Designing Positive Behavior Change Experiences: a Systematic Review and Sentiment Analysis based on Online User Reviews of Fitness and Nutrition Mobile Applications

Francisca Pimenta
francisca.pimenta@iti.larsys.pt
ITI/LARSyS
Funchal, Portugal

Laís Lopes
laissantoslopes@gmail.com
ITI/LARSyS
Funchal, Portugal

Frederica Gonçalves
frederica.goncalves@iti.larsys.pt
ITI/LARSyS and University of Madeira
Funchal, Portugal

Pedro Campos
pcampos@uma.pt
ITI/LARSyS and University of Madeira
Funchal, Portugal

ABSTRACT

While mobile devices have become ubiquitous, illnesses derived from poor lifestyle habits are on the rise. However, our understanding of design mechanisms that induce healthier behavior change through mobile devices is still limited. Using the BCT Taxonomy, and online user reviews as an indicator of experience satisfaction, we make a three-folded contribution to designing interactive systems for behavior change: (i) a systematic review of applications for physical activity and healthier eating habits, coding BCTs; (ii) sentiment analysis performed on 20492 review sentences of these apps; and (iii) design implications regarding the implementation features for each BCT cluster, considering the highest-scored features in terms of sentiment analysis. Positive expressions referred to the framing/reframing technique. Contrarily, negative expressions were mostly related to reward and threat. Findings from this study can be used to benchmark interactions between users and behavior change interfaces, and provide design insights to support positive user experiences.

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in HCI*; *Human computer interaction (HCI)*;

KEYWORDS

behavior change systems, mHealth, mobile apps, persuasive technologies, text analysis, user experience, user reviews

ACM Reference Format:
Francisca Pimenta, Laís Lopes, Frederica Gonçalves, and Pedro Campos. 2020. Designing Positive Behavior Change Experiences: a Systematic Review and Sentiment Analysis based on Online User Reviews of Fitness and Nutrition Mobile Applications. In *19th International Conference on Mobile and Ubiquitous Multimedia (MUM 2020), November 22–25, 2020, Essen, Germany.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3428361.3428403

1 INTRODUCTION

The World Health Organization reported that cardiovascular diseases (CVDs) are the leading cause of mortality worldwide, accounting for 31% of all deaths in 2016. Effective prevention strategies for most CVDs include tackling physical inactivity and unhealthy diet choices [54].

Behavioral change interventions addressing these risk factors could drastically reduce overall mortality from non-communicable diseases such as CVDs [53]. The widespread adoption of mobile phones highlights a significant opportunity to impact health behaviors globally. Mobile health (mHealth) applications are more accessible and reach a broader audience than traditional face-to-face interventions [44]. To illustrate, fitness and nutrition are the most popular categories of downloaded applications from the leading app stores. Despite their popularity, findings from previous research suggest that while these apps are used by many, about half of the users eventually abandon them [33]. A quite low adoption rate for these apps suggests a struggle to sustain interest after the novelty effect has worn off [34].

Concerning the mechanisms that drive behavior change, it is apparent that the inclusion of a theoretical foundation enhances the effectiveness of health interventions [23]. Despite ample theory on behavior change, studies have shown that the majority of mHealth solutions often lack a foundation in behavioral change theory, or have implemented a limited number of behavior change techniques (BCTs) [13], [31]. A BCT is *"an observable, replicable, and irreducible component of an intervention designed to alter or redirect causal processes that regulate behavior; that is, a technique is proposed to be an "active ingredient" (e.g., feedback, self-monitoring, and reinforcement)"*. BCTs can be used alone or in combination. [41].

Data is also limited regarding user perceived satisfaction on specific behavioral change features of fitness and nutrition apps. Studies exploring associations between the number of BCTs implemented—an indicator of likely efficacy—and user perceived satisfaction, rely on user ratings rather than user reviews (a possible proxy

MUM 2020, November 22–25, 2020, Essen, Germany

ACM ISBN 978-1-4503-8870-2/20/11.
https://doi.org/10.1145/3428361.3428403

for health benefits) [6]. Reviews submitted to app stores allow users to provide feedback that may contain information related to the features they value the most. They allow to qualitatively account for the users' different opinions at a much more detailed level than ratings. Reviews also pose an easy and cheap way to gather data from a large sample of end-users. For the purpose of this paper, a *review* is a piece of text written by an ordinary end-user, self-reporting their experience of a mobile application, in their own words, in the wild. In order to process information from a large pool of reviews, we focused on extracting feature information from expressions classified within review text as "positive" or "negative" through sentence-level sentiment analysis. These expressions were coded to a standardized taxonomy for behavior change techniques.

Our goal is to better understand how fitness and nutrition applications' features support behavior change and which popular features associated to which BCT clusters elicit a positive response from users. In this paper we present a systematic review of applications for physical activity and healthier eating habits, coding their features to BCTs; a sentiment analysis performed on 20492 review sentences of these apps; and, based on these, we draft pragmatic guidance for application designers regarding the implementation features for each BCT cluster, considering user perceived satisfaction. We discuss our intervention at the intersection of behavior change theory and user feedback.

2 RELATED WORK

2.1 mHealth Behavior Change Interventions

A study by Ting-Ray Chang et al. [8] showed that the main reasons that people use a mobile application to support healthier behavior change were related to its perceived value. Some of the factors that led people to engage with these applications include attractiveness, value, ease-of-use, trust, social support, diffusiveness, as well as fun and excitement. The study's participants appreciated some motivational features and thought that these applications could help in changing their habits.

A previous study has established that top-ranked physical activity apps mention a limited number of behavior change techniques in their descriptions. It has also revealed the existence of two types of apps, according to implemented behavior change techniques: educational and motivational. The most common techniques found in physical activity apps were educational and involved providing information or demonstrations of specific physical activities [11]. Knowledge about how to perform a desired behavior is a necessary step to behavior change because it contributes to task self-efficacy, facilitating the formation of intentions to be physically active. However, information is usually not sufficient to drive behavior change. Users are more engaged with a fully interactive application than an information-based application [20]. Therefore, the inclusion of behavior change techniques for bridging the intention—behavior gap, would be beneficial to provide further motivational support.

2.2 Reporting Behavior Change Interventions

Reporting interventions with clarity and detail is crucial for the design of solutions optimized to motivate healthier behaviors. Nevertheless, given the fast advancements of mHealth research, it may be challenging to develop a large evidence base for individual mobile apps. Without a clear and shared understanding of how to best describe the content of implementation interventions, a number of risks emerge: using the same content description to represent different types of content; using different terms to represent the same content [10]; using levels of description that are not sufficiently specific to allow replication [41]; repetition/reinvention without progress [49]; and missed opportunities to draw on techniques used effectively in other settings. Comprehensive classification systems with agreed definitions could address these issues. Thus, taxonomies of behavior change techniques provide a useful tool to report, replicate and synthesize behavior change interventions based on potentially effective techniques [39].

Michie et al. [41] developed the BCT taxonomy v1 (BCTTv1) [2]. BCTTv1, on which we relied to code the interventions featured in this paper, lays the foundation for the systematic specification of behavior change interventions, increasing the chances of identifying the active ingredients and the conditions under which they are effective. Its 93 techniques are grouped according to similarity of active ingredient. The groups are as follows: *goals and planning*, *feedback and monitoring*, *social support*, *shaping knowledge*, *natural consequences*, *comparison of behavior*, *associations*, *repetition and substitution*, *comparison of outcomes*, *reward and threat*, *regulation*, *antecedents*, *identity*, *scheduled consequences*, *self-belief*, and *covert learning*. A study by Wood et al. [52] has evaluated the application of BCTTv1 to code BCTs in intervention descriptions and demonstrated the reliability of the classification system.

Michie et al. [40] offer clear support for including techniques from the goals and planning and the feedback and monitoring groups of techniques. Interventions including self-monitoring and at least one of four other self-regulatory techniques were significantly more effective than interventions not including these techniques. However, the success of BCTs also depends on factors such as how they are delivered, the targeted behavior, and context. A myriad of reviews by researchers from different disciplines have also examined the effectiveness of mHealth interventions when targeting one specific behavior (e.g. increasing physical activity) [5], [7], [38]. Zhao et al. [55] explored health-related behavioral change across a broader range of health issues. Previous reviews have also identified appropriate BCTs and BCT combinations for use in the health setting to facilitate the development of more effective gamified interventions [16]. Considering many BCTs occur together in a given intervention, it is difficult to pinpoint which BCTs or BCT combinations promote user perceived satisfaction.

2.3 Text Mining

Text mining is the process of analyzing text to extract information that is useful for particular purposes [51]. In other words, looking for patterns in text. While the effectiveness of interventions for behavior change has been extensively covered, user perceived satisfaction regarding features to support behavior change is a less explored topic. Text mining is a valuable tool to analyze data such as user reviews. Researchers from several fields have resorted to text mining techniques to analyze large text datasets [24], [26]. Liu et al. [36] conducted an analysis to compare the underlying trends in CHI community between 1994 and 2013. Their study identifies the evolution of major themes in the discipline and highlights individual topics as popular, core or backbone research topics within HCI. However, it remains difficult to automatically summarize the large amount of feedback for popular apps and make sense out of it [21]. Especially for app designers and developers, the timely understanding of user experience and integration into release cycles is crucial. Fu et al. [21] proposed Wiscom, a system that can analyze tens of millions of reviews in mobile app stores at three different levels of detail. Their system is able to automatically find inconsistencies in reviews, identify reasons why users like or dislike a given app, provide a view of how users' reviews evolve over time and identify users' concerns and preferences of different types of apps. The work of Chen et al. [9] went in a similar direction, with a system that prioritized and presented the groups of most "informative" reviews to developers.

2.3.1 Understanding User Reviews. Mining online user reviews has also been studied previously as a method to obtain user experience information [27]. Unlike typical methods that measure user experience, which usually consist of structured studies designed to gather particular pieces of information (e.g. questionnaires and focus group studies), spontaneous experience reports (i.e. online user reviews) disclose experiences that users themselves consider to be meaningful [32]. Pagano et al. [42] found that the information in textual feedback could be classified into four groups: community (references to other feedback or apps, questions, explanations and recommendations addressed to other users); requirements (feedback disclosing the user expectations of the app's performance and improvement requests); rating (praise and dispraise); and user experience (feature information and descriptions of the app in action, including use cases where the application has proven helpful). Shipman et al. [45] also explored the practices of reviewers. Their research found that there are five common reasons that motivate users to write reviews: a desire to share a positive experience; a desire to warn people about a negative experience; a desire to contribute to community knowledge; a perceived need to bolster or detract from an individual's online reputation; or when other (non-monetary) incentives are offered in exchange for reviews. However, other researchers [48] have found that users are more likely to leave longer reviews when they rate an app poorly.

2.3.2 Sentiment Analysis in User Reviews. Sentiment analysis is useful to uncover the emotional tone of users' reviews, which often include user sentiments about specific features and descriptions of their experiences with such features [25]. Research in this area is already well established. Pang et al. [43] classify movie reviews as "positive" or "negative" at full-text level. This approach provides information regarding whether a review is positive or negative, but lacks granularity. Pulse [22] is a prototype system that goes a step further by working at sentence-level. It was tested by being applied to a dataset of car reviews in order to identify topic and associated sentiment, and was proven to allow its users to explore large quantities of text both "at a glance" or at a finer level. The work of Guzman et al. [25] identified fine-grained app features in the reviews and extracted the user sentiments about these features, giving them a general score across all reviews. Research by Luiz et al. [37] proposed a similar framework that extracted relevant information (e.g., information about functionalities, bugs, requirements, etc) from reviews of apps and analyzed the associated sentiment. This work provides designers and app developers a richer understanding of user feedback than a star rating strategy.

3 METHOD

3.1 Data Collection

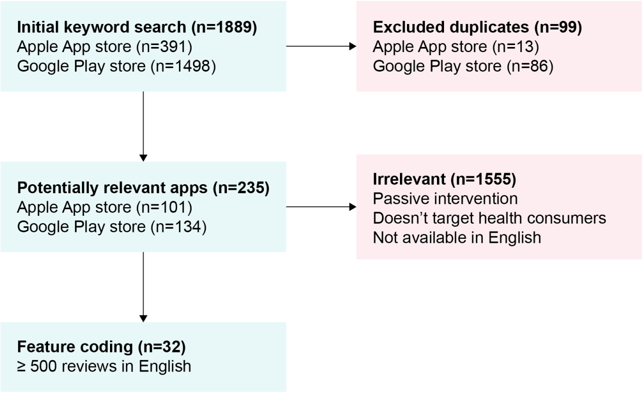
Initial keyword search (n=1889)
Apple App store (n=391)
Google Play store (n=1498)
Excluded duplicates (n=99)
Apple App store (n=13)
Google Play store (n=86)
Potentially relevant apps (n=235)
Apple App store (n=101)
Google Play store (n=134)
Irrelevant (n=1555)
Passive intervention
Doesn't target health consumers
Not available in English
Feature coding (n=32)
≥ 500 reviews in English

Figure 1: Flowchart of the application selection process.

3.1.1 Keyword Search. To retrieve data from the app stores, we employed a similar process to other reviews [3]. We chose terms that would potentially retrieve motivational apps (i.e., those that included well-established behavior change techniques for bridging the intention-behavior gap such as *action planning*). "Pedometer", for instance, would potentially retrieve step counting apps that only provided information on how many steps the user has taken. On the other hand, "pedometer game" retrieved apps with additional features such as the ability to set a daily goal of steps. On September 11, 2019, we crawled all categories of Apple App Store and Google Play Store for the terms "exercise motivation", "habit forming", "physical activity", "stepcounter game" and "pedometer game" using pre-existing scripts [17], [18]. The search returned 391 results from Apple App Store and 1498 results from Google Play Store. After removing duplicates, we were left with 378 results from Apple App Store and 1412 results from Google Play Store. Applications were considered potentially relevant based on their title and store description if they met the following criteria:

- Active interventions that engage users in the process of behavior change;
- Target health consumers (i.e., those who self-identify as needing support to be less sedentary and practice healthier eating habits), rather than health care and physical education professionals;
- Be available in English.

The exclusion criteria were as follows:

- Rely heavily on self-reporting;
- Passive interventions that only provide non-tailored instructions or advice (e.g., premade workout programs, premade diet programs, educational content, information intended for performance optimization of athletes);

- Only track progress related to specific sports or require equipment;
- General habit-forming applications;
- Outcome-oriented applications (e.g., lose weight);
- Provide only non-contextual notifications (e.g., fitness quotes applications);
- Duplicate (e.g., free and premium versions with no difference in features).

The first filtering was performed by two researchers (co-authors) based on the title of the application, description and languages available. The next stage was to download and use the applications which features were not clear in the description, at least until the point that one of the above criteria were not satisfied. After filtering, 101 results from Apple App Store and 134 from Google Play Store remained to analyze. We then selected the applications that had at least 500 reviews written in English, leaving us with 32 apps (Table 1).

3.1.2 Acquisition of App Store Reviews. To determine which behavior change features were positively or negatively received by users, our approach was to qualitatively analyze online user reviews. For acquiring an extensive sample of app store user reviews for the selected apps we employed web scraping techniques. We scraped 500 user reviews from each app of our sample of applications—totaling 16000 reviews—on December 3, 2019, using Appfollow [1], an online platform that provides automated worldwide review gathering. However, we cannot assume that a review focuses on a single topic; individual sentences within a review may express different opinions. For instance, if a reviewer rates an application 4 out of 5, the review is likely to contain mostly positive remarks about the application, and a few negative remarks. While mobile app reviews often contain useful user feedback, manual analysis of those reviews is challenging due to their large volume and noisy-nature. Therefore, the review corpus was fed to the user research platform Dovetail [15], where a sentence-level sentiment analysis was performed. This led to a dataset of 20492 review sentences that belonged to the selected apps. It should be noted that the analyzed dataset does not include all reviews. Instead, this work only included reviews that did not merely provide a star rating (we ignored the ratings accompanying the reviews because ratings reflect on review-level, and our analysis targets specific feature mentions within reviews), but also contained an English review text. In addition, due to the time-consuming nature of this endeavor, we obtained the first 500 reviews that were sorted in the default order ("newest first") by Appfollow. While the problem of selecting a comprehensive set of reviews has been addressed by systems that rank reviews according to their estimated helpfulness, such approaches do not account for the fact that the top reviews may be redundant, repeating the same information, or presenting the same positive (or negative) perspective [47]. Therefore, the newest reviews potentially provide the least biased information regarding the last available version of each application. Because of this approach, we had a slightly more complete picture of apps with close to 500 reviews, since the scraping process was likelier to catch all of their available reviews. However, especially for very popular apps with many reviews, comments repeat, and thus do not add as much information as reviews for less popular apps [28].

3.2 Analysis

3.2.1 Popular Keywords and Terms from Review Sentences. The 20492 review sentences previously tagged as "positive" or "negative" were then further mined using Sketch Engine online app [35], a tool for text analysis that allows to search for common themes within the positive and the negative sentences. We first reviewed two lists of single-words ordered by frequency for the positive and the negative tags. Considering that the single-word lists did not enlighten us regarding the context of each word, we also used the term extraction tool to extract keywords. Keywords are terms (sometimes multi-words) that are typical of a corpus or define its content or topic. They provide us a better understanding of the reviews' content. Keywords of the two corpora were compared, identifying what is unique in the positive tags compared to the negative tags. This process returned two keyword lists: the most characteristic terms of the positive tags list and the most characteristic terms of the negative tags list. Adverbs, verbs, pronouns, conjunctions and prepositions were not considered, as these did not provide relevant information. We focused on structures containing adjectives + nouns, which are the target of user opinions. The authors then continued to filter the lists, excluding generic terms that had no possible connection to applications' features (e.g., "app") and keywords that had less than 5 hits.

3.2.2 Feature Coding. The 32 applications from our search yields were then further examined by two researchers. Behavior change features found in these apps were iteratively listed. For the purpose of this paper, we define *feature* as a notable property of a device or software application. The two researchers then independently coded the features using the BCTTv1. Behavior change techniques were classified as either present or absent. Where both researchers identified the BCT as present or absent, agreement was recorded and where one researcher identified the BCT but the other did not identify the BCT, disagreement was recorded. After both researchers finished coding, a consensus conference was held [12]. The frequency of individual techniques and BCT clusters included in the apps was counted. We also noted the number of BCTs and BCT clusters included per application.

The Kappa statistic was used to assess the interrater reliability of the coding. There was substantial agreement between two coders (.833).

4 RESULTS

4.1 Text Mining

The dataset shows a clear skew towards positive reviews: positive expressions comprise 68% of the 20492 tagged expressions. StepsApp Pedometer had the highest%age of positive tags (95%), while VeryFitPro had the lowest (23%).

Considering the entire corpus of tags (positive and negative), there are some words that may refer to applications' features and are common to both lists, suggesting trends regarding what users usually point out in written reviews. Some of the most mentioned words in both lists are related to techniques from the identity cluster, particularly the framing/reframing technique, e.g., "game" (n=1958); to the shaping knowledge and repetition and substitution clusters, e.g., "workout" (n=493); to the goals and planning cluster, e.g.,

Table 1: Selected applications.

Application	Genres	App store
StepsApp Pedometer & Step Counter	Health & Fitness	Google Play Store
Keep Trainer - Workout Trainer & Fitness Coach	Health & Fitness	Google Play Store
Standland	Health & Fitness, Games, Simulation	Apple App Store
Pedometer for walking - Step Counter	Health & Fitness	Google Play Store
PUMATRAC - Fitness Training, Workouts & Running	Health & Fitness	Google Play Store
Wokamon - Fitness Game	Health & Fitness, Games, Simulation, Adventure	Apple App Store
Accupedo Pedometer - Step Counter	Health & Fitness	Google Play Store
Pocket Plants	Simulation	Google Play Store
Workout Diary - Trainings plan - Fitness tracker	Health & Fitness	Google Play Store
Orna: The GPS-RPG	Role Playing	Google Play Store
Zombies, Run! (Free)	Health & Fitness	Google Play Store
StepUp Pedometer Step Tracker: Step Up Fitness!	Health & Fitness	Google Play Store
Yodo - Cash for walking & running	Health & Fitness	Google Play Store
Walkr: Fitness Space Adventure	Adventure	Google Play Store
Noom: Health & Weight	Health & Fitness	Google Play Store
StepBet: Walk, Get Active, Win	Health & Fitness, Games	Apple App Store
Health & Fitness Tracker with Calorie Counter	Health & Fitness	Google Play Store
Fitness RPG - Gamify Your Pedometer	Health & Fitness	Google Play Store
winwalk pedometer - walk, run, sweat & win rewards	Health & Fitness	Google Play Store
Lympo - Walk. Run. Earn.	Health & Fitness	Google Play Store
Pedometer Step Counter - Fitness Tracker	Health & Fitness	Google Play Store
Sweatcoin Pays You To Get Fit	Health & Fitness	Google Play Store
Fit Radio Workout Music & Coach	Health & Fitness	Google Play Store
Sworkit Fitness – Workouts & Exercise Plans App	Health & Fitness	Google Play Store
Run An Empire	Health & Fitness, Strategy, Games, Role Playing	Apple App Store
Walkroid - simple pedometer	Health & Fitness, Strategy, Games, Role Playing	Google Play Store
Pedometer 2.0	Health & Fitness, Strategy, Games, Role Playing	Google Play Store
Pokémon GO	Adventure	Google Play Store
One You Active 10 Walk Tracker	Health & Fitness	Apple App Store
LifeCoin - Rewards for Walking & Step Counting	Health & Fitness	Google Play Store
Harry Potter: Wizards Unite	Adventure	Google Play Store
VeryFitPro	Health & Fitness	Google Play Store

"goal" (n=278); to the feedback and monitoring cluster, e.g., "track" (n=181); and to the reward and threat cluster, e.g., "money" (n=270). While "game", "workout", "goal" and "track" are markedly positive, "money" has more negative than positive tags.

Popular keywords from the positively scored tags are mostly related to the identity cluster, particularly the framing/reframing technique, e.g., "great game" (n=135), "fun game" (n=67), "cute game" (n=42), "amazing game" (n=27), "awesome game" (n=27), "little game" (n=19), "story line" (n=17), "nice game" (n=12), "cool game" (n=11), "wonderful game" (n=9), "fantastic game" (n=8), "great story" (n=8), "enjoyable game" (n=7), "fun little game" (n=7), "idle game" (n=7), "favorite game" (n=6), "good story" (n=6) and "interesting story" (n=5). The list also includes terms that refer to the goals and planning cluster, e.g., "good workout" (n=12), "great program" (n=11), "step goal" (n=8), "home workout" (n=7), "good program" (n=6), "great step counter" (n=5), "activity level" (n=5); to the repetition and substitution cluster, e.g., "exercise routine" (n=5); to the social support cluster, e.g., "great community" (n=6), "good reminder" (n=5); and to the reward and threat cluster, e.g., "reward system" (n=5). Keywords of the negative tags list refer mostly to technical issues during application usage, not mentioning specific features.

4.2 Feature Coding

This review and sentiment analysis identifies BCT clusters that are consistently implemented in mobile applications for physical activity and healthier eating behavior change. Feedback and monitoring techniques were present in 25 of the reviewed applications, goals

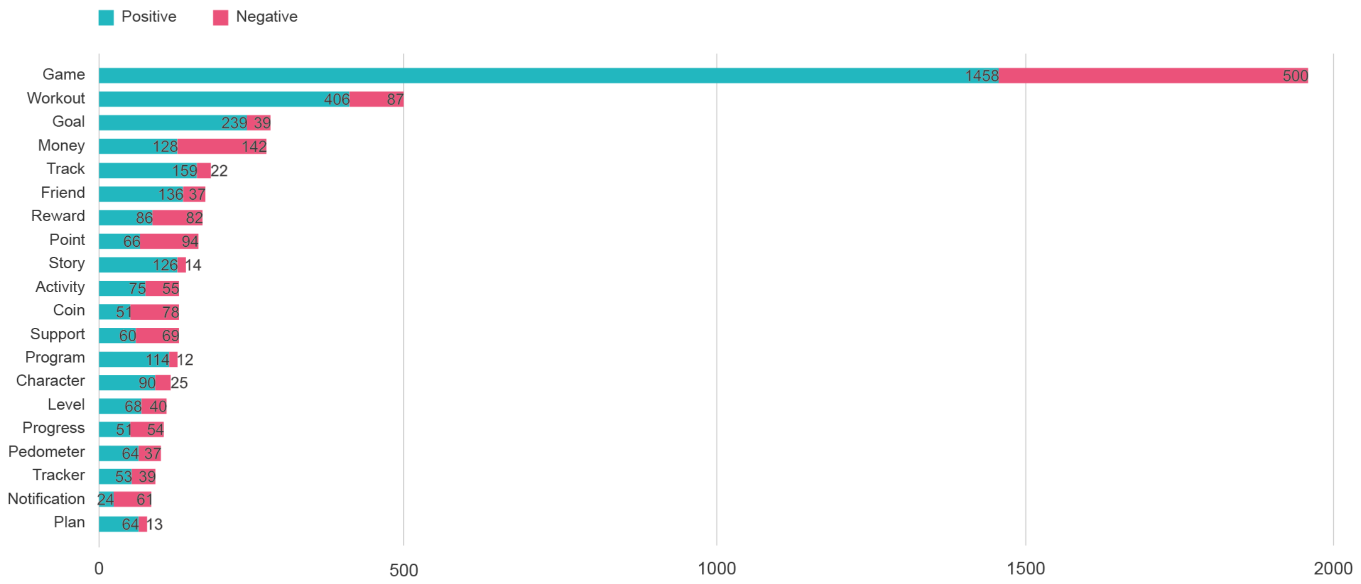

Figure 2: Frequent words potentially related to application features, and corresponding sentiment analysis.

and planning techniques in 19, and reward and threat techniques in 18. Previous studies also found techniques that fall into these clusters to be commonly used in apps targeting physical activity and healthy eating [50]. Comparison of behavior techniques were included in 13 applications, social support techniques in 11, shaping knowledge and identity techniques in 8, repetition and substitution techniques in 7, associations techniques in 5, regulation techniques in 3, and natural consequences and covert learning techniques in 1. Comparison of outcomes, antecedents, and self-belief clusters were not recruited.

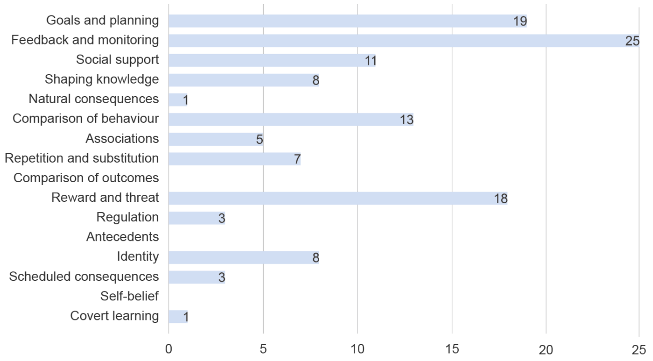

Figure 3: Frequencies of BCT clusters identified in the 32 reviewed applications.

Our review highlights variation in the use of behavior change techniques; however, all applications reviewed include at least 1 recognized BCT. The average number of BCTs included per app is 7 (range: 1- 13). Regarding BCT clusters, all applications reviewed included at least 1 and the average number of BCT clusters per app is 4 (range: 1-7). The most common BCT clusters are feedback and monitoring (n=25), featured in 78% of the reviewed apps; goals and planning (n=19), featured in 59% of the apps; and reward and threat (n=18), featured in 56% of the apps (Figure 3).

5 DISCUSSION

Methodological guidelines for intervention reporting emphasize describing intervention content in detail. In this paper we applied the Behavior Change Techniques Taxonomy version 1 (BCTTv1) to a sample of fitness and nutrition mobile apps for behavior change, assessing the capacity and utility of this taxonomy for characterizing active ingredients. We also reported on the sentiment analysis of particular behavior change features of these mobile apps, through online user reviews.

Overall, the data suggest that users respond well to fitness and nutrition mobile applications with features intended to promote and support behavior change.

Gamification features seem to leave a strong impression, as they are mentioned very often in reviews. Popular term “game” (n=1958) was coded to the identity cluster, specifically the framing/reframing technique, which is *“the deliberate adoption of a perspective or new perspective on behavior (e.g., its purpose) in order to change cognitions or emotions about performing the behavior”* [2]. We considered gamification features belonged to this cluster, as they work by framing the desired behavior like a game rather than a task. However, “game” is a broad term and, taken out of context, does not yield very meaningful information. It can refer, for instance, to the inclusion of game-like elements based on reward and threat (e.g., conditional rewards); more information would be needed to code it reliably.

Identity-based features may be undervalued (i.e., they are featured in less than half of the reviewed applications, but present high sentiment analysis scores, as can be seen in Figure 2). Although the use of game-like elements can consist solely of point attribution systems as users act in accordance with the desired behavior, users seem to be fonder of storytelling. To further understand why

Table 2: Relation between BCT clusters, the online review terms with the most positive sentiment analysis score, and examples of implementation features.

BCTTv1 cluster	Most positive word(s)	Implied feature example
Goals and planning	goal; mission	Set a daily goal of steps tailored to the user's level of activity. Set up a mission, specifying the context, duration and intensity of the activity.
Feedback and monitoring	track	Record activity and provide stats. Provide personalized feedback from a coach based on the user's in-app activity and progress.
Social support	community; group; coach; friend; team	Provide access to a dedicated forum. Allow the user to join or create teams in-app to win extra points.
Shaping knowledge	program, plan, routine, workout	Set a meal plan and provide healthy recipes with instructions on how to make them.
Comparison of behavior	community, group, friend, team	Display videos of workout routines. Include in-app leaderboard.
Associations	reminder	Send notifications prompting the user to stand when they have been inactive for too long.
Repetition and substitution	program, plan, routine	Provide tools for scheduling workouts for each week.
Reward and threat	bonus, achievement, reward	Inform the user that they will win achievements as they comply with the desired behavior.
Identity	storyline; story; character; game	Create an alternative narrative to motivate the user to perform the target behavior.
Self-belief	activity	Show user stats over time, highlighting improvements. Display a list of achievements.

some of these words are mentioned, we also checked the keywords lists. Representative keywords of the positive list include terms such as “story line”, “great story” and others that confirm a liking for intricate narratives. These findings suggest that such features may present a high intrinsic motivation building potential. In a world where mobile devices have become ubiquitous, people expect mobile application interventions to provide rich, immersive experiences. Therefore, crafting narratives as the backbone of applications seems to please users. Some reviews specifically report narratives to be motivating:

“I usually hate running and I get tired really easily but this app motivates me and keeps me going. The story line keeps getting more intense and interesting and keeps your adrenaline high! I'm just using the free version for now.”

”Fantastic app great story can't wait to run tomorrow.”

Social support techniques also seem to be well received. These would translate into features that foster interpersonal relationships and the sense of belonging to a group; features that help users feel supported and cope with challenges and setbacks. For instance, the following was reported:

“So much amazing content, hugely motivational and the gateway to a great community of supportive runners of every ability. The best running app there is!”

”There is a great community to support you.”

Other features that were welcomed by users were having access to a workout program with instructions and being able to set up a tailored step goal (features that employ techniques from shaping knowledge, repetition and substitution, goals and planning, and feedback and monitoring clusters).

Features based on goals and planning and feedback and monitoring techniques are already quite mainstream in the mHealth for behavior change setting. Nevertheless, there is an opportunity to take advantage of shaping knowledge and repetition and substitution techniques and deliver applications with a focus on educational content, as users reported satisfaction regarding tools that prompt the planning of a routine and deliver instructions. In light of previous findings [20], educational content should, however, be delivered in the form of active interventions, as research has shown that users are more engaged with a fully interactive application than an information-based application.

As expected, some of the most common words mined from the online user review corpus—“money” (n=270), “reward” (n=168), “point” (n=160) and “coin” (n=129)—relate to the very frequently implemented reward and threat cluster. Reward and threat techniques consist of arranging for the delivery of a reward as a consequence of performing the desired behavior, or the removal of such reward if the individual engages in unwanted behavior. An example of a design feature in the mobile app setting is, for instance, awarding the user an achievement for each day that they walk a certain amount of steps. Even though reward and threat techniques are quite common in the reviewed applications, they seem to elicit mixed feelings from users. Possibly because they consist of external reinforcement mechanisms, which means that the users' behavior is extrinsically motivated. Extrinsically motivated users will exercise to receive a reward, while intrinsically motivated users will exercise because they find it fun and satisfying. Extrinsically motivated users will continue to comply with a behavior, but it might not be in itself rewarding, therefore yielding poor user perceived satisfaction. Although external rewards were negatively scored, they can be useful tools in interventions targeted at users that have little initial interest in increasing their physical activity and stick to healthier eating habits, or lack the skills to do so. Designers should include them in this setting of interventions with caution, assessing user feedback.

This work shows that mobile applications for behavior change in the physical activity and healthy eating domain often deliver features based on feedback and monitoring, goals and planning, and reward and threat behavior change techniques. However, features based on other clusters, such as identity, social support and shaping knowledge, are well received by users. We highlight BCT clusters that are not implemented as often but leave users feeling happier.

Table 2 summarizes these findings, outlining design feature suggestions for each BCT cluster, inferred from the most positively scored related term.

5.1 Ethical Concerns

Albeit their popularity, technologies that aim to shape the users' behavior pose some ethical concerns. Are these types of interventions manipulative? Numerous frameworks for ethics (e.g., [4], [14], [19], [29], [46]) have been explored in the HCI community. Methods and research approaches such as critical design, reflective design and adversarial design take on an ethical standpoint. Berdichevsky and Neuenschwander [4] were among the first researchers to propose guidelines, and suggest that users should only be persuaded of something they themselves would consent to be persuaded to do. While these efforts have been shown to be effective in the research context, they have less clear implications for design practitioners [30]. Thus, we see an opportunity to connect theory and practice. Analyzing user reviews may be a viable method to probe users regarding their opinion on persuasive mobile apps and move towards a practical application of the research conducted so far.

5.2 Strengths and Limitations

There are several strengths of this work. Firstly, we conducted a systematic review using broad search terms to increase the probability of identifying all eligible mobile applications. Secondly, intervention contents were reliably coded used a standardized taxonomy for behavior change techniques. Thirdly, we used a quite large review corpus (identifying a set of "important" reviews may lead to non-representative results). This corpus was analyzed through automated feedback mining, eliminating biases related to human analysis. Lastly, we present replicable suggestions of design features inferred from the highest scored terms of online user reviews, and link these to BCT clusters. From a methodological perspective, this research introduces a new method that leverages text mining techniques to reveal design implications from online user-generated content in the physical activity and nutrition setting. This experimental work process could also be deployed to inform the design of applications for behavior change in other domains.

There are a few limitations associated with this work. Regarding user feedback, the average user is underrepresented among reviewers, and thus reviews may not always report the typical user experience. However, satisfaction extremes are well represented, rendering user reviews a viable source of information to enlighten fitness and nutrition app developers as well as health practitioners and researchers about situations where the product under review performs good or bad [27]. In addition, literature has shown that in general, the review length of mobile apps seems to be influenced by the category of the app. The median of review length in the health & fitness category was 100 characters, twice as long as the median for games [48]. Concerning our findings, this bias might mean we might encounter fewer quality comments regarding applications categorized as games. Regarding the text mining protocol, the current analysis is subject to text mining disadvantages such as a possible lack of accuracy factoring for spelling errors and other inconsistencies. Alas, this may sacrifice the narrative structure of reviews, potentially leaving out meaningful information about what exactly led the user to positively (or negatively) review a certain feature. Furthermore, automated text mining removes some biases, but can't detect spam or fake reviews. This analysis considers BCT clusters (groups that aid the recall of BCTs), leaving out the differentiation between specific techniques and therefore may lack granularity.

6 CONCLUSION

Much has been written about the fields of behavior change and text mining on their own. Combined, however, they provide a powerful tool to explore user feedback regarding mHealth behavioral interventions. Online user reviews are capable of providing data regarding BCTs for promoting engagement with health-protective behavior, their mode of delivery (i.e., features of mobile applications), and the sentiment of the experience they provide to end-users. When deciding what BCTs to include in an intervention, assessment of likely user perceived satisfaction should be taken into consideration to deliver positive and meaningful user experiences. We build on the BCTTv1 and link BCT clusters, target behavior, user feedback and implied design features. Future work will aim at exploring the replicability of the proposed method in other application domains to evaluate user perceived satisfaction regarding design features for behavior change.

ACKNOWLEDGMENTS

This research was partially funded by IDERAM through grants no. M1420-01-0247-FEDER-000019 and M1420-01-0247-FEDER-000038.

REFERENCES

[1] [n.d.]. AppFollow. https://appfollow.io

[2] [n.d.]. BCT Taxonomy (v1): 93 hierarchically-clustered techniques. https://digitalwellbeing.org/wp-content/uploads/2016/11/BCTTv1{_}PDF{_}version.pdf

[3] Mohammad A Al-Ramahi, Jun Liu, and Omar F El-Gayar. 2017. Discovering Design Principles for Health Behavioral Change Support Systems: A Text Mining Approach. *ACM Trans. Manage. Inf. Syst.* 8, 2–3 (jun 2017). https://doi.org/10.1145/3055534

[4] Daniel Berdichevsky and Erik Neunschwander. 1999. Toward an ethics of persuasive technology. *Commun. ACM* (1999). https://doi.org/10.1145/301353.301410

[5] Paulina Bondaronek, Ghadah Alkhaldi, April Slee, Fiona L. Hamilton, and Elizabeth Murray. 2018. Quality of publicly available physical activity apps: Review and content analysis. https://doi.org/10.2196/mhealth.9069

[6] Paulina Bondaronek, April Slee, Fiona L. Hamilton, and Elizabeth Murray. 2019. Relationship between popularity and the likely efficacy: An observational study based on a random selection on top-ranked physical activity apps. *BMJ Open* (2019). https://doi.org/10.1136/bmjopen-2018-027536

[7] Judit Bort Roig, Nicholas Gilson, Anna Ribera, Ruth Contreras Espinosa, and Stewart Trost. 2014. Measuring and Influencing Physical Activity with Smartphone Technology: A Systematic Review. *Sports medicine (Auckland, N.Z.)* 44 (feb 2014). https://doi.org/10.1007/s40279-014-0142-5

[8] Ting-Ray Chang, Eija Kaasinen, and Kirsikka Kaipainen. 2012. What influences users' decisions to take apps into use? https://doi.org/10.1145/2406367.2406370

[9] Ning Chen, Jialiu Lin, Steven C.H. Hoi, Xiaokui Xiao, and Boshen Zhang. 2014. AR-miner: Mining informative reviews for developers from mobile app marketplace. In *Proceedings - International Conference on Software Engineering*. https://doi.org/10.1145/2568225.2568263

[10] Heather Colquhoun, Jennifer Leeman, Susan Michie, Cynthia Lokker, Peter Bragge, Susanne Hempel, K. A. McKibbon, Gjalt Jorn Y. Peters, Kathleen R. Stevens, Michael G. Wilson, and Jeremy Grimshaw. 2014. Towards a common terminology: A simplified framework of interventions to promote and integrate evidence into health practices, systems, and policies. *Implementation Science* (2014). https://doi.org/10.1186/1748-5908-9-51

[11] David E Conroy, Chih-Hsiang Yang, and Jaclyn P Maher. 2014. Behavior Change Techniques in Top-Ranked Mobile Apps for Physical Activity. *American Journal of Preventive Medicine* 46, 6 (2014), 649–652. https://doi.org/10.1016/j.amepre.2014.01.010

[12] Karen L Courtney, Marcy Antonio, Ashley Garnett, and Judith T Matthews. 2019. Applying the Behavior Change Technique Taxonomy to Mobile Health Applications: A Protocol. *Stud Health Technol Inform.* 257 (2019), 64–69.

[13] Logan T. Cowan, Sarah A. van Wagenen, Brittany A. Brown, Riley J. Hedin, Yukiko Seino-Stephan, P. Cougar Hall, and Joshua H. West. 2013. Apps of Steel: Are Exercise Apps Providing Consumers With Realistic Expectations?: A Content Analysis of Exercise Apps for Presence of Behavior Change Theory. *Health Education and Behavior* (2013). https://doi.org/10.1177/1090198112452126

[14] Janet Davis. 2009. Design methods for ethical persuasive computing. In *ACM International Conference Proceeding Series*. https://doi.org/10.1145/1541948.1541957

[15] Dovetail Research Pty. Ltd. [n.d.]. Dovetail. https://dovetailapp.com

[16] E. A. Edwards, J. Lumsden, C. Rivas, L. Steed, L. A. Edwards, A. Thiyagarajan, R. Sohanpal, H. Caton, C. J. Griffiths, M. R. Munafò, S. Taylor, and R. T. Walton. 2016. Gamification for health promotion: systematic review of behaviour change techniques in smartphone apps. , e012447 pages. https://doi.org/10.1136/bmjopen-2016-012447

[17] Facundoolano. 2018. app-store-scraper. https://github.com/facundoolano/app-store-scraper

[18] Facundoolano. 2018. google-play-scraper. https://github.com/facundoolano/google-play-scraper

[19] B. J. Fogg. 2003. *Persuasive Technology: Using Computers to Change What We Think and Do*. https://doi.org/10.1016/B978-1-55860-643-2.X5000-8

[20] Jill Freyne, Emily Brindal, Gilly Hendrie, Shlomo Berkovsky, and Mac Coombe. 2012. Mobile Applications to Support Dietary Change: Highlighting the Importance of Evaluation Context. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems (CHI EA '12)*. Association for Computing Machinery, New York, NY, USA, 1781–1786. https://doi.org/10.1145/2212776.2223709

[21] Bin Fu, Jialiu Lin, Lei Li, Christos Faloutsos, Jason Hong, and Norman Sadeh. 2013. Why people hate your app. https://doi.org/10.1145/2487575.2488202

[22] Michael Gamon, Anthony Aue, Simon Corston-Oliver, and Eric Ringger. 2005. Pulse: Mining customer opinions from free text. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/11552253_12

[23] Karen Glanz and Donald B. Bishop. 2010. The role of behavioral science theory in development and implementation of public health interventions. https://doi.org/10.1146/annurev.publhealth.012809.103604

[24] Shantanu Godbole, Indrajit Bhattacharya, Ajay Gupta, and Ashish Verma. 2010. Building Re-Usable Dictionary Repositories for Real-World Text Mining. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM '10)*. Association for Computing Machinery, New York, NY, USA, 1189–1198. https://doi.org/10.1145/1871437.1871588

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

[26] Nathan Harmston, Wendy Filsell, and Michael P.H. Stumpf. 2010. What the papers say: Text mining for genomics and systems biology. https://doi.org/10.1186/1479-7364-5-1-17

[27] Steffen Hedegaard and Jakob Grue Simonsen. 2013. Extracting usability and user experience information from online user reviews. In *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/2470654.2481286

[28] Johannes Huebner, Remo Manuel Frey, Christian Ammendola, Elgar Fleisch, and Alexander Ilic. 2018. What people like in mobile finance apps – An analysis of user reviews. In *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3282894.3282895

[29] Pasi Karppinen and Harri Oinas-Kukkonen. 2013. Three approaches to ethical considerations in the design of behavior change support systems. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-642-37157-8_12

[30] Raymond Kight and Sandra Burri Gram-Hansen. 2019. Do ethics matter in persuasive technology?. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-030-17287-9_12

[31] Azar K.M.J., Lesser L.I., Laing B.Y., Stephens J., Aurora M.S., and Burke L.E. 2013. Mobile applications for weight management: Theory-based content analysis.

[32] Hannu Korhonen, Juha Arrasvuori, and Kaisa Väänänen-Vainio-Mattila. 2010. Let users tell the story. https://doi.org/10.1145/1753846.1754101

[33] Paul Krebs and Dustin T Duncan. 2015. Health App Use Among US Mobile Phone Owners: A National Survey. *JMIR mHealth and uHealth* (2015). https://doi.org/10.2196/mhealth.4924

[34] Vanessa R. Lerch, Sharon T. Steinemann, and Klaus Opwis. 2018. Understanding fitness app usage over time: Moving beyond the need for competence. In *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3170427.3188608

[35] Lexical Computing CZ s.r.o. [n.d.]. Sketch Engine. https://www.sketchengine.eu

[36] Yong Liu, Jorge Goncalves, Denzil Ferreira, Bei Xiao, Simo Hosio, and Vassilis Kostakos. 2014. CHI 1994-2013. https://doi.org/10.1145/2556288.2556969

[37] Washington Luiz, Felipe Viegas, Rafael Alencar, Fernando Mourão, Thiago Salles, Dárlinton Carvalho, Marcos Andre Gonçalves, and Leonardo Rocha. 2018. A feature-oriented sentiment rating for mobile app reviews. In *The Web Conference 2018 - Proceedings of the World Wide Web Conference, WWW 2018*. https://doi.org/10.1145/3178876.3186168

[38] Ulrik Lyngs, Kai Lukoff, Petr Slovak, Reuben Binns, Adam Slack, Michael Inzlicht, Max Van Kleek, and Nigel Shadbolt. 2019. Self-control in cyberspace: Applying dual systems theory to a review of digital self-control tools. In *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3290605.3300361

[39] Susan Michie, Charles Abraham, Craig Whittington, John McAteer, and Sunjai Gupta. 2009. Effective Techniques in Healthy Eating and Physical Activity Interventions: A Meta-Regression. *Health Psychology* 28, 6 (2009), 690–701. https://doi.org/10.1037/a0016136

[40] Susan Michie, Dean Fixsen, Jeremy Grimshaw, and Martin Eccles. 2009. Specifying and reporting complex behaviour change interventions: The need for a scientific method. *Implementation science : IS* 4 (feb 2009), 40. https://doi.org/10.1186/1748-5908-4-40

[41] Susan Michie, Michelle Richardson, Marie Johnston, Charles Abraham, Jill Francis, wendy Hardeman, Martin P Eccles, James Cane, and Caroline E Wood. 2013. The Behavior Change Technique Taxonomy (v1) of 93 Hierarchically Clustered ...: EBSCOhost. *ann. behav. med* (2013). https://doi.org/10.1007/s12160-013-9486-6

[42] Dennis Pagano and Walid Maalej. 2013. User feedback in the appstore: An empirical study. In *2013 21st IEEE International Requirements Engineering Conference, RE 2013 - Proceedings*. https://doi.org/10.1109/RE.2013.6636712

[43] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification Using Machine Learning Techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10 (EMNLP '02)*. Association for Computational Linguistics, USA, 79–86. https://doi.org/10.3115/1118693.1118704

[44] Matthew Price, Erica K. Yuen, Elizabeth M. Goetter, James D. Herbert, Evan M. Forman, Ron Acierno, and Kenneth J. Ruggiero. 2014. mHealth: A mechanism to deliver more accessible, more effective mental health care. *Clinical Psychology and Psychotherapy* 21, 5 (2014), 427–436. https://doi.org/10.1002/cpp.1855

[45] Frank M. Shipman and Catherine C. Marshall. 2013. Are user-contributed reviews community property? Exploring the beliefs and practices of reviewers. In *Proceedings of the 5th Annual ACM Web Science Conference, WebSci'13*. https://doi.org/10.1145/2464464.2464473

[46] Jilles Smids. 2012. The voluntariness of persuasive technology. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-642-31037-9_11

[47] Panayiotis Tsaparas, Alexandros Ntoulas, and Evimaria Terzi. 2011. Selecting a comprehensive set of reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2020408.2020440

[48] Rajesh Vasa, Leonard Hoon, Kon Mouzakis, and Akihiro Noguchi. 2012. A preliminary analysis of mobile app user reviews. In *Proceedings of the 24th Australian Computer-Human Interaction Conference, OzCHI 2012*. https://doi.org/10.1145/2414536.2414577

[49] Kieran Walshe. 2009. Pseudoinnovation: the development and spread of healthcare quality improvement methodologies. *International Journal for Quality in Health Care* 21, 3 (apr 2009), 153–159. https://doi.org/10.1093/intqhc/mzp012

[50] Thomas L. Webb, Judith Joseph, Lucy Yardley, and Susan Michie. 2010. Using the Internet to promote health behavior change: A systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. *Journal of Medical Internet Research* 12, 1 (jan 2010). https://doi.org/10.2196/jmir.1376

[51] Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal. 2016. *Data Mining: Practical Machine Learning Tools and Techniques*. https://doi.org/10.1016/c2009-0-19715-5

[52] Caroline E Wood, Michelle Richardson, Marie Johnston, Charles Abraham, Jill Francis, Wendy Hardeman, and Susan Michie. 2015. Applying the behaviour change technique (BCT) taxonomy v1: a study of coder training. *Translational behavioral medicine* 5, 2 (jun 2015), 134–148. https://doi.org/10.1007/s13142-014-0290-z

[53] World Health Organization. [n.d.]. Global Action Plan for the Prevention and Control of NCDs 2013-2020.

[54] World Health Organization. 2017. Cardiovascular diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[55] Jing Zhao, Becky Freeman, and Mu Li. 2016. Can mobile phone apps influence people's health behavior change? An evidence review. https://doi.org/10.2196/jmir.5692