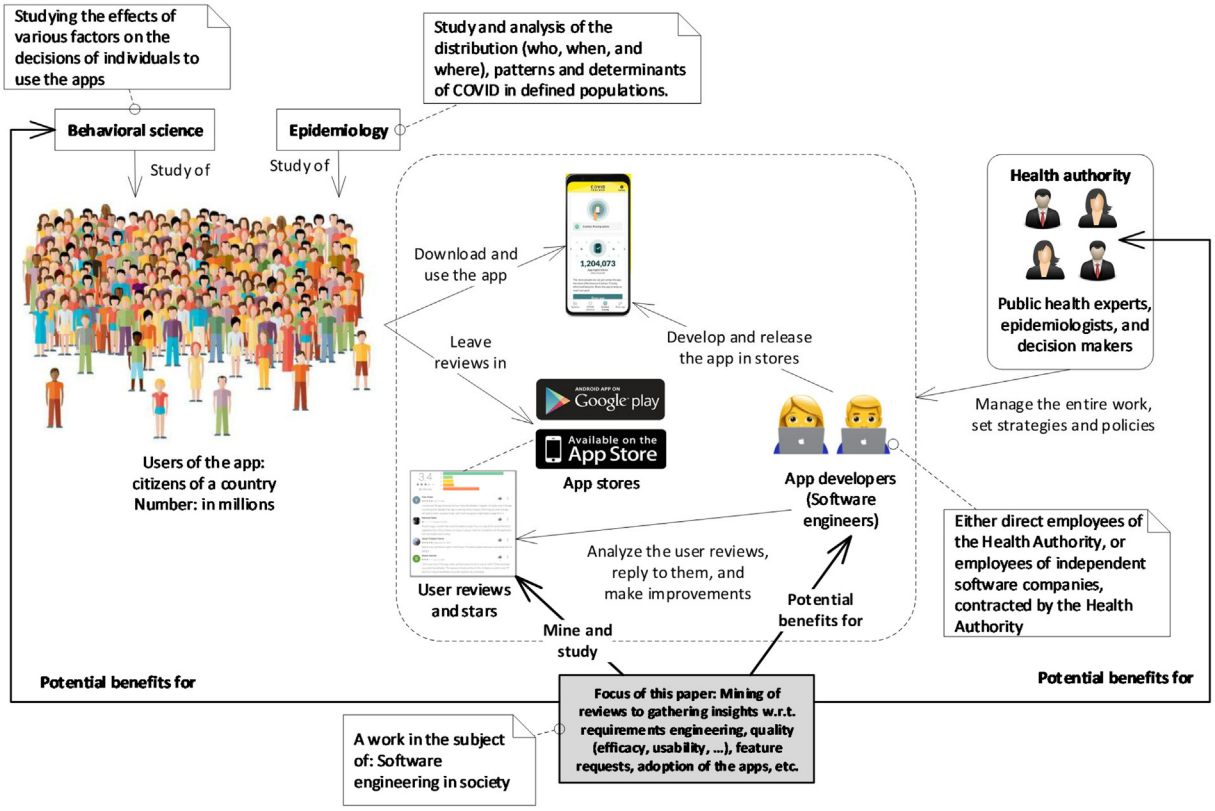


Fig. 3. Research context of this study, including the key stakeholders: app users (the public), app software engineers, public-health experts, and decision-makers.

While our current paper has some similarity in objectives with those articles, we look at more apps (nine), and also, the nature and scale of our analyses are different (more in-depth) compared to the analyses, reported in the above blog posts and paper.

3. Research context, method, and questions

We discuss next the context of our research, the research method, and the research questions of our study. We then discuss our dataset and the tool that we have used to extract and mine the app reviews.

3.1. Context of research

It is important to understand and clarify the research context (scope) of our work. To best present that, we have designed a context diagram as shown in Fig. 3.

In the center of the study are the contact-tracing apps and the reviews entered by users in the app stores, who are citizens of a given country, and their number is often in the millions. A team of software engineers develops and releases the app in stores.

A team of public-health experts and decision-makers who work for a country's Health Authority (e.g., ministry) manage the entire work (project), set strategies and policies related to the apps and their release to the public. Software engineers are either direct employees of the Health Authority or employees of an independent software company, which is contracted by the Health Authority to develop and maintain the app.

The focus of this paper is to mine the user reviews and gather insights, with the aim of providing benefits for various stakeholders: the software engineering teams of the apps, public-health experts, decision-makers, and also the research community in this area.

We should mention that the involved teams of software engineers may already read and analyze the user reviews, sometimes replying to them in the app stores, and make improvements in their apps accordingly. However, those software engineers often only focus on reviews of their own apps. Our study extends and takes a different angle on the issue by considering several apps and analyzes the various trends in the reviews of those apps. We will discuss the apps under study and how we have selected (sampled) them from among all worldwide contact-tracing apps in Section 3.3.

While our focus in this work positions this work in the area of software and software engineering in society (Kazman and Pasquale, 2019), we also show in Fig. 3 two related fields (behavioral science and epidemiology), which we reviewed for relevant literature related to contact-tracing apps, in Section 2.5. Our analysis in this paper (Section 4) could provide potential benefits to researchers and practitioners in those fields as well.

3.2. Research method and research questions

The research method applied in this paper is an "exploratory" case study (Runeson and Höst, 2009). As defined in a widely-cited guideline paper for conducting and reporting case study research in software engineering (Runeson and Höst, 2009), the goals of exploratory studies are "*finding out what is happening, seeking new insights and generating ideas and hypotheses for new research*", and those have been the goals of our study.

For data collection and measurement, we used the Goal-Question-Metric (GQM) approach (Basili, 1992). Stated using the GQM's goal template (Basili, 1992), the goal of the exploratory case study reported in this paper is to understand and to gain insights into the user reviews (feedback) of a subset of COVID contact-tracing apps from the point of view of stakeholders of

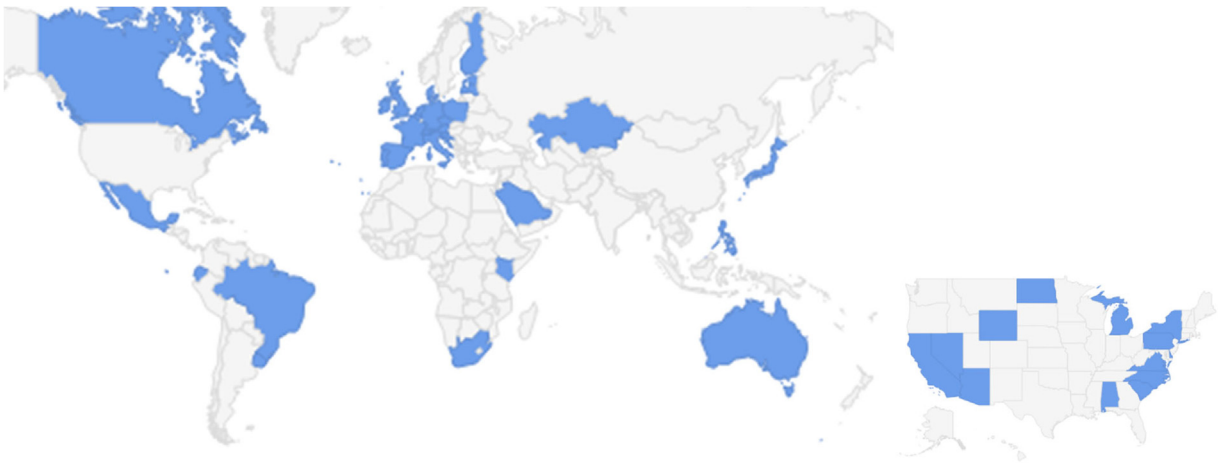


Fig. 4. The 35 countries and 15 US states which have developed contact-tracing apps.

these apps (e.g., app developers, decision-makers, and public health experts).

Based on the above goal and also given the types of user review data available in app stores, we derived the following research questions (RQs):

- RQ1: What ratios of users are satisfied/dissatisfied (happy/unhappy) with the apps?
- RQ2: What level of diversity/variability exists among different reviews and their informativeness?
- RQ3: What are the key problems reported by users about the apps?
- RQ4: By looking at the "positive" reviews, what aspects have users liked about the apps?
- RQ5: What feature requests have been submitted by users in their reviews?
- RQ6: When comparing the reviews of Android versus the iOS versions of a given app, what similarities and differences could be observed?
- RQ7: Is there a correlation between the number of app downloads and the country's population size?
- RQ8: Are there correlations between the number of reviews and the country's population or the number of downloads? And also, what ratio of app users has provided reviews?
- RQ9: What insights can be observed from the trends of review volumes and their sentiments over time?

An important aspect of our research method is the data analysis technique, which is mainly data mining. As we discuss in-depth in Section 3.4, we have selected and used a widely used commercial app-review data mining and analytics tool.

3.3. Apps under study: Sampling a subset of all worldwide contact-tracing apps

As discussed in Section 1, according to a regularly updated online article[22] in the grey literature, more than 78 countries and regions have developed so far (or are developing) contact-tracing apps. At least five other open-source contact-tracing implementations have been developed, based on the Apple-Google Exposure Notification API,[23] e.g., the apps by MIT and MITRE Corporation. They could be, in principle, reused and adapted by any country/region's healthcare agency. We show in Fig. 4 the 35 countries and 15 US states that have developed and published contact-tracing apps. These data have been taken from the above online article[22] (as of mid-September 2020).

Analyzing user reviews of "all" those 50+ apps would have been a major undertaking, and thus, instead, we decided to sample a subset of all worldwide contact-tracing apps, including nine apps. Also, to make the assessments more comparable, we limited the sampling to European countries by selecting the four apps developed in the British Isles and five apps from mainland Europe. We selected the apps developed for England and Wales (parts of the UK), the Republic of Ireland, Scotland, Northern Ireland, Germany, Switzerland, France, Austria, and Finland.

Table 2 lists the names, key information (such as first release dates and versions since the first release), and descriptive statistics of both Android and iOS versions of the nine selected apps. Each app can easily be found in each of the two app stores by searching for its name. Let us note that all data used for our analysis in this paper was gathered on September 17, 2020. We discuss in the next section the tools we used to extract and mine the data in this paper (including those shown in Table 2).

In terms of the number of downloads, we did not find any publicly available exact metrics in the app stores. Google Play Store provides approximate download counts in the form of, for example, 100,000+ (meaning 100,001–500,000). Apple App Store does not provide any exact nor estimate of download counts for the iOS apps.

An interesting point in Table 2 is that some apps have had many versions since the first release, and some only had a few. Each app has received anywhere between only 63 (NHS COVID) to 20,972 reviews (Corona-Warn Germany), and counting. It is interesting to see that, in all cases, the Android apps have received more reviews compared to iOS apps. This seems to align with the general trend in the app industry, as reported in the grey literature: "*Android users tend to participate more in reviewing their apps*"[24] and "*Android apps get way more reviews than iOS apps*".[25]

We will conduct and report some correlations analysis in Section 4.7 on some of the metrics shown in Table 2.

We should mention that in the writing phase of this paper (in November 2020), we heard the news that France launched[26]

22 www.xda-developers.com/google-apple-covid-19-contact-tracing-exposure-notifications-api-app-list-countries/

23 google.com/covid19/exposurenotifications

24 https://medium.com/@takuma.kakehi/we-need-app-reviews-but-we-need-to-ask-at-the-right-time-e2916b126c8e

25 medium.com/@chiragpinjar/why-android-apps-get-way-more-reviews-than-ios-apps-30c5b9e7ee71

26 www.healthcareitnews.com/news/emea/france-launches-new-contact-tracing-app-tousanticovid

Table 2
The sampled apps and their descriptive statistics (*: As discussed in the text, all data used for our analysis in this paper were gathered on Sept. 17, 2020)

App	OS	First release date	# of downloads	Versions since first release	Reviews*		
					# of reviews (as of our analysis)	Avg. stars of reviews	Ratio of Android to iOS reviews
1-StopCOVID NI	Android iOS	July 28, 2020	100,000+ –	2	195 97	3 2.5	2.01
2-NHS COVID (ENG)	Android iOS	August 13, 2020	100,000+ –	2	174 63	1.9 2.3	2.76
3-Protect Scotland (SCO)	Android iOS	September 10, 2020	100,000+ –	3	573 110	4 4	5.21
4-COVID Tracker Ireland (IE)	Android iOS	June 19, 2020	500,000+ –	3	1,463 274	2.9 3.1	5.34
5-Corona-Warn Germany (DE)	Android iOS	June 25, 2020	5,000,000+ –	8	20,972 6,772	2.7 2.3	3.10
6-SwissCovid (CH)	Android iOS	June 18, 2020	500,000+ –	9	1,370 652	3.1 3.1	2.10
7-StopCovid France (FR)	Android iOS	June 6, 2020	1,000,000+ –	10	2,397 241	2.6 2.1	9.95
8- Stopp Corona Austria	Android iOS	Mar 27, 2020	100,000+ –	10	1,961 599	2.4 2	3.27
9-Finland Koronavilkku (FI)	Android iOS	August 31, 2020	1,000,000+ –	3	1,276 236	3.4 3.3	5.41

a new contact-tracing app, named *TousAntiCovid* (literally translates to: "All Anti Covid") in late October 2020, which replaced the previous app, StopCovid, in the app stores. However, the review data that we had fetched using our chosen analytics tool (AppBot, as discussed in the next section) was until mid-September, so our analysis is on the France' StopCovid app, and the dataset had integrity w.r.t. that app.

3.4. Tool used to extract and mine the app reviews, and the dataset

We wanted to use an automated approach to extract, mine, and analyze the apps' user reviews. We came across the Google Play API,[27] which provides a set of functions (web services) to get such data. At the same time, we found that there are many powerful online tools that do the job of fetching the review data from app stores and even include useful advanced features such as text mining, topic analysis, and sentiment analysis (Guzman and Maalej, 2014) on review texts. The large number of such tools indicate the fact that an active market for app review "analytics" is emerging. There are also various research-prototype tools for mining user requirements and feedbacks, such as github.com/openreqeu and github.com/supersede-project. However, we found that there are only a handful of out-of-the-box working approaches, and commercial tools with high usability.

We came across a high-quality candidate tool to extract and mine the app reviews, i.e., a commercial tool named AppBot (appbot.co). The tool provides a large number of data-mining and sentiment analysis features. For example, as we will use in Section 4.1, AppBot uses an advanced method, based on AI and Natural Language Processing (NLP), to assign one of the four types of sentiments for each given review: positive, neutral, mixed, and negative sentiment. Also, as we will discuss in 4.5, another feature of AppBot is to automatically distinguish reviews that contain "Feature requests" submitted by users among all reviews of an app.

To do the above analysis, there have been specific papers that have proposed (semi-) automated techniques, which could be somewhat seen as the competitors for commercial App-analytics tools, such as AppBot. For example, a paper by Maalej and Nabil (2015) introduced several probabilistic techniques to classify app reviews into four types: bug reports, feature requests, user experiences, and ratings. The approach uses review metadata such as the star rating and the tense, as well as text classification, NLP, and sentiment analysis techniques.

Other papers have proposed or used sentiment techniques to classify each review, e.g., into a positive or negative review, just like what the AppBot tool does. For example, the authors of Guzman and Maalej (2014) used NLP techniques to extract the user sentiments about apps' features.

In summary, to make our choice of tools/techniques to extract and mine the app reviews, we could either use the approaches presented in the above papers or the commercial tool AppBot. To make our tool choice, we tried the AppBot tool on several apps in our selected pool and observed that the tool works well and its outputs are precise. Also, the fact that "24 of Fortune-100 companies" (according to the tool's website: appbot.co), e.g., Microsoft, Tweeter, BMW, LinkedIn, Expedia and New York Times are among the users of the tool, were strong motivations for us in favor of the AppBot tool over the techniques presented in the above papers. In addition, almost all techniques presented in the above papers had no publicly available tool support, and if we had to choose them, we had to develop new tools, which was clearly extra work, for which we saw no reason. Thus, we selected and used AppBot for all the data extraction and data mining.

However, we were still curious about the precision of the analyses (e.g., sentiment-analysis algorithm) done by AppBot. We initiated personal email communication with the co-founder of AppBot, asking about the precision of the analyses by the tool. The reply that we received was: "*We [have] trained our own sentiment analysis so it worked well with app reviews. Here's the details of our algorithm:*

- *Developed specifically for short forms of user feedback, like app reviews*
- *Understands the abbreviations, nuanced grammar and emoji*
- *Powered by machine learning*
- *Over 93% accuracy*

[27] developers.google.com/android-publisher/api-ref/rest/v3/reviews

Table 3
Sampling methods applied to address the research questions.

Research Question	Sampling methods
RQ1: What ratios of users are satisfied/dissatisfied (happy/ unhappy) with the apps?	All reviews were analyzed to generate sentiment charts of all reviews of each app.
RQ2: What level of diversity/variability exists among different reviews and their informativeness?	All reviews were analyzed to generate box plots of reviews text length. Then, we used "stratified" random sampling to choose a few reviews for discussions in paper text: choosing a few "long" reviews (in terms of text length) and a few "short" reviews to discuss the issues via examples.
RQ3: What are the key problems reported by users about the apps?	All reviews were analyzed via word-cloud visualizations. Then, we used "stratified" random sampling to choose a few reviews for discussions: choosing a few reviews by random from each group of key topics (words) appearing the most in reviews.
RQ4: By looking at the "positive" reviews, what aspects have users liked about the apps?	The set of "positive" reviews for each app was derived using the AppBot tool. We also calculated the ratios of positive reviews among all reviews of an app, and report in the paper (Section 4.4). Then, we used random sampling to choose a few reviews for discussions in paper text.
RQ5: What feature requests have been submitted by users in their reviews?	The set of " feature request" reviews for each app was derived using the AppBot tool. Then, we used random sampling to choose a few reviews for discussions in paper text.
RQ6: When comparing the reviews of Android versus the iOS versions of a given app, what similarities and differences could be observed?	All reviews were analyzed to generate charts. To compare the problems reported for each OS version, we used random sampling to choose a few reviews for discussions in paper text.
RQ7: Is there a correlation between the number of app downloads and the country's population size?	All reviews were analyzed to generate charts.
RQ8: Are there correlations between the number of reviews and the country's population or the number of downloads? And also, what ratio of app users has provided reviews?	All reviews were analyzed to generate charts.
RQ9: What insights can be observed from the trends of review volumes and their sentiments over time?	All reviews were analyzed to generate charts.

- *Trained on over 400 million records*"

We thus were quite satisfied that the tool that we were going to use has high quality and high precision in the analyses and results that it produces.

From another perspective, we are followers of the "open science" philosophy and reproducible research, especially in empirical software engineering (Fernández et al., 2019), and we believe that empirical data generated and analyzed in any empirical software engineering study should be provided online (when possible) for possible use by other researchers, e.g., for replication and transparency. By following that principle, we provide all the data extracted and synthesized for this paper in the following online repository: www.doi.org/10.5281/zenodo.4059087. Since to download the raw review data and analyze them using the commercial tool AppBot, we acquired a paid license for it, we cannot share the raw dump of all review data for all the apps in the above online repository, but instead, we share in there the aggregated statistics that we have gathered from the raw review data. Interested readers can easily acquire a license for the tool (AppBot) and download the raw data.

We also think that some readers may be interested in exploring the dataset and reviews on their own and possibly conducting further studies like ours. To help with those, we have recorded and provided a brief (10-minute) video of live interaction with the dataset (to be analyzed in this paper) using AppBot, which can be found in youtu.be/qXZ_8ZTr8cc.

3.5. Sampling method for each RQ

Given the considerable size of review datasets of the apps, for the case of certain RQs in our study (Section 3.2), we had to choose and apply suitable sampling methods to be able to systematically address each RQ (in Sections 4.1–4.8). In Table 3, we present the sampling methods that we applied to address the RQs.

When planning and applying the following sampling approaches, we benefitted from sampling guidelines in software engineering research (Baltes and Ralph, 2020) and general literature about sampling (Henry, 1990).

4. Results

We present the results of our analysis by answering the RQs of our study. Note that our study comprehensively looks at nine RQs of our study (as raised in Section 3.2) in detail through the next nine sub-sections, and thus, this section is quite extensive in terms of size and depth. Since most RQs are quite "independent" from each other, the reader can read each of the following result sub-sections independently and does not have to read all of them sequentially and in full, from this point on.

4.1. RQ1: What ratios of users are satisfied/ dissatisfied (happy/ unhappy) with the apps?

Our first exploratory RQ (analysis) was to assess the ratios of users, which, as per their reviews, have been happy or unhappy with the apps.

"Stars" (a value between 1–5) are the built-in rubric of app stores (both the Google Play and the Apple App Store) which let users mention their level of satisfaction or dissatisfaction with an app when they submit their review. This feature is also widely used in many other online software systems, such as online shopping (e-commerce) web applications, including Amazon. For the case of online shopping and also paid mobile apps, the number "stars" on a product (or app) often strongly impacts the choice of

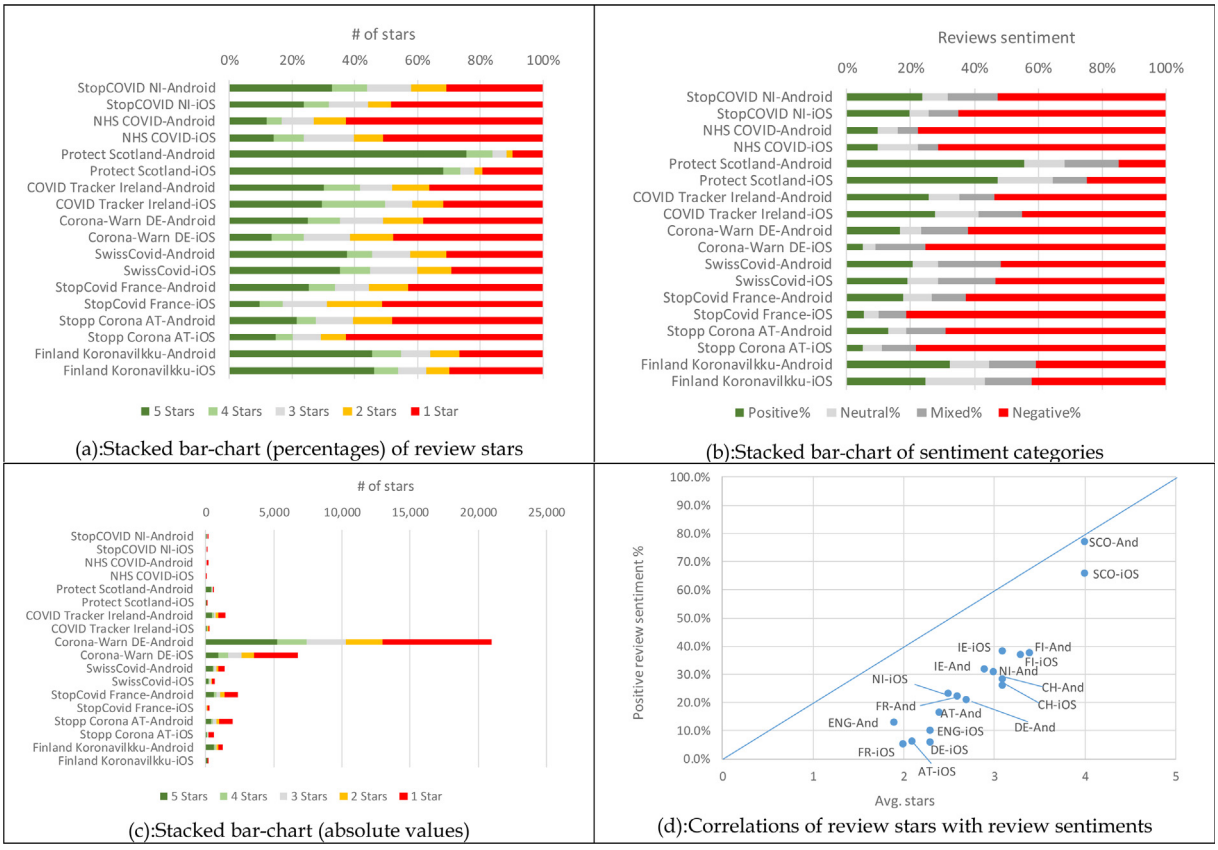


Fig. 5. Distribution of review stars, and review sentiment categories.

other users whether to buy a product (or app) or not (Hu et al., 2019), a relationship also seen in the levels of adoption of apps in mobile app stores (Genc-Nayebi and Abran, 2017).

A user can choose between 1 to 5 stars when s/he submits a review. Another more sophisticated way to derive users' satisfaction with an app is to look at the semantic tone of the review text, e.g., when a user mentioned in her/his review: "*I really like this app!*" that would clearly mean her/his satisfaction with the app. On the other hand, a review text like: "*the app crashed on my phone several times. Thus, it is not a usable app.*", implies the user's dissatisfaction with the app. Making broad use of this, especially on longer textual reviews, can have some limitations when automatically analyzed, but the majority of reviews can have sentiment successfully detected (Genc-Nayebi and Abran, 2017).

In the NLP literature, automatic identification of the semantic tone of a given text is referred to as *sentiment* analysis (Liu, 2012). Sentiment analysis refers to the use of NLP to systematically quantify the affective state of a given text. A given text can have four types of sentiments (Liu, 2012): positive, negative, neutral, and mixed. A positive sentiment denotes that the text has a positive tone in its message. "Neutral" sentiment implies that there is no strong sentiment in the text, e.g., "*I have used this app*". A text is given the "mixed" sentiment when it is conflicting sentiments (both positive and negative).

Our chosen data-mining tool (AppBot) supports the above four types of sentiments for each given review: positive, neutral, mixed, and negative sentiment. To classify the sentiment for a given review, AppBot calculates and provides a sentiment score of each review (a value between 0%–100%).

We show in Fig. 5 the distribution of stars as entered by the users in reviews and also the distribution of reviews' sentiment categories. We show both a 100% stacked bar and a stacked bar of absolute values for the stars. As we can see, since the German Corona-Warn app has received many more reviews compared to the others in the set, it has overshadowed the others in the stacked-bar figure.

We can see from these charts and also the average stars of each app (Table 2) that the users are generally dissatisfied with the apps under study, except the Scottish app. We furthermore averaged the stars from the mean score of Android and iOS versions of each app, e.g., for *StopCOVID NI*, this resulted in 2.75 (average of 3 and 2.5). Based on this metric, the *Protect Scotland* app is the highest starred (4/5), and *NHS COVID* is the least starred (2.1/5). The average of stars for all the other apps ranges between these two values. We should note that we have not installed nor tried any of the apps, and thus all our analyses are purely based on mining user reviews.

One very interesting consideration is what factors have led to the Scottish app be ranked the highest in terms of stars. Reviewing a subset of its reviews revealed that the app seems easy to use and is quite effective, e.g., one user said: "*Brilliant app. It collects zero personal data, no sign ups, no requirement to turn on location, nothing! All you have to do is turn on Bluetooth, that's it.*".[28] Of course, more in-depth assessment and comparison of the apps are needed to be done.

We were expecting that stars and the reviews' sentiments would have correlations, i.e., if a user has left a 1 star for an app, s/he has most probably had also left a negative (critical) comment in the review, and vice versa. We show in Fig. 5 a scatter-plot of those two metrics, in which 18 dots correspond to the 18 apps under study. The Pearson correlation coefficient of the two measures is 0.93, showing a strong correlation.

28 bit.ly/ScottishAppAPositiveReview

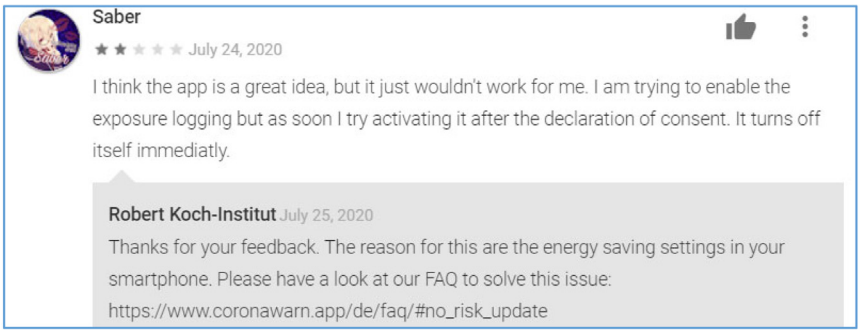


Fig. 6. A user review and the reply by the development team for the German app.

When comparing the average stars with the positive reviews sentiment percentages in Fig. 5-(d), the sentiment percentages seem to be consistently more negative (have lower values in the Y-axis). By analyzing a subset of the dataset (reviews), we observed that many reviews are similar to the following phrase/tone: "*I like the app, but rant rant rant*", and then the user has entered 4 stars, for example. The "rant" (complain) part could be quite harsh, thus causing the textual reviews sentiment score to fall down.

It is also interesting to see in Fig. 5-(d) that, generally, the dots of the two OS versions of each app are relatively close to each other in this scatter-plot, meaning that users have independently scored both versions of each app in quite similar levels. In some cases, the iOS version of a given app has a slightly higher average star value than the Android version, and it was the other way around for the other apps. A uniform relationship could not be observed.

Lesson learned/recommendation: The users are generally dissatisfied with the apps under study, except the Scottish app. Future studies could look into what factors have made the Scottish app be different than others in the pool of apps under study. That could a research question (RQ) to be studied by researchers in future works.

4.2. RQ2: A large diversity/variability in reviews and their informativeness

In addition to using the AppBot tool for automated text mining and sentiment analysis of the large set of reviews, it was important to read a subset of reviews, to actually get a sense of the dataset. For example, we browsed through the large list of 27,000+ reviews of the German Corona-Warn app. The Google Play Store provides a "like" button to let users express whether they found a given review "helpful". We found a few such reviews, such as the following[29]:

"Solid user interface and good explanation of the data privacy concept. Surprisingly well done, I was expecting it to be more cumbersome. Edit: It would be good if we could see how many tokens the app has collected in the last 14 days. This would make the app more attractive to open and raise the confidence in that it actually works. Also interesting metric would be to know how many users have been warned by the app. This has been released to the public (I believe 300 notifications so far). Unfortunately, I cannot find the Android system settings which apparently shows the number of contacts collected. Either its not available easily or I just can't find it. Anyway - I think it would be great if the app could show this information rather than asking the user to search for information in the settings". (translated automatically from German by AppBot, which uses the Google Translate API) (see Fig. 6).

Many reviews, including the above one, were feature requests, and some of them could indeed be useful for the development team for improving the app. Many reviews were also replied by the development team in a careful way, which was refreshing to see. For example, there was the following thread in one of the reviews[30] for the German app:

From the above example review, we can realize that the apps should be designed as simply as possible, since typical citizens ("laymen") are not often "technical" people, and we cannot assume that they will review the online FAQ pages of the app to properly configure it.

Lesson learned/recommendation: Contact-tracing apps should be designed as simple as possible (for usability), as we cannot expect layperson citizens to review the online FAQ pages of the app to properly configure it, especially for a safety-critical health-related app.

Lesson learned/recommendation: Developers of the apps can and should engage directly with reviews and reply, not only gaining insight into the most commonly raised concerns but also answering the questions in public view. This can even provide a positive "image" of the software engineering team behind the app in public view (in terms of accountability, responsiveness, and being open to feedback).

Essentially, similar to any other mobile app, reviews could range from short phrases such as "*Not working. Weird privacy settings*",[31] which are often not useful nor insightful for any stakeholder, to detailed objective reviews (like the one discussed above), which are often useful.

One way of analyzing the diversity/variability of reviews was to measure each review's length in words. We gathered those data for five of the nine apps (as examples) and provided the boxplots of both OS versions of those five example apps in Fig. 7. Since we observed that there are many "outlier" data points in the box plots, we provide the plots with and without outliers. For the readers who are less familiar with boxplots, we provide a conceptual example in Fig. 7 about the meaning of the boxes in boxplots and lines in it. More details about boxplots and their

29 bit.ly/ADetailedReviewInGermanApp

30 play.google.com/store/apps/details?id=de.rki.coronawarnapp&hl=en&reviewId=gp%3AAOqpTOHpV4mTMXZCOblZUe32-fVVSRjMY2PH9Jtrd1nKfkzZo7__2Pidwr0Ex_6W7P5FYRCAwrakyhYauBALyw

31 bit.ly/AShortReviewInGermanApp

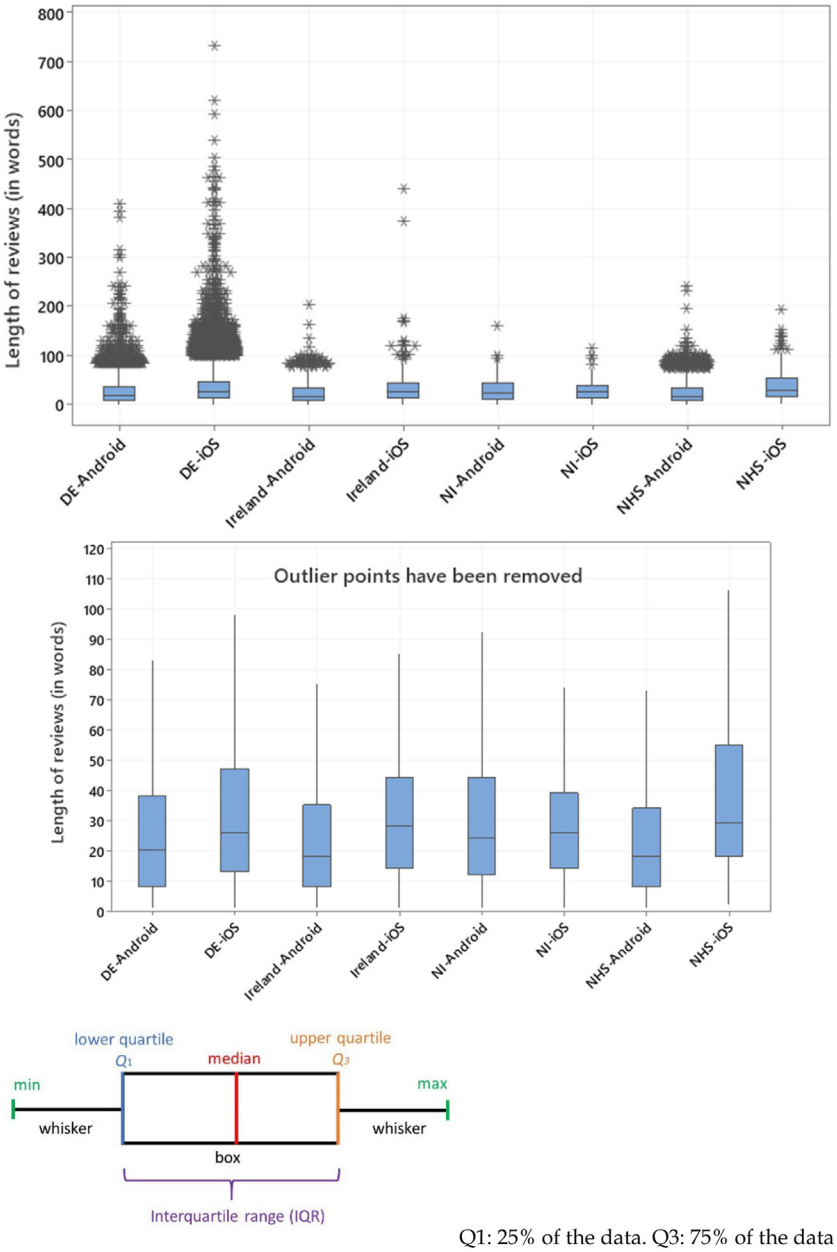


Fig. 7. Boxplot showing the distribution of textual "length" of reviews (in words) for five example apps. With and without "outlier" data points.

terminology can be found in the statistics literature (Potter et al., 2006).

As we can see in Fig. 7, for all five apps, the bulk of reviews are relatively short in length. The German app, on both Android and iOS platforms, has a more noticeable collection of longer comments which is particularly evident for the iOS version. The slight increase of the median may be due to linguistic differences, but the relatively large number of longer (>200 word) reviews implies a great degree of user engagement in commenting on these apps, especially on iOS, or could have cultural/social root causes, e.g., it could be German users often tend to provide "detailed" (extensive) feedbacks.

As another insight, we found that a proportion of reviews included error messages or crash reports. For example, for the German app (Corona-Warn) again, a user mentioned in her/his review[32]: *"After a few days the app stopped working. Several error messages appeared including 'cause: 3' and 'cause: 9002'. Tried to troubleshoot it by checking the Google services version, deleting the cache, reinstalling the app etc".*

When we interpret this review, it is logical to conclude that it is quite impossible for a layperson to deal with such errors and error messages, given the nature and full public outreach of the app. Thus, we wonder whether such error messages and crashes have been one of the several reasons why the apps under study have been rated quite low in reviews overall. By reading more reviews, we observed that many users (citizens) with some technical (IT) background had taken various steps to make the apps work, e.g., reinstalling them, etc. However, we believe that, for a layperson, taking such troubleshooting steps is out of the

[32] bit.ly/GermanAppExampleReviewWithErrorMessages

question, and such a person would usually ignore and remove the app, and possibly would leave a harsh review for it in the app store, and submit a low score for the app.

Lesson learned/recommendation: Just like any other mobile app, user reviews for contact-tracing apps range from short phrases such as "*Not working*", often not that useful nor insightful, to detailed objective reviews that could be useful for various stakeholders. Thus, if any stakeholder (e.g., the app's development team) wants to benefit in a qualitative way from the reviewers, they need to filter and analyze the "informative" reviews.

4.3. RQ3: Problems reported by users about the apps

As another "core" RQ of our study, we wanted to identify the main issues (problems) that users have reported about. Having received anywhere between 63 and 20,972 reviews (and counting) as of this writing for each app, the nine apps had in total 39,425 review comments. Of course, manual analysis of such a large and diverse textual feedback was not an option. The AppBot tool provides various features such as sentiment analysis (Guzman and Maalej, 2014) and critical reviews to make sense of large review text datasets. We show the outputs of word-cloud visualization for all the nine apps in Appendix A (Fig. 33). We also include the AppBot tool's user-interface in Appendix A, as a glimpse into how the tool works.

For generating word clouds based on reviews, AppBot provides six types of options to filter review subsets: interesting reviews, popular reviews, critical reviews, trending up reviews, trending down reviews, and new reviews. AppBot has a sophisticated NLP engine to tag views under those six categories, for example: "*The Popular tab shows you the 10 words that are most common in your reviews. This helps you to identify the most common themes in your app reviews*"[33]; and "*the Critical tab is a quick way to find scary stuff in your reviews. This can help isolate bugs and crashes, so you can quickly locate and fix problems in your app faster*"[2]. Since we are interested to know about the problems reported by users about the apps, to generate the word-cloud visualizations show in Appendix A, we have filtered by "critical" reviews.

For apps of non-English-speaking nations, e.g., Germany and France, unsurprisingly, almost all reviews were in their official languages, and we used the Chrome browser's built-in translate feature to see the review texts in English. For readers wondering about the original reviews in the original languages, we also show the word-clouds of two example apps (StopCovid France and Stopp-Corona Austria) based on their original review data.

Let us consider the *COVID Tracker Ireland* app as an example. As we can see in its word-cloud, "battery, "draining", and "uninstall" is among the most "critical" words. The fact that these apps make regular usage of Bluetooth signals leads to high battery usage, and this issue has been widely discussed in many online sources.[34] Furthermore, the terms "work" and "update" appear prominently with negative sentiment in the word-cloud of app reviews from Germany and Finland. By reading a subset of those reviews, we observed that it has not been obvious for many users of those apps how to use those apps properly (how to get them to "work").

In Appendix A, words in a word cloud are colored according to their sentiments in reviews. AppBot provides four types of sentiments: positive (green labels in the word-cloud), negative (red), neutral (grey), and mixed (orange). There is another useful feature in AppBot: when we click on each word in the cloud, all the reviews containing that word are listed.

Lots of insights can be gained from the word clouds, word sentiments, and also by live interaction with the dataset in the AppBot tool (we invite the interested readers to do so). As discussed in Section 3.4, we have posted an online video of live interaction with the dataset in youtu.be/qXZ_8ZTr8cc

Of course, comparing these data and findings for two different contact-tracing apps should be done with a "grain of salt", since their contexts (users' demographics, software features, and requirements) are quite different. For example, England's NHS COVID app has a feature to allow users to scan a QR code, as the government has asked shops to request shoppers to do so when entering shops. Many of the reviews for this app are about issues with that feature (see the word "code" in the word cloud), a feature that apparently does not exist in the other apps.

Among the word clouds, we can visually notice that the *Protect Scotland* app's word cloud shows an overall positive picture, with lots of green (=positive) sentiments. In the rest of the word clouds, red (negative) sentiments are the majority.

Lesson learned: Problems reported by users about each of the apps are quite different from one another. However, there are still some issues reported by users for most apps, e.g., high battery usage.

One word cloud in Appendix A (the second one from the top in Fig. 33) belongs to all data: when all the reviews of all apps are analyzed (n = 39,425 reviews). This word cloud shows that the reviews have a negative sentiment towards the functioning ("is working" in the word cloud) of the German app. Furthermore, there seem to be major issues with the Bluetooth handshake protocol.

In the next three sub-sections, we look at three examples apps (countries) and their specific problems, as reported in user reviews. We select the two apps with the highest number of reviews: the German app (27,744 reviews, combined for both OS apps) and the French app (2638 reviews). The case of several UK apps is also interesting since UK is a nation with four regions, for which three different apps have been developed: "StopCOVID NI" for Northern Ireland, "NHS COVID-19" app for England and Wales, and "Protect Scotland" for Scotland. We also analyze the case of the UK and its apps next.

4.3.1. Problems reported about the German app

As visualized in the word-cloud in Fig. 33, one of the frequent words with negative sentiments for this app is "*funktioniert*" (German), meaning "works" (in English), which has appeared in 5550 negative reviews.

As discussed in Section 4.3, there is a useful feature in AppBot: when we click on each word in the cloud, all the reviews containing that word are listed (as shown in Fig. 8). To ensure reproducibility of our analysis and for the interested reader, we show in Fig. 8 the steps for retrieving the "critical" (negative) reviews in which a certain keyword ("*funktioniert*" in this example) is mentioned, using the AppBot tool. As we can see in Fig. 8, the term "*funktioniert*" (German), = "works" (English), has appeared in 5550 reviews in the time window under study (April–September 2020).

We looked at a random subset of that large review set (5550 records) which contained the keyword "works". It turned out that most of the negative reviews with the keyword "works", were conveying the message that the app does not work and were

33 support.appbot.co/help-docs/using-words-page/

34 www.lancasterguardian.co.uk/health/coronavirus/nhs-test-and-trace-app-shouldnt-drain-your-battery-or-affect-your-privacy-2982165

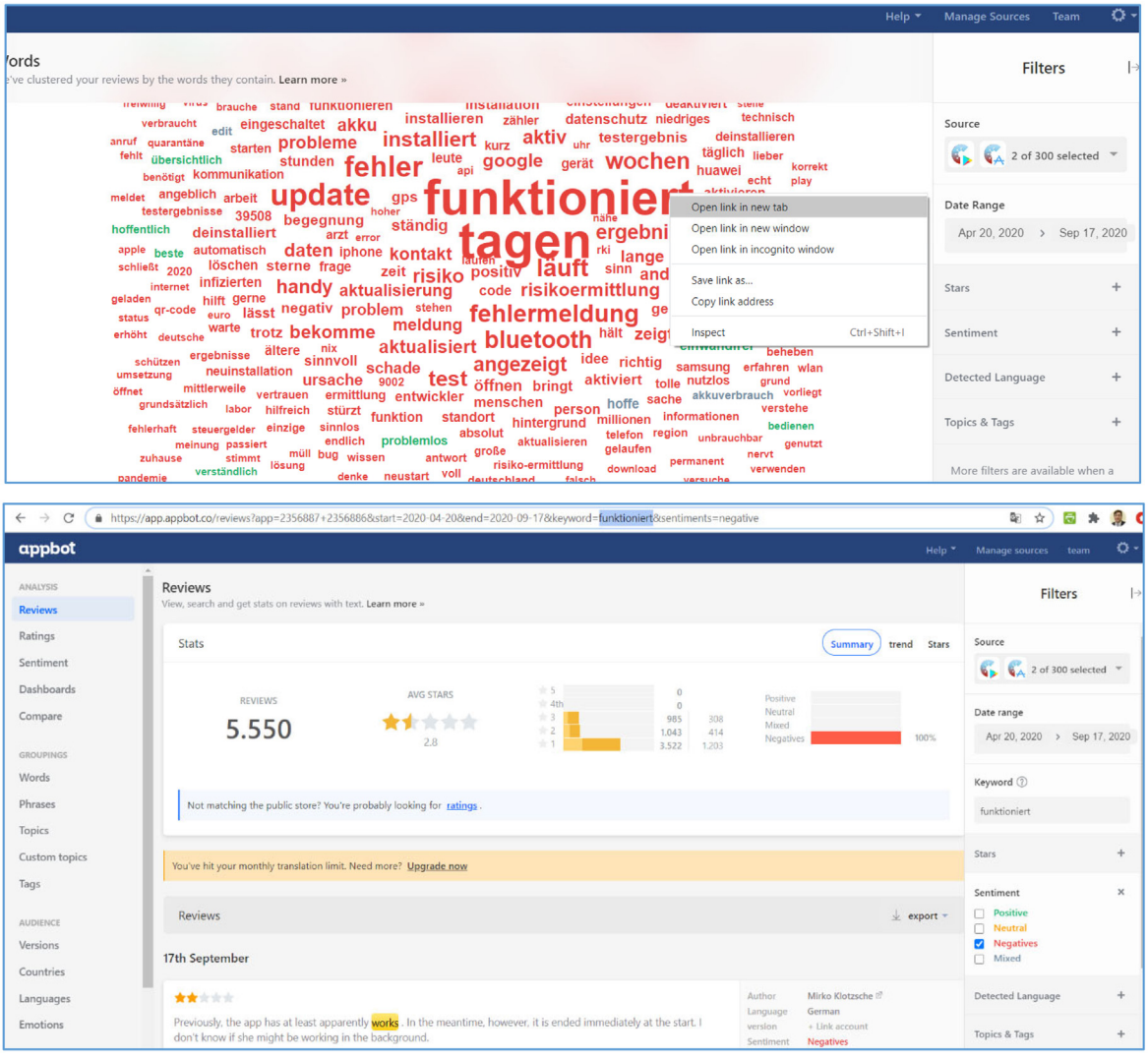


Fig. 8. Retrieving the "critical" (negative) reviews of the German app in which a certain keyword is mentioned, using the AppBot tool.

actually a sort of bug reports (two more examples are shown Figs. 9 and 10).

Lesson learned/recommendation: For the German app, a substantial number of reviews are about the app not working, which can be seen as bug reports. But unfortunately, since most users are non-technical people, informative and important components of a bug report (e.g., steps to reproduce the defect). However, fortunately, developers of the apps have more information attached to reviews, e.g., the device and the app version that the user has installed. Thus, in some cases, it would be hard, or even impossible, for the app's software engineering team to utilize those reviews as bug reports. A recommendation could be that in the app itself (e.g., in its "tutorial" screens), explicit messages are given to the users, asking them that, if they wish to submit bug reports as reviews, they should include important components of a bug report (e.g., steps to reproduce).

We also noticed that many of the reported issues were about the app not working on certain mobile devices. One example was as follows (translated from German by AppBot):

This example review, and many other reviews that we looked at, implied occurrence of sporadic (intermittent) app crashes for specific mobile device models. Such challenges are quite common in industry and have been studied in software engineering, e.g., in Joorabchi et al. (2013). We see above that the development team has replied to this review, mentioning that they will contact the user when they have more information, but there is no newer follow-up reply about the solution. It quite is possible that the development team has fixed some of those issues in the upcoming updated versions (patches).

Lesson learned/recommendation: A large number of cross-(mobile) device issues have been reported for the German and other apps too. This denotes inadequate cross-device testing of the apps, possibly due to the rush to release the apps to the public. Given the nature of the apps, and since the apps could be installed on any mobile device model/version by any citizen, the development and testing teams should have taken extra care in cross-device development and testing of the apps. There are many sources both in the academic literature (Husmann et al., 2016; Nebeling et al., 2015) and also grey literature[35] on this issue, which the development and testing teams can benefit from.

35 www.google.com/search?q=mobile+app+%22cross+device%22+testing

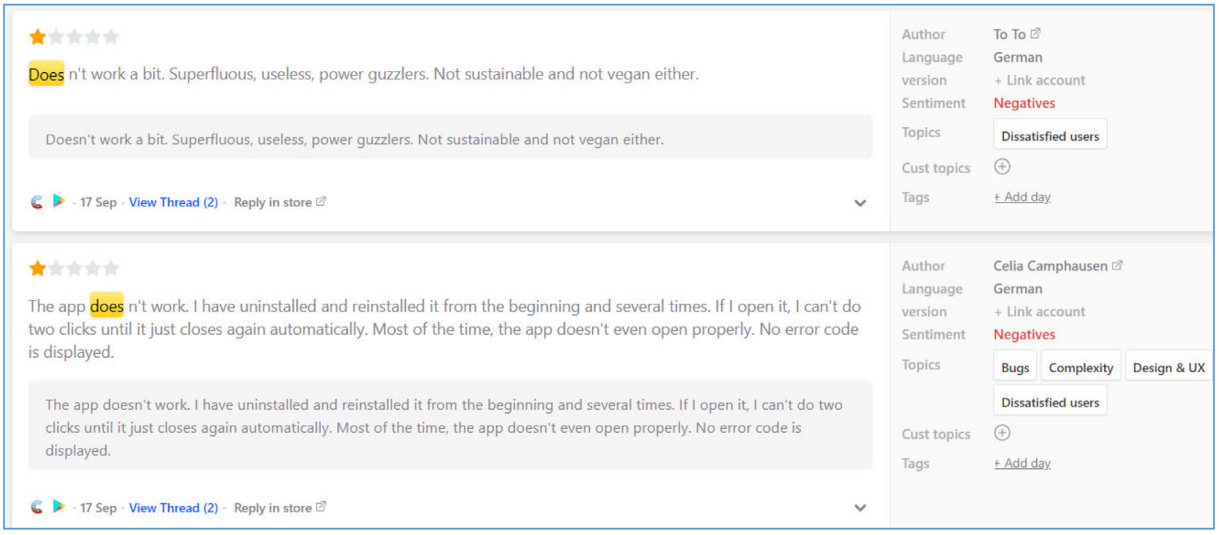


Fig. 9. Two critical (negative) reviews mentioning problems with running the German app.

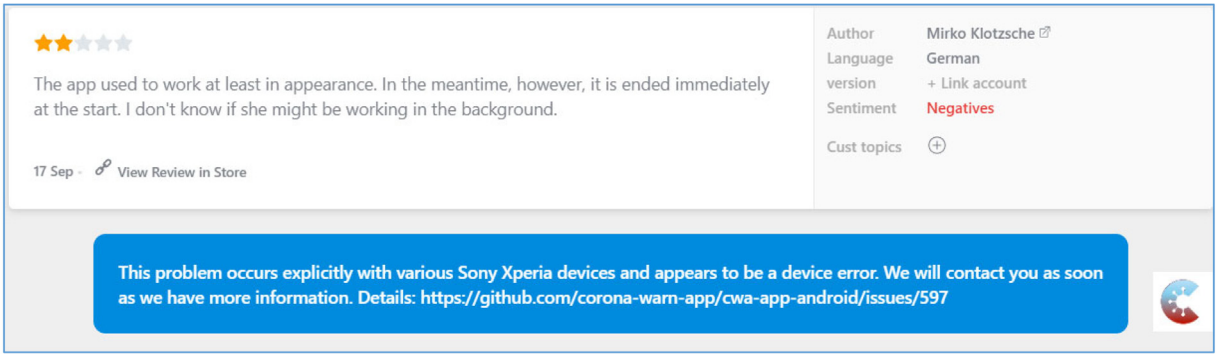


Fig. 10. A critical (negative) review mentioning sporadic (intermittent) app crashes.

As visualized in the word-cloud of Fig. 33, another frequent keyword within negative sentiments for the German app is "Tagen" (German), which has been translate to "meet" by AppBot (it uses Google Translate), but the correct translation should actually be "days" when we looked at the full sentences in the reviews. This keyword has appeared in 4465 negative reviews, and three examples are shown in Fig. 11. Two of the example reviews indicate that a specific functionality (called "Risk assessment") was not available over several days. The third example in Fig. 11 is a crash report.

Lesson learned/recommendation: We see that, for the German app, a specific functionality (called "Risk assessment") did not work for many users for several days. Such a malfunction usually gives a negative perception to users about an app, even if the other features of the app do work properly. It is thus logical to recommend that app developers should not include a feature in the app release if they predict or see from reviews that the feature does not work for certain users or on certain times/days.

As another issue type, 2273 negative reviews mentioned the keyword "Bluetooth". Two example reviews from that large set are shown in Fig. 12. Both these example reviews are bug reports, but again without important information (e.g., phone model/version and steps to reproduce the defect) to trace and fix the bug.

As another issue type, 1264 negative reviews mentioned the keyword "battery" ("Akku" in German). Two example reviews from that large set are shown in Fig. 13. Related to that issue, there have been a lot of discussions in the media (such as[36]) and also apps' support pages[37] about the high battery usage. Thus public (users) and media have complained about the issue. In response to this, the Android team has apparently made improvements[38] to the Apple-Google Exposure Notification API, i.e., *"In contrast to Classic Bluetooth, Bluetooth Low Energy (BLE) is designed to provide significantly lower power consumption"*.

Many users have reported the battery usage of the apps on their phones, e.g., a YouTube video[39] shows the battery usage screenshot of an iPhone on which Switzerland's SwissCovid is running. The video showed that, on a time period of 10 days, the app ("Exposure Logging" or "Exposure Notification" service on iPhone) had consumed only 4% of the battery (shown in Fig. 14). Also, many bug reports have been filed in the German app's GitHub repository about its high battery usage, e.g.[2], in which screenshots of battery usage have been submitted. We show two of those screenshots in Fig. 14 (they are in German, but the usage ratio is clearly understandable), along with a bug report in which the systematic bug report items have been provided, e.g., Describe the bug, Expected behavior, Steps to reproduce the issue.

36 www.hitc.com/en-gb/2020/09/25/nhs-covid-19-app-battery-usage-explained-will-the-app-drain-your-battery
37 github.com/corona-warn-app/cwa-app-ios/issues/671
38 developer.android.com/guide/topics/connectivity/bluetooth-le
39 www.youtube.com/watch?v=yEXiH4UIvSk

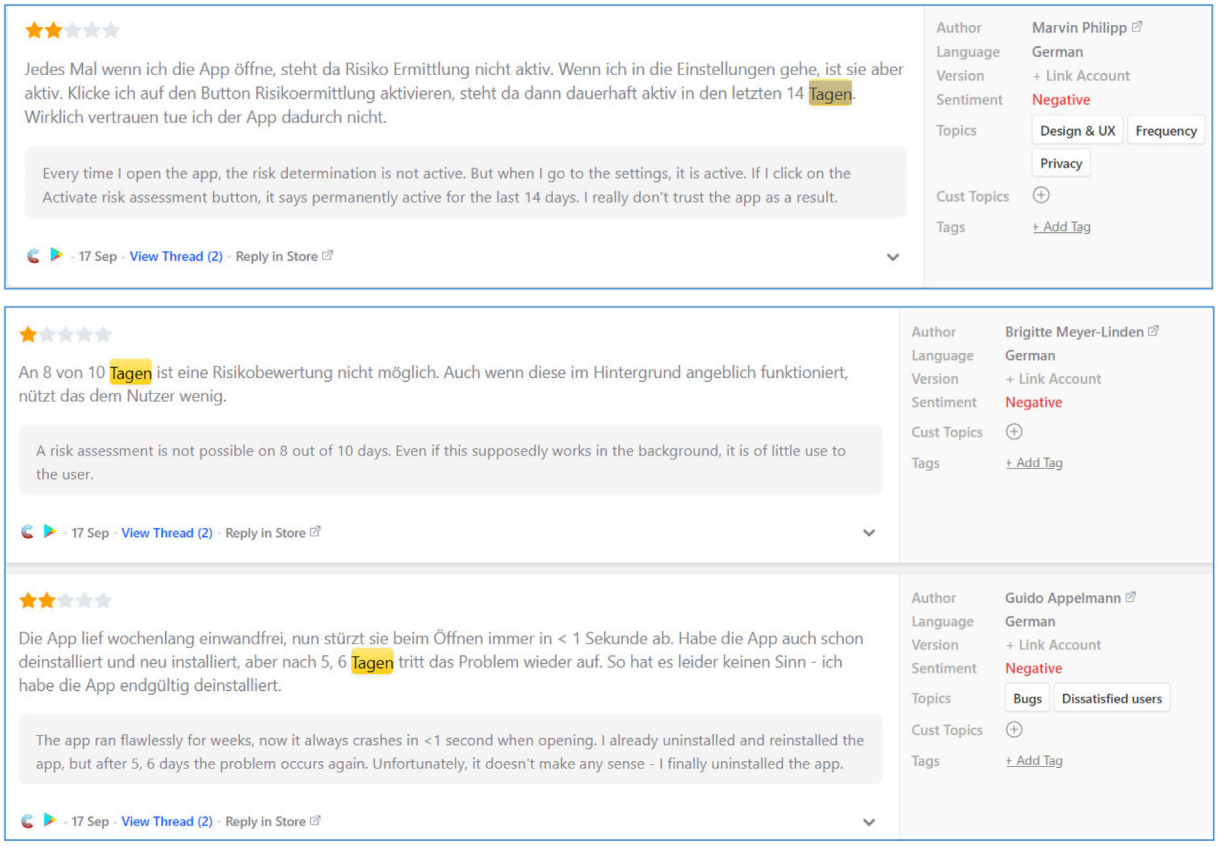


Fig. 11. Several critical (negative) reviews of the German app, mentioning problems with the keyword "days".

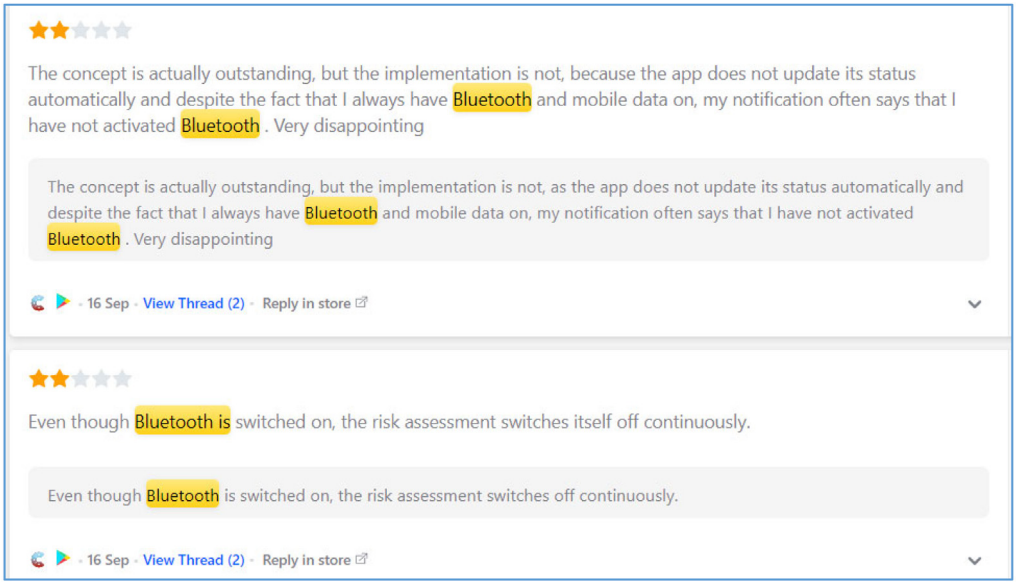


Fig. 12. Two critical (negative) reviews of the German app, mentioning problems with the keyword "Bluetooth".

The first author of the paper also installed the StopCOVID NI app on his iPhone, and let it run for more than a week non-stop. We provide a screenshot from the battery usage screen of his phone in Fig. 14, in which the "Exposure Notification" service has consumed only 4% of the battery, which we consider a reasonable power consumption (not high). But we should mention that he moved out of his home a few times only during that week, and he barely came close to anyone. Thus, the app did not have to exchange information with other uses who had the StopCOVID NI app on their phones.

We also found some discussions[40] in an online forum about the UK NHS app, in which one user mentioned: "*Our IT director reported massive battery drain after installing the app as in 70% to 15% on his journey home (three trains). I wonder if the drain comes*

40 se23.life/t/nhs-track-and-trace-mobile-app/15699/24

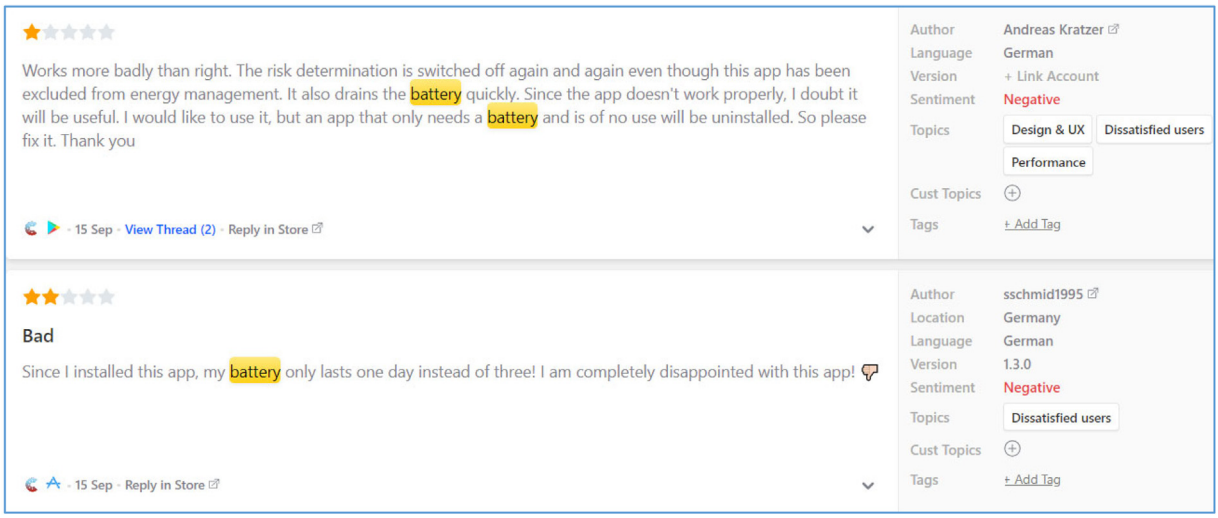


Fig. 13. Two critical (negative) reviews of the German app, mentioning problems with the keyword "battery".

from the amount of contacts you have with other people - if you are sat at home it has no contacts to ping but on trains etc. there might be hundreds". Thus, just like any other mobile app, we can observe that the type of usage and movement of the user in different environments could indeed impact the battery usage of the app.

4.3.2. Problems reported about the French app

As visualized in the word-cloud of Fig. 33, for the French app, two of the frequent words with negative sentiments for this app are "application" and "l'application", referring to "app", which are trivial terms. When translated to English, the other frequent terms are "activate", "install", and "Bluetooth". We discuss a small randomly chosen subset of those reviews next.

237 reviews critical reported problems with the keyword "activate" for the French app. Two example reviews from that set are shown in Fig. 15. We also include the permanent links to the reviews for traceability. These users have reported serious problems with activating the app, which is unfortunate.

Lesson learned/recommendation: We see rather trivial issues in the apps, i.e., users have to "activate" multiple times, instead of just once. We would have hoped that the test teams of the apps had detected and fixed those trivial issues before release.

130 reviews critically reported problems with the keyword "install" for the French app. Two example reviews from that set are shown in Fig. 16.

Lesson learned/recommendation: The first example review of Fig. 16 denotes issues w.r.t. internationalization (language settings). It is important that a given app automatically switches to the home country's language since some non-English users will feel odd if they see a sudden switch from their native language to English in the app's GUI.

Exactly 300 critical reviews reported problems with the keyword "Bluetooth" for the French app. Two example reviews from that set are shown in Fig. 17. The first example review is about high battery drainage of Bluetooth, like all other apps in the pool.

The second example review in Fig. 18 is about the incompatibility of the app on old phones. The second example review also raised an important issue: a large ratio of elderly are known to not have the latest smartphones or even how to install and use apps like these on their phones. In fact, a paper has been published on this very subject, entitled: "*COVID-19 contact tracing apps: the 'elderly paradox'*" (Rizzo, 2020).

Lesson learned/recommendation: High battery drainage of Bluetooth has also been reported for the French app.

4.3.3. Problems reported about the three apps in the UK

For the four regions of the UK, three apps have been developed: *NHS COVID-19* for England and Wales, *StopCOVID NI* for Northern Ireland, and *Protect Scotland* for Scotland. We review next a subset of the common problems reported for all three and then review a subset of issues reported for each of them.

Common problems reported for all three apps:

One major issue reported by users is the lack of "interoperability" between the apps, i.e., if a user from one region, using that region's app, visits another part of the UK, the app will not record the contact IDs in the new region and in case of entering a positive COVD result, the app will not notify those contacts. This issue has been reported in a large number of reviews, e.g.:

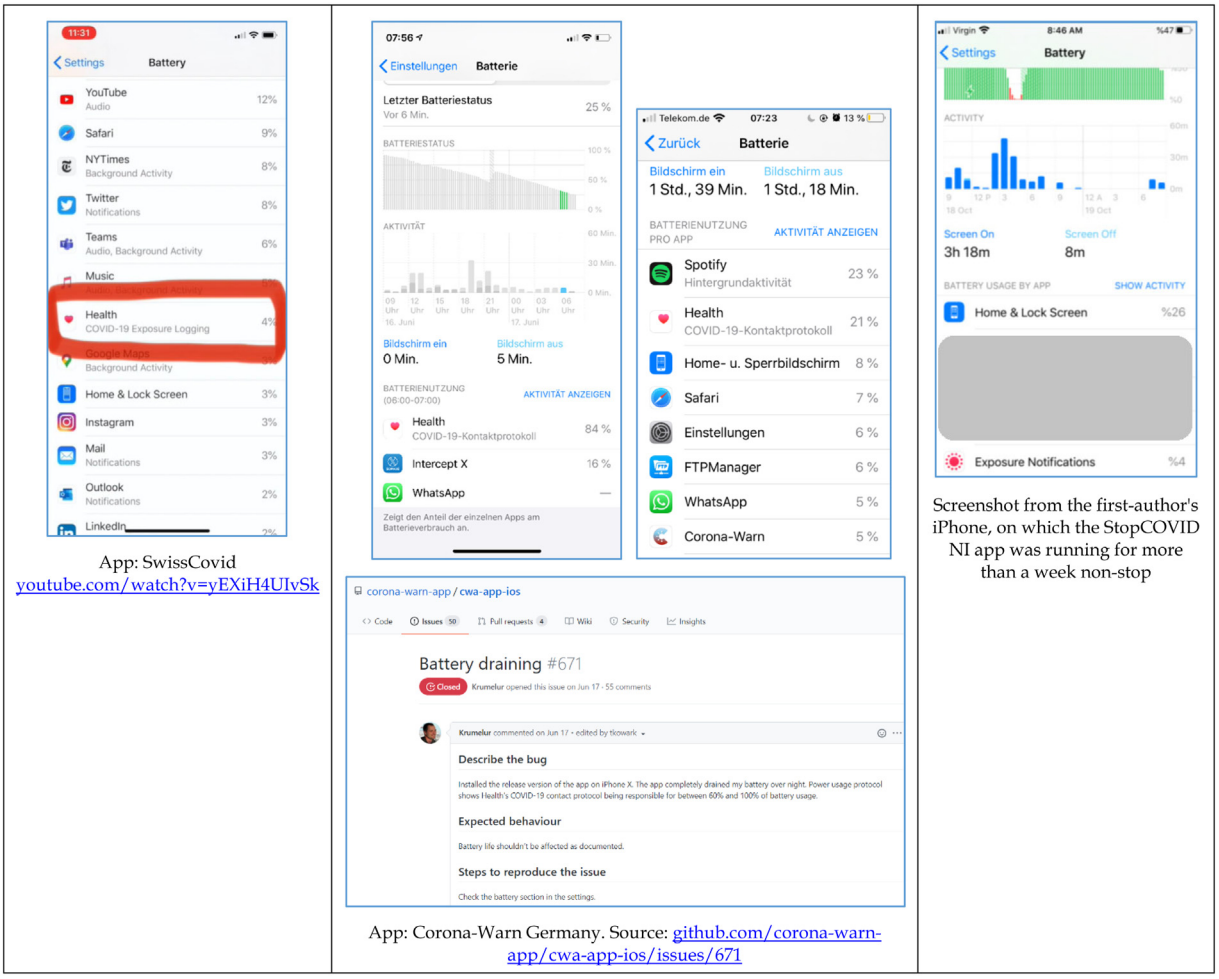
- *"Complete and utter waste of space. Only works if I come into contact with someone else using the same backstreet application, who has managed to get tested without being turned away, and inputs a code into their app. If I bump into someone from England, Wales, Ireland, or anywhere else for that matter with COVID-19 then this app does diddly squat - What's the point??"*[41]
- *"it's not linked to apps used in other parts on the UK, again a missing feature."*[42]
- *"Live in Scotland and work in England. Only one app will work at a time. Do I choose the NHS Covid or Protect Scotland version!!"*[43]

Also, a number of users, understandably, compared the features of the three apps and complained about the case of a given app not having the feature provided by another UK-based app. An example review:

41 appbot.co/apps/2437310-protect-scotland/reviews/1957885880/ or directly in Google Play store: play.google.com/store/apps/details?id=gov.scot.covidtracker&hl=en_GB&reviewId=gp%3AAOqpTOHVDLw1vCASIapQQmiyem1xyhXCw4SQBcOPdRXy0v1YPz95_hcZ5CUz7kWFe8004v5TbARTEjYyuEJoGw

42 appbot.co/apps/2437310-protect-scotland/reviews/1961498492

43 appbot.co/apps/2437310-protect-scotland/reviews/1960556725



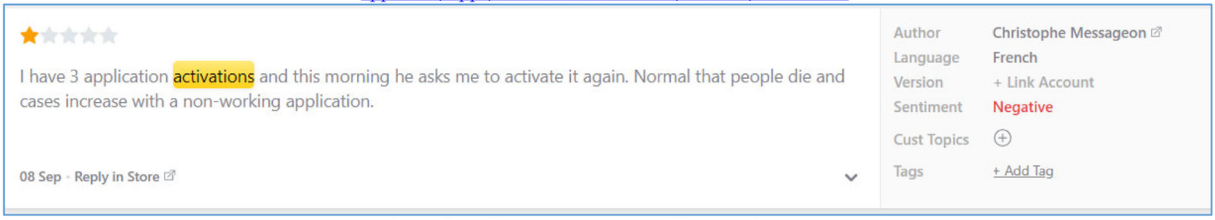
App: SwissCovid
youtube.com/watch?v=yEXiH4UIvSk

App: Corona-Warn Germany. Source: github.com/corona-warn-app/cwa-app-ios/issues/671

Screenshot from the first-author's iPhone, on which the StopCOVID NI app was running for more than a week non-stop

Fig. 14. Screenshots and one bug report submitted by users about battery usage of the apps.

appbot.co/apps/2370437-tousanticovid/reviews/1950620865



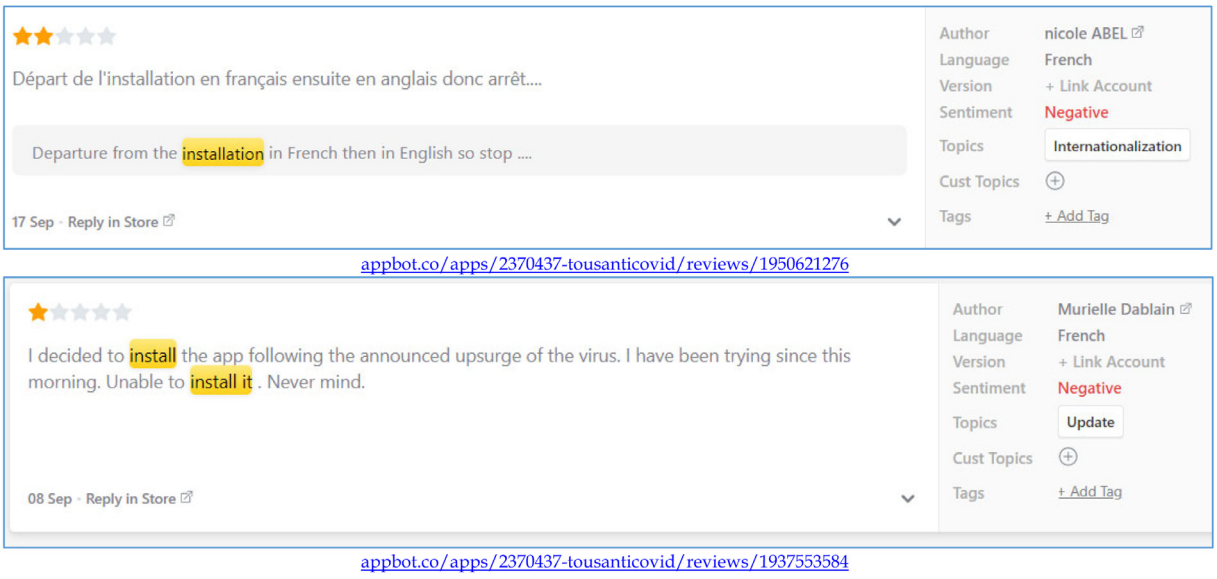
appbot.co/apps/2370437-tousanticovid/reviews/1937553799

Fig. 15. Two critical reviews of the French app, mentioning problems with the keyword "activate".

- "*Looks great, easy to use but oh how they missed out some useful features such as a [NHS] Covid-19 [app's] alert state notifier, scanning business QR Codes, etc. so user's data needn't be handed over in pubs, etc*".[44]

Many users reported having problems installing the apps, e.g., 10 reviews of the 573 *Protect Scotland Android* app. Some of those installation problems were due to having older phone models, but we still saw several reviews reporting newer (phone models not being able to install the apps, e.g., "*Waste of time have tried to install numerous times got the very latest Samsung S20 and it doesn't*

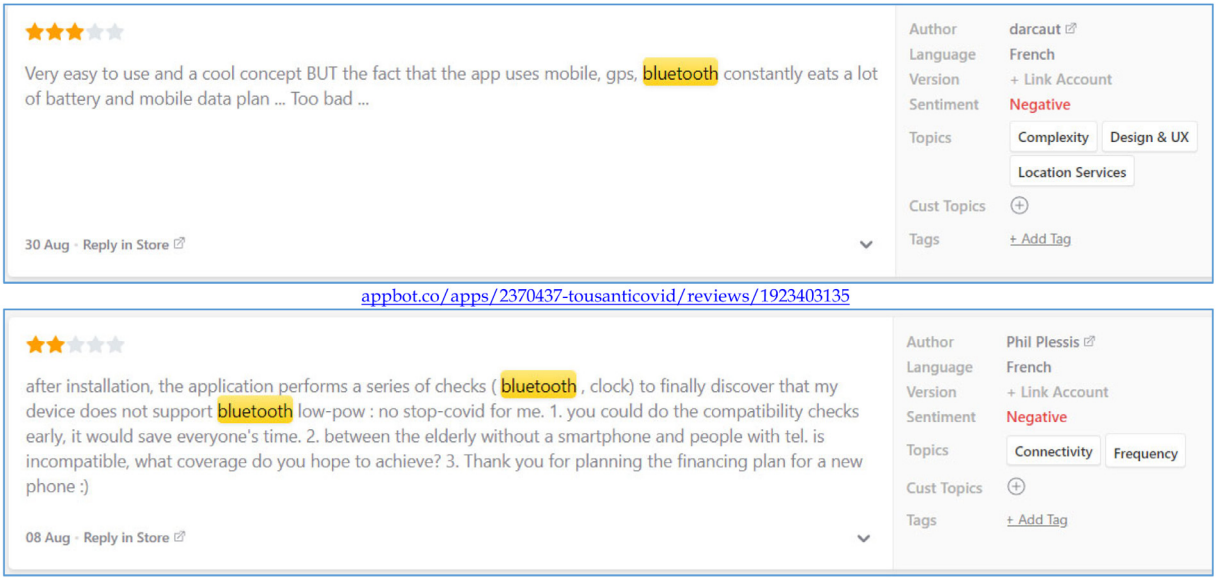
44 appbot.co/apps/2437310-protect-scotland/reviews/1961498492



appbot.co/apps/2370437-tousanticovid/reviews/1950621276

appbot.co/apps/2370437-tousanticovid/reviews/1937553584

Fig. 16. Two critical reviews of the French app, mentioning problems with the keyword "installation".



appbot.co/apps/2370437-tousanticovid/reviews/1923403135

appbot.co/apps/2370437-tousanticovid/reviews/1896028609

Fig. 17. Two critical reviews of the French app, mentioning problems with the keyword "Bluetooth".

install on my phone".[45] This raises the issue of the development team not doing adequate installation testing of the app using the latest phone models. There are indeed many advanced commercial testing tools on the market, e.g., Testinium (testinium.com), to conduct that testing efficiently. As discussed in Section 2.3, the first author of the current paper served as a consultant to the development team of the StopCOVIDNI app and conducted an inspection of test plans and test cases of the app. One of the comments that he had made was indeed installation testing of the app on multiple phone models using such test tools.

Another common issue that we noticed for the UK apps was the lack of response by apps' development teams to almost all reviews in the app store (only the Google Play store allows replies to reviews). This was in contrast to the case of some other apps, e.g., the German app, whose development team has been proactive in replying and communicating with users directly via the review threads.

Lesson learned/recommendations: The development team of all apps should be proactive in replying to user reviews, and filtering informative reviews, and getting more information (e.g., steps to reproduce the defects/problems) from them, e.g., by direct replies to the reviews in app stores.

StopCOVID NI app:

As visualized in the word-cloud in Fig. 33, one of the frequent words with negative sentiments for this app is "notifications" which has appeared in 12 negative reviews, e.g.:

45 appbot.co/apps/2437310-protect-scotland/reviews/1960556689

46 appbot.co/apps/2392818-stopcovid-ni/reviews/1950428670

47 appbot.co/apps/2356887-corona-warn-app/reviews/1950881368

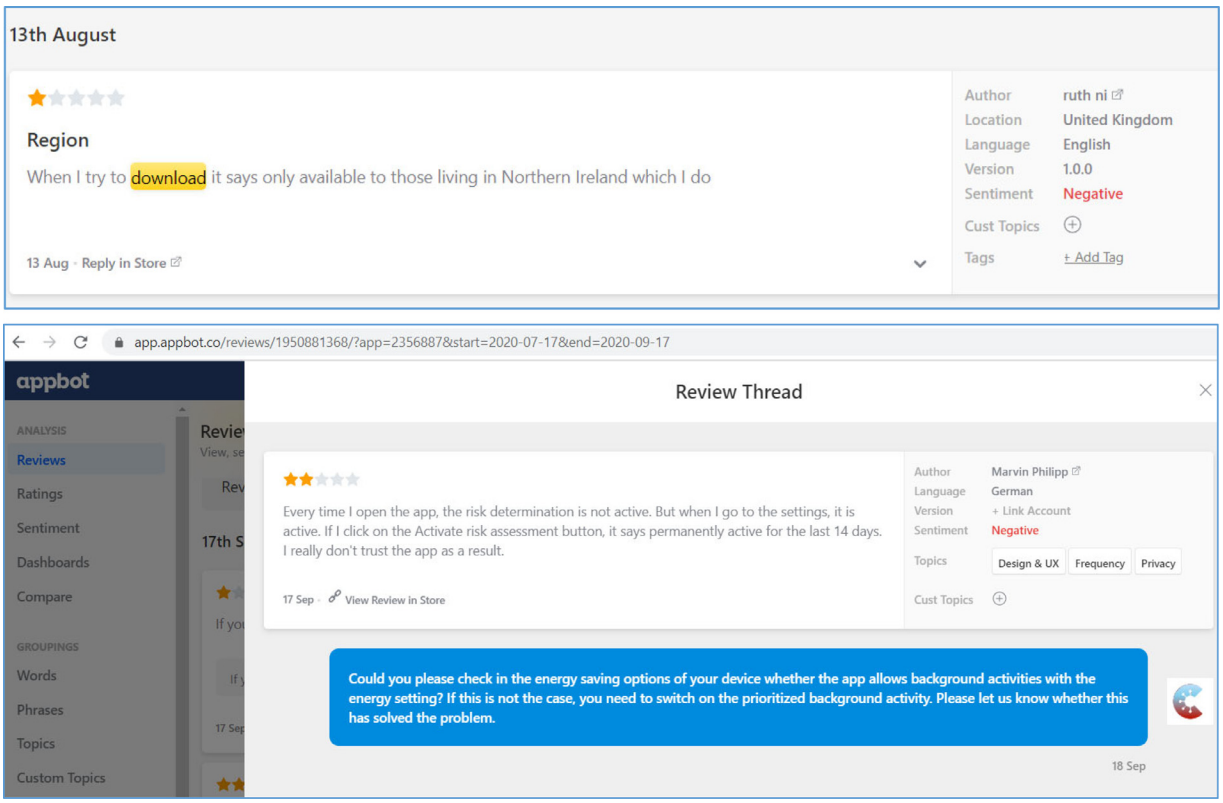


Fig. 18. Comparing a review[46] without any reply from the development team for the StopCOVID NI app and a review[47] with a reply from the development team for the German app.

- *"Want to get people to uninstall it? Don't produce audible notifications you haven't been exposed this week at 6am on a Fri morning, waking people up"*[48]
- *"I am getting a warning that exposure notifications may not work for the area I am in. As this is Northern Ireland I am unclear why it is saying this. The exposure log does not appear to have made any checks since early August. This does not give confidence that the app is working properly. I do hope the designers are reading these reviews as this appears to be a recurring issue"*[49]
- *"I was keen to install and safe. I have iphone7 with latest update. And like so many others, I get the exposure notification error. Can't select to turn on exposure notifications. Useless. Very disappointing"*[50]
- *"Hopefully the app does what it says on the tin - but I get another error message that says "Exposure Notifications Region Changed", followed by "COVID-19 Exposure Notifications may not be supported by "StopCOVID NI" in this region. You should confirm which app you are using in Settings". I am using an iPhone 11 Pro Max running iOS 13.5.1. so I have no confidence that the app is working properly at present".*[51]

Lesson learned/recommendations: There seem to be rather trivial usability issues with some of the apps (e.g., the case of exposure notification errors in the NI app). This raises the question of the inadequate usability testing of the apps and the possibility of releasing them in a "rush".

Another frequent word with negative sentiments for this app is "download", which appeared in 16 negative reviews, e.g.:

- *"When I try to download it says only available to those living in Northern Ireland which I do"*[52]
- *"The app just tells me I need to live in NI to use it. I do. I deleted and downloaded again. Same problem"*[53]
- *"I just downloaded this and used the 'share this app' function to all my contacts in N Ireland and the link doesn't work!! Not a good start for the app and doesn't build my confidence that any other part of the app works! I am now getting multiple messages from people asking what is the link for. Very disappointing"*[54]

Some randomly sampled negative reviews, under the category of the "download" issue, were:

- *"Tried to download on an elderly relative's Samsung phone but the app isn't compatible. Nowhere can I find a list of compatible devices or Android versions. Sadly the app won't help the most vulnerable"*[55]

48 appbot.co/apps/2392818-stopcovid-ni/reviews/1950428641
49 appbot.co/apps/2392818-stopcovid-ni/reviews/1950428665
50 appbot.co/apps/2392818-stopcovid-ni/reviews/1950428700
51 appbot.co/apps/2392818-stopcovid-ni/reviews/1950428741
52 appbot.co/apps/2392818-stopcovid-ni/reviews/1950428670
53 appbot.co/apps/2392818-stopcovid-ni/reviews/1950428675
54 appbot.co/apps/2392818-stopcovid-ni/reviews/1950428858
55 appbot.co/apps/2436851-stopcovid-ni/reviews/1950429102

- *"What is the point of urging people to install this app to stop the spread of covid 19 yet when the app is not working for some people the developers don't even bother to fix or reply to the email that they ask people to send if there is a problem. Google tried their best to resolve the matter immediately and also notified the developer yet 5 days past and nothing"*[56]
- *"The app does not seem to work correctly unless automatic battery optimisation is switched to manually allow app to run in background. Settings -> Battery -> App launch -> StopCOVID NI. Might also be under Applications -> StopCOVID NI -> Battery Optimisation depending on version. Once I switched this I went from 5 checks over 10 days to 8 checks in a single day"*[57]
- *"As others have said, the app does not properly run in the background as intended - the app needs to be open and the phone unlocked. Good idea in theory, however poor execution, going forward this app will be useless without correction."*[58]

NHS COVID-19 app:

As visualized in Fig. 33, one of the frequent words with negative sentiment for this app is "code" which has appeared in 153 of the 341 negative-sentiment reviews for this app. This phrase does not refer to source code, but to a QR code which is used in the app (see the real photo example of a restaurant with the QR code in Fig. 19). There were a great number of criticisms, and the followings are only some examples:

- *Well, as a business we are directed to register for track and trace. Having registered for a QR code and subsequently printed said code. I thought, in good naval tradition, 'Lets give it a test before we put the poster up'. So download the app from Play Store. Scanned the code and a message pops up 'There is no app that can use this code'. Next move, open the application. What do we find!! Currently only for NHS Volunteer Responders, Isle of White and Newham residents'. What is the point of publicizing this if it does not have basic functionality? Measure twice cut once MrX Also there should be an option for no Star as it appropriate for this application!*[59] → Poor alignment of publicity timing
- *QR location doesn't seem to work for me. Used a standard QR reader on my phone and it took me straight to venue but the QR reader in the app said QR code not recognized.*[60] → Poor testing of that module software
- *It is not working. My daughter has been working within 2 meters of someone all week who tested positive yesterday. She's had no notification and they both have the app. The other person has put the code in she received at 5pm yesterday and daughters had no alert.*[61] → Raises serious concerns about efficacy and quality of the software
- *After months in the planning, this app has been such a let down. At first I was receiving notifications, but when I clicked to open and read them, they disappeared. Last week I developed symptoms and updated the app. My husband has the app and, 6 days later he still hasn't received a notification that he's been near someone with symptoms. Yesterday I updated this app with my positive test result code, and still - nothing on my husband's phone. I have no confidence in this app at all!*[62] → Raise serious concerns about efficacy and quality of the software
- *"I move about a lot with my job. I can't change the post code to the area I'm in unless I uninstall the app"*[63] → The need for better software requirements engineering
- *Keeps asking for post code but doesn't give any where to put it in*[64] -> Poor usability and UX/UI design
- *I booked a test through the app. Test is negative. SMS and Email with negative result. No code in SMS or email. App still counting down isolation. I don't want to go out with phone and potentially give false positives to loads of people.*[65] -> Poor integration of the software in the healthcare (business) processes
- *This app requires Bluetooth and location to be on all the time, goodbye battery life. Some qr codes won't scan and no manual input option as a fall back. The alerts are useless as well, it tells you that someone in my area has been confirmed to have Covid, but doesn't tell you if you had been to a place they checked in, so is a pointless notification*[66] -> Poor quality

Confusion about the QR code aspect of the app has also been covered in many news articles, e.g.,[67] which stated that: "*Some residents [of Newham, UK] reported that the QR code throws up an error message in the app or simply takes too long to scan, causing queues to enter a shop — hardly ideal in these times of social distancing*". An app reviewer on Google Play said[68]: "*Although the app looks good, if I can't use the QR scanner, it defeats the object of the app's purpose*".

Decoding the software engineering issues from the above issues could give us some insights: *"QR code throws up an error message in the app"* denotes that not enough testing has been done on all possible QR codes; and *"simply takes too long to scan"* denotes that not enough performance testing has been done.

Lesson learned/recommendations: Some of the reviews provide insights on software engineering issues of the apps, e.g., not enough testing has been done on all possible types of QR codes, and not enough performance (load) testing has been done.

Protect Scotland app:

As visualized in the word-cloud in Fig. 33, one of the words with negative sentiments for this app is "people". By inspection of the comments by the AppBot, we saw 18 comments having this term, e.g.:

- *"Think about this... What's the point unless 100% of people have this app? I could be in a supermarket with 100 people. One person has Covid-19 in said Supermarket, but is the only one who does not have the app. That person inflects several people, but they won't know where they caught it - because that one person didn't have the app"*[69]: the user stresses the need for wide adoption of the app, which is a valid issue.

Another frequent term with negative sentiment was "Google", which was related to the confusion about the updates to the Google Play Services and Google Exposure Notification API. Most (layman/nonprofessional) users, understandably, could not figure out how to update and check their updates. Two example reviews were:

56 app.appbot.co/apps/2436851-stopcovid-ni/reviews/1950429005/
57 appbot.co/apps/2436851-stopcovid-ni/reviews/1950429052
58 appbot.co/apps/2436851-stopcovid-ni/reviews/1950429071
59 appbot.co/apps/2411517-nhs-covid-19/reviews/1951839432
60 appbot.co/apps/2411517-nhs-covid-19/reviews/1994520039
61 appbot.co/apps/2411517-nhs-covid-19/reviews/1993224335
62 appbot.co/apps/2411517-nhs-covid-19/reviews/1993228781
63 appbot.co/apps/2411517-nhs-covid-19/reviews/1993233064
64 appbot.co/apps/2411517-nhs-covid-19/reviews/1991681430
65 appbot.co/apps/2404650-nhs-covid-19/reviews/1993813691
66 appbot.co/apps/2411517-nhs-covid-19/reviews/1989859049
67 www.wired.co.uk/article/nhs-covid-app-trial-newham
68 app.appbot.co/apps/2411517-nhs-covid-19/reviews/1951842716/
69 appbot.co/apps/2437310-protect-scotland/reviews/1951729752

Fig. 19. A QR code to be scanned with the NHS COVID-19 app before entering a venue in London, UK.
Source: www.xinhuanet.com/english/2020-09/25/c_139396283_6.htm.

- "*Keeps telling me I need to update Google Play Services for the app to work - even though it is fully up to date. So app not usable*".[70]
- "*I can't get it to work, getting message 'google exposure notification api not available on this device' even though it's running Android 11*".[71]

Lesson learned/recommendation: Especially for Android phones, the update mechanism of the OS and its components (e.g., APIs) should be "seamless" (automatic) since we cannot expect all users to have the "technical" skills to do such tasks properly.

Just like for other apps, there were also multiple reviews about high battery usage and other issues related to when the phone's Bluetooth is on, e.g.:

- *"This app requires Bluetooth to be permanently on. A real battery killer, I also get bombarded by nearby devices that see my Bluetooth is on. Google location is accurate to around 1 m, why is this not enough? Uninstalled until something better comes up"*.[72]

4.4. RQ4: A glimpse into positive reviews and what users like about the apps

While the focus of our analysis so far has been mostly on negative comments, it is important to realize that the reviews of the apps are all not negative, as many positive reviews have also been reported. When looking, as we are, at multiple apps in different countries, the positive reviews can be very important to determine what has been done well on one app that could be generalized into others. One way to see a bird's-eye view of positive reviews is to look at review sentiment categories, as shown in Fig. 5-(b). The ratios of positive reviews, among all reviews of an app, based on sentiment analysis, range between 5% (for the StopCovid France-iOS app) to 56% (for the *Protect-Scotland* Android app).

Using the AppBot tool, it is possible to drill down into a specific app and gain an understanding of sentiment both over time and overall through an easy-to-use user interface (UI) as shown in Fig. 20 (the case of the *Protect-Scotland* Android app). While it is true that the overall sentiment for this app is highly positive, the time-based chart shows the bulk of reviews were made on the 10th and 11th of September immediately after the 10th September launch. Three days after launch, the number of reviews per day had fallen below 50 and, by day four, below 25, a pattern which has continued. Such a large number of positive reviews being made immediately post-release will skew any overall mean or median measurement of overall sentiment. It is also interesting to consider how accurate these reviews will be given the short period of use the users must have had before posting them in the app store, for example, how would battery usage or alert efficacy be measured in the 48-hours immediately post-launch? How does this fairly measure any future updates?

Lesson learned/recommendations: When considering review sentiments, time-boxing should be used to look at specific updates or recent opinions as well as all-time data to allow mitigation of any lasting effects of large numbers of reviews in a short period, such as launch.

The second pane in Fig. 20 shows the ability of AppBot to produce the textual data filtered as needed, shown is the Protect Scotland-Android App reviews filtered for positive sentiment only in the date range September 9th to September 17th 2020. The use of this feature allows the quick gathering of overall positive statements and can avoid the biases mentioned above with date filtering.

In the sections below, we sample a few positive reviews of several apps below and interpret the findings from those samples.

4.4.1. Sampling positive reviews of the Protect Scotland app

For *Protect Scotland* Android app, 319 of the 573 reviews (55.7%) had positive sentiments, according to the AppBot tool. For the *Protect Scotland* iOS app, 52 of the 110 reviews (47.3%) had positive sentiments.

70 appbot.co/apps/2437310-protect-scotland/reviews/1951721703
71 appbot.co/apps/2437310-protect-scotland/reviews/1951721414
72 appbot.co/apps/2437310-protect-scotland/reviews/1951722760

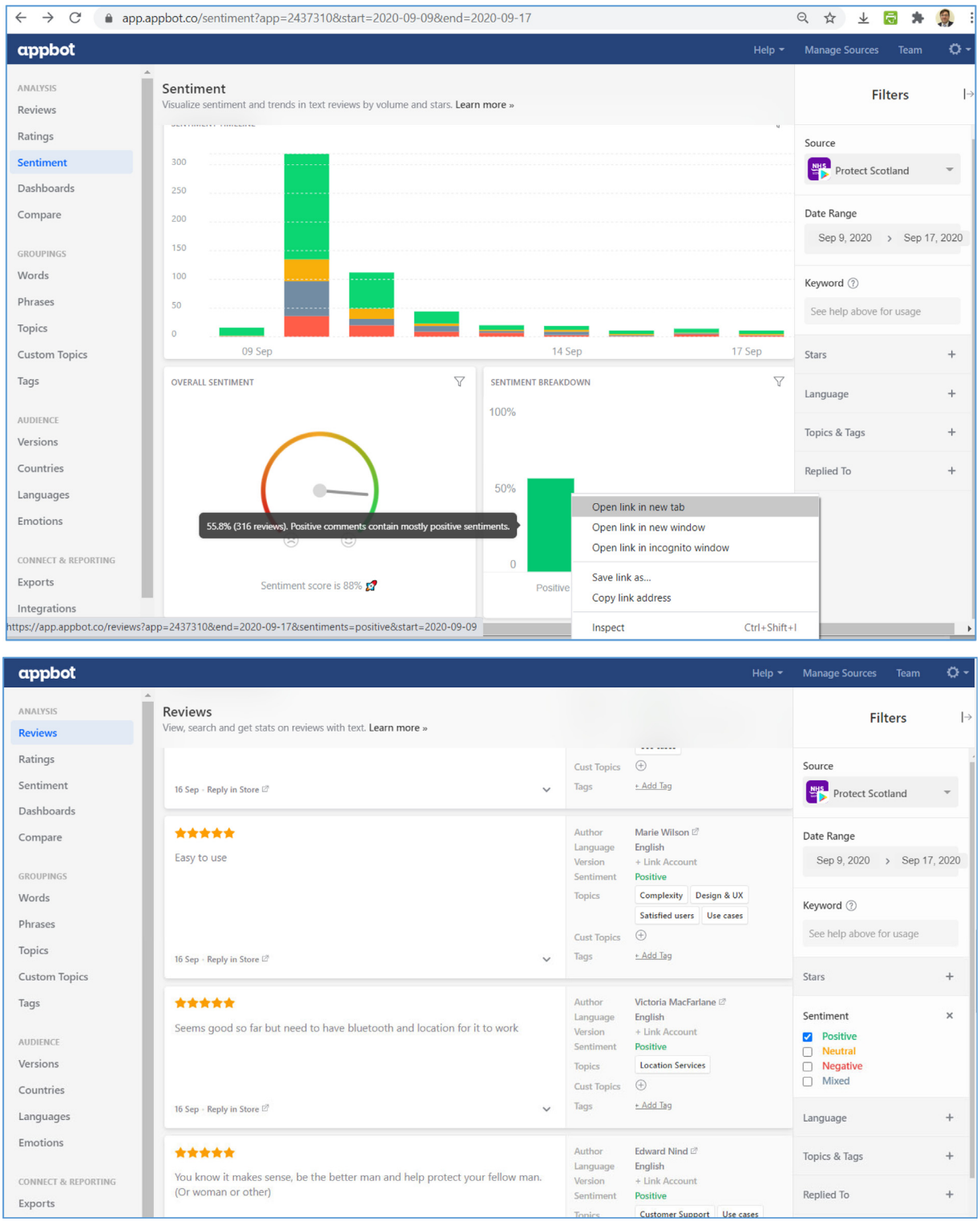


Fig. 20. A simple way to view the "positive" reviews of the *Protect-Scotland* Android app in the AppBot tool.

As expected, a large majority of reviews are entered by citizens (layman mobile phone users), in which users have expressed their non-technical feedback on the apps. We show a small randomly selected subset below. Each of the following items is exactly one review comment for the Protect Scotland-Android app. Note that, for brevity, we do not include the URLs for each of these reviews, but they can be found in our online dataset (linked from Section 3.4):

- *"Useful and anonymous*
- *Super easy and not needing internet/data just Bluetooth no details needed brilliant*
- *Very easy to set up, happy to play my part in hopefully returning to our normal way of life soon, if this app helps that in anyway then it's fine by me*
- *Simple and easy to use.*
- *Excellent privacy and will help Scotland. Easy to use.*

- *The more people that use this App the more effective it is. It is tracing and will help reduce spread and lockdown.*
- *Plain and simple, and uses the Apple/Google model. Well done Scotland.*
- *Easy to download and activate*
- *Very simple to download, just remember to keep your Bluetooth on when going out*
- *Seamless and invisible. Well designed app we all need.*
- *Super easy layout and quick to set up.*
- *I love that it is confidential and keeps you informed of any exposure.*
- *Just what is needed*
- *Great idea, hopefully everyone who can download it does soon!*
- *Easy to install*
- *Well done Scot.Gov*
- *Great app and can helpful to get info.. recommend to download this app*
- *Finally...*
- *Really good idea! Hopefully more people will download it to make it more worth while!*
- *Good idea*
- *This app is visually appealing, easy to use and gives peace of mind*
- *Very simple and transparent. Get it downloaded*
- *Impressed with the app and also the clear messaging about how it works and what it does.*
- *Fingers crossed I never have to use it.*
- *Really simple to use and the more people that use it the more effective it becomes.*
- *Very easy to set up and gives users genuine privacy. Just install it!*
- *Takes seconds to set up, simple and easy to understand.*
- *So easy to install*
- *Very easy to set up*
- *Easy to download and use*
- *Simple and easy to use. Impressed by the privacy and security controls. Encourage everyone to download it".*

As we can see in the above randomly sampled subset, some users have liked the good usability of the Protect Scotland app, while some others have commended the seemingly well-designed privacy and security controls of the app. Many users have been "optimistic" in their reviews by mentioning terms such "hope" and "hopefully" in their reviews, hoping that more people will use the app. The term "hopefully" appears three times in the small subset above, and the string "hop" appeared in 45 of 573 reviews (7.8%) for the Protect Scotland-Android app.

We found that almost all positive comments were not "technical", but rather were in "surface" level, i.e., users just praised the ease of installation and use of the app. But given the complex nature and function of the app (sending notifications to recent contacts in the case of COVID test being positive), very few users could actually see whether the app is really doing what it is supposed to; as discussed in the quality, efficacy and accuracy aspects of these apps (Section 2.4). To clarify our point, let us take for instance a regular (non-health) mobile app such as a typical "Notes" mobile app. It is much easier for a user (layman) to figure out if such an app is working or not, e.g., by creating some example notes (to do's) and then checking in another day in the app, if those notes are showing up properly.

Although a contact-tracing app can be easy to install and easy to use, and even have a nice UI, there is no guarantee for its "core" feature to work properly. This reminds us of the French news article[73] that we reviewed in Section 1, as it had reported that, as of mid-August 2020, "*StopCovid [the French app] had over 2.3 million downloads [of a population of 67 million people] and only 72 notifications were sent [by the app]*". As we can imagine and also as reported in the grey literature[74]: "*A faulty proximity tracing app could lead to false positives, false negatives*", which stresses the fact that these apps are safety-critical. Some reviews actually made explicit notes about this very important issue: "*Obviously I can't know how well the tracing works, but everything else about the app is great. It clearly explains all you need to know about how it works (especially in relation to privacy) and the interface itself is simple and effective. A very well made app*".

Returning back to the type of positive (and mostly optimistic) statements mentioned in positive reviews, some reviews were mentioning using this app to save lives, e.g., "*We've all got to do this guys! Save your granny's life!*". Some other positive reviews of the Scottish app praised its possible integrability with other European apps: "*Glad it's up and running using the template adopted by other European nations, allowing potential international use in the future*".

4.4.2. Sampling positive reviews of the StopCovid France app

For *StopCovid France* Android app, 424 of the 2397 reviews (17.7%) had positive sentiments, according to the AppBot tool. For *StopCovid France* iOS app, 13 of the 241 reviews (5.4%) had positive sentiments. As discussed above, the ratio of positive reviews, among all reviews of the French app, based on sentiment analysis, was one of the lowest among all the nine apps in the dataset.

Note that almost all reviews for the French app were in French, and thus we used the Google Translate tool to use them in this paper. Some of the positive reviews in the French reviews dataset were insightful, e.g., "*Assuming it's working it's a great app, you don't have to do anything. Just have it running. However some way to know its working would be nice. Even just a counter of how many other apps it's discovered in that day, just to know its doing something*". We can treat this review as a feature request by users, a topic that has been studied in focus in research papers (Iacob and Harrison, 2013; Maalej and Nabil, 2015). The review also refers to the unclear efficacy of the app: "*Assuming it's working*". Many other users also expressed their uncertainty on whether the app actually does its core feature, e.g., "*Apparently it works. And it's transparent. Nothing to say ... It would be good from an epidemiological point of view if it was used more as an app!*", which also referred to the important epidemiological aspects (Guttal et al., 2020) of the app and the need for its wide usage.

Another feature request was: "*Works very well. An improvement would be to know the number of phones using the cross app in the week, month, day. It will help everyone to perceive the usefulness of their action by installing this app*".

Similar to the other apps in the study, some users reported in their review their satisfaction with how the app is preserving their privacy: "*Perfect compliance with the GDPR [EU's General Data Protection Regulation] so nothing to fear. Let's all activate this app to help fight the virus and prevent our businesses from being closed!*". Some other comments were happy of how usage of this app could have made their lies easier in lockdowns, e.g., "*Thanks to this application, I no longer have to close high schools*" (Note that we did not find any online news article linking usage of contact tracing apps to ease of school closure in France).

Another positive and detailed comment that we came through was: "*The choice of activating or not is free and Bluetooth is activated at the same time. I find that the battery holds up very well. Data protection is fully explained. The only problem is that there*

73 www.lefigaro.fr/secteur/high-tech/stopcovid-2-3-millions-de-telechargements-et-seulement-72-notifications-envoyees-20200819

74 www.eff.org/deeplinks/2020/04/challenge-proximity-apps-covid-19-contact-tracing

aren't enough people using it. It should be publicized, and then it to be really effective", which again commented on the wider public usage epidemiological aspects (Guttal et al., 2020).

Furthermore, there were reviews which were somewhat thank-you feedbacks to the development team, e.g., "*Thanks to the developers for removing the permanent notification!*", which referred to an apparently annoying notification mechanism which was in the app in its previous versions, and apparently has been fixed based on earlier user feedback.

Another review reported an interesting feature of the app: "*Good application example when you are at the bakery, there is a message that says you have to put on a mask and stay 1 meter [apart]*".

Many bug reports were informally mentioned in some of the reviews: "*Like a lot of people, I put my phone in airplane mode at night, but you have to reactivate the app every day. It's annoying and above all we forget ... Otherwise the application is very useful*".

4.4.3. Sampling positive reviews of the StopCOVID NI app

For the *StopCOVID NI* Android app, 46 of the 195 reviews (23.5%) had positive sentiments, according to the AppBot tool. For the *StopCOVID NI* iOS app, 19 of the 97 reviews (19.6%) had positive sentiments. We discuss a few insightful examples of those comments below.

- Positive feedback about the app and its features:
  - *"the app seems to be working well in the background and notifies me on a weekly basis for the searches it has done."*
- Some reviews expressed that the app is not easy to be found in app stores:
  - "*Excellent App, giving it 4 stars as it was difficult to find on play store, ended up looking it up on Google*"
  - *"Very easy to setup, but hard to find in app store".*
- Some feature requests:
  - *"Great app, glad it's out now. Would like to be able to find out just how many have downloaded it though".*
  - *"… only fault I find is it would be good if you could make it so you can put your post code in and find out how many is in your area that would be a big plus"*
  - *"App for southern Ireland gives a few overall general statistics, this gives no information at all".*

Lesson learned/recommendation: The apps must be clearly identifiable and searchable in app stores to maximize the number of users downloading it.

Lesson learned/recommendation: Where possible, some feedback (such as statistics about COVID cases in the region and also the number of close-by phone IDs recorded in the past) should be provided as a feature of the app, to encourage users that the app is working to emphasize the pro-social and individual benefit it is having.

4.5. RQ5: Feature requests submitted by users

Each submitted review can have different messages in it, e.g., error report, feature requests, user just mentioning her/his satisfaction or dissatisfaction with the app. If a review's text contains some form of suggestions for new features, it can be used as a 'feature request' by the app's development team in further improving it. Mining feature requests from app reviews and also from Twitter data have been recognized as a form of "requirements engineering" (elicitation) in the software engineering community, as several papers have been published in this area, e.g., Iacob and Harrison (2013), Jha and Mahmoud (2019, 2017), Williams and Mahmoud (2017), Guzman et al. (2017), Lu and Liang (2017), Maalej et al. (2019) and Nayebi et al. (2017). The topic has also been referred to as "crowd-based" requirements engineering (Groen et al., 2015).

Given the large number of reviews in general, and also in our study (our dataset has 39,425 records), however, pinpointing and extracting only the reviews which contain 'feature requests' manually is not an option. As discussed in Section 3.3, AppBot has a useful feature of to filter reviews to show only those with "Feature requests" submitted by users for an app. We show an example of using this feature for the case of the German *Corona-Warn* app in Fig. 21. This filtering feature is listed under a "Topic and tags" dropdown list, in which the items have been generated by applying the "Topic modeling" technique (Wallach, 2006), which is an NLP technique, by the AppBot tool.

As two representative examples from our set of nine apps, we look at feature requests for the cases of the German *Corona-Warn* app and the *COVID Tracker Ireland* app, next.

4.5.1. The case of German Corona-Warn app

The suggestion for a new feature, in the example review shown in Fig. 21, is the following phrase: "*... it would be nice if I could add all my [COVID] tests [in the app] and have a test history*" (even the original English text is highlighted in yellow by the AppBot tool for easy finding). The app's development is quite active and actively replies to almost all comments when needed, e.g., see the "*View Thread (2)*" link just below the example review shown in Fig. 21. In this case, the app's development replied by saying that (English translation from German): "*We have the wish-list repository on Github, in which comparable suggestions for expanding the app have already been created, see* github.com/corona-warn-app/cwa-wishlist".

Once we had the 781 feature-request-reporting reviews for this app, we were interested in grouping them to actually see a refined list of suggested features by users. Our chosen tool (AppBot) did not have such a feature. In fact, we saw the need for some form of thematic analysis (or qualitative coding) to group and combine that subset of reviews. We looked into the literature but did not see a ready-to-use technique for this purpose, and of course, we did not have the time resources to do "manual" qualitative coding of the reviews. We thus raise the need for such a technique and analysis to future works.

Nevertheless, we think it is worthwhile to mention in the following several insightful feature requests given by users for the German app. For full traceability, we also provide the "permanent" links to each review in the AppBot's database:

- *Would be great if you can see how many people you exchanged the keys with.* appbot.co/apps/2356887-corona-warn-app/reviews/1900405383
- *I think the app is generally well done and as a computer scientist, I would also like to praise the public interaction (including OpenSource)! The only thing I would still like would be if I could see how many relevant encounters have taken place in the last 14 days. So it might be a little more transparent that the app is actually doing something in the background.* appbot.co/apps/2356887-corona-warn-app/reviews/1896344794
- *It would be great if you could get more information such as infection numbers and spread at district level.* appbot.co/apps/2356887-corona-warn-app/reviews/1822553545

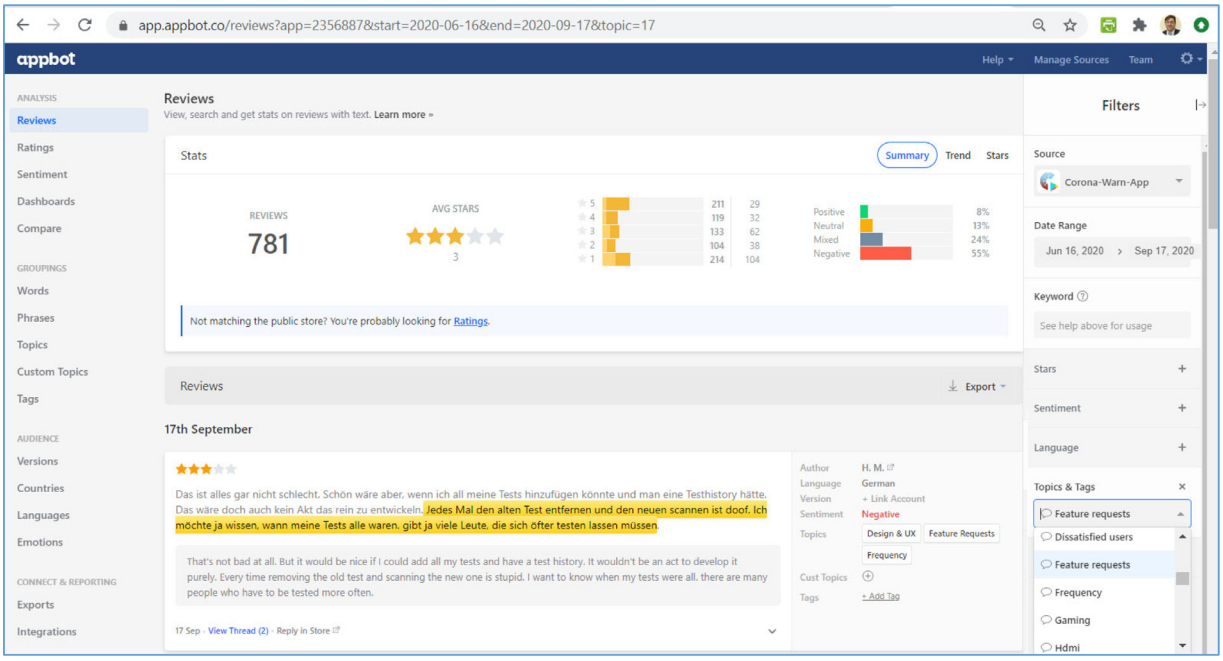


Fig. 21. The useful feature of AppBot to filter reviews to only see the "Feature requests" submitted by users for an app (example in this screenshot: the German app).

- *Improvements are always possible, e.g., how many encounters there were with other app users and the like.* appbot.co/apps/2356887-corona-warn-app/reviews/1818108607
- *More info would be nice. The risk assessment is all well and good, but it would be better if you could see how many other app users you had contact with* appbot.co/apps/2356887-corona-warn-app/reviews/1809389199
- *It would be great if this app were also available for tablets. We are always on the road with an iPad on business. This way, a possible infection could also be tracked after meetings.* appbot.co/apps/2356887-corona-warn-app/reviews/1805884538
- *This is a great app, but the need for internet is annoying because I don't have a lot of mobile data and it would be better if you could use the app without internet.* appbot.co/apps/2356887-corona-warn-app/reviews/1804643241
- *Kudos to the activists and developers who made it that way. Every little helps in the fight against Covid19. Download it and activate the anonymised decentralised logging of physical proximity to other users of this app! One suggestion: the nerd in me wants to see a list of the beacons detected on my device.* appbot.co/apps/2356887-corona-warn-app/reviews/1803093210
- *Good, but I would have liked optional GPS tracking. If, for example, Corona breaks out on the train and you are not directly next to a person, you can still be informed based on the location if you want* appbot.co/apps/2356887-corona-warn-app/reviews/1798228592
- *Works great, but it would be awesome if I could check how many people I've met that also had the app installed (not only those who are/were infected)* appbot.co/apps/2356887-corona-warn-app/reviews/1798010216

Even with this small sample set of reviews (above), we can see that the new feature of seeing how many encounters there were with other app users, is a common feature request as mentioned by users, at least for the German app.

On the other hand, we were curious to see the precision of the tool in identifying reviews with feature requests, as we observed several reviews being incorrectly classified as feature requests. To do this, we looked at a random subset (more than 50) of those 781 reviews, and manually identified whether they did not have obvious feature requests in them. We list several of those cases below, along with our hypothesis of why we think the AppBot tool has incorrectly classified them as feature requests. We should note that this observation raises questions on the precision of the AppBot's feature to filter reviews by those containing feature requests.

1. *The app has been telling me in the weekly overview for several weeks that I had 223 checks within the last 14 days. How can it be that I am presented with the same number 223 week after week? I am out and about every day so there must be fluctuations. So something doesn't work there and that makes the whole thing very questionable.* appbot.co/apps/2356887-corona-warn-app/reviews/1946942728
2. *The app has only caused one problem for me so far: I have never removed my test result, now it has probably been removed from the databases and the app tries forever to retrieve it, but I cannot add a new test. Please fix.* appbot.co/apps/2356887-corona-warn-app/reviews/1941257952
3. *Would be even better if more people participated!* appbot.co/apps/2356887-corona-warn-app/reviews/1859275190
4. *Uncomplicated and (for me) calming. Since I use public transport a lot for work, I would like to be warned, as well as others, if you had contact* appbot.co/apps/2356887-corona-warn-app/reviews/1799464996

Reviews #1 and #2 above should have been classified as bug reports. Review #2 even has clear words in it to hint to the classifier tool that it is a bug report, e.g., "*Please fix*" and "*The app has only caused one problem for me*". One possible reason on why we think AppBot has incorrectly classified review #3 as a feature request is the phrase "*Would be even better*" in it. For review #4, it could be the phrase "*I would like to*".

Lesson learned/recommendation: A variety of insightful feature requests have been provided by users, e.g., by user of the German app: How many encounters there were with other app users (how many people you exchanged the keys with); infection numbers and spread at district level; can the app be used without internet? As a form of "iterative" requirements engineering" (elicitation) (Iacob and Harrison, 2013; Jha and Mahmoud, 2019, 2017; Williams and Mahmoud, 2017; Guzman et al., 2017; Lu and Liang, 2017; Maalej et al., 2019; Nayebi et al., 2017) or "crowd-based" requirements engineering (Groen et al., 2015), the app's software engineering teams are encouraged to review those feature requests and select a subset to be implemented.

Lesson learned/recommendation: While AppBot's feature to filter reviews to see feature requests only is a useful feature, we found many example reviews which AppBot incorrectly classified as feature requests. We realize that an NLP/AI-based algorithm has been used to do that classification, and such an algorithm will have limited precision, but still, there is a need to improve such algorithms by developers (vendors) of App review analytics tools, such as AppBot.

4.5.2. The case of COVID tracker Ireland app

Using the "Topic modeling" feature (Wallach, 2006) of AppBot (just like what we did in Fig. 21 for the German app), we used the same tool feature to review and analyze the feature requests submitted by users for the *COVID Tracker Ireland* app, as shown in Fig. 22.

There are various feature requests (inside reviews), such as the first one in Fig. 22: "*Would be better if we just saw trends instead of daily figures*". The app's software engineering teams could review those comments and consider implementing those features if there are good reasons to do so. For this particular comment, the team does not have to "replace" the old feature ("*daily figures*" [of COVID cases]) with the new feature ("*trends*") but could add the new feature, and users can choose to use them alternatively as options, inside the app UI.

The last review shown in Fig. 22 implies that many users find this particular app not that useful, and such reasons lead them to uninstall the app. The user also rightly mentions that the app "*gives you information you can get anywhere*", e.g., media, news sites, etc.

Lesson learned/recommendation: Many users have cast doubts on the usefulness of the apps, i.e., they do not provide most of the "right" and much-needed features that many users are looking for. Thus, using "crowd-based" requirements engineering (Groen et al., 2015) techniques for these apps are critically needed.

4.5.3. Comparing the ratios of app reviews which are feature requests

It would be interesting to look into the ratios of app reviews, which are feature requests, and compare the ratios of all the nine apps under study. In the literature, there have been studies on such ratios, e.g., Pagano and Maalej (2013) showed that only about 7% of app reviews are feature requests. As we can see in Fig. 21, when we filter all the reviews in the time window under search (from the app's first release until Sept. 17, 2020), 781 of all 20,972 reviews (or 3.7%) have been identified as those containing suggestions for new features. We calculated these metrics for both OS versions of all nine apps, and data are shown in Table 4. The ratios are between 2.1% to 10.2%, with an average of 5.1%, which is not too far from the 7% ratio as reported by Pagano and Maalej (2013).

4.6. RQ6: Comparing the reviews of Android versus the iOS versions of the apps: Similarities and differences

Given the wealth of the information in the dataset, we found it a good opportunity and area of interest to compare the reviews of Android versus the iOS versions of the apps, and to observe the similarities and differences (if any). For the sake of space in this paper, we selected three of the most insightful aspects, which we report next:

- Popularity of each OS app version as measured by the average number of 'stars'
- Sentiment of reviews for the Android versions of the apps versus their iOS versions
- Problems reported for each OS version

4.6.1. Popularity of each OS app version as measured by the average number of 'stars'

We take the average number of 'stars' (as reported by app stores) for each of the two OS app versions of all the nine apps under study, and visualize the values as a scatter (XY) plot in Fig. 23.

While a clear correlation can be seen here, most apps do show a slight disparity between Android and iOS ratings, as seen more broadly with other apps in the literature (Guzman and Maalej, 2014), though in no cases is this more than 0.5 stars. For most apps, star ratings of the Android version are slightly higher than the iOS version, except the England & Wales and the Republic of Ireland apps, in which star ratings are slightly higher on iOS than on Android.

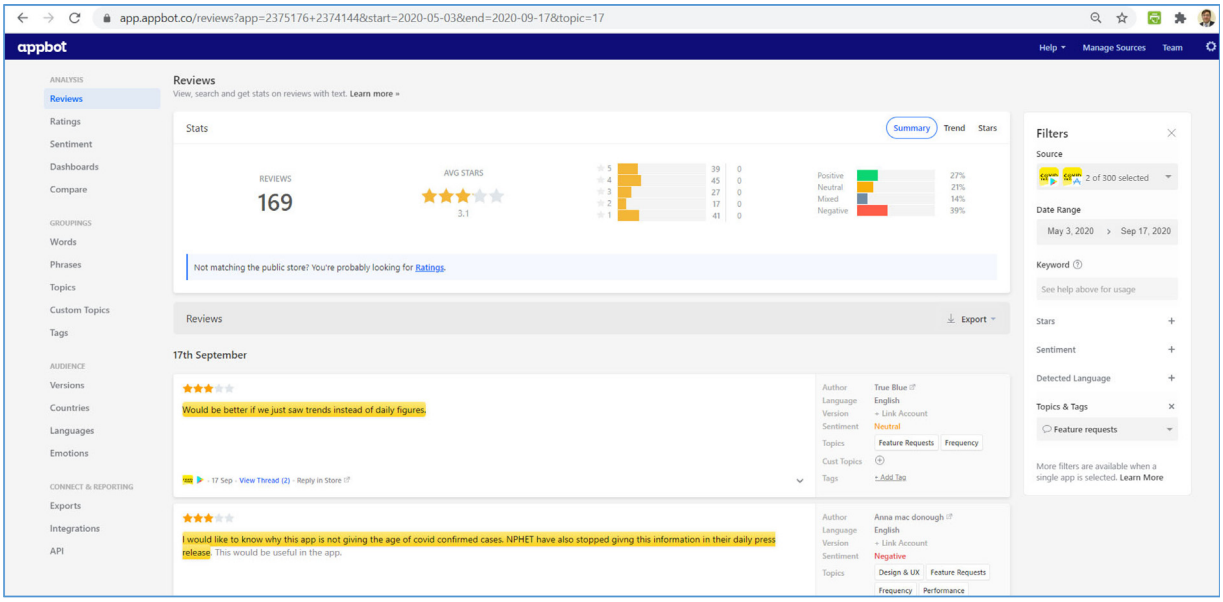
Also, as seen in Fig. 23, in both platforms, the ranking order of the apps is much the same, with only some mid-table positions being different. The highest and lowest-ranked apps, Protect Scotland and England & Wales, respectively, are clearly the highest and lowest regardless of platform.

One caveat to this analysis is that the average star rating, reported by app stores, takes no account of the volume of users, which may be worth considering in future work, especially in cases where there is a notable disparity between the OS versions. Outside of the scope of this work also is a consideration of how the apps were developed for different platforms, specifically to what extent did they share a codebase or interface. While the correlation implies that there is a great degree of similarities between country apps on different platforms, it would be interesting to further examine the differences at a technical level and see how this may relate to platform disparity.

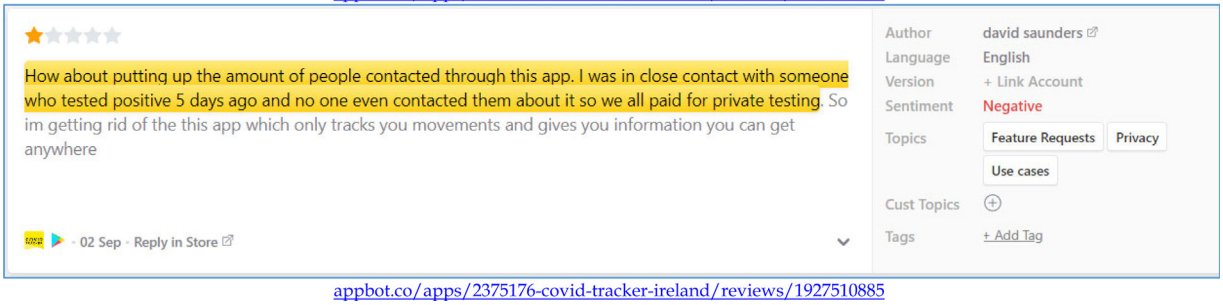
Lesson learned and recommendation for future work: It would be interesting to examine the differences among the apps and also their two OS versions at a technical level, e.g., their code-base, software architecture.

Of interest to note here is that the relatively close nature of the rankings on both platforms, and with some crossover between which has a better score for individual apps, implies that the underlying decentralized services provided by Google for Android and Apple for iOS, which all the decentralized apps will use, make little or no difference to user perception.

Lesson learned: There is a clear correlation between user (star) ratings on different platforms, which shows that the underlying Google or Apple decentralized technology makes little difference in user perception compared with the frontend/OS implementation. Also, since most probably, the Android and iOS versions of each app have the same features (we did not check/compare those in detail), users' satisfaction (stars) of either version of a given app are quite similar.



appbot.co/apps/2375176-covid-tracker-ireland/reviews/1950639714
appbot.co/apps/2375176-covid-tracker-ireland/reviews/1950411937



appbot.co/apps/2375176-covid-tracker-ireland/reviews/1927510885

Fig. 22. Reviews a subset of the "Feature requests" submitted by users for the *COVID Tracker Ireland* app.

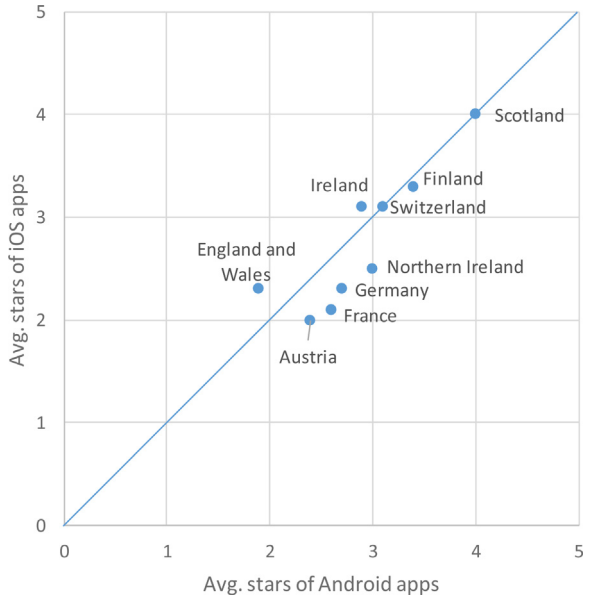


Fig. 23. Popularity of Android versions of the apps versus their iOS versions, for each country, as measured by the average 'stars'.

Table 4
Ratios of reviews which are "feature requests".

App	OS	Reviews		% of reviews with feature requests
		Total # of reviews (in our analysis time-window)	# of reviews with feature requests	
1-StopCOVID NI	Android iOS	195 97	18 4	9.2% 4.1%
2-NHS COVID (ENG)	Android iOS	174 63	12 5	6.9% 7.9%
3-Protect Scotland (SCO)	Android iOS	573 110	19 5	3.3% 4.5%
4-COVID Tracker Ireland (IE)	Android iOS	1,463 274	121 28	8.3% 10.2%
5-Corona-Warn Germany (DE)	Android iOS	20,972 6,772	781 269	3.7% 4.0%
6-SwissCovid (CH)	Android iOS	1,370 652	34 15	2.5% 2.3%
7-StopCovid France (FR)	Android iOS	2,397 241	101 18	4.2% 7.5%
8- Stopp Corona Austria	Android iOS	1,961 599	42 25	2.1% 4.2%
9-Finland Koronavilkku (FI)	Android iOS	1,276 236	34 8	2.7% 3.4%

4.6.2. The sentiment of reviews for the Android versions of the apps versus their iOS versions

As we discussed in Section 3.2, it was interesting to observe that, in the case of all nine apps, the Android apps have received more reviews compared to iOS apps. This seems to align with the general trend in the app industry, as reported in the grey literature: "*Android users tend to participate more in reviewing their apps*"[75] and "*Android apps get way more reviews than iOS apps*".[76]

Apart from the "volume" of reviews received for each OS version, we were curious to compare the "sentiment" of reviews between the two OS versions. As discussed in Section 4.1, AppBot calculates and provides a sentiment score of each review (a value between 0%–100%). The higher this value, more positive the review sentiment, meaning that the review text has a positive tone in its message. AppBot also provides a single aggregated value from all reviews of an app, which we gathered and visualized as a scatter plot in Fig. 24.

Comparing the sentiment analysis in Fig. 24 with the equivalent comparison of star ratings in Fig. 23 shows a clear difference in how sentiment analysis has detected comments compared with simple ratings. While again, as in Section 4.6.1, most of the apps show a correlation between platforms, there are some interesting differences here. For example, while the apps of both France and Austria have a very low (∼6%) number of positive reviews on iOS, they still have low but many times higher levels of positive reviews on Android (∼22%). Based on the sentiment analysis, the Scottish app remains the most highly regarded app by some degree, with 60%–80% positive reviews depending on the platform and no other app having >40% positive reviews. Compared with star ratings, the sentiment analysis does show a difference, however, in the most negative end, with the German app garnering the lowest proportion of positive reviews on iOS (<5%). The app for England & Wales is still the worst reviewed on Android (∼12% positive), with Germany just ahead (∼17%).

Seeing a disparity in general between iOS and Android reviews again, we searched in the grey literature (online sources) to see if there were any discussions or reported evidence on why Android app reviews are slightly more "positive" than iOS app reviews. However, we could not find any. We think it is worthwhile to investigate this issue in future studies. It is also clear from sampling reviews that those, especially towards the longer end, are more nuanced and will discuss both positive and negative aspects of the app, perhaps making an overall judgment hard to make especially using automated sentiment analysis.

Lesson learned/recommendation: The sentiment analysis of apps can provide more complex granular output compared to just the "star rating", but there seems to be an inherent negative bias, especially on Android, which should be further investigated in future studies to better understand the phenomenon. A possible future Research Question (RQ) would be: Why is there an inherent negative bias in Android versions of an app compared to the iOS version?

4.6.3. Problems reported for each OS version

In our dataset, we separated the reviews of each OS version and then fed them into the AppBot's sentiment analysis and word-cloud visualization, similar to what we had done in Section 4.2. As representative examples, we selected five of the nine apps. We mainly selected the apps with most reviews, to ensure that the sentiment analysis will have enough data to provide reliable and meaningful results.

We show in Table 5 the word-cloud visualization, enriched with sentiment results (by the color of texts), for each OS version of five example apps.

Visually, it is possible to see some differences and similarities, e.g., for the case of the *Protect Scotland* app, the term "install" seems to occur many times in the "positive" sentiment in the reviews of its Android version compared to the iOS version. By a closer inspection, i.e., clicking on each term in the word-cloud in AppBot's UI and reviewing a sample of the corresponding reviews,

75 https://medium.com/@takuma.kakehi/we-need-app-reviews-but-we-need-to-ask-at-the-right-time-e2916b126c8e

76 medium.com/@chiragpinjar/why-android-apps-get-way-more-reviews-than-ios-apps-30c5b9e7ee71

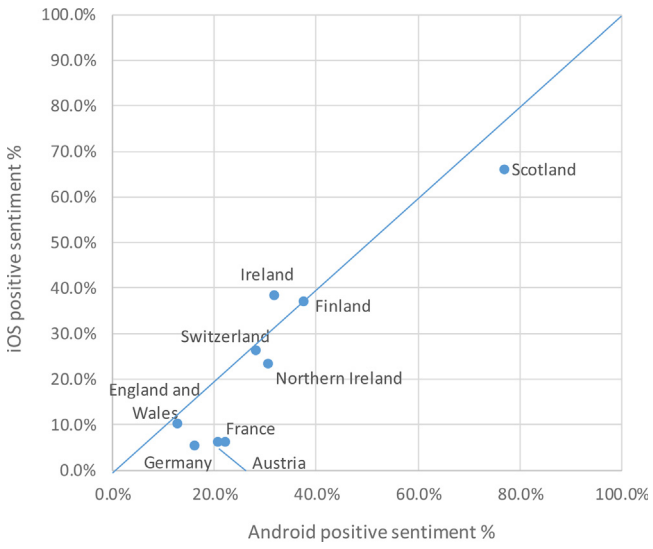


Fig. 24. Sentiment of reviews for the Android versions of the apps versus their iOS versions, for each country.

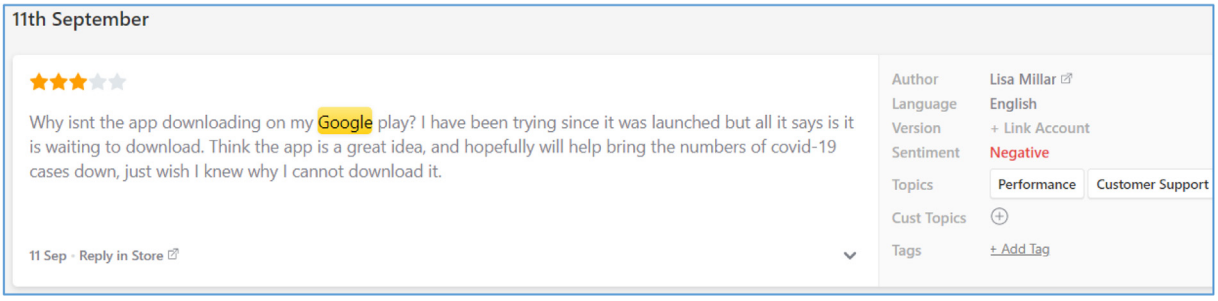


Fig. 25. A user feedback from an Android user of the *Protect Scotland* app.

we found that many users have said something along the lines of "*Easy to install*".

Recommendation for future work: It would be worth investigating in the future what differentiates the Android app users from the iOS app users to have different opinions about the app installation.

Again for the *Protect Scotland* app, there is red (negative) sentiment for the term "Google" in the Android version, and clearly, that term does not occur in the iOS version reviews, since Android decides are strongly associated with Google and its services, e.g., Google Play app. One user mentioned[77] the review shown in Fig. 25.

Although visual analysis (manually) of the word clouds in Table 5 could provide interesting insights, we were keen to find a way to more systematically and numerically analyze the similarities and differences of the OS versions of each app. We reviewed the text mining literature and found that there are indeed advanced methods to compare "semantic" similarities/differences of two large bodies of text, e.g., one widely used metric is *semantic overlap*, which is based on a concept called *semantic folding* (*fingerprinting*) (Webber, 2015).

There is an online tool,[78] which provides an easy-to-use implementation of the semantic-overlap metric. We fed the entire reviews of each OS of each app to this tool, and the values are shown in the last column of Table 5. For the readers interested in how the cortical.io tool works, we show in Fig. 26 a screenshot of the tool, comparing the semantic of the texts in the review dataset of the two OS versions of the *Stopp-Corona* Austria app.

The semantic-overlap measures, shown in Table 5, range between 45% (Stopp-Corona Austria) to 86% (COVID Tracker Ireland). A low semantic-overlap measure for an app could have various possible root causes, e.g., (1) the Android and the iOS versions of the app may have different features, thus leading to different user opinions; (2) there could be "platform" issues in either of the versions, which could cause negative or positive reactions (reviews) from users, e.g., in the example review of the *Protect Scotland* app shown in Fig. 25, the user had problems downloading the app via Google Play.

We should mention that, as shown in Table 5, for the Corona-Warn German app, the cortical.io tool did not generate any output after letting it run for a long time (no response), possibly due to the large dataset size of German app reviews.

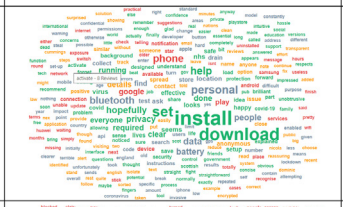
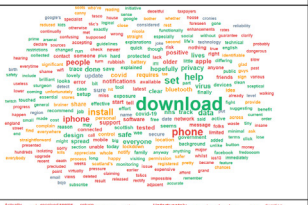
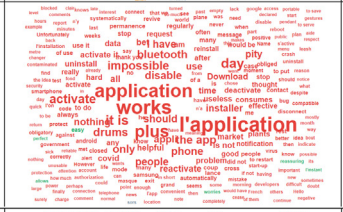
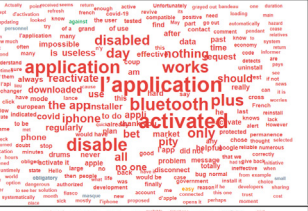
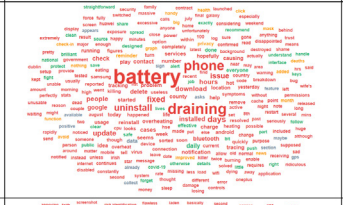
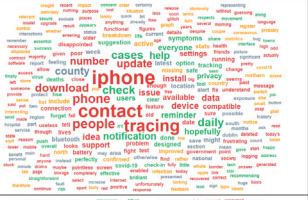
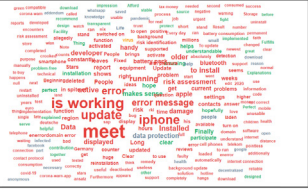
Lesson learned/recommendation: The semantic-overlap measures between the two OS versions of the apps ranged between 45% and 86%. Possible root causes for low or high similarity should be studied in future works.

4.7. RQ7: Correlation of app downloads with country population

One would expect to see a correlation between the number of app downloads with each country's population size. Such an analysis would also provide us with a measure of usage (penetration) of the apps in each country.

77 appbot.co/apps/2437310-protect-scotland/reviews/1951719103
78 www.cortical.io/freetools/compare-text

Table 5
Word-clouds of reviews for both OS versions of five example apps.

Country / app	Android app	iOS app	Semantic overlap
Protect Scotland			82%
StopCovid France			47%
Stopp-Corona Austria			45%
COVID Tracker Ireland			86%
Corona-Warn Germany			The cortical.io tool did not respond, due to large data size

We visualize the correlation of the number of downloads with country populations as a scatterplot in Fig. 27. We also show the linear trend lines in the chart. As discussed in Section 3.2, to get the number of downloads, we interpolated the number of downloads from Play Store's estimate, e.g., we averaged the Protect Scotland app's estimated download count of 100,000+ (meaning 100,001–500,000) to 300,000.

The Pearson correlation coefficient between the two metrics is 0.651. Thus, we can say that there is a reasonable correlation between the two metrics, i.e., for a country with a larger population, as one would expect, there are more downloads. However, the cases of Finland (FI) and Germany (DE) are special (in a positive way) since they are above the trend line.

On the other hand, countries such as England and France are below the trend line and raise some questions, e.g., one wonders why there have been relatively fewer downloads in those countries. One could analyze such questions based on the "social fabric" (Hayden, 2006) of each country and also using different social metrics.

Although we are not social scientists, given our limited knowledge of how societies work and how citizens relate themselves to the societies that they live in, we believe that number of downloads, to some extent, portray, in a macro-scale in the context of a country, the level of its citizens' involvement (engagement) in society and the social responsibility. For these attributes, there have been many studies, advances, and metrics (indices) in social sciences. We were able to find data for two such relevant metrics: (1) Trust In Public Institutions index (TIPI), from an OECD dataset[79]: a value between 0 (no trust at all, in public institutions) to 10 (complete trust); and (2) Civic Engagement Index (CEI): an indicator between 0-100, calculated mainly based on voter participation which some argue being "*the best existing means of measuring civic and political engagement*".[80]

We show the two correlations in Fig. 28 as two scatter plots: correlations of the number of downloads (estimated), normalized by the population size, with two above social metrics (TIPI and CEI). The Pearson correlation coefficients in the two correlation charts are 0.455 and −0.166, as also embedded in the charts. The former shows a moderate correlation, while the latter indicates a very weak negative correlation, which is somewhat surprising (discussed in detail in the following).

We discuss the most important/interesting observations from the charts of Figs. 27 and 28.

- The Finnish app, German app, and English app ("ENG" for England in the charts) could be considered "exceptions" in the charts. The Finnish app is the "best" performing, while the English (covering England + Wales) app is the worst performing in all charts of Figs. 27 to 28.

79 ourworldindata.org/trust
80 www.oecdbetterlifeindex.org/topics/civic-engagement/

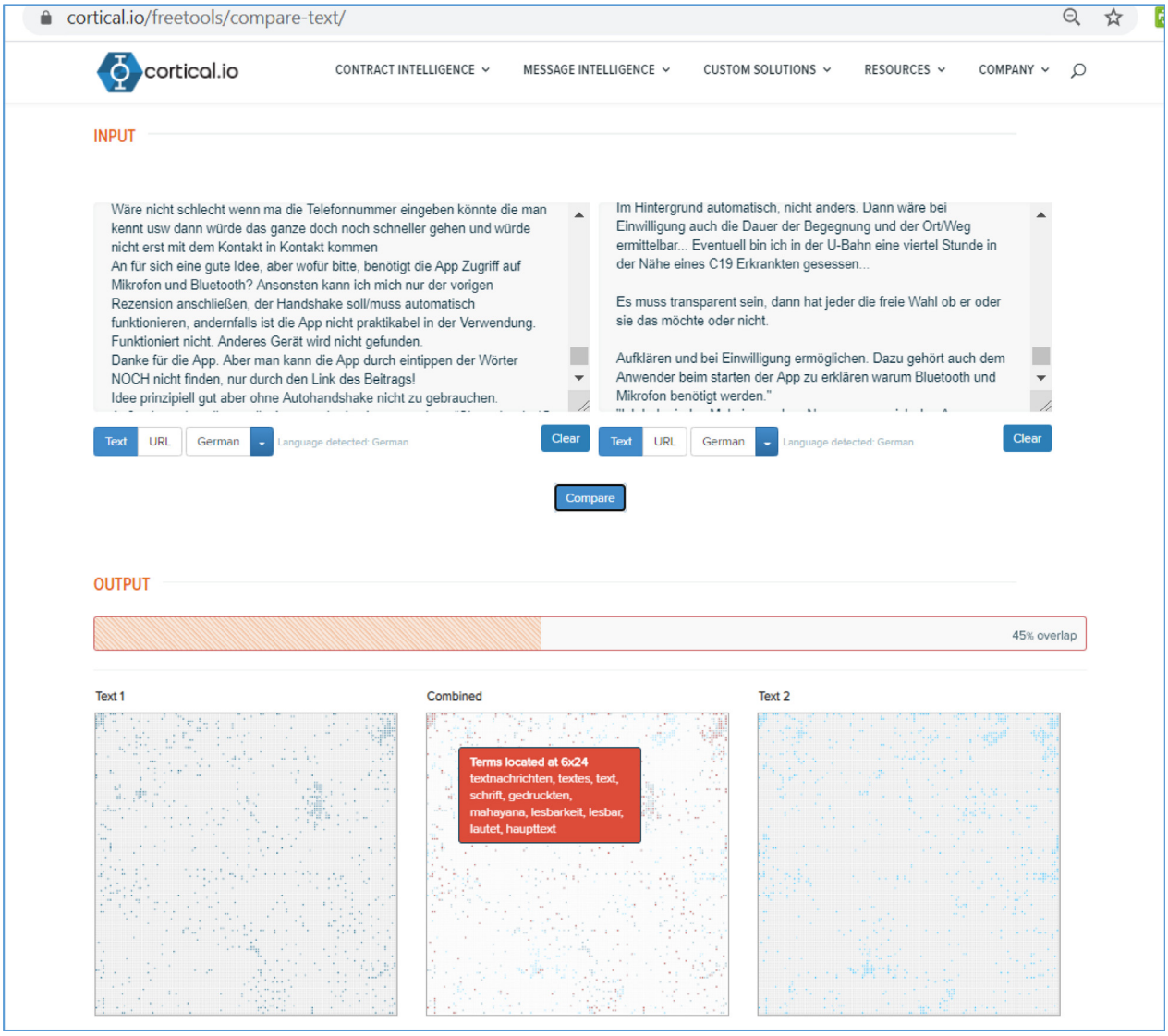


Fig. 26. A screenshot of the cortical.io tool, comparing the semantic of the texts in the review dataset of the two OS version of the *Stopp-Corona* Austria app.

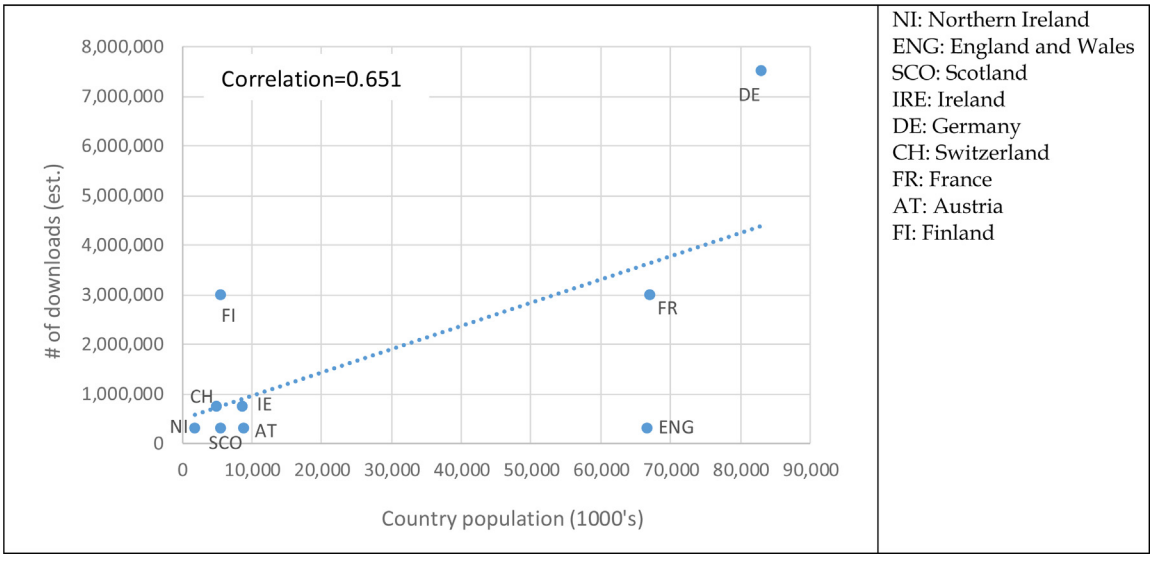


Fig. 27. Number of downloads versus country populations (1000's) (estimated); including the linear regression lines.

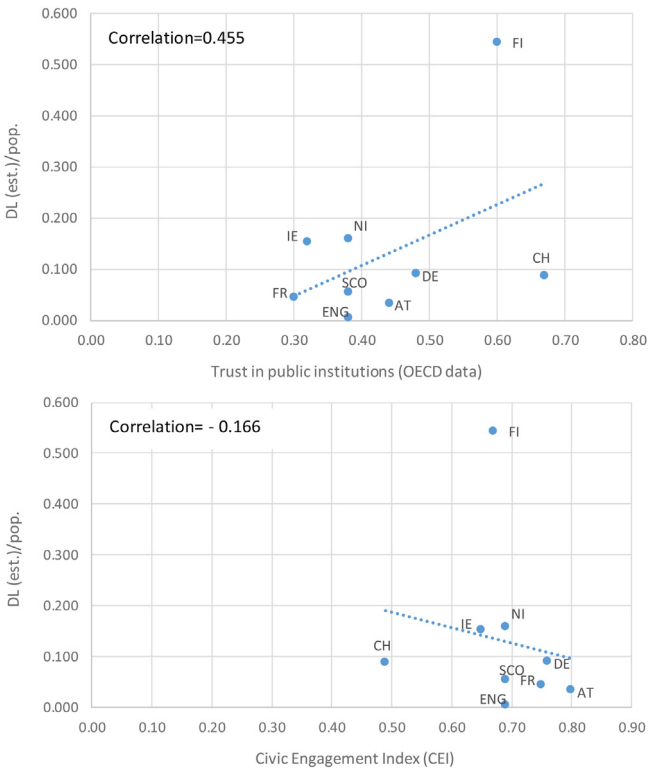


Fig. 28. Correlations of the number of downloads with two social metrics: Trust In Public Institutions index (TIPI), and Civic Engagement Index (CEI).

- Finland ranks second in the TIPI index and first in the normalized measure of downloads to population ratio (DL/pop.). Among the nine data points, this value ranges between 0.005 (for NHS app) to 0.544 (for Finland Koronavilkku app), meaning that, for those two countries, respectively, about 5 of each 1000 citizens and 1 of every two citizens have downloaded the app. This is such a diverse variation in usage (penetration) of the apps when comparing those countries. Finland's "*good*" performance with such a high download ratio has also been covered in many news articles.[81,82]
- Higher relative download ratios of German and Finnish apps (Fig. 27) could be due to a variety of reasons, e.g., those governments have taken more proactive measures to do more publicity for the apps in their countries or had "encouraged" their citizens to download and use the app. Again, going in-depth into these important issues is outside the scope of our paper since they related to the behavioral science, social aspects, and epidemiologic aspects of the apps, and further work on these issues is needed, similar to the papers published already on these topics (Farronato et al., 2020; Nicholas et al., 2020) (as reviewed in Section 2.5).
- Switzerland (country code: CH) has the highest rank in the TIPI index, but the download to population ratio of the Swiss app is quite low. To find out what has possibly led to such a low download volume, we did a news search for the Swiss app and immediately found news articles that the legal actions and even a referendum has been "*launched against SwissCovid app*".[83,84] It could be that such events have led the public to think again before installing the app.

Of course, we should interpret correlation data with caution since, as it is well known in the statistics and general scientific literature that: "*Correlation does not imply causation*" (Aslam, 2015; Ksir and Hart, 2016), which refers to the inability to legitimately deduce a cause-and-effect relationship between two variables solely based on an observed association or correlation between them.

Lesson learned/recommendation: There is a moderate correlation between the number of downloads normalized by the population size and the *Trust In Public Institutions index* (TIPI). This seems to denote that the more trust a country's population, as a whole, has on their government, the higher the ratio of app downloads, and expectedly the higher the use.

4.8. RQ8: Correlation of the number of reviews versus country population and also the number of downloads

Similar to the previous RQ, one would expect to see a correlation between the number of app reviews with each country's population size. We visualize the correlation of the number of reviews with population data as a scatter plot in Fig. 29. The distribution of data points in Fig. 29 is quite similar to the distribution of data points in Fig. 27, implying that the number of downloads and reviews have had similar trends.

A higher number of reviews, relative to each country's population, could imply a variety of possible factors: (1) whether the population of one country tends to be more "outspoken" (or critical) of public (governmental) activities, in this case: contact-tracing apps, than other countries; (2) whether the apps of certain

81 qz.com/1898960/whats-behind-finlands-contact-tracing-app-success-user-privacy/

82 uk.reuters.com/article/us-health-coronavirus-finland-app/one-in-four-finns-downloaded-covid-19-tracing-app-in-four-days-idUKKBN25U20H

83 lenews.ch/2020/07/24/referendum-launched-against-swisscovid-app/

84 lenews.ch/2020/10/13/initiative-against-swiss-covid-app-fails/

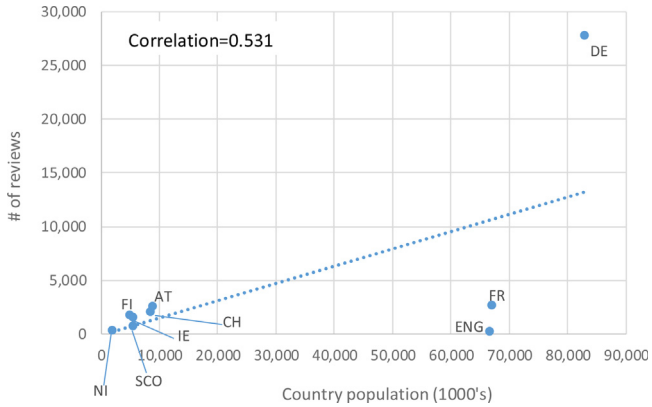


Fig. 29. Number of reviews versus country populations (1000's); including the linear regression line.

countries have really high or really low quality, as it seems that users tend to submit reviews in those extreme cases (according to comments mentioned in most reviews); and (3) Higher or more active engagement with national digital technologies (do citizens "care" to provide feedback to apps as a type of digital technologies?).

As the last analysis in this section, we had the data to calculate the download-to-review ratios of the apps, i.e., to determine the ratio of users (people who have downloaded the apps) who have left reviews. We show the data as a scatter plot in Fig. 30. We can see, for example, that the highest and lowest performers in this metric are: Austrian app (one review per 153 users), and the Finnish app (one review per 2351 users), respectively. The German app is the second-highest, with one review per 358 users. Similar to the above discussions about similar metrics, root-cause analysis of this last indicator would also need to analyze the behavioral, social, and epidemiologic aspects of the apps, and we leave that to further works.

Lesson learned: The apps from the German-speaking countries Austria and Germany have the highest number of reviews per number of users.

4.9. RQ9: Trends of review volumes and their sentiments over time

Another insightful/interesting aspect of the review dataset that we found to be worth analyzing was the trends of review volumes and their sentiments over time. The AppBot tool again already has a feature to get such trend-charts easily and then we could do our interpretation/analysis on the charts.

We show those charts for all the nine apps under study, in Appendix B (Fig. 34). In the AppBot tool, there is even a way to select either or both of the OS versions of the app and have the chart generated separately. For the case of three apps (the Austrian, UK, and Irish apps), chosen as examples, we provide both charts for the two OS versions (Android and iOS). For the sake of brevity and space, for the other apps, we provide only the charts for the iOS versions only.

One important piece of information in this timeline analysis is the release date of each version of a given app. The Apple App Store provides the version history (release date of each version) of a given app and AppBot uses that data to include those time instances in the generated charts; see the example of the Austrian app in Appendix B (Fig. 34). The Google Play Store does not provide that information for its hosted apps.

As expected, the first release date of apps of different countries is different since each country's response to COVID and its decision for developing and releasing contact-tracing apps have been made independently. Back in Section 3.1, we included the first release dates in Table 2, and also discussed the number of releases after that date until the data-extraction date of our study (September 17, 2020) (see Fig. 31).

We discuss the most interesting and insightful observations in Fig. 34 below:

- Trends of review volumes are changing throughout the time horizon. Often for the case of most apps under study, a large volume of mostly negative reviews has been recorded in the first few days/weeks of its first release, and then a decline has occurred. Does this imply that users have lost interest in the apps over time? This reminds us of the well-known "hype cycle" (Fenn and Raskino, 2008). As a matter of fact, various reports have linked the uptake of contact-tracing apps to hype cycle, e.g., a report by MIT Technology Review[85] mentioned that: "*If contact-tracing apps are following Gartner's famous hype cycle, it's hard to avoid the conclusion they are now firmly in the ;trough of disillusionment'. Initial excitement that they could be a crucial part of the arsenal against covid-19 has given way to fears it could all come to nothing, despite large investments of money and time*". A news article in the American NBC News, came with this title: "*Covid apps went through the hype cycle. Now, they might be ready to work*"[86] . Trends of review volumes also show this cycle trend to some extent. But of course, we should clarify that review volumes are not necessarily the "usage" trends, but there may be some relationship between the two. In early days and weeks of an app, it is often the case that governments conduct massive publicity campaigns via TV ads, social media, and newspapers to "encourage" their citizens to install and use the apps. And, citizens indeed respond very proactively to such calls in the initial days, e.g., it was reported[87] by the UK Department (Ministry) of Health and Social Care that about six million people downloaded the NHS contact-tracing app on the first day it was launched. Out of those six million people, as Fig. 34 (in Appendix B) shows, about 3000 users left reviews in the Android version, and about 900 users left reviews in the iOS version of the app.
- Some reviews have come before a given app was even released, e.g., for the Swiss app, the first review came on June

85 www.technologyreview.com/2020/08/10/1006174/covid-contract-tracing-app-germany-ireland-success/

86 www.nbcnews.com/tech/tech-news/covid-apps-went-through-hype-cycle-now-they-might-be-n1242249

87 www.digitalhealth.net/2020/09/nhs-covid-19-app-downloaded-10-million-times-since-launch/

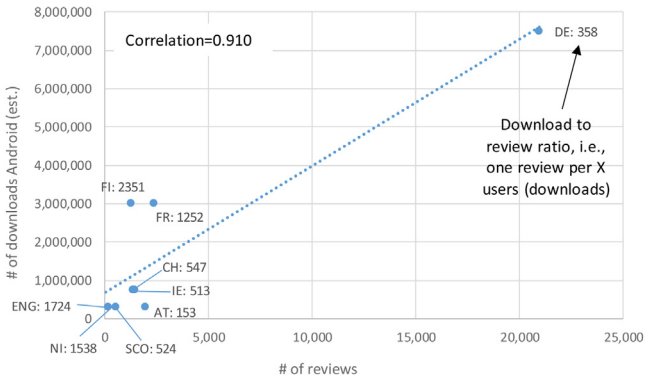


Fig. 30. Download to review ratios of the apps.

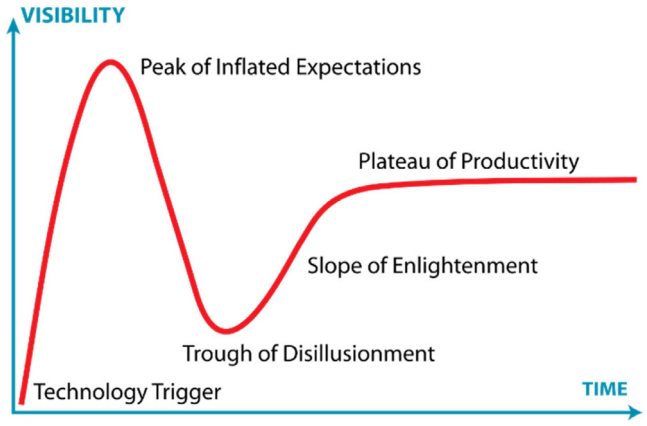


Fig. 31. The Gartner "hype cycle" (Fenn and Raskino, 2008).

2nd; however, the app was officially launched[88] on June 25th. By looking into that first review[89] and its author, we found that that review was written by a software engineer, and it could be guessed that he was in the development team of the app.

- The UK (NHS) app was officially "launched" in the nation to the citizens on September 24, 2020. But on its Apple app store page,[90] we see that it was first released (version 3.0) on Aug. 13, 2020, more than 40 days before the official launch date. That period was actually a "trial" (test) period[91] in the UK's Isle of Wight, in the London borough of Newham, and among NHS volunteers. The fact that the app was downloadable from the app store but not working for regular people caused many confusions among citizens, and many negative reviews have come in due to that in those 40 days, e.g., "*I've been asked to download this app by my local leisure centre. The app says it is only for Isle of Wight and Newnham residents and is asking for a code. I cant use it*",[92] and "*The BBC & the NHS both say that the app is ready. The app says that's only a test for the Isle of Wight. Which is it?? FAIL!!!*"[93]
- We also noticed the review "bursts" after the apps' first release and wondered about their underlying factors. For example, for the Stopp-Corona Austria app, we can notice in Appendix B (Fig. 34) a peak in review volume on June 26th.
  - We hypothesized that bursts in the review volumes could be due to the release of new versions. But with the release of each version, we cannot notice much change in the ongoing review volumes. This could be since most mobile phones do the installation of app updates automatically and seamlessly, so layman users do not notice the new versions on their phones. As a related line of work, there have been many studies in the Mining Software Repositories (MSR) community, analyzing similar software artifacts, e.g., bugs and feature requests, in the time horizon. For example, e.g., a paper by Garousi (2009) analyzed the "arrival" pattern of when issues (bugs and feature requests) have been submitted and then "closed" (addressed) by the development team. We thus find out that these trends are *quite different* than conventional software systems, for which with new versions, sometimes more bug reports are entered in their bug repositories (Garousi, 2009).
  - As a second possible factor, we hypothesized that changes (especially jumps) in the review volumes could be due to media coverage of the apps, i.e., if citizens hear more about the apps in the news, they would install/use them more on a given day, and then possibly leave reviews for the app in the app stores. We found that this hypothesis hold for a few cases. For example, for the Austrian app, by doing a Google search for "*stopp corona app österreich 26th June*", we immediately found a news article: "*Corona app: The update*

88 www.bag.admin.ch/bag/en/home/das-bag/aktuell/news/news-25-06-2020.html
89 appbot.co/apps/2348215-swisscovid/reviews/1822036574/
90 apps.apple.com/us/app/id15204276633.0
91 www.digitalhealth.net/2020/08/covid-19-new-trial-nhs-contact-tracing-app/
92 app.appbot.co/apps/2411517-nhs-covid-19/reviews/1957894169/
93 app.appbot.co/apps/2411517-nhs-covid-19/reviews/1953464033/

for the automatic handshake is here"[94] (the article is in German). The news article discussed an important issue: "*A new and improved version of the Stop Corona app of the Austrian Red Cross is now available in the stores. After the update, the automatic digital handshake now works on all devices with the mobile operating systems iOS (Apple) and Android (Google)*". By reviewing some of the reviews recorded on that day, we clearly saw that many (34 reviews) were recorded on that single day, some of which were:

- *"Thanks for today's update. A very useful and good app.*
- *The official interface is now used, so an anonymous handshake can be carried out in the background, finally! Many thanks to the developers.*
- *All of our neighbors [countries] also rely on the API from Apple & Google and look almost identical with the exception of a few design differences. The design is really successful and intuitive".*

Lesson learned: Bursts in the review volumes seem to be not strongly correlated with new releases (versions) of the apps but instead with more news coverage of the apps.

- For the cases of the two OS versions of the Ireland app, we notice slight differences in review volume "peaks" (bursts) between them (the timeline around early to mid-August). A noteworthy situation developed in that time frame and was widely covered in the media, as we discuss next.
  - According to an Irish news article,[95] "*From August 7 to August 12, more than 150,000 uninstalls [of the COVID Tracker Ireland app] had been reported [mainly due to high battery usage of the app]*". According to another article,[96] "*a Google Play Services update caused the app to rapidly drain handset batteries for a two-day period earlier this month [August]*". On August 10th, another news article[97] reported that Google would "*launch fix for battery drain affecting Covid Tracker Ireland app*". Proactive communication was made to the public by the Irish health ministry on this, e.g., see the Tweet shows in Fig. 32.
  - A few days later, the health ministry publicized that the update has been installed on the phones by mentioning that: "*Google have informed us that the majority of Android phones in Ireland have been updated but it may take a day or so for every Android phone to receive the update. We would encourage anyone that uninstalled the app at the weekend to download it again over the next few days.*"[98]
  - As per our analysis of review texts, most of the reviews submitted in the few days around August 10th, for the Android version of this app, are related to the high battery drainage of the app.
  - A software-engineering observation (takeaway message) from this issue is that even if the app itself was not modified in that time period, but an automated software update of the Google Play Services did impact the performance and battery usage of the app. This reminds us of the highly complex nature of developing these apps and the need for proactive analysis of updates and their dependencies.

Lesson learned/recommendations: There could be unexpected inter-dependencies among the apps and various aspects of the mobile OS that they are running on. Updates to the OS could adversely impact a given app and could easily cause major dissatisfaction by the app users. Thus, the development team should work with OS vendors (in the above example case, Google is behind the Android OS) to prevent such chaotic situations.

Another insightful aspect is how review emotions (sentiments) change over time for a given app. A related work has been done by Martens and Johann (2017), in which the authors identified four reoccurring patterns for how review emotions (sentiments) change over time for mobile apps:

- the Consistent Emotion pattern: where the sentiment of users only slightly varies around a specific value. This value can either be consistently negative, neutral, or positive.
- the Inconsistent Emotion pattern: sentiments of different users vary considerably
- the Emotion Drop/Jump pattern: where the sentiment of users, e.g., suddenly drops due to bugs introduced or features removed in-app updates. Vice versa, an emotion jump can be introduced by fixes or feature requests implemented in app updates.
- the Steady Decrease/Increase pattern: where the overall satisfaction of users slowly decreases or increases due to changes introduced in-app updates.

While, due to space limitations of this paper, we are not including investigations of such patterns on the nine apps under study, such analysis could be useful in future works.

5. Implications and discussions

Now that we have presented the results of our analysis in Section 4 extensively, we take a step back and look at our original goal in this study, which has been to gain insights into the user reviews of contact-tracing apps to find out what end users think of COVID contact-tracing apps and the main problems that users have reported. As a reminder, our focus has been to assess the "software in society" aspects of the apps based on their users' reviews, and thus we need to clarify again the scope of our work and the many related issues that are important but outside this paper's scope: public health aspects of the app, behavioral science (Anon, 2020), and epidemiology.

Next, we present the implications of the results. Afterwards, we discuss the limitations of our work.

5.1. Implications for various stakeholders

As discussed in Section 1, our study and its results could provide implications, benefits, and insights to various stakeholders: (1) software engineering teams of the apps, (2) decision-makers, and public health experts who manage the development

94 www.roteskreuz.at/news/datum/2020/06/26/stopp-corona-app-das-update-fuer-den-automatischen-handshake-ist-da/

95 www.irishexaminer.com/news/arid-40058456.html

96 uk.reuters.com/article/us-health-coronavirus-ireland-apps/active-irish-covid-19-tracing-app-users-drop-on-battery-problem-hse-idUKKBN25N1PA

97 www.siliconrepublic.com/enterprise/battery-drain-covid-tracker-ireland-app-google

98 www.irishtimes.com/business/technology/hse-says-fix-for-covid-tracker-app-rolled-out-to-all-android-users-1.4326646

99 twitter.com/HSELive/status/1292510112184229889.

Fig. 32. A tweet[99] by the Irish health ministry about an issue in their app.

and adoption of those apps in the society, (3) developers of app-review mining tools, and (4) other SE researchers.

Also, let us recall from Section 3.2 that our research method in this paper has been "exploratory" case study (Runeson and Höst, 2009), whose goals are defined as: "*finding out what is happening, seeking new insights and generating ideas and hypotheses for new research*", and those have been the goals of our study. Thus, throughout our work, we gathered and derived the implications, benefits and insights, as we highlighted throughout Section 4.

To "package" the lesson learned, recommendations, and implications that we discussed and derived in different parts of Section 4 into one single picture, we synthesize and present them in Table 6. For full traceability, we also provide the section number from which each implication and benefit has been derived so that readers can go back and read the details.

As the table shows, most of the evidence-based implications/benefits are for the software engineering teams of the apps, but we believe there are valuable suggestions to other stakeholders too.

5.2. Limitations and potential threats to validity

In this section, we discuss limitations and potential threats to the validity of our study and the steps we have taken to minimize or mitigate them. The threats are discussed in the context of the four types of threats to validity based on a standard checklist for validity threats presented in Wohlin et al. (2000): internal validity, construct validity, conclusion validity, and external validity.

Internal validity: Internal validity is a property of scientific studies that reflects the extent to which a causal conclusion based on a study and the extracted data is warranted (Wohlin et al., 2000). A threat to internal validity in this study lies in the selection bias (i.e., selection of the nine apps under study). As discussed in Section 3.3, analyzing user reviews of "all" the 50+ worldwide apps would have been a major undertaking, and thus, instead, we decided to sample a set of nine apps. Future studies could analyze other apps of other countries and compare the trends/findings.

Construct validity: Construct validity is concerned with the extent to which the objects of study truly represent the theory behind the study (Wohlin et al., 2000). In other words, the issue relates to whether we actually analyzed the issues that we had originally intended to assess (as per our RQs raised in Section 3.2). We defined the RQs clearly and, as discussed in Section 3.2, for data collection and measurement, we used an established approach: Goal-Question-Metric (GQM) (Basili, 1992). Some research questions (RQ7 and RQ8) rely on clearly defined numbers, i.e., the number of app downloads and countries' population sizes. The other RQs rely on sentiment analysis that has been performed based on the well-established and mature tool AppBot. Furthermore, for instance, in the case of the French and German app, the reviews were automatically translated, which could cause issues with respect to the construct validity. However, the author team includes researchers who speak French and German. These authors performed a review of the translated word clouds, and in case of issues with translations, which were in general rare, corrections of the translated reviews were made.

Conclusion validity: The conclusion validity of a study deals with whether correct conclusions are reached through rigorous and repeatable treatment (Wohlin et al., 2000). We, as a team of three researchers together, analyzed nine apps. The conclusions for the different apps were drawn by different authors of the team and cross-checked by the other authors. Also, the analysis of the nine apps overall leads to the situation that the approach of drawing conclusions was refined and performed iteratively in several iterations, which provides an additional step of quality control.

External validity: External validity is concerned with the extent to which the results of this study can be generalized (Wohlin et al., 2000). The study has clearly defined context, i.e., to analyze user reviews (feedbacks) of a subset of the COVID contact-tracing apps, both for Android and iOS. The study does not intend to generalize to other contact-tracing apps. However, we have only analyzed the data for apps from nine countries, i.e., from England and Wales, the Republic of Ireland, Scotland, Northern Ireland, Germany, Switzerland, France, Austria, and Finland. In order to make the assessments more comparable, we limited the sampling to European countries by selecting the four apps developed in the British Isles and five apps from mainland Europe. Given that there are country-specific differences in mobile app user behavior (Lim et al., 2014), this is a threat to generalizability over arbitrary countries (about 78 countries and regions have, so far, developed COVID contact-tracing apps). However, we think the results provide interesting insights across different countries in Europe, and our research approach can in the future be applied to further countries.

Table 6
Evidence-based lessons learned, recommendations, and implications derived by our study for various stakeholders.

Lesson learned/ recommendations/ implications	Based on discussions in section number…	Stakeholders			
		Software engineering teams of the apps	Decision-makers, behavioral scientists, and public health experts	Developers (vendors) of app-review analytics/ mining tools	SE researchers
The users are generally dissatisfied with the apps under study, except the Scottish app. This issue is perhaps the clearest and the most important message of our study, which should be investigated by stakeholders.	4.1	x	x		
Future studies could look into what factors have made the Scottish app be different than others in the pool of apps under study. That could a research question (RQ) to be studied by researchers in future works.	4.1				x
Contact-tracing apps should be designed to be as simple as possible to operate (for usability), as we cannot expect layperson citizens to review the online FAQ pages of the app to properly configure it, especially for a safety-critical health-related app.	4.2	x			
Developers of the apps can and should engage directly with reviews and reply, not only gaining insight into the most commonly raised concerns but also answering the questions in public view. This can even provide a positive "image" of the software engineering team behind the app in public view (in terms of accountability, transparency, responsiveness, and being open to feedback).	4.2	x			
Just like any other mobile app, user reviews for contact-tracing apps range from a short phrase such as "*Not working*", often not that useful nor insightful, to detailed objective reviews, which could be useful for various stakeholders. Thus, if any stakeholder (e.g., the app's development team) wants to benefit in a qualitative way from the reviewers, they need to filter and analyze the "informative" reviews.	4.2	x	x		
A common issue for most apps is high battery usage (drainage). Software engineers should use various heuristics and approaches to minimize battery usage. Mobile phone users are sensitive about battery usage, and any app that uses a high amount of battery would be likely to be uninstalled by mobile users.	4.3	x			
For the German apps, a substantial number of reviews are about the app not working, which can be seen as bug reports. But unfortunately, since most users are non-technical people, informative and important components of a bug report (e.g., phone model/version and steps to reproduce the defect) are not included in the review. Thus, it would be quite impossible for the app's software engineering team to utilize those reviews as bug reports. A recommendation could be that in the app itself (e.g., in its "tutorial" screens), explicit messages are given to the users, asking them that, if they wish to submit bug reports as reviews, they should include important components of a bug report (e.g., phone model/version and steps to reproduce).	4.3.1	x			
A large number of cross-(mobile) device issues have been reported for the German and other apps too. This denotes inadequate cross-device testing of the apps, possibly due to the rush to release the apps to the public. Given the nature of the apps, and since the apps could be installed on any mobile device model/version by any citizen, the development and testing teams should have taken extra care in cross-device development and testing of the apps. There are many sources both in the academic literature (Husmann et al., 2016; Nebeling et al., 2015) and also grey literature[a] on this issue, which the development and testing teams can benefit from.	4.3.1	x			

(*continued on next page*)

Table 6 (*continued*).

Lesson learned/ recommendations/ implications	Based on discussions in section number…	Stakeholders			
		Software engineering teams of the apps	Decision-makers, behavioral scientists, and public health experts	Developers (vendors) of app-review analytics/ mining tools	SE researchers
Certain features of a given app did not work for many users for several days, e.g., for the German app, a functionality called "Risk assessment." Such a malfunction usually gives a negative perception to users about an app, even if the other features of the app do work properly. It is thus logical to recommend that app developers should not include a feature in the app release if they predict or see from reviews that the feature does not work for certain users or on certain times/days.	4.3.1	x			
We see a rather trivial issue in the apps, i.e., users have to "activate" multiple times instead of just once. We would have hoped that the test teams of the apps had detected and fixed those trivial issues before release.	4.3.2	x			
It is important that a given app automatically switches to the home country's language since some non-English users will feel odd if they see a sudden switch from their native language to English in the app's GUI.	4.3.2	x			
The development team of all apps should be proactive in replying to user reviews, and filtering informative reviews and getting more information (e.g., steps to reproduce the defects/problems) from them, e.g., by direct replies to the reviews in app stores.	4.3.3	x			
There seem to be rather trivial usability issues with some of the apps (e.g., the case of exposure notification errors in the NI app). This raises the question of the inadequate usability testing of the apps and the possibility of releasing them in a "rush".	4.3.3	x			
Some of the reviews provide insights on software engineering issues of the apps, e.g., not enough testing has been done on all possible types of QR codes, and not enough performance (load) testing has been done.	4.3.3	x			
For Android phones, the update mechanism of the OS and its components (e.g., APIs) should be "seamless" (automatic), since we cannot expect all users have the "technical" skills to do such tasks properly.	4.3.3	x			
The apps must be clearly identifiable and findable in app stores to maximize the number of users downloading it.	4.4.3	x			
Where possible, some feedback (such as statistics about COVID cases in the region and also the number of close-by phone IDs recorded in the past) should be provided as a feature of the app to encourage users that the app is working to emphasize the pro-social and individual benefit it is having.	4.4.3	x	x		
A variety of insightful feature requests has been provided by users, e.g., by use of the German app: How many encounters there were with other app users (how many people you exchanged the keys with); infection numbers and spread at district level; can the app be used without internet? As a form of "iterative" requirements engineering" (elicitation) (Iacob and Harrison, 2013; Jha and Mahmoud, 2019, 2017; Williams and Mahmoud, 2017; Guzman et al., 2017; Lu and Liang, 2017; Maalej et al., 2019; Nayebi et al., 2017) or "crowd-based" requirements engineering (Groen et al., 2015), the app's software engineering teams are encouraged to review those feature requests and select a subset to be implemented.	4.5.1	x	x		

(*continued on next page*)

6. Conclusions and future work

The initial exploratory analysis of COVID contact-tracing app reviews reported in this paper is only a starting point. As the COVID pandemic has paralyzed most of the life and businesses around the globe, contact-tracing apps, if managed well, may have the potential to help bring the COVID outbreak under control. It is vital for governments and health authorities to develop and offer effective apps that all citizens can use.

Table 6 (*continued*).

Lesson learned/ recommendations/ implications	Based on discussions in section number…	Stakeholders			
		Software engineering teams of the apps	Decision-makers, behavioral scientists, and public health experts	Developers (vendors) of app-review analytics/ mining tools	SE researchers
Given the nature of the COVID pandemic, the governmental policies and guidelines regularly change, and thus, the contact-tracing apps have been regularly updated/adapted to those changes. This is related to the widely-discussed issue of changing/unstable software requirements. Thus, SE researchers are encouraged to work on such issues related to contact-tracing apps.	4.5.1				x
While AppBot's feature to filter reviews to see feature requests only is a useful feature, we found many example reviews which AppBot incorrectly classified as feature requests. We realize that an NLP/AI-based algorithm has been used to do that classification, and such an algorithm will have limited precision, but still, there is a need to improve such algorithms by developers (vendors) of App review analytics tools, such as AppBot.	4.5.1			x	
Many users have cast doubts on the usefulness of the apps, i.e., they do not provide most of the "right" and much-needed features that many users are looking for. Thus, using "crowd-based" requirements engineering (Groen et al., 2015) techniques for these apps are critically needed.	4.5.2	x	x		
It would be interesting to examine the differences among the apps and also their two OS versions at a technical level, e.g., their code-base, software architecture.	4.6.1				x
The sentiment analysis of apps can provide more complex granular output compared to just the "star rating," but there seems to be an inherent negative bias, especially on Android, which should be further investigated in future studies to better understand the phenomenon. A possible future Research Question (RQ) would be: Why is there an inherent negative bias in Android versions of an app compared to the iOS version?	4.6.2	x	x		x
The semantic-overlap measures between the two OS versions of the apps ranged between 45% to 86%. Possible root causes for low or high similarity should be studied in future works.	4.6.3				x
There is a moderate correlation between the number of downloads normalized by the population size and the *Trust In Public Institutions index* (TIPI). This seems to denote that the more trust a country's population, as a whole, sin their government, the higher the ratio of app downloads, and expectedly the higher the use. Behavioral scientists can possibly investigate this issue in more detail.	4.7.1		x		
There could be unexpected inter-dependencies among the apps and various aspects of the mobile OS that they are running on. Updates to the OS could adversely impact a given app and could easily cause major dissatisfaction by the app users. Thus, the development team should work with OS vendors (in the above example case, Google is behind the Android OS) to prevent such chaotic situations.	4.7.2	x			

[a] www.google.com/search?q=movbile+app+%22cross+device%22+testing.

An important issue that we realized during our analysis is the need to compare the different features of the apps. It is fair to say that two main basic use cases for a contact-tracing app are: (1) scan for nearby phones which are located within 2 meters for at least 15 min and record their phone tokens (keys), (2) if the phone user notifies the app that s/he is COVID positive, the app should notify the recorded phones via the stored keys. Although a detailed analysis of different apps' features is outside the scope of our work, we realized that some apps have much more features beyond those two basic features, e.g., a recent version of the German app includes a feature[100] showing the number of tokens (phone IDs) the app has collected. At least for the Northern Ireland (NI) app that the first two authors are familiar with, there is no such a feature. The number of features of a given app and how well they work, of course, could impact the widespread usage and popularity of the app and also to help bring the COVID outbreak under control.

[100] github.com/corona-warn-app/cwa-wishlist/issues/5

As we analyzed and reported in this paper, mining user reviews of contact-tracing apps seem like a useful analysis towards providing insights to various stakeholders, e.g., app developers, decision-makers, public health experts, researchers, and the public. Of course, such an analysis and software engineering aspects can only provide some pieces of the "big picture". Therefore, as discussed in Section 2, sharing the data and collaborations with other important disciplines, including public health and behavioral science (Anon, 2020) shall be conducted. As discussed in Section 1, the first author has been a member of the Expert Advisory Committee for the StopCOVID NI app, in which he has discussed and will continue discussing the results of this study with public health and behavioral science experts.

Furthermore, more collaboration between various stakeholders of the apps (e.g., software engineering teams of the app, decision-makers, and public health experts) is needed. Further work is required in the areas of software engineering, requirements engineering, public health, and behavioral science to improve the public adoption, software quality, and public perception of these apps.

Based on our study, we can see several promising directions for future research in this area:

- First, it would be of interest to reverse engineer the list of features in each app and compare them with each other. Beyond the analysis performed with AppBot, we see the potential for thematic analysis (or qualitative coding) to group feature-request-reporting reviews into a list of suggested features by reviewers. The results can support requirements engineering of refined/ better COVID contact-tracing apps or apps for other pandemics in the future.
- As discussed in Section 4.6.2, our data and analysis showed Android app reviews were slightly more "positive" than iOS app reviews. We could not find any discussions or reported evidence in the academic or grey literature about this phenomenon and think it is worthwhile to investigate it.
- Future studies could analyze other apps of other countries and compare the trends/findings.
- It would also be interesting and insightful to assess the correlations between users' opinions and security/privacy issues. As mentioned in Khalid et al. (2014), complaints related to privacy affect ratings negatively. It is also a common belief that privacy affects the user adoption rate as well.
- Future studies to explore whether it is necessarily the case that apps with more features have worse ratings and vice versa.
- It will be worthwhile to introduce automated approaches and tools to identify and refine the requirements of these apps in a shared pool that can help stakeholders across the board also across apps and countries. An effort to understand requirements and issues across apps and countries may foster solutions to issues that prevent people from installing or keeping these apps installed.
- Rigorous evaluation of the precision of AppBot's outputs and, possibly, using another app analytics tool for making the similar analysis and cross-comparisons of the results
- Investigating how the review emotions (sentiments) change over time for the nine apps under study, based on the four reoccurring patterns identified by Martens and Johann (2017)

CRediT authorship contribution statement

Vahid Garousi: Conceptualization, Methodology, Validation, Investigation, Data curation, Writing – original draft, Visualization, Project administration. **David Cutting:** Validation, Investigation, Writing - review & editing. **Michael Felderer:** Validation, Investigation, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Word-cloud visualizations of the problems reported by users about the apps

See Fig. 33.

Appendix B. Trends of review volumes and their sentiments over time for the apps

See Fig. 34.

Appendix C. Further related work

C.1. Grey literature on software engineering of contact-tracing apps

Unlike academic (peer-reviewed), in the grey literature (such as news articles and technical reports), there are plenty of articles on the software engineering aspects of contact-tracing apps.

An interesting related news article was entitled: "*UK contact-tracing app launch shows flawed understanding of software development*".[101] The article argued that: "*In a pandemic, speed is critical. When it comes to developing high-quality software at speed, using open-source is essential, which other nations were quick to recognize*". The article also criticized the approach taken by the UK healthcare authorities in developing their app from scratch: "*Countries such as Ireland, Germany, and Italy used open-source to build [develop] their own applications months ago. Sadly the UK did not follow suit, and wasted millions of pounds and hours of resources trying to build its own version.*"

Some other related developments include a news article[102] reporting that developers world-wide have found and reported a large number of defects in the England's open-source contract-tracing app. The peer reviews can be found on the GitHub page[103] of the app. The news article[104] went on to say that: "*… Developers have scrutinized every line of code and raised 27 issues on its Android version and 17 on the iOS version*", a summary of which can be found in Fig. 35 (reproduced from the news article). The major concerns raised by software developers included: (1) the app storing the timestamps of contacts and every move of the user with GPS data on a central server; and (2) storage of Google Analytics tracking data which could help to identify users, and thus invalidate the entire idea of app usage being anonymous. In its privacy guidance, the NHSX app had promised: "*the app will not be able to track your location and it cannot be used for monitoring whether people are self-isolating or for any law enforcement purposes*". The news article argued that: "*New shortcomings in the NHSX contact-racing app could further limit effectiveness and scare away users*", which we think is a fair assessment.

Several companies, who have been involved in various software engineering aspects of contact-tracing apps, have also shared (published) grey literature materials (e.g., blog posts). For example, a large software company named ExpleoGroup (with presence in 25+ countries) was hired by the Irish Health Service

101 www.verdict.co.uk/contact-tracing-app-launch/

102 eandt.theiet.org/content/articles/2020/05/developers-find-new-flaws-in-source-code-of-nhs-contract-tracing-app

103 github.com/nhsx/COVID-19-app-iOS-BETA/issues?q=

104 eandt.theiet.org/content/articles/2020/05/developers-find-new-flaws-in-source-code-of-nhs-contract-tracing-app

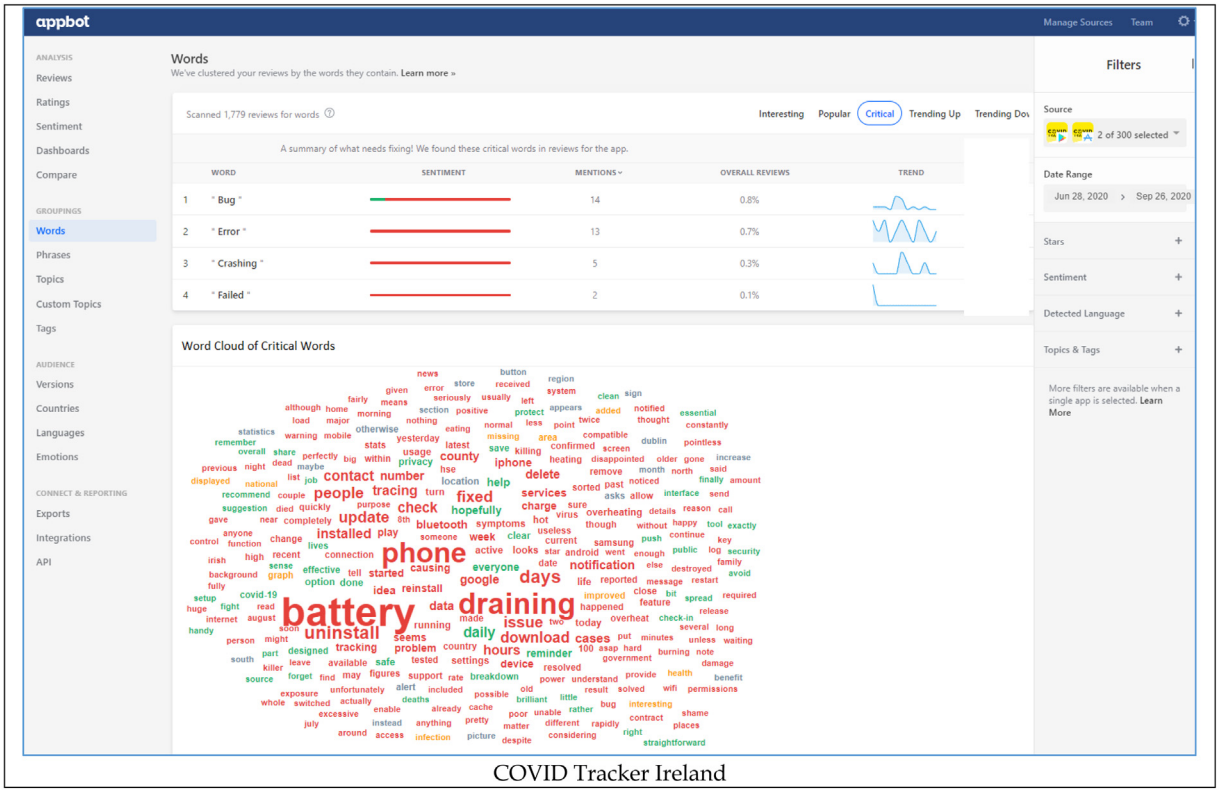


Fig. 33. Word-clouds for the app reviews of all nine apps in which phrases are color codes based on sentiment analysis: positive (green), negative (red), neutral (grey), and mixed (orange).

Executive (i.e., Health Ministry) to conduct quality assurance and testing of the Irish "COVID Tracker" app.[105] The company published a blog post[106] on July 2020 about its test strategy for the app. The post discussed details of how the app was functionally tested, which was mostly manual, but in real-life settings, carrying out approximately 3400 individual software tests, with work effort totaling 4727 human-hours. In addition to functional testing, other types of testing were also conducted, according to the blog post: (1) Performance testing; (2) Exposure Notification Distance Testing: As devices have varying Bluetooth strengths, Expleo tested varying attenuation values that manage Bluetooth strength to maximize compliance with requirements; (3) Graphical user interface (GUI) testing: Expleo tested the look and feel of the app, ensuring user-friendly navigation; the correct and reliable function of all buttons; and that all content and text matched requirements.

The first two authors of this paper have also collaborated with the ExpleoGroup. For example, the first author provided consulting, for the StopCOVIDNI app, by conducting code review and inspection of test plans and test cases. Some of his contributions are discussed in an online technical report.[107] serving as the testing "Closure Report" of the StopCOVIDNI app, published by ExpleoGroup.

The company which has developed both the apps for the Republic of Ireland and Northern Ireland is NearForm.[108] NearForm has published a blog post[109] in which it has discussed about bringing "privacy by design" to contact-tracing apps. The post includes many details about testing, including privacy- and security testing, e.g.: *"Intensive, repetitive testing is crucial when it comes to privacy, and was a core element of the entire development process for this app. Because of the close collaboration among everyone involved, we were able to test the contact tracing app continually across multiple cohorts throughout both the design stage and app development"*.

There have been many other news articles on software engineering aspects of the apps, e.g., for the case Australia's app, it was reported[110] that, "*… developers have highlighted ongoing problems with the [Australian] contact -racing app being able to exchange Bluetooth handshakes with iPhones if the iPhone screen is locked*".

C.2. Formal and grey literature on overall quality issues of contact-tracing apps

In addition to formal and grey literature on software engineering and software quality aspects of these apps, there are many sources (in both literature types) on "quality" issues (not specific to software). For example, a technical report (Leith and Farrell, 2020) by two Irish researchers conducted an evaluation of Google/Apple Exposure Notification API for proximity detection in a commuter bus. The assessment focused on wireless networking aspects of the issue by measuring "attenuation", i.e., the loss of transmission signal strength measured in decibels (dB). Many media articles have reported various criticisms, using such reports as

105 covidtracker.gov.ie

106 expleogroup.com/news/expleo-announces-its-vital-role-in-ensuring-success-of-covid-19-contact-tracing-app

107 covid-19.hscni.net/wp-content/uploads/2020/07/Expleo-StopCOVIDNI-Closure-Report-V1.0.pdf

108 www.bbc.co.uk/news/uk-northern-ireland-53599514

109 www.nearform.com/blog/bringing-privacy-by-design-to-contact-tracing-apps/

110 www.theguardian.com/australia-news/2020/jun/17/covid-safe-app-australia-covidsafe-contact-tracing-australian-government-covid19-tracking-problems-working

Fig. 33. (*continued*).

sources, e.g., Irish Times published an article entitled: "*Precision of tracing apps in doubt after TCD study*".[111]

German Deutsche Welle (DW) News agency published a video with the following title: "*Coronavirus tracing apps: False hope and hidden dangers?*".[112] The Australian news agency, ZDNet, reported that: "*COVIDSafe's [Australia's contact-tracing app] problems aren't Google or Apple's fault despite government claims*".[113]

There are many reports and news articles in the US as well, e.g., a US-based nonprofit organization published a comprehensive article with the following title: "*The challenge of proximity apps for COVID-19 contact tracing*".[114] Among the many arguments included in it was the following: "*Questions about quality, efficacy and accuracy may compound Americans' existing wariness toward tracking technologies like contact-tracing apps. Yet for these apps to work, they need to be adopted by most of the population: Their benefit increases exponentially with the number of users. This presents a circular problem: The effectiveness of these apps will inevitably influence whether people are willing to install them, while the number of people who install the app will directly influence its effectiveness*". Another indeed insightful discussion was: "*Reliable applications of this sort typically go through many rounds of development and layers of testing and quality assurance, all of which takes time. And even then, new apps often have bugs. A faulty proximity*

111 www.irishtimes.com/news/ireland/irish-news/covid-19-precision-of-tracing-apps-in-doubt-after-tcd-study-1.4247865

112 www.youtube.com/watch?v=pYFc5W8E91w

113 www.zdnet.com/article/covidsafes-problems-arent-google-or-apples-fault-despite-government-claims/

114 www.eff.org/deeplinks/2020/04/challenge-proximity-apps-covid-19-contact-tracing

Fig. 33. (*continued*).

tracing app could lead to false positives, false negatives, or maybe both, which stresses the fact that these apps are safety-critical". Furthermore, various metrics have been reported in support of efficacy of these apps: For the Irish app, a news article[115] reported that: "*A total of 308 users registered positive tests in the app's first seven weeks of operation, generating almost 700 close contact alerts, a proportion of whom subsequently tested positive for COVID-19*".

C.3. Behavioral science, social science and epidemiologic science of the apps

The use of contact tracing as a means of controlling infectious disease is long established (Davidson, 1996) and has been seen recently for example in Ebola outbreaks (Browne et al., 2015). The techniques used for contact tracing have however been largely centralized and focused on manual data collection by "contact tracers". Where technology such as apps have been previously used, they have been aids for the contact tracers to record data and/or systems for centralized use, analysis, and visualization of generated data (Sacks et al., 2015; Danquah et al., 2019). An actual contact-tracing app, one that actually performs rather than merely supports the process, measuring contacts and handling notifications is a novel innovation in public health.

While the *potential* benefits of such an app are generally agreed by epidemiologists, the critical success factor is adoption. For an app to be effective, it is estimated that half the population must both install it and have the app active, a significant challenge especially in countries where use of such an app is optional (Trang et al., 2020).

Some literature identifies that while contact-tracing apps face the same challenges as any technology platform, they also have additional challenges based around the sensitive health-related nature of their work and trust issues in governments. For example, Farronato et al. (2020) identifies that platform failure is common place and most often because "*[the platform] never build[s]*

115 uk.reuters.com/article/us-health-coronavirus-ireland-apps/active-irish-covid-19-tracing-app-users-drop-on-battery-problem-hse-idUKKBN25N1PA

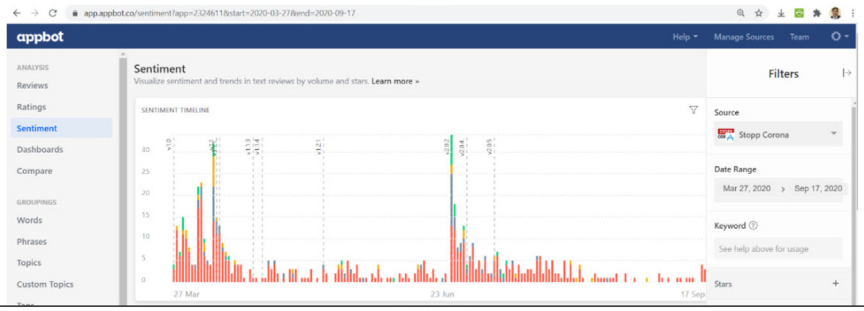
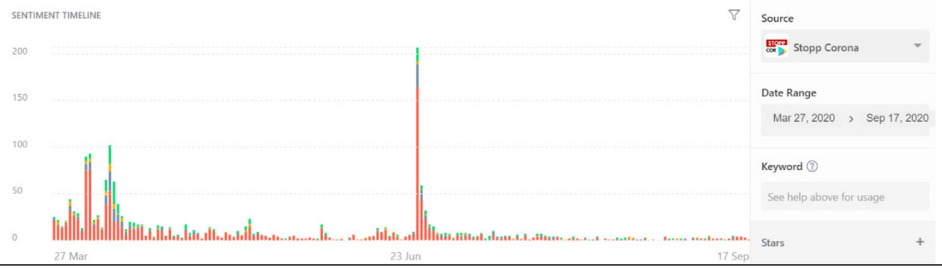
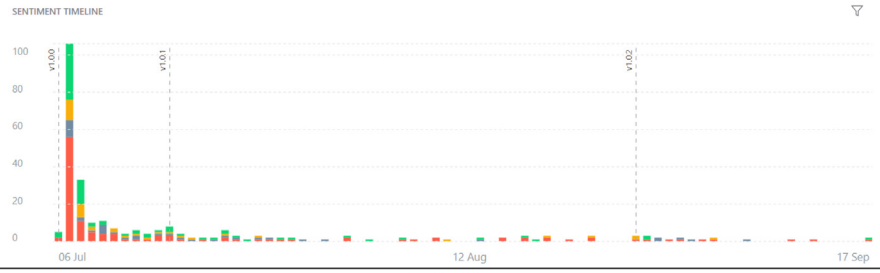
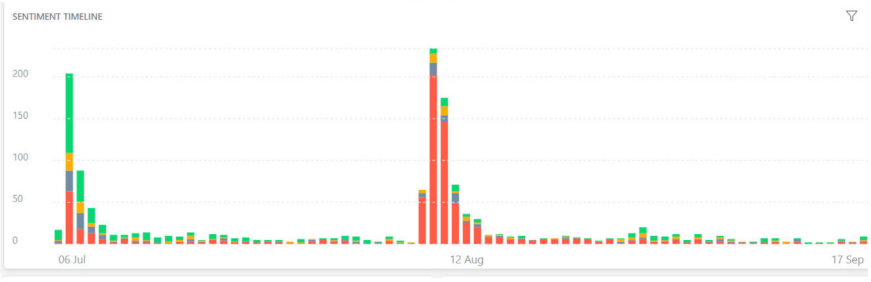
App	Trends of review volumes and their sentiments over time
Stopp-Corona Austria	iOS:  Android: 
COVID Tracker Ireland	iOS:  Android 

Fig. 34. Trends of review volumes and their sentiments over time for the apps.

a critical mass of engaged users" before citing some very public examples including Google+ and iTunes Ping. The study Farronato et al. (2020) hypothesized that most-contact tracing apps will also fail unless significant revision is made to their design and implementation arguing that the possible approach of mandating installation would be very poorly received in liberal democracies, the only possible successful outcome being the widespread optional use through clear demonstration of the value to the individual or community that could be provided. Generally, it seems the barrier to adoption is behavioral rather than technical, with the vast majority of the target audience having a device suitable to install the relevant app (Blom et al., 2021).

Throughout much of the literature is the common concept that potential users must be "sold" on using the app, either through seeing a clear individual benefit or encouragement to behave in a "pro-social" manner for societal benefit, while having their concerns clearly addressed (Trang et al., 2020; Farronato et al., 2020; Walrave et al., 2020).

The most common concern and therefore barrier to adoption raised repeatedly by potential users in different countries was around privacy (Nicholas et al., 2020; Altmann et al., 2020; Redmiles, 2020). While in most countries development was around the decentralized model a lack of information, clarity and transparency was seen to hinder public acceptance (Nicholas et al., 2020; Walrave et al., 2020).

Another paper (Horvath et al., 2020) argued that citizens' concerns about data privacy and data security breaches may reduce the adoption of COVID-19 contact-tracing apps, making them less effective. The paper implemented a choice experiment (conjoint experiment) where participants indicate which version of two contact-tracing apps they would install, varying the apps' privacy-preserving attributes. Citizens do not always prioritise privacy and prefer a centralised National Health Service (NHS), the UK's national health service, system over a decentralised system. In a further study asking about participants' preference for digital-only vs human-only contact tracing, the study found a mixture

Fig. 34. (*continued*).

of digital and human contact tracing is supported. The paper suggested COVID-19 and trust in a national public health service system mitigate respondents' concerns about privacy.

In Kukuk (2020), a theory named the *Unified Theory of Acceptance and Use of Technology (UTAUT)* has been applied to evaluate the COVID tracing apps based on a survey with students from Germany and the Netherlands. Performance expectancy and perceived credibility have been determined to have a significant impact on the intention to use a contract-tracing app in the user base under study. Apart from receiving notifications about possible infections, current contract-tracing apps appear to not provide a clear benefit to the user and are perceived as somewhat privacy-invading. Furthermore, contact-tracing apps might turn out to be a failure, as the study (Kukuk, 2020) finds a low intention to use such apps. We apply a different approach, i.e., analyzing app reviews, to investigate the sentiment about COVID contract-tracing apps.

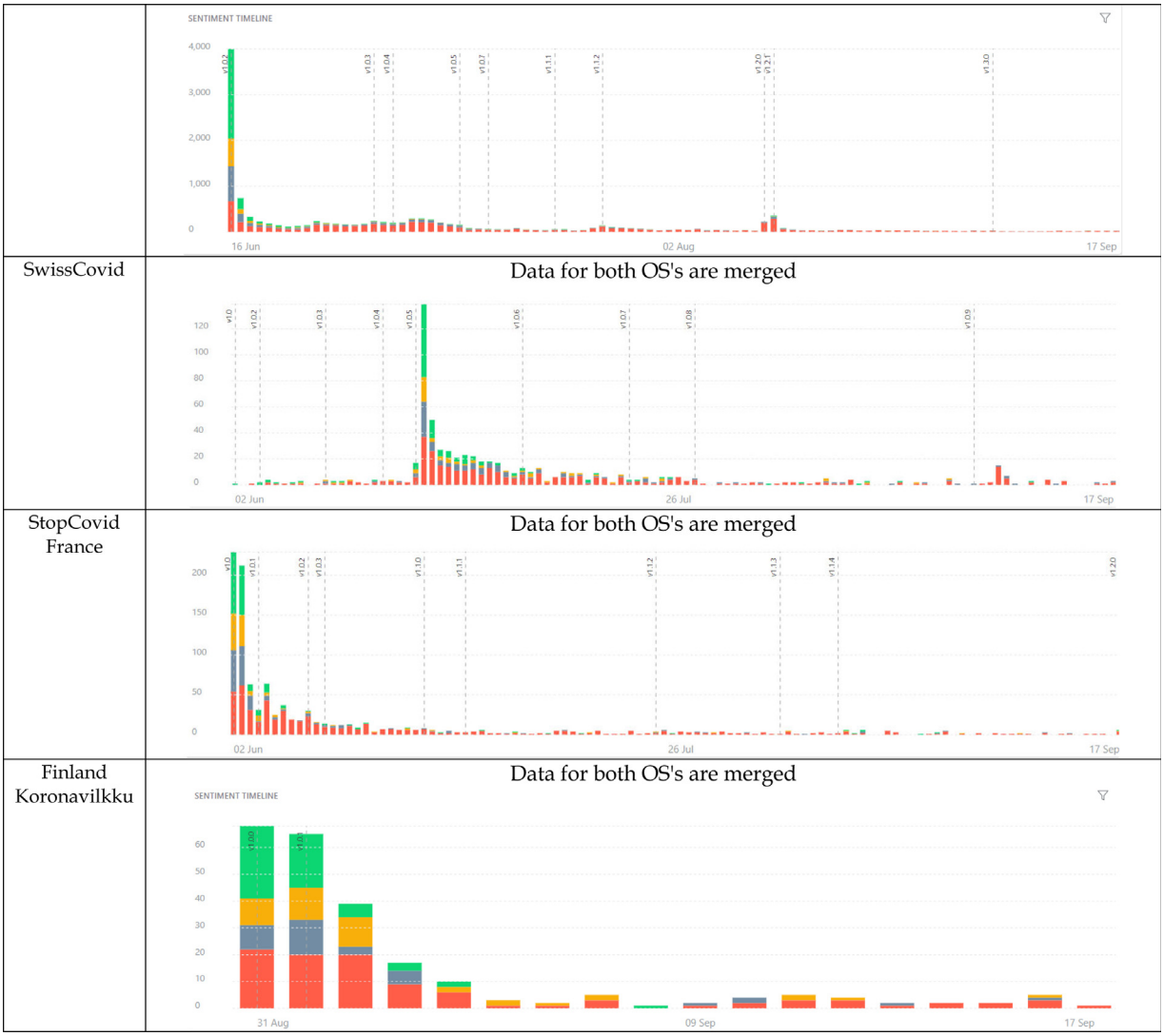


Fig. 34. (*continued*).

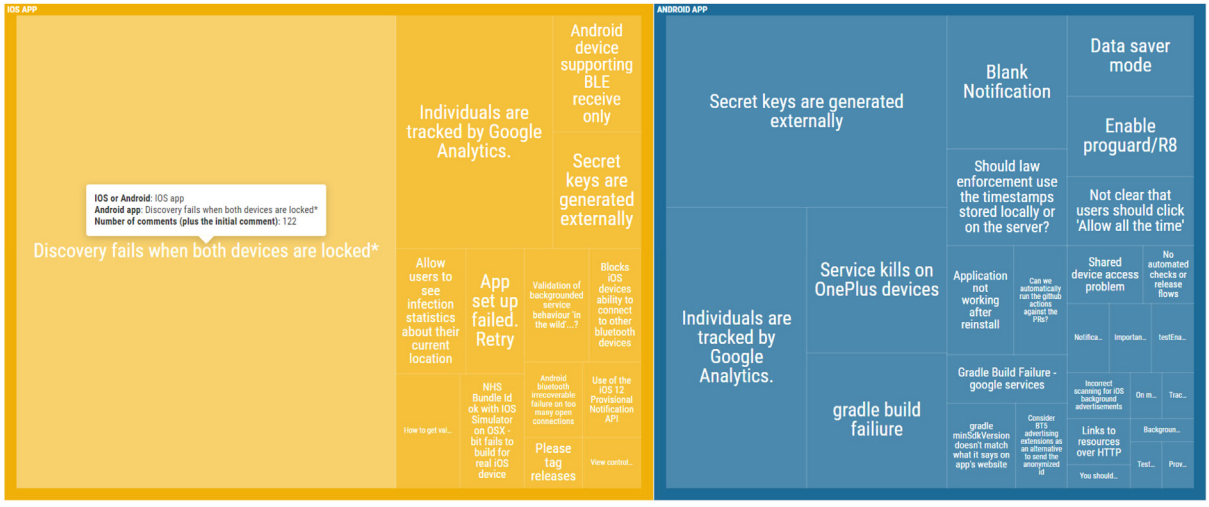


Fig. 35. Issues raised for the NHS COVID-19 iOS and Android BETA app versions on Github. Sized by number of comments posted. *Source:* Reproduced from: flo.uri.sh/visualisation/2515424/embed.

References

Ahmed, N., et al., 2020. A survey of covid-19 contact tracing apps. IEEE Access 8, 134577–134601.

Altmann, S., et al., 2020. Acceptability of app-based contact tracing for COVID-19: Cross-country survey study. JMIR MHealth UHealth 8 (8), e19857.

Anon, 2020. The British Psychological Society, Behavioural Science and Success of the Proposed UK Digital Contact Tracing Application for Covid-19. Technical Report.

Aslam, A., 2015. Research ideas: Correlation does not imply causation. Br. Dent. J. 219 (2), 49.

Baltes, S., Ralph, P., 2020. Sampling in software engineering research: A critical review and guidelines. arXiv preprint arXiv:2002.07764.

Basili, V.R., 1992. Software Modeling and Measurement: The Goal/Question/Metric Paradigm. Technical Report, University of Maryland at College Park.

Blom, A.G., et al., 2021. Barriers to the Large-Scale Adoption of the COVID-19 Contact-Tracing App in Germany.

Braithwaite, I., Callender, T., Bullock, M., Aldridge, R.W., 2020. Automated and partly automated contact tracing: a systematic review to inform the control of COVID-19. The Lancet Digital Health.

Browne, C., Gulbudak, H., Webb, G., 2015. Modeling contact tracing in outbreaks with application to Ebola. J. Theoret. Biol. 384, 33–49.

Budd, J., et al., 2020. Digital technologies in the public-health response to COVID-19. Nature Med. 1–10.

Cho, H., Ippolito, D., Yu, Y.W., 2020. Contact tracing mobile apps for COVID-19: Privacy considerations and related trade-offs. arXiv preprint arXiv:2003.11511.

Danquah, L.O., et al., 2019. Use of a mobile application for ebola contact tracing and monitoring in Northern Sierra leone: a proof-of-concept study. BMC Infect. Dis. 19 (1), 810.

Davidson, R., 1996. 'searching for Mary, Glasgow': Contact tracing for sexually transmitted diseases in twentieth-century Scotland. Soc. Hist. Med. 9 (2), 195–214.

Farronato, C., Iansiti, M., Bartosiak, M., Denicolai, S., Ferretti, L., Fontana, R., 2020. How to get people to actually use contact-tracing apps. Harvard Business Review Digital Articles.

Fenn, J., Raskino, M., 2008. Mastering the Hype Cycle: How To Choose the Right Innovation At the Right Time. Harvard Business Press.

Fernández, D.M., Graziotin, D., Wagner, S., Seibold, H., 2019. Open science in software engineering. arXiv preprint arXiv:1904.06499.

Garousi, V., 2009. Evidence-based Insights about Issue Management Processes: An Exploratory Study. In: Proceedings of the International Conference on Software Process. ICSP, pp. 112–123.

Genc-Nayebi, N., Abran, A., 2017. A systematic literature review: Opinion mining studies from mobile app store user reviews. J. Syst. Softw. 125, 207–219.

Gomez, C., Oller, J., Paradells, J., 2012. Overview and evaluation of bluetooth low energy: An emerging low-power wireless technology. Sensors 12 (9), 11734–11753.

Groen, E.C., Doerr, J., Adam, S., 2015. Towards crowd-based requirements engineering a research preview. In: International Working Conference on Requirements Engineering: Foundation for Software Quality. Springer, pp. 247–253.

Guttal, V., Krishna, S., Siddharthan, R., 2020. Risk assessment via layered mobile contact tracing for epidemiological intervention. medRxiv.

Guzman, E., Ibrahim, M., Glinz, M., 2017. A little bird told me: Mining tweets for requirements and software evolution. In: 2017 IEEE 25th International Requirements Engineering Conference. RE, IEEE, pp. 11–20.

Guzman, E., Maalej, W., 2014. How do users like this feature? a fine grained sentiment analysis of app reviews. In: IEEE international requirements engineering conference. pp. 153–162.

Hayden, F.G., 2006. Policymaking for a Good Society: The Social Fabric Matrix Approach To Policy Analysis and Program Evaluation. Springer Science & Business Media.

Henry, G.T., 1990. Practical sampling. Sage.

Hoepman, J.-H., 2020. A critique of the google apple exposure notification (GAEN) framework. arXiv preprint arXiv:2012.05097.

Horvath, L., Banducci, S., James, O., 2020. Citizens' attitudes to contact tracing apps. J. Exp. Political Sci. 1–13.

Hu, H., Wang, S., Bezemer, C.-P., Hassan, A.E., 2019. Studying the consistency of star ratings and reviews of popular free hybrid android and iOS apps. Empir. Softw. Eng. 24 (1), 7–32.

Husmann, M., Spiegel, M., Murolo, A., Norrie, M.C., 2016. UI testing cross-device applications. In: Proceedings of the 2016 ACM International Conference on Interactive Surfaces and Spaces. pp. 179–188.

Iacob, C., Harrison, R., 2013. Retrieving and analyzing mobile apps feature requests from online reviews, In: Working conference on Mining Software Repositories. pp. 41–44.

Jha, N., Mahmoud, A., 2017. Mining user requirements from application store reviews using frame semantics. In: International Working Conference on Requirements Engineering: Foundation for Software Quality. Springer, pp. 273–287.

Jha, N., Mahmoud, A., 2019. Mining non-functional requirements from app store reviews. Empir. Softw. Eng. 24 (6), 3659–3695.

Joorabchi, M.E., Mesbah, A., Kruchten, P., 2013. Real challenges in mobile app development, In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement. pp. 15–24.

Kazman, R., Pasquale, L., 2019. Software engineering in society. IEEE Softw. 37 (1), 7–9.

Khalid, H., Shihab, E., Nagappan, M., Hassan, A.E., 2014. What do mobile app users complain about? IEEE Softw. 32 (3), 70–77.

Ksir, C., Hart, C.L., 2016. Correlation still does not imply causation. Lancet Psychiatry 3 (5), 401.

Kukuk, L., 2020. Analyzing Adoption of Contact Tracing Apps using UTAUT. University of Twente.

Leith, D.J., Farrell, S., 2020. Measurement-Based Evaluation of Google/Apple Exposure Notification API for Proximity Detection in a Commuter Bus. Technical Report, Trinity College Dublin.

Li, J., Guo, X., 2020. COVID-19 contact-tracing apps: A survey on the global deployment and challenges. arXiv preprint arXiv:2005.03599.

Lim, S.L., Bentley, P.J., Kanakam, N., Ishikawa, F., Honiden, S., 2014. Investigating country differences in mobile app user behavior and challenges for software engineering. IEEE Trans. Softw. Eng. 41 (1), 40–64.

Liu, B., 2012. Sentiment analysis and opinion mining. Synth. Lect. Hum. Lang. Technol. 5 (1), 1–167.

Lu, M., Liang, P., 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. pp. 344–353.

Maalej, W., Nabil, H., 2015. Bug report, feature request, or simply praise? on automatically classifying app reviews. In: IEEE International Requirements Engineering Conference. IEEE, pp. 116–125.

Maalej, W., Nayebi, M., Ruhe, G., 2019. Data-driven requirements engineering-an update. In: 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice. ICSE-SEIP, IEEE, pp. 289–290.

Martens, D., Johann, T., 2017. On the emotion of users in app reviews. In: IEEE/ACM International Workshop on Emotion Awareness in Software Engineering. SEmotion, pp. 8–14.

Martin, T., Karopoulos, G., Hernández-Ramos, J.L., Kambourakis, G., Nai Fovino, I., 2020. Demystifying COVID-19 digital contact tracing: A survey on frameworks and mobile apps. Wirel. Commun. Mob. Comput. 2020.

Martuscelli, C., Heikkilä, M., 2020. Scientists cast doubt on effectiveness of coronavirus contact-tracing apps. https://www.politico.eu/article/scientists-cast-doubt-on-the-effectiveness-of-contact-tracing-apps/, Last (Accessed October 2020).

Morales-Ramirez, I., Perini, A., Guizzardi, R.S., 2015. An ontology of online user feedback in software engineering. Appl. Ontog. 10 (3–4), 297–330.

Nayebi, M., Marbouti, M., Quapp, R., Maurer, F., Ruhe, G., 2017. Crowdsourced exploration of mobile app features: A case study of the fort McMurray wildfire. In: IEEE/ACM International Conference on Software Engineering: Software Engineering in Society Track. IEEE, pp. 57–66.

Nebeling, M., Husmann, M., Zimmerli, C., Valente, G., Norrie, M.C., 2015. XDSession: integrated development and testing of cross-device applications, In: Proceedings of the 7th ACM SIGCHI Symposium on Engineering Interactive Computing Systems. pp. 22–27.

Nicholas, S., Armitage, C., Tampe, T., Dienes, K., 2020. Public attitudes towards COVID-19 contact tracing apps: a UK-based focus group study.

Pagano, D., Maalej, W., 2013. User feedback in the appstore: An empirical study. In: 2013 21st IEEE International Requirements Engineering Conference. RE, IEEE, pp. 125–134.

Potter, K., Hagen, H., Kerren, A., Dannenmann, P., 2006. Methods for presenting statistical information: The box plot. In: Visualization of Large and Unstructured Data Sets, Vol. 4. pp. 97–106.

Redmiles, E.M., 2020. User concerns & tradeoffs in technology-facilitated contact tracing. arXiv preprint arXiv:2004.13219.

Rekanar, K., et al., 2020. Sentiment analysis of user feedback on the HSE contact tracing app. Pre-print https://www.researchsquare.com/article/rs-96174/v1.

Rizzo, E., 2020. COVID-19 contact tracing apps: the 'elderly paradox'. Public health.

Runeson, P., Höst, M., 2009. Guidelines for conducting and reporting case study research in software engineering. Empir. Softw. Eng. 14 (2), 131–164.

Sacks, J.A., et al., 2015. Introduction of mobile health tools to support Ebola surveillance and contact tracing in guinea. Glob. Health Sci. Prac. 3 (4), 646–659.

Scherr, S.A., Elberzhager, F., Meyer, S., 2019. Listen to your users–quality improvement of mobile apps through lightweight feedback analyses. In: International Conference on Software Quality. Springer, pp. 45–56.

Scudellari, M., 2020. COVID-19 digital contact tracing: Apple and google work together as MIT tests validity. IEEE Spectr. 13.

Stoyanov, S.R., Hides, L., Kavanagh, D.J., Zelenko, O., Tjondronegoro, D., Mani, M., 2015. Mobile app rating scale: a new tool for assessing the quality of health mobile apps. JMIR MHealth UHealth 3 (1), e27.
Sun, R., Wang, W., Xue, M., Tyson, G., Camtepe, S., Ranasinghe, D., 2020. Vetting security and privacy of global COVID-19 contact tracing applications. arXiv preprint arXiv:2006.10933.
Trang, S., Trenz, M., Weiger, W.H., Tarafdar, M., Cheung, C.M., 2020. One app to trace them all? Examining app specifications for mass acceptance of contact-tracing apps. Eur. J. Inf. Syst. 1–14.
Wallach, H.M., 2006. Topic modeling: beyond bag-of-words. In: Proceedings of the international conference on Machine learning. pp. 977–984.
Walrave, M., Waeterloos, C., Ponnet, K., 2020. Adoption of a contact tracing app for containing COVID-19: A health belief model approach. JMIR Public Health Surv. 6 (3), e20572.
Wang, S., Ding, S., Xiong, L., 2020. A new system for surveillance and digital contact tracing for COVID-19: spatiotemporal reporting over network and GPS. JMIR MHealth UHealth 8 (6), e19457.
Webber, F.D.S., 2015. Semantic folding theory and its application in semantic fingerprinting. arXiv preprint arXiv:1511.08855.
Wen, H., Zhao, Q., Lin, Z., Xuan, D., Shroff, N., 2020. A study of the privacy of covid-19 contact tracing apps. In: International Conference on Security and Privacy in Communication Networks.
Williams, G., Mahmoud, A., 2017. Mining twitter feeds for software user requirements. In: 2017 IEEE 25th International Requirements Engineering Conference. RE, IEEE, pp. 1–10.
Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B., Wesslén, A., 2000. Experimentation in Software Engineering: An Introduction. Kluwer Academic Publishers.
Zarandy, A., Shumailov, I., Anderson, R., 2020. BatNet: Data transmission between smartphones over ultrasound. arXiv preprint arXiv:2008.00136.
Zhao, Q., Wen, H., Lin, Z., Xuan, D., Shroff, N., 2020. On the accuracy of measured proximity of bluetooth-based contact tracing apps In: International Conference on Security and Privacy in Communication Networks.

Vahid Garousi is an Associate Professor in Queen's University Belfast, and the Director of Bahar Software Engineering Consulting Corporation. His areas of research and practice include empirical software engineering.

David Cutting is a Lecturer in Queen's University Belfast. His areas of research include software evolution, and recovery of design information.

Michael Felderer, is a Professor at the University of Innsbruck, Austria and a Guest Professor at the Blekinge Institute of Technology, Sweden. His areas of research and practice include empirical software engineering.